INTERCONNECTS FOR FUTURE TECHNOLOGY GENERATIONS—CONVENTIONAL CMOS WITH COPPER/LOW–κ AND BEYOND

A Dissertation
Presented to
The Academic Faculty

By

Ahmet Ceyhan

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
in
Electrical and Computer Engineering

School of Electrical and Computer Engineering
Georgia Institute of Technology
December 2014

INTERCONNECTS FOR FUTURE TECHNOLOGY GENERATIONS—CONVENTIONAL CMOS WITH COPPER/LOW–κ AND BEYOND

Approved by:

Dr. Azad Naeemi, Advisor
Associate Professor, School of ECE
Georgia Institute of Technology

Dr. Saibal Mukhopadhyay
Associate Professor, School of ECE
Georgia Institute of Technology

Dr. Jeffrey A. Davis
Associate Professor, School of ECE
Georgia Institute of Technology

Dr. Yogendra Joshi
Professor, School of ME
Georgia Institute of Technology

Dr. Muhannad Bakir
Associate Professor, School of ECE
Georgia Institute of Technology

Date Approved: December 2014

To my parents, Belgin and Cumhur Ceyhan

who taught me patience and persistence

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my thesis advisor, Professor Azad Naeemi, for his unwavering guidance, constructive criticism and insightful suggestions throughout my studies at Georgia Tech. His invaluable insight in the design and potential problems of future interconnect technologies has directly shaped the work in this thesis. It has been a real privilege to work with him and I will cherish all the lessons that I have learned from him, both during my career and in my personal life.

I feel very lucky that I had the great pleasure of working with the wonderful people of the Nanoelectronics Research Lab. I would like to thank Shaloo for always being a positive influence on all of us and bringing down the walls between our cubicles by organizing events to get us together outside of the academic environment. I really enjoyed discussing potential research topics with Vachan and I am grateful to him for taking the time to brainstorm with me because many of our discussions have led to practical ideas to further my research and some have even become publications. I also wish to thank Nick for always putting a smile on everyone's faces, being a trustworthy friend to me and always happily extending his full assistance on many occasions, sometimes even before I asked. I must also thank Anant, Omar, Chenyun, Sou-Chi, Phillip, Sourav, Divya, Rouhollah and Ramy for the invaluable feedback they have given me during many of my presentations.

I am thankful to Professor Jeff Davis and Professor Saibal Mukhopadhyay for being on my proposal committee, reading my thesis and sharing their useful ideas about the direction of my research. I would also like to extend my sincere thanks to Professor Muhannad Bakir and Professor Yogendra Joshi for agreeing to be on my dissertation committee and for their insightful discussions and questions about the results of my research.

I would like to thank my friend Kerem for our many conversations, which have led to my decision to pursue a doctorate degree. I'm also grateful to my friends Kemal, Erdem, Selcuk, Giray and Baris, for being my brothers away from my family. You have made some of the really hard times during the last five years seem more tolerable with your support and encouragement.

I am extremely grateful to my parents for all their sacrifices, all the long hours of phone conversations in the middle of the night and never making me feel alone as I struggled to achieve my goals. I am deeply indebted to them for their patience during this time and I hope that I have made them proud with this dissertation.

The completion of this dissertation would not have been possible without one person beside me. Lena, the most important support, both direct and indirect, that I have received has been from you. You have stayed up with me before deadlines, taken care of me during the worst sickness of my life, and embraced my goals as your own. I cannot begin to express my thanks to you for all the positivity you brought into my life and the energy you have given me in times of despair. Your steadfast support and profound belief in me has been and will continue to be my greatest strength at every step as I achieve my goals one by one.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

BACKGROUND AND MOTIVATION

The exponential growth of the electronics industry has been guided by continued dimensional scaling of silicon–based CMOS technology for over four decades. Numerous companies have pursued smaller and faster transistors for years because miniaturization of transistors has enabled significant improvements in the transistor performance and power; higher transistor density for improved functionality, complexity and performance of microchips; and reduction in the cost for a single transistor. At the center of these advancements has been Moore's Law, which, combined with Dennard's guidelines for classical scaling introduced in 1974 [1], has determined the industry target to double the number of transistors on a microchip approximately every 18–24 months.

In recent years, the semiconductor industry has needed many innovative material and device–structure solutions to overcome significant threats to continued dimensional scaling. During the last decade, limitations to scaling started with the challenges in reducing the thickness of the gate dielectric material. This challenge stemmed from fundamental quantum laws that governed quantum mechanical tunneling of electrons from the gate to the channel [2, 3]. Even though the gate dielectric scaling stopped for a few technology generations in effort to keep gate leakage current under control, transistor performance improvement was still maintained thanks to the introduction of the revolutionary strained–silicon technology [4, 5]. The gate dielectric scaling problem was eventually resolved by replacing SiO_2 with a high–κ dielectric material [6], which allowed increasing the physical thickness of the gate oxide to reduce the probability of electron tunneling, while providing a thinner electrical equivalent for better electrostatic control of the channel, and improved transistor performance. Also, the polycrystalline silicon

gate was replaced with a metal gate because the poly gate was not compatible with the high–κ material [7]. Finally, the 22–nm technology generation announced the revolutionary departure from planar CMOS by introducing fully–depleted tri–gate transistors [8], which utilize the vertical dimension to extend the electrostatic control of the gate to three sides of a fin for improved performance at a smaller supply voltage and reduced short channel effects.

Besides smaller and faster transistors, the semiconductor industry requires fast and dense interconnects to manufacture high–performance microchips. Interconnect performance; however, degrades with dimensional scaling [9]. Resistance increases as the dimensions get smaller and the total capacitance increases due to the high density of interconnects. Therefore, the number of metal layers has gradually increased over the years [10], providing the possibility to route fine–pitch interconnects for high density at some metal layers, and wider and thicker interconnects for improved delay at other metal layers. In the last decade, Aluminum (Al) has been replaced by Copper (Cu) to improve the resistance–capacitance (RC) delay of interconnects because Cu offers increased conductivity compared to Al [11], and has a higher resistance to electromigration [10]. Furthermore, in effort to reduce the capacitance associated with interconnects, which directly determines both the interconnect RC delay and the interconnect dynamic power dissipation, progressively lower–κ dielectric materials have been introduced in many generations of technology [10]. These new materials, new processes, and the increase in the number of metal layers have enabled interconnect scaling for various technology generations. The 22–nm technology node comprises 9 Cu layers with an ultra–low –κ dielectric material [8].

All of these innovative solutions in the last decade have come to reality as a result of enormous investments in research and development. Even though utilizing the vertical dimension in both the device and the chip levels is expected to

govern the technological advancements in the near future, the semiconductor industry is expected to continue facing major challenges to continue scaling during the next decade. One of these challenges is to extend the use of 193–nm immersion lithography tools to ultra–scaled technology nodes through optimized multiple–patterning and computational–lithography techniques, until extreme ultraviolet light (EUV) lithography, which makes use of light at a wavelength of 13 nm, is ready. This dissertation focuses on another major challenge, namely interconnects, which still constitute significant limitations to the performance of microchips despite the aforementioned innovations.

The research pipeline of the semiconductor industry involves increasingly radical potential solutions to carry technology advancement through dimensional scaling to beyond conventional CMOS. Many companies encourage and conduct research on emerging device and interconnect technologies, such as carbon–based devices [12, 13] and interconnects [14, 15], nano–electromechanical systems (NEMS) [16], optical or photonic interconnects [17, 18], and even non–charge–based systems [19], to extend Moore's Law to beyond–2020 technology generations. However, any device technology that offers advantages in performance, power dissipation or ease in dimensional scaling will have to be complemented with an interconnect technology that offers similar trades. Therefore, all this research has to be centered around interconnects, which have become a highly complex problem in terms of performance and energy as well as reliability and cost.

The aim of this dissertation is to investigate the energy/performance limitations of the existing Cu/low–κ interconnect technology for use in future ultra–scaled integrated circuits down to 7–nm technology node, and to evaluate the opportunities that arise for emerging novel interconnect technologies from the materials and process perspectives. This research also aims to analyze the impact of various emerging interconnect technologies on the performances of emerging post–CMOS

devices, and to quantify the realistic circuit– and system–level benefits that these devices can offer.

1.1 The Interconnect Problem

1.1.1 The Early Interconnect Problem is Back

The early electronic equipments comprised only a few dozen components, which could be interconnected using hand–soldering techniques [20]. As the electronic systems became more complex; however, this manufacturing procedure of manually assembling circuits with discrete components quickly became costly, bulky, and unreliable. The exponential increase of the number of interconnections with the increasing number of circuit components was a major limiting factor for making more complex electronic systems. As an attempt to simplify this manufacturing process from the interconnect perspective, Danko and Abrahamson announced the Auto–Sembly process in 1949 [21], which would later evolve into the standard printed circuit board fabrication process. The worldwide pursuit of a method to reduce the cost, improve the performance, and reduce the size and weight of electronic equipments gave its fruits in late 1950's with the announcement of the integrated circuit (IC) [20]. Texas Instrument's Jack Kilby came up with a method to integrate a transistor, a capacitor, and a resistor on the same semiconductor material and connected them with soldered wires. Fairchild Semiconductor's Robert Noyce independently formed two transistors with three diffusion regions on a common substrate, using one of the transistors as a pair of diodes; the junctions as capacitors; metal leads over an oxide layer as resistors where required; and planar interconnections [20]. The substantial cost reduction in producing electronic equipment enabled by the transition from interconnecting discrete transistors to integrated circuits has led to tremendous research and development to achieve integration on increasingly larger scales [22]. Based on the observation that the number of components roughly doubled every year during the first seven years

of integrated semiconductor technology, Gordon Moore stated in his famous 1965 paper [23] that the number of components on a lowest–cost semiconductor chip grows exponentially in time. The economical growth of the semiconductor industry has been driven by his prediction that the semiconductor technology will double its effectiveness every 18 months.

Since then, the number of transistors on microchips has increased continuously with minimum feature size scaling as shown in Figure 1, resulting in the ability to integrate more functionality and complexity in logic products. This advancement has enabled the semiconductor industry to offer a wide range of logic products with different features and performances. However, as the number of transistors on a chip increased, so did the number of interconnections that are required to maintain communication between two points on a chip. Today, the semiconductor industry faces a modified version of the aforementioned early interconnect problem. To route the tremendous number of wires on a microchip in the same footprint, extra metal layers have been implemented. Interconnecting billions of transistors in integrated circuits has become a highly complex problem and a major threat to improving the integrated circuit performance at each new technology node.

1.1.2 Interconnect Latency Problem from the Resistance–Capacitance Perspective

As mentioned in Section 1.1.1, modern electronic chips have a multilevel interconnection network comprising metals with different dimensions. Short interconnects that carry signals between transistors that are relatively close to each other, within a certain functional block, are routed at local interconnect levels with fine pitches for high density. As interconnects get longer, they are made wider and thicker to reduce the associated resistance per unit length; hence delay. Therefore, the multilevel interconnection architecture is not only a requirement for routability, but also

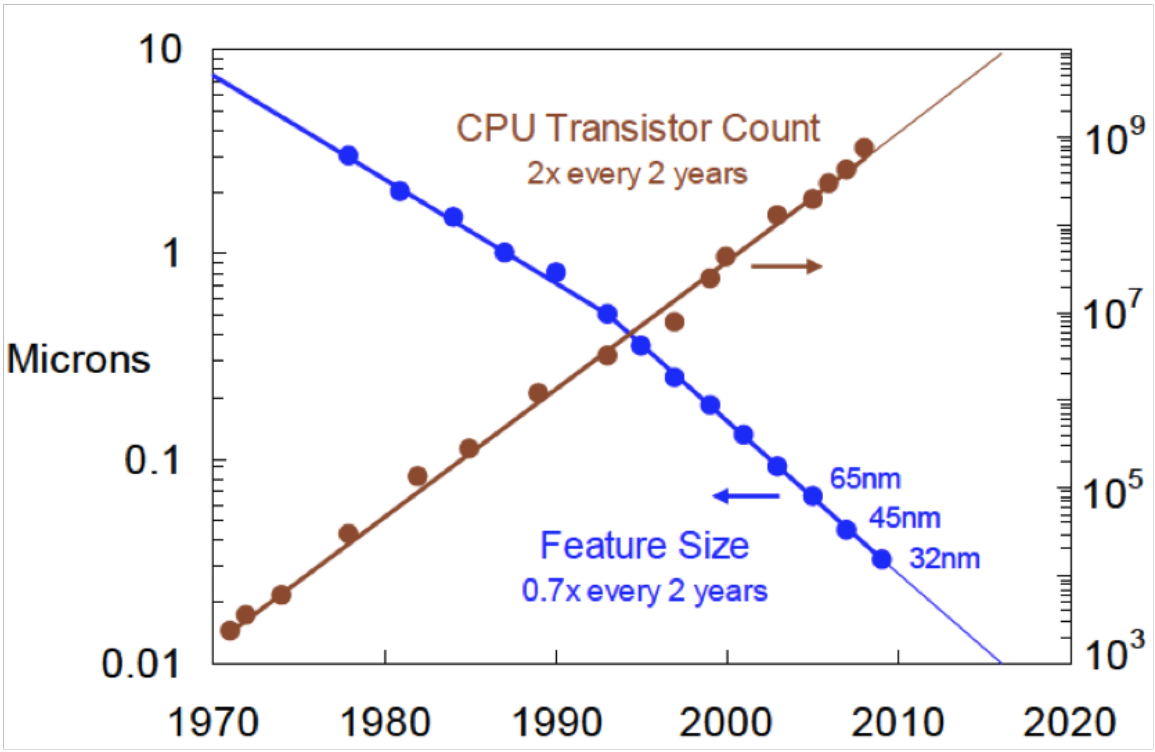

Figure 1: The number of transistors on a microchip increases and the minimum feature size decreases at each technology node. Adapted from [10].

a partial solution to the interconnect latency problem [24].

The aggressive dimensional scaling of the local metal level (M1), which causes the resistance to increase, and the increasing number of interconnects constitute an ever increasing resistive and capacitive load to the system. As the intrinsic device performance is improved with dimensional scaling, the impact of this load on the circuit speed becomes more pronounced. Some projections of the ITRS update in 2011 [25] are tabulated in Table 1 to illustrate the severity of the interconnect latency problem. It is observed that there is a quick reduction in the interconnect length at which the interconnect delay becomes equal to the intrinsic delay of an NMOS transistor.

The intrinsic latency of an *RC*–limited interconnect is proportional to,

$$\tau \propto \rho\epsilon \frac{L^2}{HT}. \tag{1}$$

Table 1: Interconnect technology parameter projections related to the latency of interconnects extracted from the 2011 update of ITRS [25]. Calculated metrics are indicated with the * sign.

	2015	2020	2025
M1 half pitch (nm)	21	12	7
Aspect Ratio	1.9	2	2.2
Cu resistivity ($\mu\Omega \cdot cm$)	6.61	9.74	15.02
Barrier/cladding thickness for Cu M1 wiring (nm)	1.9	1.1	0.6
*Resistance per unit length for M1 wires, r ($\Omega/\mu m$)	101	434	1750
NMOS intrinsic delay, $\tau = CV/I$ (Multi–gate, MG) (ps)	0.32	0.19	0.12
Capacitance per unit length for M1 wires, c (pF/cm)	1.8–2	1.6–1.8	1.5–1.8
*Distributed RC delay of 1mm M1 wire, $\tau_{int} = 0.4rcL^2$ (ps)	7676	29512	115500
*Length at which $\tau_{int} = \tau$, (μm)	6.5	2.5	1

This expression shows that the interconnect delay can be reduced by: (1) reducing metal resistivity (ρ) using new materials, (2) scaling insulator permittivity (ϵ), (3) reducing the interconnect length (L) using novel architectures, and (4) reverse scaling metal height (H) and insulator thickness (T). A variety of solutions have materialized in order to mitigate the global interconnect problem over the years. Some of these include: switching to the Cu/low–κ interconnect technology to introduce a lower $\rho\epsilon$ product, using many core architectures to reduce the maximum global interconnect length, and reverse scaling. Switching to three–dimensional integration offers the opportunity to reduce the length of the longest global interconnect as well. Another approach to solving the global interconnect scaling problem is changing the physical means of interconnection by introducing on chip optical interconnects [26, 27, 28]. Even though some of these solutions have in turn introduced other problems, such as router power dissipation in many–core architectures, it is undeniable that the nature of the global interconnect problem has changed as a result of these advances.

Furthermore, there is a radical change in local level interconnect behavior at sub–20 nm technology nodes. At such small dimensions, the resistivity of Cu interconnects has increased significantly due to size effects, such as sidewall and

grain boundary scatterings, and line edge roughness (LER) [29], which will be described in more detail in this thesis. As the dimensions of within core interconnects scale with technology and the L^2/HT term in equation (1) is kept almost constant with technology scaling, metal resistivity becomes the dominating factor in determining the interconnect intrinsic latency. This radical change in Cu interconnect limitations for ultra–scaled future technology nodes motivates looking at alternative interconnect technologies that can replace Cu at the local metal levels, where Cu wire dimensions are small. Carbon–based interconnects have long been considered as a promising alternative for future nanoscale interconnects due to their long mean free path (MFP), high current carrying capability and high thermal conductivity. Despite major technological progress in fabricating such interconnects and the rising opportunities in terms of energy and performance as studied in this dissertation, there are still many major challenges that must be overcome before they can become commercially viable options.

1.1.3 Interconnect Dynamic Power Dissipation Problem

The system–level fruits of making faster, less power–hungry transistors and faster, denser interconnects at each technology generation are illustrated in Figure 2. It is shown that the system frequency has increased, and the single–thread performances of microprocessors have improved with technology scaling for three decades. What is not shown in this figure is that this improvement is not simply the result of making faster transistors and increasing the clock frequency, but also the result of micro–architectural advancements that were enabled by the rapid increase in the transistor density with technology scaling. These advancements provided the opportunity to design microprocessors that exploit instruction level parallelism in pipelined architectures for increased throughput. Increasingly larger cores were built with higher frequency and higher power using faster and smaller transistors at each technology generation for years. However, the power consumption

of chips eventually became a major limitation to building larger cores. Figure 2 demonstrates that the increase in clock frequency slowed to keep power dissipation of microchips under control. Today, to manage power dissipation challenges and to continue increasing the performance of microprocessors, it has become necessary to implement multi–core structures for parallel computation. Careful consideration of power management will continue to be a major issue in the pursuit of extending Moore's Law to future multi– and many–core structures.

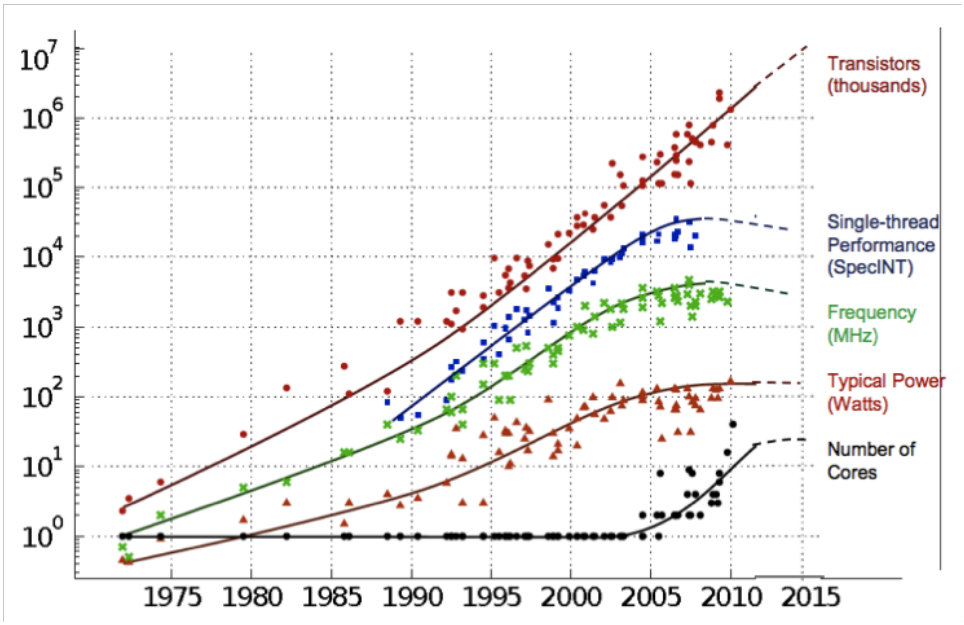

Figure 2: Microprocessor trend data: The changes in the transistor count, single–thread performance, frequency, power, and number of cores are plotted for the past 35 years. Adapted from [30].

A significant portion of the power dissipated in a microprocessor is due to the dynamic power dissipated in interconnects. An interconnect power analysis study performed on a microprocessor designed for power efficiency, consisting of 77 million transistors, and fabricated in the 0.13 μm technology in 2004, revealed that interconnects account for 50% of the total dynamic power dissipation [31]. Furthermore, as the interconnect dimensions are scaled, and the interconnect density is increased, the total capacitance associated with interconnects increases as well.

Lower–κ dielectric materials can reduce this capacitance; hence, the interconnect power. Table 2 compares interconnect and device dynamic power dissipations at three different technology nodes to underline the significance of the interconnect power dissipation problem.

Table 2: Interconnect technology parameter projections related to the dynamic power dissipation associated with interconnects extracted from the 2011 update of ITRS [25]. Calculated metrics are indicated with the * sign.

	2015	2020	2025
M1 half pitch (nm)	21	12	7
Aspect Ratio	1.9	2	2.2
Capacitance per unit length for M1 wires, c (pF/cm)	1.8–2	1.6–1.8	1.5–1.8
NMOS dynamic power indicator per device width , $E = CV^2$ $(fJ/\mu m)$	0.42	0.25	0.15
*M1 wire dynamic power indicator per length, $E_{int} = C_{int}V^2$ $(fJ/\mu m)$	0.1216	0.079	0.057
*Length at which $E_{int} = E$ for a minimum–width NMOS, (in unit of minimum device width)	3.45	3.16	2.63

Comparisons between interconnect and transistor delay/energy that are shown in Tables 1 and 2 are performed assuming multi–gate CMOS device and conventional Cu/low–κ interconnect technology projections. Both device and interconnect parameter projections are industry targets for continued Moore's Law, which may require many innovations to achieve. Therefore, emerging post–CMOS devices that meet these parameter requirements will also suffer from the same limitations imposed by the conventional Cu/low–κ interconnect. However, most of the current emerging device research is focused on speeding up or reducing the power consumption of a single device. A simple comparison of the intrinsic gate delay or the dynamic power indicator between a novel device technology and Si–CMOS will not reveal the complete picture of the promise that the new device holds. Interconnect aspects of novel devices have to be studied for a better understanding of the benefits they may offer.

1.2 Organization of the Thesis

The rest of this thesis is organized as follows. The limitations of the Cu/low–κ interconnect technology for future FinFET technology nodes are studied in two different methodologies in chapters 2 and 3. In chapter 2, compact models are used to study the impacts of various interconnect technology parameters on system performance and power dissipation based on stochastic wiring distributions [32]. In chapter 3, multiple predictive cell libraries down to the 7–nm technology node are constructed to enable early investigation of the electronic chip performance using commercial electronic design automation (EDA) tools with real chip information. Rising opportunities for carbon–based interconnects at future technology nodes are studied in Chapter 4, where various single–walled carbon nanotube (SWNT) interconnect architectures are benchmarked against the existing Cu/low–κ technology and major technology requirements for SWNT interconnects to outperform Cu are identified. Chapter 5 extends the study in Chapter 2 to another high–performance device technology, namely CNFETs, and evaluates the realistic circuit– and system–level benefits of using CNFETs by optimally designing multilevel interconnect networks for these devices. In Chapter 6, we pair various device technologies, such as CNFETs, FinFETs, TFETs and sub–threshold CMOS, with both conventional Cu/low–κ and emerging carbon–based interconnects to compare the impact of different interconnect technology parameters on device performance at the circuit–level. Finally, the main contributions of this thesis are summarized in Chapter 7 and potential future directions in research are defined.

CHAPTER 2

COPPER/LOW–κ INTERCONNECT TECHNOLOGY DESIGN AND LIMITATIONS FOR FINFET CMOS

In this chapter, we investigate the performances of conventional Cu/low–κ multilevel interconnect networks (MINs) for FinFETs at the 20– , 16– , 14– , 10– , and 7–nm technology nodes corresponding to the even years between 2012 and 2020, respectively. This study captures the impacts of interconnect variables such as size effect parameters, barrier/liner bilayer thickness, and aspect ratio on the design and performance of the MIN of a logic core. Our results indicate that the number of metal levels for a high–performance chip increases by as large as 34% due to size effects and this value can go up to 76% considering issues in barrier/liner thickness scaling at the 7–nm technology node. At this node, increasing the aspect ratio of interconnects from 2 to 3 can improve wire delay and save 2 metal levels at the cost of 35% more power dissipation. A $\pm$20% wire width variation induces wire delay variations of -20% and +44% at minimum–width wires. Designing the MIN considering this variation increases the required wire area by 4% in the worst case.

2.1 Introduction

As briefly described in Chapter 1, one of the major challenges that the semiconductor industry is expected to face in the pursuit of further miniaturization of the minimum feature size in the next decade is the degrading interconnect performance. Interconnects limit the performance of integrated circuits (IC) because they add extra delay to critical paths, dissipate dynamic power, and impose reliability concerns due to electromigration and time–dependent dielectric breakdown (TDDB). Furthermore, variations in the interconnect features during manufacturing give

rise to variations in circuit performance. These limitations become increasingly restrictive with dimensional scaling. Electron scatterings at wire surfaces and grain boundaries, and line edge roughness (LER) cause the effective resistivity of Cu to increase rapidly, resulting in larger interconnect latency [33]; the total interconnect capacitance increases due to high number and density of interconnects, causing interconnect power dissipation to account for more than 50% of the dynamic power of the chip [31]; ensuring metal and dielectric reliability becomes challenging due to smaller dimensions and weaker mechanical properties of low–κ dielectric materials [34]; and interconnect process variations make it increasingly difficult to predict circuit performance at the design stage [35].

Search of innovative material solutions to the degrading performance of interconnects in a microchip due to scaling has not been fruitful yet and the conventional Cu/low–κ interconnect technology may be the only option for the future ultra–scaled technology generations. Therefore, it is important to be able to predict the impacts of Cu/low–κ interconnect parameters on the overall performances of future ICs.

Today, to route all the interconnects on a chip and to ensure manufacturability while meeting various performance constraints, a substantial amount of effort has to be devoted to both design and process optimizations. The aim of this chapter is to perform a design–driven interconnect process optimization for high–performance Cu/low–κ MINs for FinFETs by determining the interconnect pitches of different metal levels, considering the impacts of interconnect resistivity increases due to size effects, repeater insertion, and via blockage.

There can be two complementary approaches to attack this problem. The system behavior can be simulated using compact mathematical models from the material level all the way to the architectural level such that the impact of various

technology parameters on the overall performance parameter that is being optimized can be studied and multiple design options can be explored quickly to narrow down the design space for the designer. Alternatively, the real system can be designed for multiple scenarios using the design flow for taping out an actual chip to monitor the changes in the performance parameter and make decisions depending on the outcome. The advantage of the former approach is that it can save significant design time and allow for investigating a broader range of design options in a limited amount of time whereas the advantage of the latter approach lies in providing the real chip information, which increases its accuracy. In this chapter, we concentrate our efforts on the former analysis methodology.

In the past, numerous system–level optimization tools have optimized various aspects of the wiring hierarchy for better performance, low–power operation or smaller chip size based on stochastic wiring distribution models [32]. Prior work on the impact of interconnect resistivity increase due to size effects on an MIN has shown that the increase in the number of metal levels and the impact on chip performance are negligible if size effects are taken into account during the design process [36]. This study, however, was based on the 2003 International Technology Roadmap for Semiconductors (ITRS) projections, where chip clock frequencies as high as 50 GHz were expected at the end of the roadmap. Clock frequencies turned out to be significantly lower than 2003 ITRS projections due to the introduction of many–core architectures to reduce power dissipation. Also, impact of LER on resistivity was ignored in this study. Furthermore, this study was performed down to the 18–nm technology node. With the sub–20–nm FinFET technologies in development, it is necessary to investigate the issues with MINs down to the 7–nm technology node.

In addition, at ultra–scaled technology generations, errors due to process variations become more pronounced. Also, wire delay variation increases due to increasing sensitivity of wire behavior on manufacturing variations. Systematic intra–die variations in metal thickness, linewidth, and inter–layer dielectric (ILD) thickness significantly affect the circuit performance and may lead to yield loss if not accounted for. In this work, the overall impact of lithography–induced variations in wire width and spacing on the MIN design are evaluated using a worst–case corner analysis to ensure reliable operation.

This chapter is organized as follows. In Section 2.2, assumptions, models and the methodology used to optimally design the MIN are described. In Section 2.3, the MIN for the logic core of a commercial microprocessor is designed and the results are compared with actual data. In Section 2.4, the impacts of various interconnect parameters and wire width variation on the design and performance of MINs are analyzed. Section 2.5 investigates the power dissipation in MINs and the consequent impact on the total logic core power. In Section 2.6, we summarize the important conclusions of this chapter.

2.2 Multilevel Interconnect Network Architecture Design Methodology

In this chapter, optimal MINs are designed using the interconnect density function, $i(l)$, described in [37] and considering only logic transistors to reduce complexity [38]. This density function gives the number of interconnects with a certain length, l, normalized to gate socket lengths as defined in [37]. The density function depends on the size of the logic gates in the system, which are assumed to be 2–input NAND gates and are sized based on a generic critical path delay equation [39] to satisfy a given clock frequency. It is assumed that each 2–input NAND gate drives

a typical fan–out, f_{out}, of 3 through average length interconnects, which gives

$$\tau = 0.7 L_d \frac{R_{NAND}}{n_{fin}} f_{out} \left(C_{NAND} n_{fin} + C_{int} \right)$$
$$+0.7 L_d R_{int} C_{NAND} n_{fin} + 0.4 L_d R_{int} C_{int} f_{out} \leq T_{clock}. \quad (2)$$

Here, L_d is the logic depth assumed to be equal to 15 throughout this work, R_{NAND} and C_{NAND} are the average drive resistance and input capacitance of a 2–input FinFET NAND gate, respectively, and n_{fin} is the number of fins of the FinFET devices. R_{int} and C_{int} are the resistance and capacitance of average length interconnects, respectively. The average wire length is calculated from the interconnect density function as described in [37].

Starting with the short interconnects, which are routed in the metal levels with the minimum interconnect pitch, the range of interconnect lengths that can be routed in a certain metal level is calculated based on a supply–demand equation given by,

$$e_{w,n} A = \chi p_n \sqrt{\frac{A}{N_{sockets}}} \int_{L_{n-1}}^{L_n} l i(l) dl. \quad (3)$$

The left hand side of this equation represents the available area for routing wires, where $e_{w,n}$ is the net wiring efficiency of the n^{th} metal level and A is the area of the logic core. $N_{sockets}$ is the number of sockets, which is calculated by dividing the number of gates by the percentage of die area that is occupied by logic transistors as defined in [40]. The right hand side represents the area required for routing wires that have lengths between L_{n-1} gate sockets and L_n gate sockets. L_n and p_n are the maximum length normalized to gate socket lengths and wire pitch of the n^{th} pair in nanometers, respectively.

The interconnect pitch of each metal level is determined assuming that the maximum RC time delays of minimum–pitch short interconnects in the lower levels and custom–pitch longer interconnects in the upper levels are 25% and 90% of

the clock period, respectively [40]. The RC time delay of interconnects without repeaters is calculated by [37],

$$\tau_{rc} = 4.4 \frac{\rho(AR,p)}{AR \cdot p^2} c_{int} L^2 \frac{A}{N_{sockets}}, \tag{4}$$

where ρ, c_{int} and AR are the resistivity, capacitance per unit length (p.u.l.) and aspect ratio of the interconnect, respectively. In this work, repeaters are inserted to minimize the energy–delay product (EDP) as described in [41]. The RC time delay of a repeated interconnect can be calculated by,

$$\begin{aligned} \tau_{rc} &= \left(\frac{0.7}{\delta} + 0.7\gamma + \frac{0.4}{\gamma} + 0.7\delta \right) \sqrt{\frac{\rho(AR,p) c_{int} R_0 C_0}{AR \cdot (p/2)^2}} L \sqrt{\frac{A}{N_{sockets}}} \\ \gamma &= \left(0.73 + 0.07 \ln \phi_{gate}\right)^2 \\ \delta &= \left(0.88 + 0.07 \ln \phi_{gate}\right)^2 \\ \phi_{gate} &= \frac{P_{dynamic}}{P_{dynamic} + P_{leakage}} \end{aligned} \tag{5}$$

R_0, C_0, $P_{dynamic}$ and $P_{leakage}$ are the resistance, capacitance, and dynamic and leakage power dissipations of the minimum size inverter, respectively. The net wiring efficiency of a certain metal level is calculated considering via blockage due to repeaters inserted in the upper level interconnects and connections to signal wires in the upper layers [42], and the power/ground via blockage [43].

$$\begin{aligned} e_{w,n} &= e_r \left(1 - e_{pgnd}\right) \left(1 - e_{via,n}\right) \\ e_{via,n} &= \sqrt{\frac{2 \left(N_{wires\_above,n} + N_{rep\_above,n}\right) \left(p_n + s\lambda\right)^2}{A}}. \end{aligned} \tag{6}$$

In this equation, e_r is the router efficiency typically assumed to be equal to 0.5, e_{pgnd} is the fraction of area used by routing power and ground wires, and $e_{via,n}$ is the via blockage factor associated with the n^{th} level. $N_{wires\_above,n}$ and $N_{rep\_above,n}$ are the number of wires and repeaters above the n^{th} metal level, respectively. λ is the design rule unit equal to half the minimum feature size and s is a via covering factor equal to 3 [42].

The MIN is initially designed without any repeater–inserted levels. Repeaters are then inserted starting from the topmost level, which accommodates the longest interconnect, and continued downwards for each metal level until the Si area available for repeater insertion is all used or repeater insertion no longer improves the chip performance.

2.3 Design Methodology Validation

In this section, the MIN is designed for a commercial 22–nm technology quad–core microprocessor that contains 1.4 billion transistors on a 160 mm^2 die [44] to validate and calibrate the models and methodology. The logic transistors are estimated to occupy approximately 30% of the total die area from die photos [45]. The total number of logic transistors on the die is estimated to be 15% of the total transistor count considering that logic transistors tend to be larger than transistors in memory arrays [46] and occupy a smaller percentage of the chip area [38]. Modeling the logic gates as two–input NAND gates, each logic core contains 13.125 million gates in 11.2 mm^2. The maximum clock frequency of the system is 3.9 GHz. The fin pitch of FinFETs and the minimum interconnect pitch are taken as 60 nm and 80 nm, respectively [8]. Interconnect technology parameters such as the aspect ratio, the barrier thickness, and inter–layer dielectric constant are taken from 2011 ITRS projections for the year 2012 [25]. Interconnect size effect parameters, namely the specularity parameter, p_{size}, that determines the fraction of electrons that scatter specularly at the wire surfaces, and the reflectivity parameter, R_{size}, which determines the fraction of electrons that are scattered backwards at the grain boundaries are assumed to be 0.2 and 0.3, respectively [47]. The interconnect size effect parameters will be described in more detail in the next section. Area of FinFET gates are calculated based on the assumptions and design rules outlined in [48]. Rent's parameters k_{rent} and p_{rent} are 4 and 0.667, respectively [32]. Table 3 shows the results

compared with actual data for the number of metal levels and the pitch for each metal level.

Table 3: Comparison of results from the MIN design methodology with actual data.

Metal Level	Simulation Results	Actual Data
M1	80 nm	90 nm
M2	80 nm	90 nm
M3	100 nm	80 nm
M4	110 nm	112 nm
M5	150 nm	160 nm
M6	202 nm	240 nm
M7	254 nm	320 nm
M8	308 nm	360 nm
M9	598 nm	14 μm

Table 3 demonstrates that the MIN design methodology predicts the same number of metal levels as the actual data with close interconnect pitch values at each metal level. The only significant difference is on M9 layer, which, in this study, is designed based on signal, power and clock wiring considerations as described in [37]. However, M9 layer is a special layer used for low–resistance power routing to minimize voltage droops in the actual design [10].

The power dissipation of each core due to logic operations in this quad–core processor is calculated to be 12.5 W with the methodology used in this analysis, giving a total of 50 W. Even though data on the percentage breakdown of power dissipation in this particular microprocessor is not available, it can be assumed based on [49] that about 80% of the power dissipation in a single core is due to logic operations and the remaining 20% for cache and I/O operations. Therefore, the total power dissipated in cores is about 62.5 W. Assuming that about 75% of the total chip power is dissipated in the cores [38], the total power dissipation of the chip can be estimated as 83.3 W, which is close to the published data of 77 W.

Having shown that our methodology for the MIN design captures most of the

issues with determining the number of metal levels and core power dissipation, it is possible to analyze the impact of dimensional scaling on the design and performance of MINs at ultra–scaled technology generations. In the next section, we use this methodology to design the MINs of logic cores implemented in five technology nodes corresponding to the even years between 2012 and 2020 on the 2011 Edition of the ITRS [25].

2.4 Impact of Various Parameters on MIN Design and Performance

2.4.1 Impact of Size Effect Parameters

As mentioned, one of the major challenges of implementing Cu interconnects at ultra–scaled future technology nodes is their increased resistivity due to surface and grain boundary scatterings. To calculate resistivity of narrow Cu wires, mathematical models in [29] are used with appropriate size effect parameter values. There are various size effect parameter values in literature that are derived from experimental results as shown in Table 4 and the resistivity value highly depends on these parameters. The last row in Table 4 is our optimistic base–line scenario that considers a single–crystal Cu structure. In this scenario, it is assumed that the resistivity increase can be overcome in part by using Cu wires with larger grain sizes, which can be manufactured using a subtractive process [50], where, unlike the current dual–damascene process, the grain size is not limited by the height or the width of a trench. The goal is to effectively eliminate the impact of grain boundary scatterings on the overall resistivity increase. The specularity parameter for this scenario is chosen as 0.72 [51].

Figure 3 illustrates how Cu resistivity is affected by various values of specularity and reflectivity parameters. It is plotted assuming minimum–width interconnects at each technology node. These interconnects experience the largest increase in resistivity in an MIN. However, they are short and the increase in their resistance is not highly critical. Slightly wider wires routed in the intermediate levels,

Table 4: Various published experimental Cu size effect parameters.

Specularity Parameter, p_{size}	Reflectivity Parameter, R_{size}	Reference
0	0.5	Shimada et al.[52]
0.4	0.5	Steinhoegl et al.[53]
0	0.43	Kitada et al. [54]
0	0.25	Plombon et al. [47]
0.2	0.3	Plombon et al. [47]
0.72	0.4	Steinhoegl et al. [51]
0.1	0.2	Chen et al. [55]
0.5	0.3	Besling et al. [56]
0.49	0.27	Steinhoegl et al. [51]
0.33	0.19	Steinhoegl et al. [51]
0.4	0.19	Steinhoegl et al. [51]
0.43	0.2	Guillaumond et al. [57]
0.25	0.13	Steinhoegl et al. [58]
0.3	0.08	Steinhoegl et al. [51]
0.72	0	Base–Line Scenario

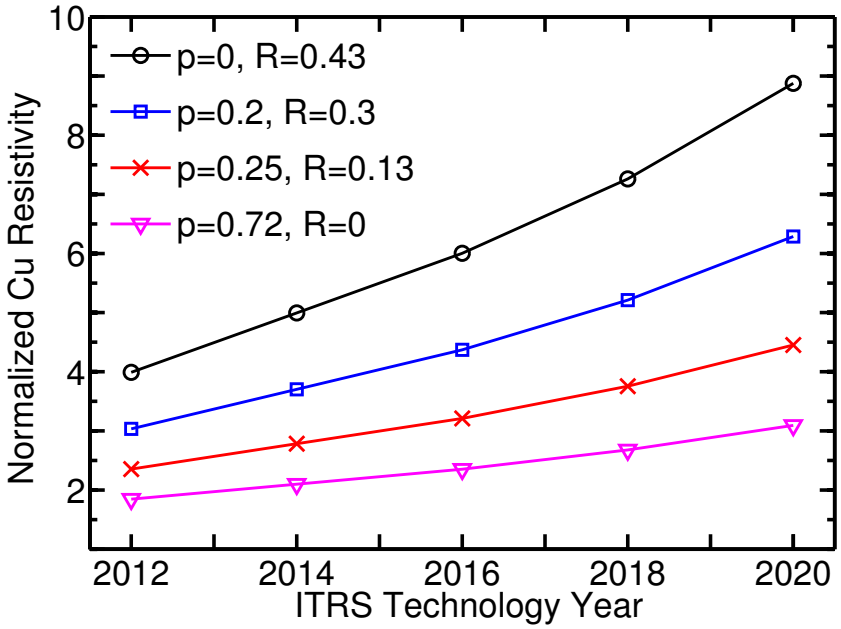

Figure 3: Minimum–size Cu wire resistivity normalized to the bulk Cu resistivity, which is 1.8 $\mu\Omega \cdot cm$. Barrier thickness and aspect ratio are taken from 2011 ITRS roadmap. Mean–free path of electrons in Cu is taken as 40 nm.

which are longer, may have a larger impact on the performance of an MIN.

The designs in this section are based on the ITRS projections [25] on the transistor count, die area, clock frequency, and interconnect technology parameters as mentioned in Section 2.3. Interconnect size effect parameters are selected from the aforementioned various experimental results in the literature [47, 58, 54] to cover a range of resistivity values. The number of cores on a chip is assumed to double every other year. Predictive technology models for FinFET devices [59] based on the industry standard BSIM–CMG model [60] are used to model the logic gates and repeaters at each technology generation. Figure 4 shows the number of metal levels and the percentage increase in the number of metal levels due to size effects with respect to the base–scenario.

In all technology nodes, the smallest number of metal levels can be achieved if single–crystal Cu interconnects with infinite grain sizes can be grown. Otherwise, the number of metal levels will increase; and this increase will worsen with scaling as shown in Figure 4. In 2020, assuming single crystal Cu, a typical change in the specularity parameter, $p_{size} = 0.4 \rightarrow 0$, would induce only $\sim 1.5\times$ increase in the resistivity of minimum size wires, which have a width of 12 nm. Assuming fully specular sidewall scatterings; however, the change $R_{size} = 0.1 \rightarrow 0.7$ induces $\sim 8\times$ increase in resistivity. Improving the LER from 40% to 20% of the linewidth reduces the resistivity by less than 15%. Hence, changes in grain boundary scatterings have a greater impact in determining the Cu resistivity and consequently the design and performance of MINs, over surface scatterings and LER. As a result, in 2020, the number of metal levels may increase by as much as 33.8% requiring 3 extra metal levels to route all interconnects.

Table 5 shows the optimized wire pitch in nanometers and the range of routed wire lengths normalized to gate socket lengths, which is 395 nm, for two sets of size effect parameters in 2012. Figure 5 plots the delay distribution function for

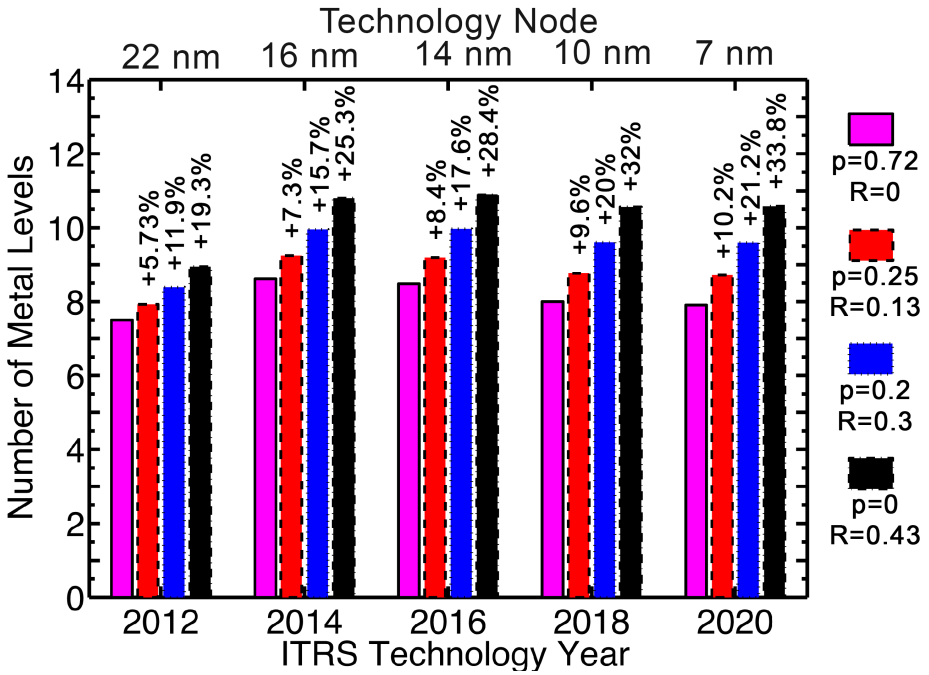

Figure 4: Number of metal levels is plotted versus the technology year considering a range of size effect parameters. Mitigating size effects can reduce the number of metal levels significantly.

these two designs. The ITRS projection for the delay of a minimum size device, τ_0, is used to divide interconnects into two groups based on their delay, τ_{RC}, as shown in Figure 5. Groups I and II comprise interconnects with τ_{RC} smaller and larger than τ_0, respectively. Table 6 shows the average wire delays in these two groups for both designs in 2012 and 2020.

It can be seen in Table 5 that interconnects routed at the first two metal levels are short enough such that the increase in their resistance is not highly critical even though they experience the largest increase in resistivity. Table 6 demonstrates that average wire delay in Group I increases by 28% due to size effects, but this delay is so small that the main limitation for routing these interconnects is the minimum size dictated by the process technology. The designs for these two levels change slightly, only due to the different via blockage caused by the upper level interconnects and repeaters. For longer interconnects, resistivity increase due to

size effects has to be compensated for by using wider and thicker interconnects, as is the case for M3 in Table 5, because they may significantly lower the maximum clock frequency of the system. Since the area available for routing interconnects is limited, this increase in width means that a smaller number of interconnects can be routed at a certain level. As a consequence, a fraction of interconnects that could be routed in M3 assuming infinite grain sizes have to be routed in upper levels. This trend increases the burden on the upper metal levels, eventually requiring extra metal levels. Similar explanations apply to future technology nodes with increasing severity.

It is shown in Table 6 that the change in average delay due to size effects in Group II is only 1% when the MIN is optimally designed considering size effects. The toy example in Table 6 emphasizes the significance of design optimization with the proper size effect parameters by showing that the average delay in Group II would have increased by 40% if the design for $p_{size} = 0.72, R_{size} = 0$ was used for $p_{size} = 0, R_{size} = 0.43$ as well.

Table 5: Multilevel interconnect network design results in 2012.

Metal Level	$p_{size} = 0.72, R_{size} = 0$ (Best)		$p_{size} = 0, R_{size} = 0.43$(Worst)	
	Pitch (nm)	$L_{min} - L_{max}$	Pitch (nm)	$L_{min} - L_{max}$
M1	80	1-49	80	1-49
M2	80	50-315	80	50-318
M3	88	316-945	102	319-831
M4	95	2681-7484	120	3244-7480
M5	143	946-1637	154	832-1405
M6	198	1638-2353	205	1405-2010
M7	255	2354-2680	257	2011-2641
M8	598	7485-16938	310	2642-3243
M9			598	7480-16938

It is now established that the change in the number of metal levels due to size effects is mainly due to the burden introduced by those wires that have small widths,

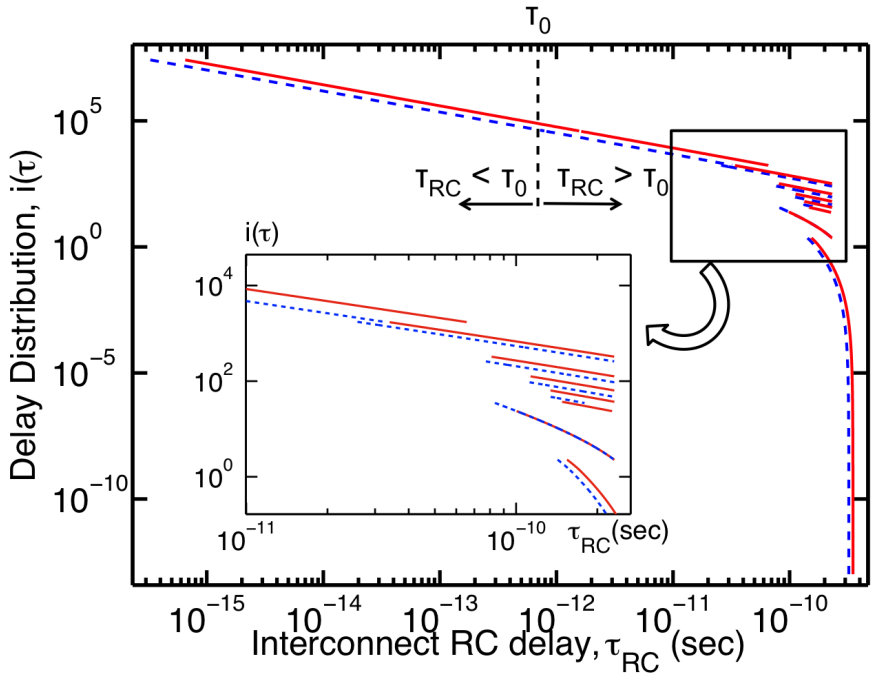

Figure 5: Interconnect delay distribution calculated for the worst case of size effects (straight line) and single-crystal Cu assumption (dashed line) in 2012. Each discontinuity corresponds to switching to a new metal level. For both cases, there are as many individual lines as the number of metal levels.

Table 6: Average interconnect delays.

2012, $\tau_0 = 0.57ps$						
Scenario	All	%	Group I	%	Group II	%
Best	1.41 ps	-	19.82 fs	-	27.08 ps	-
Worst	1.81 ps	28%	25.57 fs	29%	27.41 ps	1%
Toy	2.5 ps	77%	25.57 fs	29%	37.93 ps	40%
2020, $\tau_0 = 0.19ps$						
Best	1.01 ps	-	8.36 fs	-	12.87 ps	-
Worst	1.63 ps	61%	12.75 fs	52%	14.71 ps	14%
Toy	2.5 ps	148%	12.75 fs	52%	22.89 ps	78%

but are long enough such that their delay can put constraints on the overall system frequency. Due to the disparity between the impacts of scaling on the delay performance of interconnects and devices, the critical wire length, L_0, where the interconnect delay becomes as large as τ_0, shortens with each technology node. Combined with smaller wire widths, some of the interconnects which are longer than L_0 are thin enough to drastically suffer from size effects. Consequently, the average wire delay of Group II increases by 14% as shown in Table 6 for 2020, even when the MIN is optimized based on the proper size effect parameters and 3 extra metal levels are added.

2.4.2 Impact of Barrier/Liner Thickness

So far, it is assumed that all interconnect parameters scale according to ITRS projections including the barrier/liner thickness. The aggressive scaling of barrier/liner thickness projected by ITRS might be hard to achieve due to both manufacturing and reliability challenges. Many researchers are working on potential solutions that can reduce the barrier thickness, including using atomic layer deposition instead of sputter deposition for better control and using self–forming barrier layers [61], but the barrier/liner thickness may not be scaled down to ITRS projections at future technology nodes and its impact on performance has to be studied. Figure 6 shows the impact of barrier/liner bilayer thickness scaling on the required number of metal levels in the optimized MIN in 2020, assuming that the bilayer thickness can be extended down to $\sim 3-4nm$ [61]. The percentage values on each bar in Figure 6 represent the increase in number of metal levels taking the ITRS projections as reference for each set of size effect parameters. If the bilayer is 3.5 nm thick, the increase in the number of metal levels over the reference case is $\sim$ 12-13%. Combined with the impact of size effect parameters, the number of metal levels may increase by as much as 50%. For certain combinations of size effect parameters and barrier/liner thicknesses, it is not possible to come up with

a MIN design that meets the ITRS clock frequency projections. Therefore, research on both mitigating size effects and scaling the barrier/liner thickness are of vital importance for future technology generations.

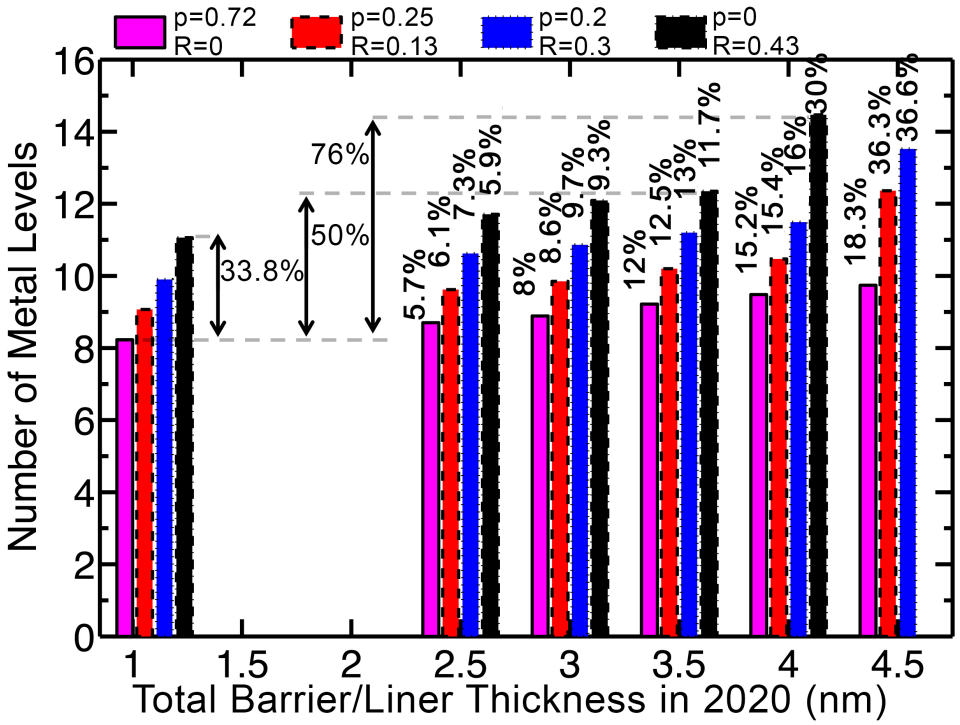

Figure 6: Number of metal levels is plotted versus the total thickness of the barrier/liner layer at the 7–*nm* technology node for various size effect parameters. The thickness of the bilayer should be scalable to 3.5 *nm*.

2.4.3 Impact of Aspect Ratio

To compensate for the increasing interconnect *RC* delay trend, one approach is to increase the aspect ratio to reduce the resistance p.u.l., as plotted in Figure 7(a) for minimum–width wires in 2020. However, increasing the aspect ratio also increases the capacitance p.u.l. as plotted in Figure 7(b). As a consequence, there can be an optimal aspect ratio for a given interconnect length as shown in Figure 7(d). For short signal interconnects that are only about 10–gate–pitch long, the aspect ratio that offers the smallest delay is as low as 1.5, whereas for 150–gate–pitch long interconnects, it is around 2.

Impact of aspect ratio on delay is determined by the relative changes of p.u.l. values of resistance and capacitance of the wire, and its length. It is assumed in Figure 7(d) that an optimally sized 2–input NAND gate drives 3 similar gates through an interconnect with a certain length. For short lengths, the resistance of the driver dominates the total wire resistance; and increasing capacitance p.u.l. as a result of increasing the aspect ratio always hurts short wires in terms of delay. For sufficiently long interconnects, for which the total wire resistance is larger than the driver resistance, increasing the aspect ratio can help until the increase in capacitance p.u.l. becomes dominant over the gain in p.u.l. resistance. In Figure 7(d), only minimum–width interconnects, which are routed in the first two metal levels, are considered assuming pessimistic size effect parameters. In a MIN, wire widths vary; hence, it is important to determine how a change in aspect ratio would impact the optimal MIN design.

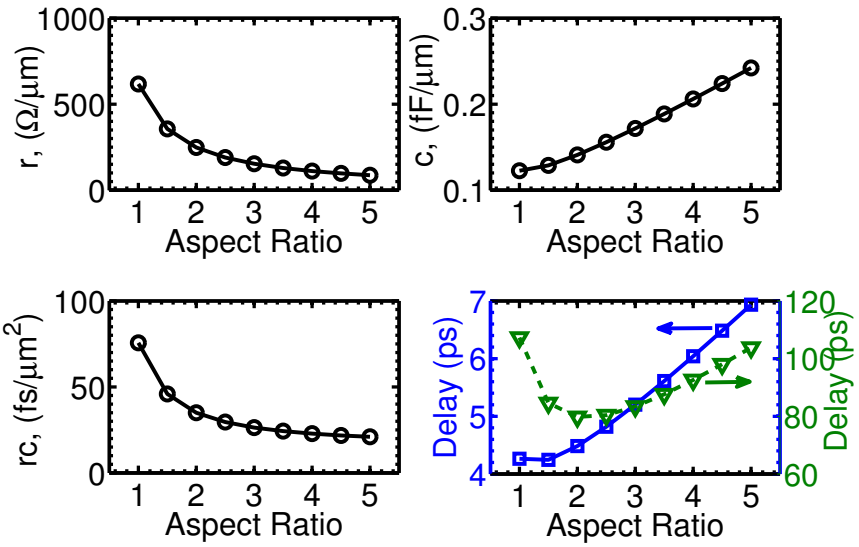

Figure 7: (a) resistance p.u.l., r, (b) capacitance p.u.l., c, (c) intrinsic interconnect rc delay p.u.l. squared, and (d) total delay assuming short (3 μm, $\sim$10 gate pitches, solid line) and longer (45 μm, $\sim$150 gate pitches, dashed line) interconnects, respectively, are plotted versus aspect ratio at the 7–nm technology node. Size effect parameters are taken as $p_{size} = 0$ and $R_{size} = 0.43$.

Designing the optimal MIN in 2020 assuming various aspect ratio values shows that the number of metal levels can be reduced initially, but will increase beyond a certain point. Comparing the designs for an aspect ratio of 2, which is the ITRS projection [25], and 3, assuming size effects parameters of $p_{size} = 0$ and $R_{size} = 0.43$, reveals that 2 metal levels can be saved by increasing the aspect ratio as shown in Table 7. This is because the increase in aspect ratio allows for slightly narrower wires to be routed at each metal level. As a result, more interconnects can be routed at a certain metal level, relieving the burden on the upper metal levels, and eventually requiring fewer number of levels. For instance, the maximum interconnect length that can be routed in M2 increases from 433 to 457 gate socket lengths. Since the number of short interconnects are very large, this increase in maximum routed interconnect length saves significant amount of area. This reduction in number of metal levels comes at the cost of 35% larger total logic core power dissipation due to the increasing interconnect power dissipation, which is caused by the extra interconnect capacitance. Therefore, an optimum design regarding the cost and power dissipation of a MIN depends on the designer's goals.

2.4.4 Impact of Wire Delay Variability

Variations in wire width and spacing affect resistance and capacitance associated with interconnects. The resistance–capacitance product and its percentage variation with width are plotted for minimum–size wires at the 7–nm technology node in Figure 8. Assuming a perfect Gaussian distribution for the width of the wire, it is equally likely to get a wire that is 44% slower or 20% faster than the nominal delay [35]. As the interconnect pitch is increased, variation in RC delay reduces.

This variation in wire delay is taken into account during the design of the MIN by introducing a variation term in the equations that are described in Section 2.2. This method assumes that the width of the longest wire that is routed in each metal level is different from the nominal width by an amount determined by the

Table 7: Multilevel interconnect network design results in 2020 showing interconnect pitch and range of interconnect lengths routed at each metal level normalized to gate socket lengths (99 nm).

Metal Level	$AR = 2$		$AR = 3$	
	Pitch (nm)	$L_{min} - L_{max}$	Pitch (nm)	$L_{min} - L_{max}$
M1	24	1-69	24	1-69
M2	25.3	70-433	24.2	70-457
M3	41	434-923	37.9	4239-11601
M4	42.3	4616-12047	39.4	458-987
M5	56.1	924-1466	54.2	988-1577
M6	70.7	1467-2040	68.5	1278-2204
M7	85.2	2041-2638	82.8	2205-2861
M8	99.6	2639-3259	97.1	2862-3546
M9	114.1	3260-3903	111.5	3547-4238
M10	128.8	3904-4573	590.3	11602-20694
M11	143.9	4574-4615		
M12	756.8	12048-20694		

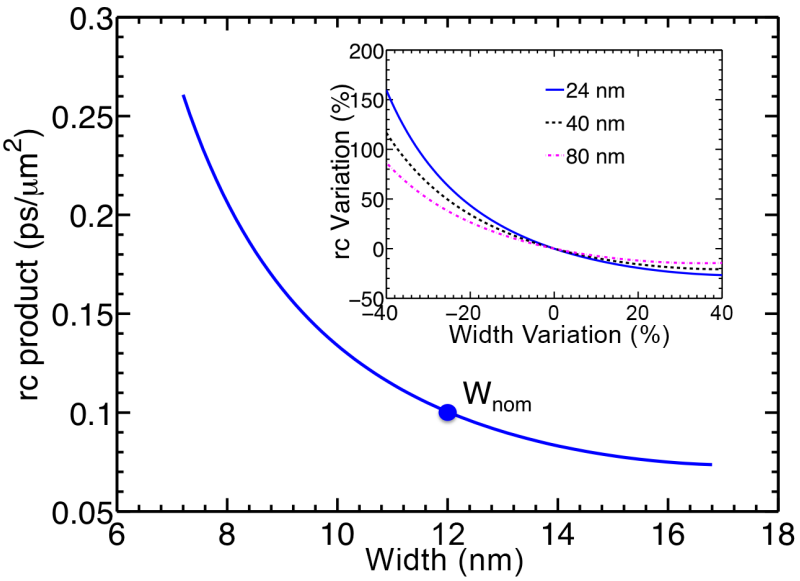

Figure 8: rc delay p.u.l squared for various width values considering an interconnect pitch of 24 nm in 2020 and size effect parameters $p_{size} = 0$, $R_{size} = 0.43$. The inset figure shows the percentage variation in rc delay versus the variation in width as a percentage of the nominal width value for various interconnect pitches.

variation parameter; and the pitch of each metal level is optimized accordingly. Although each interconnect in a metal level may have a different variation parameter, by considering the worst case for the longest interconnect, reliable operation is ensured for all the other wires. Table 8 shows two designs in 2020: Case I assumes perfect controllability of wire dimensions and Case II assumes a maximum of -20% variation in the intended metal width.

Table 8: Multilevel interconnect network design results in 2020 showing interconnect pitch and range of interconnect lengths routed at each metal level normalized to gate socket lengths (99 nm).

Metal Level	Case I		Case II	
	Pitch (nm)	$L_{min} - L_{max}$	Pitch (nm)	$L_{min} - L_{max}$
M1	24	1-69	24	1-69
M2	25.3	70-433	26.7	70-398
M3	41	434-923	39.8	4411-11858
M4	42.3	4616-12047	42.3	399-838
M5	56.1	924-1466	57.2	839-1325
M6	70.7	1467-2040	71.7	1326-1839
M7	85.2	2041-2638	85.8	1840-2373
M8	99.6	2639-3259	99.7	2374-2924
M9	114.1	3260-3903	113.6	2925-3492
M10	128.8	3904-4573	127.5	3493-4078
M11	143.9	4574-4615	141.6	4079-4410
M12	756.8	12048-20694	680.3	11859-20694

Note that the optimal wire pitch in M2 is larger in Case II to account for the -20% width variation. Increasing the pitch of M2 by 6% reduces both the wire–to–wire capacitance and resistance of the wires, but results in routing a smaller number of interconnects at this level compared to Case I. At higher metal levels, the variation in wire width has a smaller impact in wire delay. For instance, in M10, the optimal interconnect pitch is reduced by 1% even in the presence of -20% width variation because the maximum length of the interconnect in this design is smaller than in Case I. Overall, the wiring area requirement only increases by 4% due to width variations. It is important to note here that the number and pitch of each metal

level is a variable in this methodology. In an interconnect design where these parameters are constant or in an SoC design where there are many minimum size metal levels, the impact of variations would be more severe as numerous wires would require being routed at metal levels that are wider in pitch causing routability problems [35].

2.5 Power Dissipation Analysis

The total power dissipation is calculated as a sum of the dynamic and leakage power dissipated in logic gates and repeaters, and the dynamic power dissipated in wires. For logic gates and repeaters, power dissipation is given by,

$$P_{dynamic} = \frac{\alpha}{2} W_{gate} C_{gate} V_{dd}^2 f_{clock}, \tag{7}$$

$$P_{leakage} = W_{gate} V_{dd} I_{leak}, \tag{8}$$

where α is the activity factor, W_{gate} is the size of the gate normalized to the minimum size, C_{gate} is the input capacitance of the minimum size gate, V_{dd} is the supply voltage, f_{clock} is the clock frequency of the system, and I_{leak} is the average leakage current for the minimum size gate. The dynamic power dissipated in interconnects is given by the equation,

$$P_{int} = \frac{\alpha}{2} c_{int} L_{total} \sqrt{\frac{A}{N_{sockets}}} \chi V_{dd}^2 f_{clock}, \tag{9}$$

where c_{int} is the interconnect capacitance p.u.l., L_{total} is the total length of all the interconnects in the logic core in gate socket lengths, and χ is a correction factor as described in [32] and given by $4/(fan\text{–}out + 3)$. The area of the logic core, A, and the number of gate sockets, $N_{sockets}$, are used to calculate one gate socket length in meters by $\sqrt{A/N_{sockets}}$. P_{int} is directly determined by the wire capacitance, which can be reduced by introducing low–κ dielectric materials. ITRS target for the dielectric constant reduces from 2.99 in 2012 to 2.23 in 2020.

Using these power dissipation models, interconnect dynamic power dissipation and the total power dissipated in the logic cores are plotted in Figure 9 assuming various size effect parameter values. Note that the interconnect power dissipation is a significant component of the total power dissipation in the logic core. Even though individual devices are targeted to become less power–hungry at each technology generation, the overall impact of the increase in the total interconnect power, number of logic gates and repeaters in the system is a rapid increase in the total power dissipated in a logic core in the future technologies.

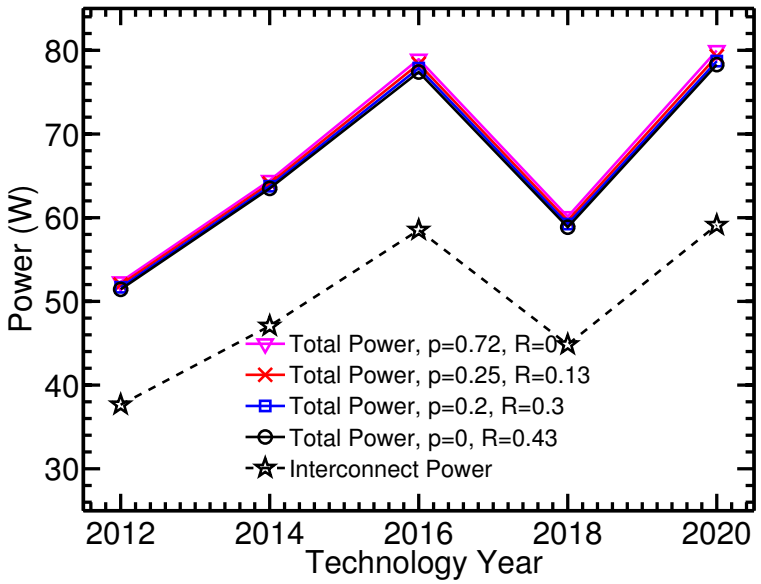

Figure 9: Interconnect and total power dissipation in the logic cores calculated from the optimal MIN design at various technology nodes considering a range of size effect parameters.

The number of transistors on a chip double every three years according to the ITRS projections, which results in the same transistor count on the chip in years 2016 and 2018. This explains the reductions in both the interconnect and total power dissipation in 2018. The interconnect power dissipation is a function of the capacitance p.u.l. and the aggregate interconnect length. The smaller core area in 2018 results in shorter total interconnect length; hence, smaller interconnect

power. Also, devices become less power–hungry at each technology generation, reducing both the dynamic and leakage power dissipation of gates. The number and size of repeaters in the system depend on size effect parameters. The total power of repeaters, however, depends only on the total capacitance of the repeated interconnects, which is determined by the aggregate length of repeater inserted wires in the system. Even though this aggregate length does depend on size effect parameters, the change in total length with size effects is observed to be small. Therefore, the impact of size effects on power dissipation is small as shown in Figure 9.

2.6 Conclusions

Optimizing the cost/area/performance of future interconnect systems is a complicated problem. In this chapter, to investigate the impact of interconnect technology parameters on the design and performances of future electronic chips, compact models are used to optimize MINs for FinFET circuits, where the wiring information of the system was obtained based on a stochastic wiring distribution model for faster simulation time. The system constraints such as the clock frequency, chip area and the maximum number of metal levels are adopted from ITRS and are used to determine the actual number of metal levels required while meeting system constraints.

The impact of Cu resistivity increase due to size effects on the individual wires and the overall design of the MIN is analyzed in detail based on the results of our optimal design methodology. It is shown that due to challenges in mitigating Cu size effects and scaling the barrier/liner thickness, the number of required metal levels may increase by 76% at the 7–nm technology node in 2020. Therefore, to meet ITRS projections for system clock frequency, area and the number of metal levels, finding scalable solutions to making thin barriers and mitigating size effects

are critical tasks.

Increasing the aspect ratio of wires decreases wire resistance p.u.l., but the total line delay also depends on the interconnect pitch, length, capacitance p.u.l., driver resistance, and size effect parameters. Although a larger aspect ratio may be beneficial in terms of the cost of the chip since our results indicate that there is an optimal aspect ratio that can reduce the required number of metal levels to route the system, this benefit comes at the expense of a larger line capacitance, which translates into larger interconnect dynamic power dissipation. Interconnects account for 60-70% of the total logic core power dissipation in the future technology nodes, which means that the total power dissipation of the chip can increase significantly with larger aspect ratio values.

In the next chapter, we use an alternative methodology, namely, using a real chip tape–out flow to evaluate the limitations of the Cu/low–κ interconnect technology for FinFETs in the future technology generations.

CHAPTER 3

ANALYSIS OF THE IMPACT OF COPPER/LOW–κ PERFORMANCE DEGRADATION ON CHIP PERFORMANCE BASED ON FULL–CHIP LAYOUTS

In this chapter, we investigate the impact of highly–scaled Cu/low–κ interconnects on the speed and power dissipation of multiple circuit blocks based on timing–closed, full–chip GDSII–level layouts with detailed routing. First, we build multiple standard cell libraries for 45– , 22– , 11– and 7–nm technology nodes and model their timing/power characteristics. Next, we pair these standard cell libraries with various interconnect files and build GDSII–level layouts for multiple benchmark circuits to study the sensitivity of the circuit performance and power dissipation to multiple interconnect technology parameters such as resistivity, barrier/liner thickness, and via resistance. We investigate the implications of slowing down interconnect dimensional scaling below 11–*nm* technology node.

3.1 Introduction

In Chapter 2 we focused on exploring various design options for the back–end–of–the–line (BEOL) architecture based on stochastic wiring distribution models [32]. In this analysis, we treated all wires equally [36, 37] and used compact models for estimating device and interconnect performance at the system level. This approach, while being extremely time–efficient, lacks the details from a real chip such as the critical path and layout information. In reality, not all wires on a chip are parts of critical paths and they are not all driven by the same type and size of drivers. In this chapter, we perform our analysis based on actual netlists and GDSII–level layouts with detailed routing instead of using stochastic models to predict wiring distribution. This methodology is more comprehensive in that it

follows the same design flow used for real chip tape–out and encompasses the diversity of interconnects in terms of length, functionality, and the type and size of drivers and receivers. The disadvantage for this approach is that it is very time consuming. To reduce simulation time, we concentrate our analysis on block–level circuits instead of a whole processor as in Chapter 2. Since interconnect size effects are more pronounced for local/intermediate–level wires, this approach is sufficient in determining the impact of interconnect parameters on the chip performance at future technology nodes.

Section 3.2 describes our design and analysis flow. Section 3.3 explains the interconnect and standard cell library preparation, introduces the interconnect scenarios that we are considering and tabulates the cell power and timing characterization results for these scenarios. In Section 3.4, we summarize our design results for multiple circuits for various interconnect technology scenarios. Section 3.5 focuses on the impact of via resistance. Section 3.6 studies an alternative path to BEOL scaling. Section 3.7 concludes this chapter.

3.2 Design and Analysis Flow

The design and analysis flow used in this chapter is illustrated in Fig. 10. The predictive libraries that are described here are created based on the Nangate 45 *nm* open cell library [62]. First, physical parameters to define device and interconnect layers for each technology node are determined based on the scaling trends projected by the ITRS [25]. Using these definitions, a library exchange format file (.lef), which has the layout information for the standard cells, and an interconnect technology file (.ict), which has the interconnect structure information, are created. These two files are used in generating a simple capacitance table file (.capTbl) to be used in early stages of the design and a more elaborate technology file (.tch) to be

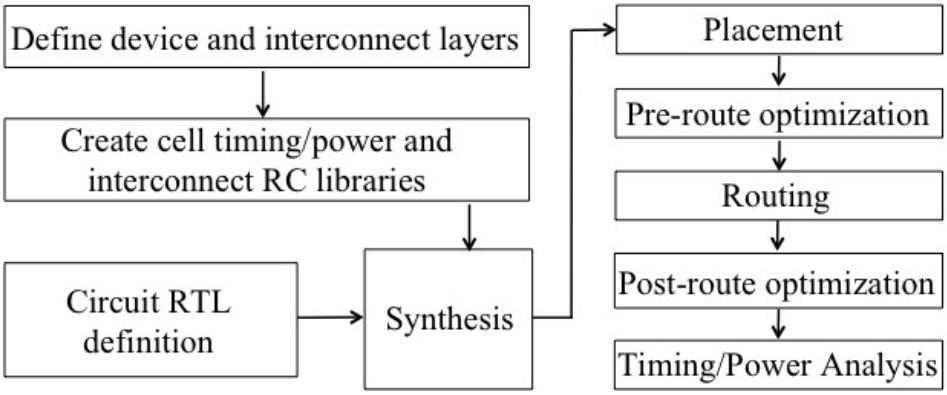

Figure 10: The overall design and analysis flow in this work.

used for accurate parasitic extraction after detailed routing. Appropriate predictive models for transistors and modified *RC*-extracted SPICE netlists for standard cells at each technology node are used to perform library characterization and generate predictive timing/power libraries (.lib, .db), which will be described in more detail later.

Using our predictive standard cell libraries, the RTL code of multiple circuit blocks are synthesized in Synopsys Design Compiler [63]. The placement, routing and optimizations are performed using Cadence Encounter [64].

3.3 Predictive Libraries

3.3.1 Interconnect Definitions

The interconnect structure and the layer dimensions are derived from the existing Nangate 45 *nm* library [62] assuming a scaling factor of roughly 0.7$\times$ at each new technology node as tabulated in Table 9. In all of the designs in this chapter, minimum number of metal levels required to route the design are used. Since we concentrate on circuit blocks instead of a full microprocessor, a maximum of 6 metal levels are used in our designs.

Table 10 demonstrates the effective resistivity values at these small dimensions calculated [29] considering the impact of size effects and the trench area lost to the

Table 9: Interconnect width (W) and thickness (T) at each technology node. All values are in nm.

	$45nm$		$22nm$		$11nm$		$7nm$	
	W	T	W	T	W	T	W	T
M1	70	130	35	65	17.4	32.5	10.8	20.2
M2:M3	70	140	35	70	17.4	35	10.8	21.8
M4:M6	140	280	70	140	35	70	21.8	43.6
M7:M8	400	800	200	400	100	200	62.2	124.5
M9:M10	800	2000	400	1000	200	500	124.4	311.2

barrier material normalized to the bulk Cu resistivity value ($1.8\mu\Omega \cdot cm$). Table 10 shows results for the same four cases of size effect scenarios as in Chapter 2. The barrier/liner thickness values are taken from the ITRS projections. Considering reliability issues at future technology generations, and challenges in scaling the barrier/liner thickness to the ITRS projected values, we also assume thicknesses of 3.5 nm, 3 nm and 2.5 nm at all metal levels of the 22-, 11- and 7-nm technology generations, respectively. These numbers are not hard numbers, but are projected to estimate the resistivity increase through a slower scaling path than the ITRS projections provided that the Cu ratio for the local metal levels are larger than or equal to 50%. The most pessimistic scenario of interconnect resistivity combines a thick barrier thickness with severe size effects.

Note that comparing the most optimistic (CASE V) and most pessimistic (CASE I) scenarios of size effects with ITRS projected barrier thickness, the effective Cu resistivity can increase by $\sim 2.95\times$ and $2.39\times$ for the local– and intermediate–level wires at the 7–nm technology node. Considering the aforementioned alternative scaling path for the thickness of the barrier/liner material can cause these values to go up to $6.52\times$ and $3.22\times$. It is therefore critical to quantify the sensitivity of circuit behavior to this resistivity change at the material level.

Table 10: Effective Cu resistivity values normalized to $1.8\mu\Omega \cdot cm$. Interconnect scenarios are listed in order of reducing severity.

Scenario	Metal Layer	Normalized Resistivity			
		$45nm$	$22nm$	$11nm$	$7nm$
CASE I	M1	-	5.1	12.98	29.47
$p = 0$, R=0.43,	M2:M3	-	5.05	12.8	28.97
thick barrier	M4:M6	-	2.67	4.73	7.75
CASE II	M1	2.81	4.49	7.75	13.29
$p = 0$, R=0.43,	M2:M3	2.79	4.44	7.67	13.13
ITRS barrier	M4:M6	1.84	2.53	3.85	5.75
CASE III	M1	2.3	3.44	5.63	9.36
$p = 0.2$, R=0.3,	M2:M3	2.29	3.41	5.56	9.24
ITRS barrier	M4:M6	1.62	2.09	2.98	4.26
CASE IV	M1	1.94	2.7	4.13	6.58
$p = 0.25$, R=0.13, ITRS	M2:M3	1.93	2.67	4.07	2.35
barrier	M4:M6	1.46	1.77	2.35	3.2
CASE V	M1	1.68	2.14	3.01	4.52
Single-crystal Cu, ITRS	M2:M3	1.67	2.12	2.96	4.44
barrier	M4:M6	1.35	1.54	1.89	2.41

3.3.2 Standard Cell Definitions

The predictive standard cell libraries that are used in this chapter are obtained by scaling the 45–nm library data [62]. For instance, the library exchange format (.lef) file for the original 45 nm library is modified using the dimensional scaling factors to generate .lef files for the predictive technology libraries. Similarly, to characterize the timing/power of the cells, the RC–extracted SPICE file from the Nangate 45 nm library is modified by: (1) changing the transistor model to the appropriate predictive models, and (2) modifying the cell internal parasitic resistance and capacitance values with appropriate scaling factors considering that the shape of the cells, hence the length and width of internal interconnects, is changed by the dimensional scaling factor.

Predictive Technology Models (PTM) for multi–gate transistors developed by

the Arizona State University [59] are used without modifying the nominal transistor parameters to characterize cells at each technology node. The cell internal capacitance values are scaled by the dimensional scaling factor assuming that the p.u.l capacitances do not change much. For instance, the internal cell capacitances in the 11–*nm* node become $\sim 0.25\times$ the original values in the 45–*nm* node. The cell internal resistance values are a function of the size effect parameters, the cross–sectional dimensions and barrier/liner material thickness.

The modified *RC*–extracted SPICE netlists for minimum size INV, NAND2 and DFF cells in the new libraries are characterized and the results are tabulated in Tables 11 and 12. The results are compared against the results for the Nangate 45 *nm* library counterparts of the cells under consideration and scaling factors are calculated for each parameter. The final timing/power libraries for all cells at each technology node are determined by modifying the original 45 *nm* library Liberty file (.lib) using an average scaling factor based on the results for the three aforementioned cells.

Note that the cell delay highly depends on the interconnect scenario at sub11–*nm* technology nodes. Considering a minimum size inverter and comparing the most optimistic and the most pessimistic scenarios for the interconnect resistivity, the cell delay increases by 18.1% and 44% at the 11– and 7–*nm* technology nodes, respectively. This moderate change is only due to within cell interconnects, which are short. The magnitude of this impact at the block level will be discussed in the next section.

3.4 Simulation Results

Using the libraries that are described in the previous section, we run full–chip layout experiments concentrating on three different categories of circuits. These three different categories of circuits are represented by an encryption circuit (AES), a

Table 11: Cell delays at various interconnect scenarios calculated at a medium input slew/output load case. Input slew=18.75ps (14.06ps for DFF), output load=0.64/0.88/1.76/3.2fF at 45/22/11/7-nm technology nodes, respectively.

Scenario	Cell	Delay(ps)			
		45nm	22nm	11nm	7nm
CASE I	INV	-	20.29	12.54	13.04
$p = 0$, R=0.43,	NAND2	-	24.33	14.57	15.55
thick barrier	DFF	-	48.37	23.4	24.34
CASE II	INV	43.55	20.28	11.62	11.6
$p = 0$, R=0.43,	NAND2	49.05	24.32	14.17	13.42
ITRS barrier	DFF	122.9	48.26	22.76	20.17
CASE III	INV	43.56	20.25	10.84	10.79
$p = 0.2$, R=0.3,	NAND2	49.05	24.32	13.74	12.71
ITRS barrier	DFF	122.82	48.08	22.62	19.66
CASE IV	INV	43.62	20.25	10.78	9.78
$p = 0.25$, R=0.13, ITRS	NAND2	49.17	24.3	13.6	12.39
barrier	DFF	122.76	47.94	22.24	18.94
CASE V	INV	43.61	20.24	10.62	9.06
Single-crystal Cu, ITRS	NAND2	49.17	24.29	13.53	11.49
barrier	DFF	122.77	47.55	21.83	18.04

Table 12: Cell characterization results for cell power, leakage, output slew and capacitance at a medium input slew/output load case as described in the caption of Table 11.

Cell Characteristics		Technology Node			
		45nm	22nm	11nm	7nm
INV	cell power (fJ)	0.445	0.203	0.064	0.074
	input cap. (fF)	0.463	0.346	0.169	0.126
	output slew (ps)	32.29	12.97	8.67	7.6
	leakage (pW)	2843	4311	3055	2438
NAND2	cell power (fJ)	0.669	0.178	0.081	0.063
	input cap. (fF)	0.523	0.233	0.116	0.084
	output slew (ps)	36.75	16.35	9.38	8.32
	leakage (pW)	4962	6019	3698	2907
DFF	cell power (fJ)	3.413	1.859	0.652	0.435
	input cap. (fF)	0.877	0.299	0.145	0.106
	output slew (ps)	35.37	11.17	4.32	3.55
	leakage (pW)	42965	42477	28832	22850

low–density parity check circuit (LDPC), and a Fast Fourier Transform (FFT) circuit. LDPC represents a wire–dominated group of circuits with a very high routing demand. FFT represents circuits with a highly regular layout. Most cells in the FFT circuits that communicate with each other are clustered together and there is a small number of connections between these smaller clusters. The third group of circuits whose regularity lie somewhere between the former two groups are represented by the AES circuit, which is a random logic circuit with a fair amount of routing demand. There are small clusters of cells within the AES circuit similar to FFT, but the communication between these smaller clusters are much higher compared to FFT. The placement and routing results for these three circuits considering a pessimistic scenario of interconnects as described in the previous section is illustrated in Figure 11.

For each design, we set the maximum target utilization to around 85%, but this number is adjusted in case of severe wiring congestions by changing the initial utilization during placement. For instance, due to the high wiring demand of the LDPC circuit, the initial utilization is lowered to 25% to increase the total footprint and have enough tracks to route the design. Similarly, the number of metal levels for each design are determined based on the wiring demand of the circuit. The minimum number of metal levels that ensures routability is used for each scenario. The minimum number of metal levels required to route each design are 4, 5 and 6 for FFT, AES and LPDC circuits, respectively.

By focusing on these three different circuits, we come up with generic conclusions regarding the impact of interconnect parameters on circuit power and performance. Timing is closed in all of the designs in this study for analyzing and comparing both the critical path delay and the power dissipation as described in more detail later. The simulation results for the AES, LDPC and FFT circuits are tabulated in Tables 13, 14 and 15, respectively.

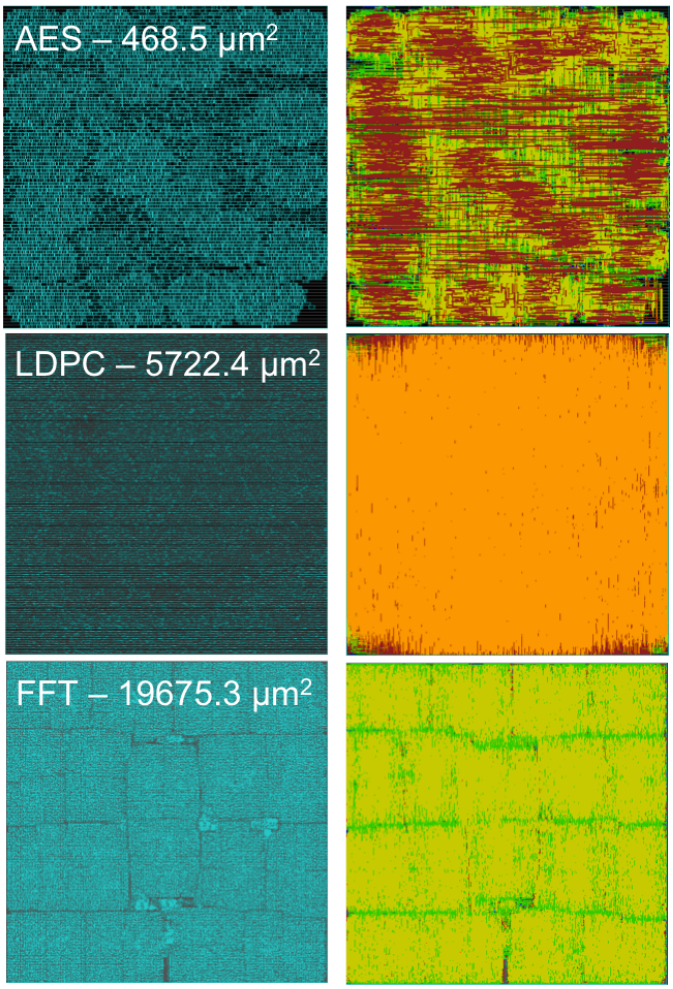

Figure 11: Placement and routing results for AES, LDPC and FFT considering a pessimistic scenario for interconnect size effects.

Table 13: Placement and routing results for all designs for the AES circuit at multiple technology generations and considering various size effect scenarios.

Circuit	Tech Node	Design Scenario	Min. Clock Period (ps)	Iso-performance Results									
				Target Period (ps)	Cell Count	Buffer Count	WL (mm)	Utilization (%)	WNS (ps)	Total Power (mW)	Net Switching (mW)	Cell Internal (mW)	Cell Leakage (mW)
AES	45nm	CASE II	714	714	17559	5121	198.1	84.9	0	18.35	9.9	8.03	0.422
		CASE V	710	714	16907	4818	200.7	81.53	0	18.05	9.802	7.849	0.403
	22nm	CASE I	236	236	20050	6538	87.66	86.22	0	20.4	9.703	10.18	0.517
		CASE II	226	236	19818	6379	86.81	86.21	+2	20.31	9.656	10.14	0.515
		CASE V	216	236	19818	6354	84.19	85.78	+7	20.15	9.51	10.13	0.511
	11nm	CASE I	164	164	17651	5725	40.95	85.93	+3	10.29	4.6	5.394	0.291
		CASE II	134	164	17257	5547	40.85	85.25	+17	9.696	4.507	4.899	0.29
		CASE III	126	164	17695	5518	44.03	86.11	+34	9.634	4.517	4.832	0.285
		CASE IV	118	164	17381	5219	41.54	85.42	+36	9.49	4.473	4.737	0.28
		CASE V	108	164	17411	5091	42.5	85.62	+41	9.396	4.46	4.671	0.265
	7nm	CASE I	202	202	17647	5769	26.53	86.83	+1	6.094	2.114	3.763	0.217
		CASE II	148	202	15908	4582	24.39	80.95	+29	5.362	1.875	3.334	0.153
		CASE III	120	202	15604	4537	24.33	80	+24	5.161	1.866	3.145	0.15
		CASE IV	110	202	14425	3855	26.72	77.22	+44	5.086	2.04	2.902	0.144
		CASE V	102	202	12382	2665	24.33	74.06	+40	4.457	1.801	2.531	0.125

Table 14: Placement and routing results for all designs for the LDPC circuit at multiple technology generations and considering various size effect scenarios.

Circuit	Tech Node	Design Scenario	Min. Clock Period (ps)	Iso-performance Results									
				Target Period (ps)	Cell Count	Buffer Count	WL (mm)	Utilization (%)	WNS (ps)	Total Power (mW)	Net Switching (mW)	Cell Internal (mW)	Cell Leakage (mW)
LDPC	$45nm$	CASE II	1260	1260	78047	28442	3927	34.619	0	178	124.7	51	2.222
		CASE IV	1100	1260	75051	26793	3825	33.341	0	167.5	117.7	47.82	2.044
	$22nm$	CASE I	620	620	60495	22092	1636	29.214	0	88.136	57.65	28.85	1.636
		CASE II	590	620	59844	18658	1642	29.034	0	86.097	57.25	27.25	1.597
		CASE V	500	620	57129	16601	1603	28.147	+2	81.76	53.82	26.54	1.405
	$11nm$	CASE I	570	570	45583	8711	796.4	27.68	0	30.28	19.81	9.67	0.798
		CASE II	390	570	43333	6987	777.4	26.86	+1	28.05	18.59	8.782	0.677
		CASE V	300	570	40975	5007	773.8	26.07	+1	26.48	17.68	8.227	0.576
	$7nm$	CASE I	680	680	50735	13744	510.37	29.76	0	19.19	10.04	8.39	0.752
		CASE II	470	680	45111	8699	519.7	27.89	0	16.96	9.79	6.597	0.567
		CASE V	280	680	39106	5178	472.9	26.22	+2	14.45	7.91	6.1	0.438

Table 15: Placement and routing results for all designs for the FFT circuit at multiple technology generations and considering various size effect scenarios.

Circuit	Tech Node	Design Scenario	Min. Clock Period (ps)	Iso-performance Results									
				Target Period (ps)	Cell Count	Buffer Count	WL (mm)	Utilization (%)	WNS (ps)	Total Power (mW)	Net Switching (mW)	Cell Internal (mW)	Cell Leakage (mW)
FFT	$11nm$	CASE I	480	480	231865	18754	1167.2	68.04	+2	154.847	61.98	88.14	4.727
		CASE II	350	480	230716	17716	1145.1	67.89	+7	153.783	61.32	87.76	4.703
		CASE V	280	480	230608	17502	1145.4	67.29	+21	150.999	59.53	86.8	4.669
	$7nm$	CASE I	590	590	236174	22881	698.65	68.439	+2	102.3	33.02	65.41	3.871
		CASE II	370	590	233350	20498	685.71	68.024	+10	100.19	32.01	64.33	3.849
		CASE V	240	590	231457	18473	678.52	67.562	+16	98.42	31.5	63.09	3.609

3.4.1 Impact of Size Effects on Critical Path Delay

To quantify the impact of the increase in wire resistivity at ultra–scaled dimensions on circuit speed, we run multiple simulations for each design to find the maximum clock frequency that each scenario of interconnect size effects can support. This is reported as the minimum clock period, which is calculated by gradually reducing the clock period until any further reduction results in a negative worst negative slack (WNS) value. For all the designs that are reported in this study, the minimum clock period value decreases if size effects can be mitigated from CASE I to CASE V in Table 10. The impact of interconnect size effects on the circuit speed increases as technology scales. At the 11–*nm* technology node and beyond, this impact increases drastically. For the AES circuit, the difference in the circuit speed comparing the most pessimistic and optimistic scenarios of size effects is as high as 52% and 98% at the 11–and 7–*nm* technology nodes, respectively. These values are 90% and 143% for the LDPC circuit, and 71% and 104% for the FFT circuit. Therefore, irrespective of the circuit size and type, there is a drastic reduction in circuit speed due to interconnect resistivity increase as dimensional scaling continues with each new technology node.

Another important conclusion from the critical path delay analysis is that the improvement in the intrinsic device speed at each new technology node translates into smaller and smaller returns in the circuit speed due to the effect of the wires. In fact, in all of the circuits that are studied here, the circuit speed degrades beyond the 11–*nm* technology node for severe size effect scenarios. Therefore, it is not enough to improve the device intrinsic properties beyond the 11–*nm* technology node to improve the circuit speed, and it is critical to mitigate size effects and find solutions to manufacture thin barrier/liner regions. For instance, the speed of the AES circuit will degrade by 10% from the 11–*nm* technology node to the 7–*nm* technology node if the interconnect size effects are as severe as CASE II for both

technology nodes. However, by mitigating size effects from CASE II to CASE IV during the shift to the 7–*nm* technology node, this circuit speed can be improved by 18% instead.

3.4.2 Impact of Size Effects on Power Dissipation

To quantify the impact of interconnect size effects on the power dissipation of our benchmark circuits, we ran iso–performance simulations for each design at the frequency that each circuit can support for all the experimental setups. This frequency corresponds to the minimum clock period value that is estimated for the simulations in CASE I during our analysis for the critical path delay. The power dissipation values are calculated based on a switching activity of 0.2 for primary inputs and 0.1 for sequential cell outputs. The three components of the total power dissipation are (1) the net switching power, which is the power dissipated in charging the interconnect capacitance and cell pin input capacitances, (2) the cell internal power, which is the power dissipated within each cell including the short circuit power, and (3) the cell leakage power. The percentage contribution of each of these components to the total power dissipation depend on the circuit. Also, unlike the critical path delay analysis results, our power dissipation analysis results indicate that the impact of interconnect size effects on total power dissipation highly depends on the circuit. However, for all of our benchmark circuits, this impact increases with technology scaling.

For the AES circuit, the total power dissipation monotonously increases as the interconnect resistivity is progressively worsened from CASE V towards CASE I. At each technology node, comparing the results for the most pessimistic and most optimistic interconnect scenarios shows that the power increases significantly at sub–11 *nm* technology nodes due to the degrading interconnect performance. The percentage increase in total power is 9.51% and 36.73% at the 11– and 7–*nm* technology nodes, respectively. Most of the change in the power occurs in cell internal

power, which is due to both the increase in the number of buffers in the system and the upsizing of some of the gates on the critical paths to meet timing constraints. In this comparison, the increase in the number of buffers is 12.45% and 116.5% at the 11– and 7–*nm* technology nodes, respectively. It is important to note that for CASE I, 1/3 of all the gates in the design are buffers. The net switching power, which is due to both the cell input capacitances and the total interconnect capacitance as defined before, is also affected by these changes through the insertion of extra input pin capacitance, but the overall impact is not as pronounced as for the cell internal power since the fraction due to the interconnect capacitance changes only slightly. The extra buffers and larger gates directly affect the change in the total cell leakage power as well, but the leakage power is a small component of the total power in this analysis.

The LDPC circuit results are similar to the AES circuit results in terms of the monotonous power dissipation increase with worsening interconnect performance. However, the power dissipation breakdown for the LDPC circuit is very different. The interconnect capacitance has a much more pronounced impact on the total power dissipation of the LDPC circuit compared to the AES circuit. Since this is a wire dominated circuit, the total interconnect capacitance is much larger compared to the total input pin capacitance. Therefore, the largest component of power is the net switching power, which is largely dominated by the interconnect power. As a result, although the interconnect distribution is not a function of the interconnect resistivity as strongly as the number of buffers or the gate sizes, any slight change in this distribution between designs has a larger impact on the net switching power; hence the total power, compared to the AES circuit. In short, due to the change in the weights of the impact of different parameters on the total power dissipation of the circuit, it is not reasonable to expect a larger power dissipation difference between interconnect scenarios for the LDPC circuit than the AES circuit

simply based on the critical path delay results. In fact, our results show that the percentage increase in total power when comparing CASE I and CASE V results for the LDPC circuit is 14.35% and 32.8% at the 11- and 7–*nm* technology nodes, respectively, which is not too different than the AES circuit results. This is true in spite of the fact that the percentage increase in the number of buffers is 73.97% and 165.4% at the 11– and 7–*nm* technology nodes, respectively. The significant difference in the impact of interconnects on the percentage change for the critical path delay for AES and LDPC circuits does not reflect to the power dissipation results in the same way due to the difference in the circuit type.

The FFT circuit is a much larger circuit compared to the AES and LDPC circuits. Therefore, the simulation time for the FFT circuit is much longer. To save simulation time, we have focused on the 11–*nm* technology node and beyond for the FFT circuit because those are the nodes where the more interesting changes occur. Having seen a monotonous change for both the critical path delay and power dissipation analyses in our previous benchmark circuits, we concentrate our efforts on only three cases of interconnect scenarios knowing that the results for the other cases will fall within the range of the results we get if we concentrate on the lower and upper extreme cases. Our results indicate that the significant change in the critical path delay is not at all translated to the results for the power dissipation in the FFT circuit. Comparing the two extreme cases, the percentage increase in total power is only 2.55% and 3.94% at the 11– and 7–*nm* technology nodes, respectively. This result is directly related to the regularity of the layout and the size of the FFT circuit. Since most of the cells that communicate with each other are placed closely by the routing tool to minimize the total wire length, which is indicated by the clear clusters of cells in Figure 11, and there are a small number of connections between these clusters, the cells on the critical path are a very small portion of this large circuit. Therefore, even at the 7–*nm* technology node, the percentage increase

in the number of buffers is only 23.9%.

3.5 Impact of Via Resistance on Performance

So far, it has been established that the interconnect RC delay increase with dimensional scaling causes the circuit performance to degrade in terms of both speed and power dissipation. In this analysis, so far, we have focused on the line resistance and have assumed optimistic values for the via resistances to isolate the impact of the line resistance on the overall system performance/power. Recently, it was shown [65] that via resistance has a significant impact on the circuit speed at the 7–nm technology node and needs to be considered in optimizing the BEOL architecture. This study is based on a circuit model considering an inverter driving a similar inverter through a variable–length, horizontal interconnect at the third metal level. In this section, we take into account the resistance increase of the via structure due to both dimensional scaling and possible misalignment issues to the underlying metal layer and investigate the impact of the via resistance on circuit performance based on timing closed GDSII–level layouts.

We used Synopsys Raphael [66] to estimate the via resistance at the 7–nm technology node for both the ideal and misaligned via structures. The simulation structure is illustrated in Figure 12. The barrier material resistivity is assumed to be $500\mu\Omega \cdot cm$ [67]. The Cu resistivity is calculated according to interconnect scenario CASE I as defined before. The horizontal run length, L_z, for the top, M_U, and bottom, M_L, metal levels are assumed to be very small to avoid any impact on the final estimated via resistance value. The misalignment length, L_{mis}, is calculated as a percentage of the ideal via width and is varied from 0 to 50% of the width value. The vertical length of the via, L_{via}, is based on the layer definitions as determined during library construction. The impact of misalignment on V1-V3 resistance values are tabulated in Table 16. Via dimensions and resistance values

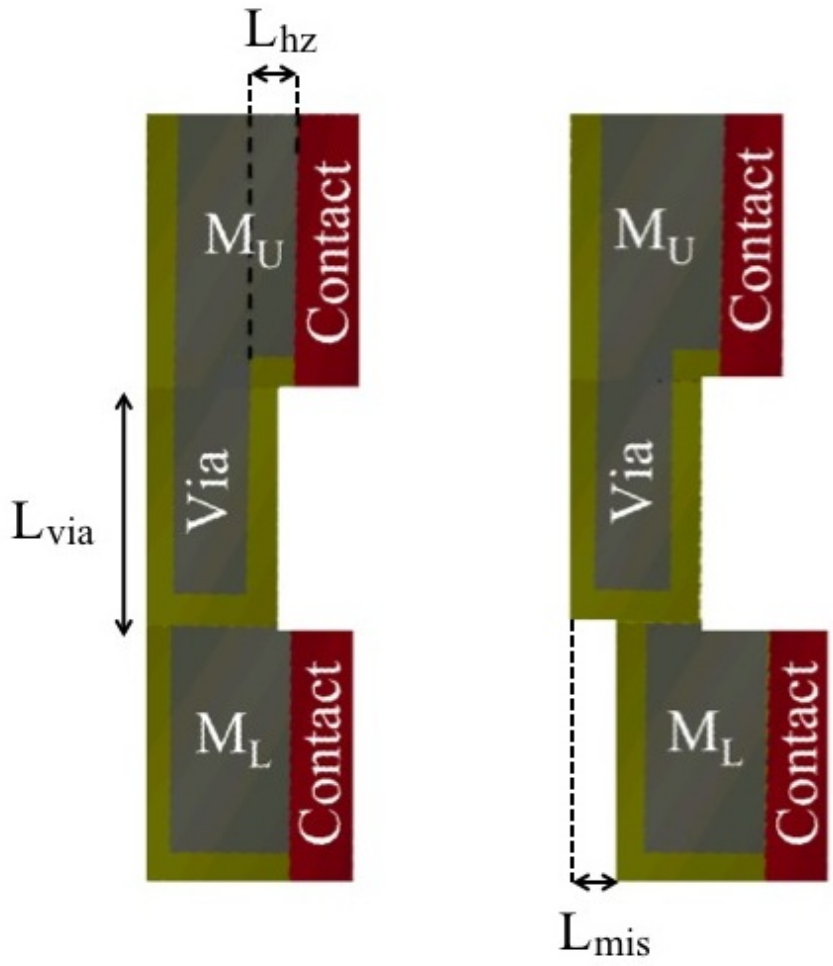

Figure 12: The simulated structures for well-aligned and misaligned via structures at the 7–*nm* technology node.

for all via layers are tabulated in Table 17 for three different cases considering very optimistic resistance values that we used to isolate line resistance changes so far, a realistic scenario for well–aligned vias and a 50% misaligned via scenario.

To perform the experiment for quantifying the impact of via resistance on the circuit performance/power at the 7–*nm* technology node, we focus on the design results for the AES circuit from the previous section under interconnect scenario CASE I. If the exact same netlist is used to recalculate the critical path delay of the circuit, we observe a 18.57% increase between CASE A and CASE B results. As

Table 16: V1-V3 resistance values at the 7–nm technology node

Misalignment (%)	0	10	20	30	40	50
Resistance (Ω)	311.4	313.9	333.21	360.6	397.9	456.2

Table 17: Via Dimensions and Resistance Values

	Width (nm)	L_{via} (nm)	Resistance (Ω)		
			Optimistic CASE A	Ideal CASE B	50% Misaligned CASE C
V1-V3	10.8	18.7	24.08	311.4	456.2
V4-V6	21.8	45.1	14.25	30.46	67.18
V7-V8	62.2	127.6	0.68	2.97	4.1
V9	124.4	311.2	0.41	0.98	1.28

a result, the WNS for this design goes from +1 ps as shown in Table 16 to -35 ps. This method of comparison is similar to the discussion in [65] as the design is not reoptimized considering the new set of via resistance values and the critical path is assumed to stay the same. For a fair comparison, however, the correct set of via resistance values have to be taken into account during the design process, so the timing–driven placement and routing can be performed more accurately for each scenario. Design results for the AES circuit, which take into account the correct via resistance values are tabulated in Table 18. The results for CASE C are a worst–case corner analysis for via misalignment.

Note that the isolated impact of the via resistance on circuit speed in this scenario is only 3.96% between CASE A and CASE B. Therefore, the timing–driven placement and routing tools can compensate for the increasing via resistance if the correct values are provided during the design process. For instance, as the via resistance is increased from CASE A towards CASE C, the number of vias per standard cell in the design reduces and the total wirelength increases. This means that the placement and routing tools work to use a smaller number of vias even though the number of standard cells in the design increases, mainly due to a larger number of buffers, while running longer wires to connect them. Therefore, it can

be concluded that the trade–off between using shorter wires to connect two points by changing the metal layer through a via and using a slightly longer wire for the same connection avoiding a via connection shifts towards the latter option as via resistance is increased. In short, the overall impact of via resistance comparing CASE A and CASE C results for the AES circuit design is to reduce the maximum circuit speed by 13.86% and to increase the total power dissipation by 13.69%.

Table 18: Placement and routing results for the AES circuit under multiple via resistance scenarios.

Design Scenario	Min. Clock Period (ps)	Iso-performance Results									
		Target Period (ps)	Cell Count	Via Count	Buffer Count	WL (mm)	WNS (ps)	Total Power (mW)	Net Switching (mW)	Cell Internal (mW)	Cell Leakage (mW)
CASE A	202	230	17457	124681	5744	24.39	+9	5.246	1.986	3.069	0.191
CASE B	210	230	17736	121695	5801	25.35	+1	5.805	2.002	3.607	0.196
CASE C	230	230	18011	121641	6177	26.73	0	5.964	2.009	3.744	0.211

3.6 Alternative Path for BEOL Scaling

Based on the discussion so far, one idea can be to follow an alternative, slower path for BEOL scaling beyond the 11–*nm* technology node. In this section, we investigate the implications of using the 11–*nm* technology node BEOL design with the 7–*nm* technology node FEOL. In other words, we assume that during the shift from the 11–*nm* technology node to the 7–*nm* technology node, the device dimensions can be shrinked and the intrinsic device performance is improved, but to avoid the significant performance degradation at both the cell and system level due to interconnects, the BEOL dimensions are not scaled. To study this scenario, we follow the same library construction flow described before and design the AES circuit with this new experimental library. The within cell interconnects (M1) is the only metal level that is scaled to the 7–*nm* technology dimensions in this analysis. This study is performed for an optimistic interconnect resistivity scenario (CASE V).

The results indicate that the major problem with this approach is the routing congestions due to the small dimensions of the cells that are being connected by wide wires. Compared to the all–11–*nm* technology node, there is a 2.67$\times$ reduction in the footprint of the circuit with this approach, while the number of pins to connect stays almost unchanged. The high pin density gives rise to the wiring congestion and design rule violations as illustrated in Figure 13. To overcome the congestion problem, multiple solutions can be tried. The design can be slowed down to reduce optimization steps including the insertion of buffers, breaking down of complex cells and upsizing of gates, all of which increase either the total pin density or the silicon area utilization. Furthermore, extra metal levels can be added or total chip area can be increased to provide more supply for the increasing routing demand. Increasing the chip area is not a preferred solution due to cost reasons. Clearly, the question at hand is an optimization problem with many parameters to consider while designing the BEOL architecture, which will directly impact the

cost/area/performance of the chip. The implications of using an 11–nm BEOL architecture with a 7–nm FEOL without changing the area of the chip or the number of metal levels compared to the all–7–nm technology node are tabulated in Table 19.

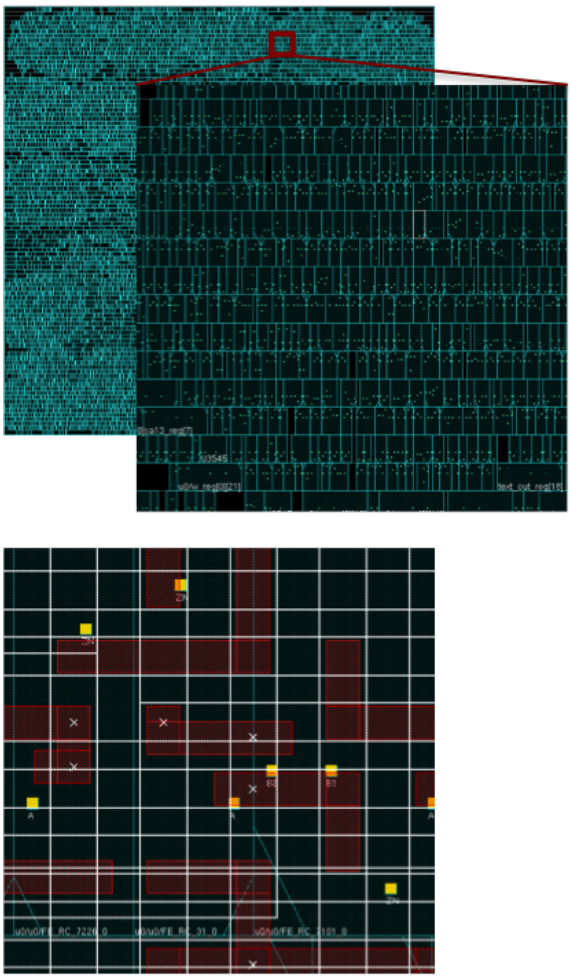

Figure 13: Placement density for the AES circuit assuming $7nm$ FEOL + $11nm$ BEOL structure and the routing congestions at M2.

Note that if the 11–nm BEOL technology is used with the 7–nm FEOL in the AES design (row 2) without changing the area and the number of metal levels compared to the all–7–nm technology (row 1), the speed of the circuit needs to be reduced to $\sim$ 0.5$\times$ its value in the original all–7–nm technology design to avoid routing congestions and design rule violations. This way, the number of buffers

Table 19: Design results for the AES circuit using the 7–*nm* technology node FEOL with 7–and 11–*nm* BEOL options with 5 metal levels.

BEOL Technology	Min. Clock Period (ps)	Buffer Count	Total Power (mW)
$7nm$	102	5527	13.41
$11nm$	200	1467	4.706
$7nm$	200	2665	4.457

in this design is much smaller, which reduces the pin density. Also, if the original all–7–*nm* technology were to be operated at this smaller frequency (row 3), it would have dissipated less power. Therefore, slowing down the BEOL scaling to slow down resistivity increase associated with the wires degrades both circuit performance and power dissipation due to congestion problems.

Another solution to overcoming congestion issues is to increase the routing capacity by adding extra metal levels. Additional metal layers will add to the cost of the chip, but may improve performance. In this study, we compare two cases: (1) add an extra local metal level at the 7–*nm* technology node local metal dimensions, (2) add an extra intermediate metal level at the 11–*nm* technology node metal dimensions. As a result, the former scenario (CASE 1) has scaled M1-M2 whereas M3-M5 are adopted from the 11–*nm* technology node BEOL structure and the latter scenario (CASE 2) has scaled M1 whereas M2-M6 are adopted from the 11–*nm* technology node BEOL structure. The results are tabulated in Table 20.

Table 20: Design results for the AES circuit using the 7–*nm* technology node FEOL with 7–and 11–*nm* BEOL options with extra metal levels.

BEOL Tech.	T_{min} (ps)	Cell Count	Buffer Count	Footprint (μm^2)	Utilization (%)
Original	102	17851	5527	469.11	86.8
CASE 1	150	17490	5398	468.47	85.6
CASE 2	180	10135	1499	467.19	65.7

Using another scaled local metal level is clearly the better option as it can provide enough routing capacity to increase the silicon area utilization such that a

large number of buffers can be inserted to increase circuit speed. However, the minimum clock period is still $\sim 50\%$ larger than the original all–7–*nm* technology node results. Adding a new intermediate level does introduce some extra routing capacity, but it is not effective enough as indicated by the lower utilization and small buffer count, which result in a slow circuit speed. Therefore, slowing down the BEOL architecture dimensional scaling to compensate for the significant resistivity increase of the wires and the performance degradation that it brings is not a trivial question. During the shift from the 11–*nm* technology node to the 7–*nm* technology node, the wire pitches of the metal levels need to be carefully optimized to maintain routability while trying to avoid performance degradation due to interconnects.

3.7 Conclusions

In this chapter, we built multiple predictive cell libraries down to the 7–*nm* technology node to enable early investigation of the electronic chip performance using commercial electronic design automation (EDA) tools. Using these libraries, we quantified the impacts of inter– and intra–cell interconnect technology parameters on the speed and power dissipation of multiple circuit blocks at future technology nodes based on GDSII–level layouts of three circuits with different wire demand and layout structures.

We showed that the line resistance increase can hinder the circuit performance improvement during the shift from the 11–*nm* technology node to the 7–*nm* technology node. We also showed that via resistance becomes a significant contributor to circuit delay at the 7–*nm* technology node, but the placement and routing tools can in part compensate for its impact if the correct via values are taken into account during design. We investigate possible issues in slowing down the BEOL

scaling below 11–*nm* technology node to alleviate the resistance increase. Our results indicate that simply slowing down the BEOL scaling to compensate for the resistance increase associated with interconnects is not an effective solution as it introduces congestion issues, which degrades for performance and power dissipation of circuits. A more effective solution would require optimizing not only the interconnect structure, but the standard cell library as well.

CHAPTER 4

OPPORTUNITIES FOR SWNT INTERCONNECTS AT THE END OF THE ROADMAP

As the results presented in Chapter 3 illustrated, the performance degradation with dimensional scaling of short local– and intermediate–level interconnects that are used to route connections within circuit blocks has an increasing negative impact on both the performance and dynamic power dissipation of ICs. In this chapter, based on these results, it is shown that the historical trend of achieving smaller interconnect latency for short local and intermediate level interconnects with technology scaling will not hold true for future technology nodes. Therefore, new opportunities that rise as a consequence of this radical change in the nature of the interconnect problem are investigated. Contrary to the previous studies, which have indicated that individual single–wall carbon nanotube (SWNT) interconnects are too resistive for high–performance CMOS applications and must be used in bundles, it is demonstrated that they can offer significant delay and energy–per–bit improvements in high–performance circuits at the end of the roadmap. Performances of various design scenarios that comprise one or a few parallel individual SWNT interconnects are compared against the performance of the conventional Cu/low–κ interconnect technology at future technology nodes using delay, energy–per–bit and EDP as metrics.

4.1 Introduction

Historically, the delay of short local and intermediate interconnects has been much smaller compared to the delay of switches and has scaled with technology. The

delay of short interconnects has been determined by the output resistance of transistors and interconnect capacitance. The length of long global interconnects, however, did not scale with technology scaling since they ran across the chip. The delay of repeated global interconnects remained constant resulting in an increasing delay trend compared to gate delays. Therefore, global interconnects were thought to be the more serious interconnect problem [68, 9, 69]. As the results in Chapter 3 indicated and as shown in this chapter, this historical trend of constraining the interconnect problem to the long interconnects at the global level will not be true for future ultra scaled technologies.

Below 20 *nm* interconnect width, the delay of short local interconnects can no longer be determined by just the output resistance of transistors and interconnect capacitance because of the aforementioned dramatic increase in metal resistivity that stems from size effects and process variations. For ultra–scaled technology nodes, the minimum size interconnect resistance p.u.l becomes comparable to that of individual metallic SWNTs. Therefore, new opportunities arise for using individual SWNTs for interconnect applications in high–performance circuits at such highly scaled technology nodes.

Carbon nanotubes (CNTs) have long been considered as a promising alternative material for future nanoscale interconnects due to their long mean free path (MFP), high current carrying capability and high thermal conductivity. Previous studies have shown that individual SWNTs are too resistive for interconnect applications in high performance chips [14, 70] and that they can potentially be used only in ultra–low power circuits [71]. Therefore, in order to reduce the high resistance associated with SWNTs, researchers have concentrated on manufacturing bundles of SWNTs that conduct current in parallel [72, 73, 74, 75, 76, 77]. Taking advantage of the fact that SWNT bundles tend to grow perpendicular to a surface [78], researchers have manufactured both vertical and horizontal bundles of CNTs.

As a result, there has been significant progress in on–chip integration of vertical bundles of CNTs as vertical interconnects and using them as vias. Growing long, dense bundles of horizontal metallic SWNTs and making reliable connections to every tube in the bundle, however, has proven to be very challenging [79, 80, 81]. Many research groups have concentrated on solving this problem. Even though the low catalytic activity of CNT synthesis is overcome in [79] and 84% catalyst activity is reported, the SWNT bundles grown in this work are very sparse. Although 87% metallic SWNT bundle is reported in [80], the density of the bundle after separation is very low and it is not suitable for large–scale integration. Recently, an electric–field induced alignment approach was taken to grow horizontal bundles of multi– and few–walled CNTs for interconnect applications [81]. The length of the CNTs in [81] is only 6 μm and it is reported that after a certain critical length, the electric–field distribution will not favor the growth of horizontally aligned CNTs. In short, manufacturing horizontal bundles of SWNTs remains as a challenging task.

On the other hand, there has been significant progress in wafer–level fabrication of perfectly aligned individual SWNTs with high densities on single crystal wafers, such as quartz and sapphire, [82, 83, 84, 85, 86, 87], and in turning semiconducting tubes to metallic by Platinum nanocluster decoration [88]. Considering these advances in manufacturing well–aligned metallic SWNT interconnects and the aforementioned change in the behavior of local/intermediate level copper interconnects, the RC delay of individual SWNT interconnects are compared with that of the conventional Cu/low–κ interconnect technology.

In this chapter, we quantify the potential improvements that can be achieved by using various single wall carbon nanotube (SWNT) interconnect designs considering the impact of broken tubes. In Section 4.2, main assumptions for Cu and

SWNT interconnect are described. Intrinsic interconnect metrics, such as the resistance, capacitance, delay and EDP of minimum–size Cu wires are compared against various SWNT interconnect designs at the future technology nodes considering the effect of broken tubes. In Section 4.3, a complete circuit analysis is presented and the impact of variation in kinetic inductance of CNTs on circuit performance is evaluated. The requirements for various SWNT interconnect metrics for outperforming copper interconnects are tabulated and compared against what has been achieved so far. This chapter is concluded in Section 4.4.

4.2 Intrinsic Interconnect Metrics

4.2.1 Assumptions and Technology Parameters

In this chapter, the conventional Cu/low–κ interconnect technology configuration shown in Figure 14(a) has been used as the reference structure in comparing the relative performances of various SWNT interconnect designs. The ITRS update of 2010 [25] is used to estimate the values of interconnect minimum pitch, aspect ratio, interlayer dielectric constant, and conformal barrier thickness. The resistance p.u.l. of copper interconnects in this work is calculated [29] assuming a 40% LER and choosing reflectivity (R) and specularity (p) parameters of 0.5 each [51]. The capacitance p.u.l. of Cu interconnects are calculated using the electrostatic simulator, RAPHAEL [66].

For SWNT interconnects, it is assumed that a bed of tubes is first laid on the substrate and then the tubes in certain regions are etched away using lithography techniques. As a consequence, there is maximum control over the pitch of SWNT interconnects, but the tubes are randomly placed as illustrated in Figure 15.

The resistances p.u.l. of the metallic tubes are calculated using the models in [14], [89]. In this section, it is assumed that the interconnect length is larger than the mean free path of electrons such that the impact of quantum resistance in the total resistance is minimum. Impact of quantum resistance is considered in the

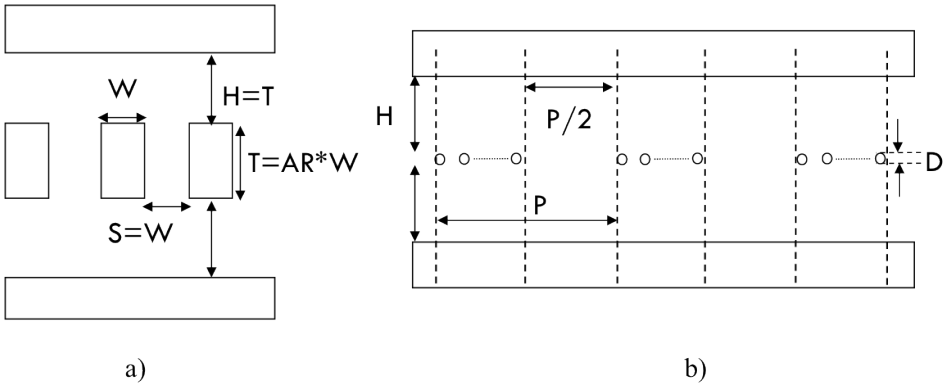

Figure 14: (a) Reference Cu interconnect configuration considered in this paper. W, T, S and H stand for the interconnect width and interconnect thickness, spacing between interconnects and the interlayer dielectric thickness, respectively, (b) Few SWNTs interconnect configuration. P stands for the interconnect pitch and D stands for the tube diameter. Tubes are assumed as randomly distributed in consecutive regions of half a pitch separated by forbidden regions of the same width.

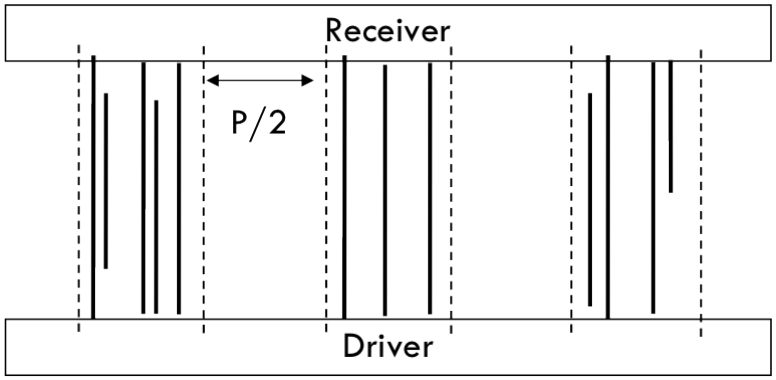

Figure 15: Top view of the SWNT interconnect configuration. Tubes are randomly placed. In this work, considering the advances in manufacturing long, dense and well–aligned SWNTs, we assume that the lengths of the tubes are homogeneous, but they may be broken at a random location along the length and the distance between consecutive tubes may vary.

next section. The mean free path of electrons in a metallic SWNT is assumed to be linearly dependent on the diameter of the tube [90]. For interconnect applications, where low–bias mean free path of electrons is of interest [89], it is assumed that a metallic SWNT with a diameter of 2 nm will have a mean free path of 2 μm [91]. To calculate the capacitance p.u.l. for various configurations of SWNT interconnects, field–solvers RAPHAEL [66] and COMSOL [92] are used.

4.2.2 Resistance, Capacitance, RC delay and EDP Trends

To investigate rising opportunities for emerging interconnect technologies, the intrinsic properties such as the resistance and capacitance associated with the new technology have to be evaluated. Compared to the resistance p.u.l. of SWNT interconnects, which is almost constant with technology, Cu interconnect resistance p.u.l. has a steep upwards trend as minimum dimensions are reduced as shown in Figure 16. As a consequence, the huge gap between the resistance p.u.l. of Cu and that of SWNT interconnects reduces quickly and vanishes at the 7.5 nm minimum wire width. Utilizing a larger number of tubes that conduct current in parallel means that this intersection will occur at an earlier node. For instance, if SWNT interconnects with three tubes are considered, their resistance p.u.l. intersects that of Cu interconnects at the 11–nm minimum wire width. Another important conclusion that can be derived from Figure 16 is that the diameter of the tubes has to be equal to or larger than 2 nm as tubes with a diameter of 1 nm are still too resistive for high performance applications. A maximum of three tubes in SWNT interconnects is considered in Figure 16 simply because there is not enough room to place more tubes at the end of the roadmap [25].

Furthermore, replacing Cu wires with SWNT interconnects with a single tube, two tubes and three tubes can reduce the average capacitance p.u.l. by 3.44×, 2.23× and 2.17×, respectively, as plotted in Figure 17.

At very large dimensions, it is expected that the total capacitance for a three–tube

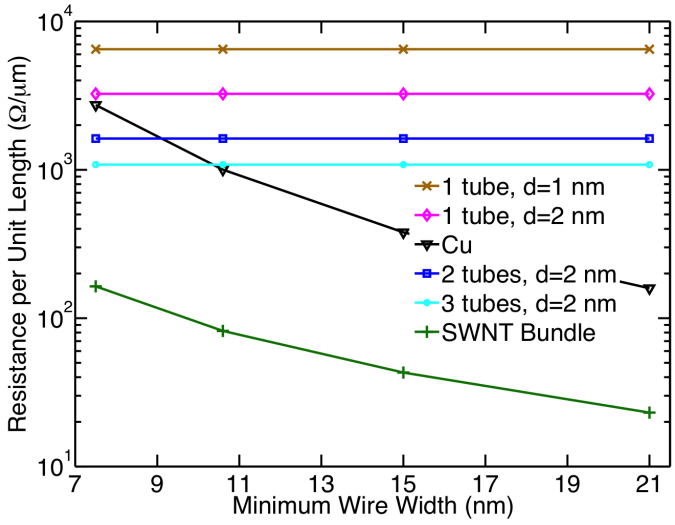

Figure 16: Comparison of the resistance p.u.l. associated with Cu interconnects and SWNT interconnects considering various number and diameter of tubes in a single layer. The resistance p.u.l. for a SWNT bundle of 1 nm diameter tubes is also shown as reference, where it is optimistically assumed that the density of metallic tubes in the cross–section of the bundle is $1/3nm^2$ [89], higher than the Van der Waals limit of only $1/4.5nm^2$.

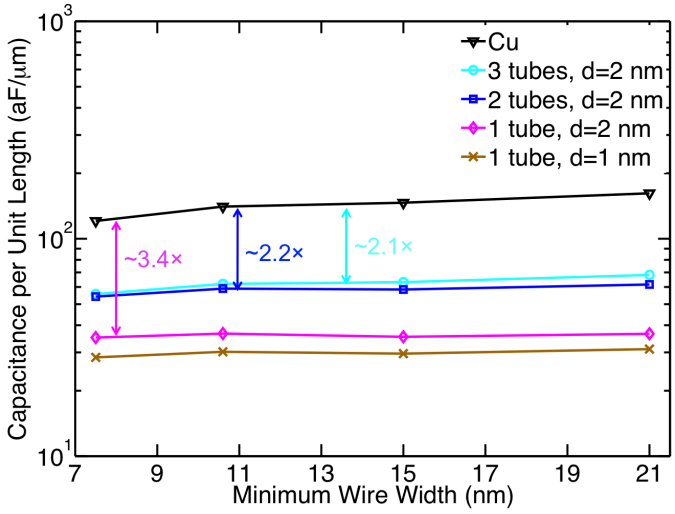

Figure 17: Comparison of the capacitance p.u.l. associated with Cu interconnects and SWNT interconnects considering various number and diameter of tubes in a single layer. The capacitance p.u.l. for a SWNT bundle is the same as the Cu interconnects [89].

design will be 1.5× larger than a two–tube design. At small dimensions, however, the degradation in capacitance p.u.l. is not linearly dependent on the number of tubes in parallel due to significant electrical shielding between the tubes. This is illustrated in Figure 17, where the capacitance p.u.l. of SWNT interconnects with two and three tubes in parallel are quite similar. As more and more tubes are placed in parallel, the capacitance p.u.l. will converge to the capacitance of a metal plate with a thickness equal to the diameter of the tubes. The results illustrated in Figure 17 are for the average capacitance of SWNT interconnects. In calculating these capacitances, it is assumed that the tubes in the design are placed as far from each other as possible, thus minimizing the effects of electrical shielding and maximizing the total capacitance.

It is demonstrated so far, that SWNT interconnects can offer similar or better resistance p.u.l. and much better capacitance p.u.l. values compared to Cu interconnects. Based on this discussion, a better comparison in terms of the performance of an interconnect technology is the *RC* product p.u.l. squared. Figure 18 demonstrates that the *RC* product of individual SWNT interconnects intersect with that of Cu interconnects at around 11–*nm* minimum wire width and SWNTs offer better *RC* products beyond this technology node. Similar to the discussion with resistances, this intersection point can be pulled to earlier technology nodes if multiple tubes are used in the design.

In addition to the *RC* delay improvements that SWNT interconnects can offer over Cu interconnects, they can reduce the energy–per–bit in a system significantly as their intrinsic capacitances are very small. In Chapter 2, it was shown that even though local interconnects in a MIN are short, they account for a significant fraction of the total interconnect power. Thus, this reduction in interconnect capacitance can potentially translate into significant savings in total interconnect power dissipation. Therefore, another important metric to consider is the EDP

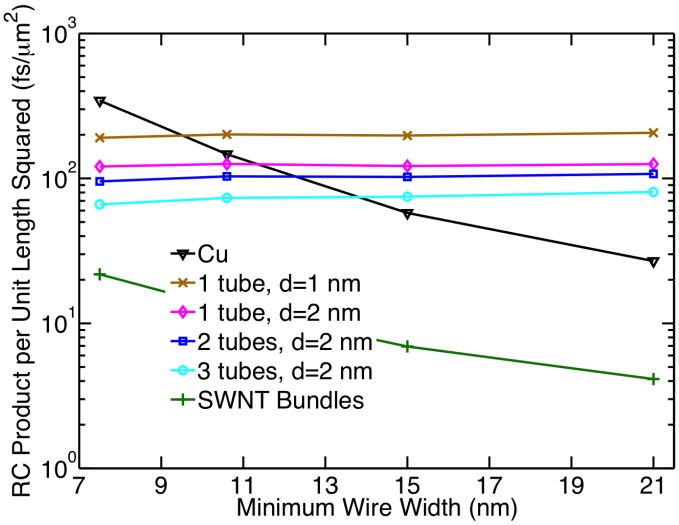

Figure 18: Comparison of the *RC* product p.u.l. squared associated with Cu interconnects, bundles of SWNT interconnects and SWNT interconnects considering various number and diameter of tubes in a single layer. The bundles are the same size as Cu interconnects and the density of metallic tubes in the cross–section of the bundle is assumed to be $1/3nm^2$ [89].

p.u.l. cubed of each of these designs. As Figure 19 illustrates, significant savings can be achieved with SWNT based interconnects at the end of the roadmap in terms of EDP as well. EDP for multiple–tube SWNT interconnect designs is only slightly worse than that of individual tubes.

So far, all tubes in a SWNT interconnect design are assumed well connected to both the driver and the receiver. However, any of these tubes may potentially be broken along the length, which results in a worse *RC* delay than expected. In Figure 20, the *RC* product of Cu interconnects is compared to the worst–case *RC* product of SWNT interconnects with two and three tubes.

It is observed that the worst case for a design with two tubes occurs when one of the tubes is broken close to the driver. Tubes that are broken at the driver side significantly slow down the line since they are not connected to the source. They introduce extra load to the system through the coupling capacitances. If a tube is broken at the receiver end, however, the line is effectively charged by the

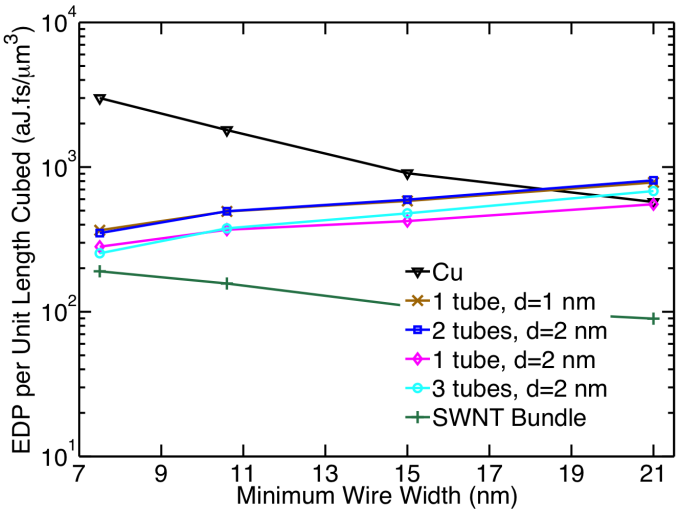

Figure 19: Comparison of the EDP p.u.l. cubed associated with Cu interconnects, bundles of SWNT interconnects and SWNT interconnects considering various number and diameter of tubes in a single layer. The bundles are the same size as Cu interconnects and the density of metallic tubes in the cross–section of the bundle is assumed to be $1/3nm^2$ [89]. SWNT interconnects with 3 parallel tubes can perform almost as good as SWNT bundles in terms of EDP.

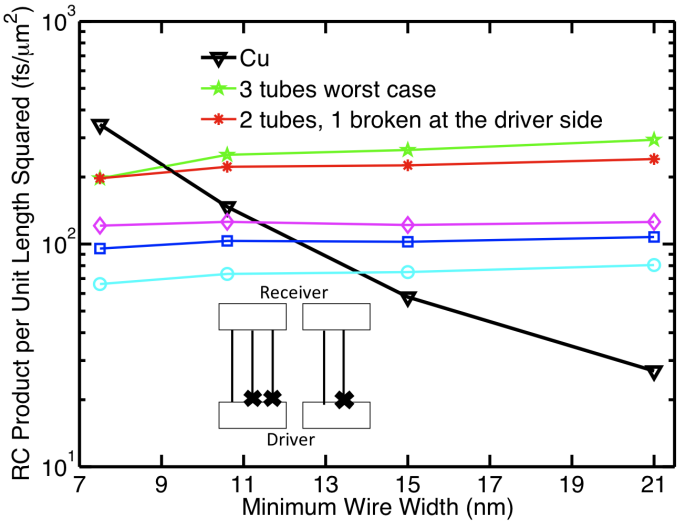

Figure 20: Comparison of the *RC* product p.u.l. squared associated with Cu interconnects and SWNT interconnects considering various number of tubes in a single layer and the effect of possibly broken tubes. Only the worst case is plotted when the impact of broken tubes are considered.

source and the loading effect is minimized. Similarly, the worst case for a design with three tubes occurs when only one of the outer tubes is well connected to both the driver and the receiver whereas the other two are broken at the driver side. Comparing Figures 18 and 20 , it can be seen that broken tubes can significantly degrade interconnect *RC* delay performance. EDP is also degraded due to broken tubes as shown in Figure 21.

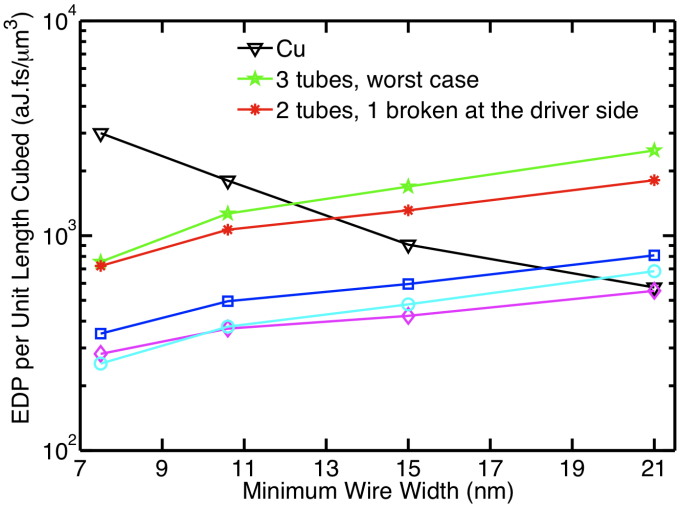

Figure 21: Comparison of the EDP p.u.l. cubed associated with Cu interconnects and SWNT interconnects considering various number of tubes in a single layer and the effect of possibly broken tubes. Only the worst case is plotted when the impact of broken tubes are considered.

In terms of either metric, however, multiple SWNT interconnect designs can outperform Cu interconnects at the end of the roadmap, even when the worst–case scenario of broken tubes is considered. If all the tubes in a design are broken at some point along their length, then the signal will not be carried from the driver to the receiver and a failure in communication will occur.

4.3 Complete Circuit Analysis

To compare the potential performances of various SWNT interconnect designs, we assume an inverter driving 3 similar inverters through an interconnect whose

length is varied. Equivalent circuit models for SWNT interconnects including quantum capacitance and kinetic inductance are utilized as presented in [14, 89]. The complete simulation circuit is illustrated in Figure 22 for a three-tube SWNT interconnect design. ITRS projections for the half pitch of the first metal level at the end of the roadmap, which is 7.5 nm in year 2024, is assumed for all comparisons.The interconnect parameters for SWNTs are tabulated in Table 21. For Cu interconnects, resistance and capacitance p.u.l. values are calculated as 2.73 $K\Omega/\mu m$ and 118.02 $aF/\mu m$, respectively. ITRS projections for the ON resistance and input capacitance are assumed in calculating driver parameters.

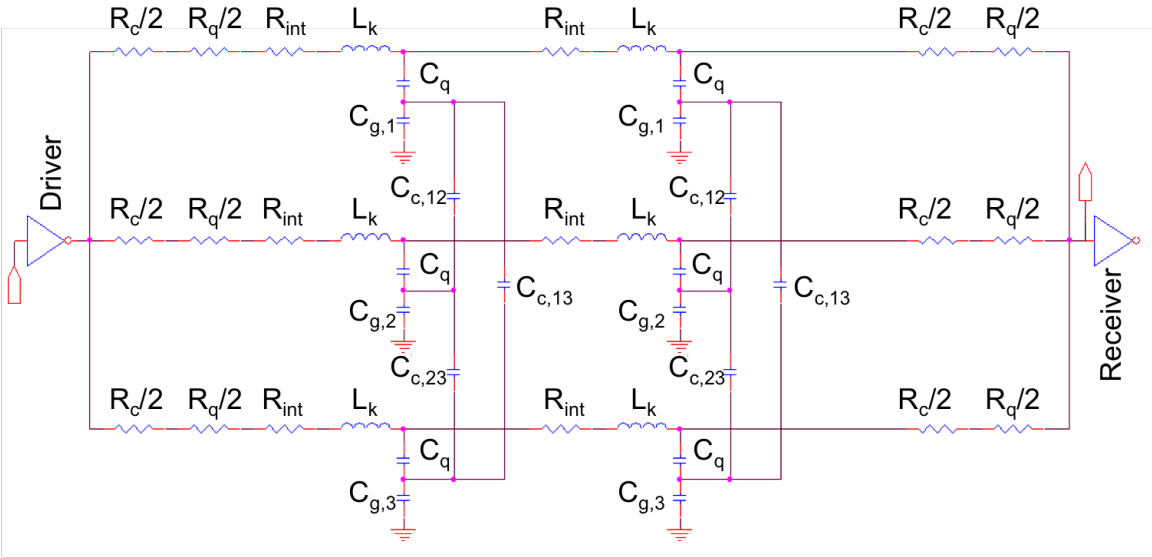

Figure 22: The schematic for the complete circuit simulated in HSPICE shown for a three-tube SWNT interconnect design.

Most gates in a high–performance circuit are larger than the minimum size. In this study, we compare interconnect performances considering 5× the minimum–size gates. As a consequence, the resistance of SWNT interconnects become the important parameter due to the large capacitive load at the receiver side. The

Table 21: Driver and interconnect parameters for high–performance circuits at the 7.5–nm technology node.

Symbol	Quantity	Value
$R_{driver}(K\Omega)$	minimum–size driver output resistance	~ 19
$R_c(K\Omega/tube)$	contact resistance	1.5
$R_q(K\Omega/tube)$	quantum resistance	6.5
$R_{int}(K\Omega/\mu m)$	interconnect resistance p.u.l.	3.25
$C_q(aF/\mu m)$	quantum capacitance	400
$L_k(nH/\mu m)$	kinetic inductance	4
$C_{in}(aF)$	minimum–size driver input capacitance	~ 14
$C_{g,n}(aF/\mu m)$	n_{th} CNT to ground capacitance	Design–dependent
$C_{c,mn}(aF/\mu m)$	coupling capacitance between m_{th} and n_{th} CNTs	Design–dependent

Design–dependent Parameter Values			
Symbol	1 tube	2 tube	3 tube
$C_{g,1}$	35.82	27.33	22.6
$C_{g,2}$	–	27.33	10.96
$C_{g,3}$	–	–	22.6
$C_{c,12}$	–	18.26	47.4
$C_{c,13}$	–	–	4.39
$C_{c,23}$	–	–	47.4

total interconnect resistance multiplied by the load capacitance is a larger component of delay than the driver resistance multiplied by the interconnect capacitance. Therefore, designs with smaller resistance p.u.l. offer better performance in terms of circuit delay. This fact is illustrated in Figure 23, where it is shown that a three–tube SWNT interconnect design offers about $5\times$ better performance compared to Cu interconnects. At short interconnect lengths, where transistor parasitics and contact resistance become important, improvements in speed and EDP are smaller. In fact, at very short lengths, the performance of SWNT interconnects are deteriorated due to the dominance of the contact resistance and transistor parasitics such that Cu interconnects may outperform SWNT interconnects.

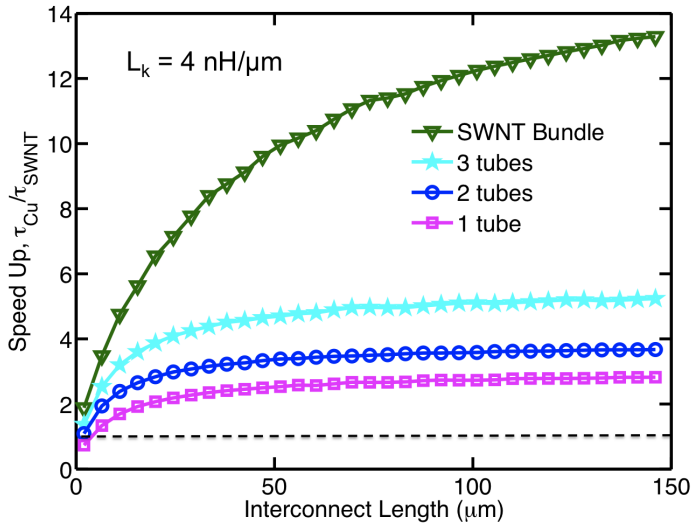

Figure 23: Speedup offered by single or few SWNT interconnect designs with various number of tubes and bundles of SWNTs as a function of interconnect length assuming that drivers and receivers are $5\times$ the minimum size. Kinetic inductance is assumed to be equal to its theoretical value, which is $8nH/\mu m$ per conduction channel.

As Figure 24 demonstrates, even though the capacitance of three–tube SWNT interconnects design is larger than that of fewer tubes, they still offer better EDP gain than other designs considering $5\times$ minimum size drivers and receivers. This

superior performance in the EDP stems from their significantly smaller RC product p.u.l. squared values made possible by significantly reduced interconnect resistances due to parallel conduction.

Compared to SWNT bundles, single– or few–SWNT interconnect designs offer smaller speedup. However, EDP gain offered by three–tube SWNT interconnect design is similar to that of bundles. In fact, at shorter interconnect lengths, bundles are outperformed by this design.

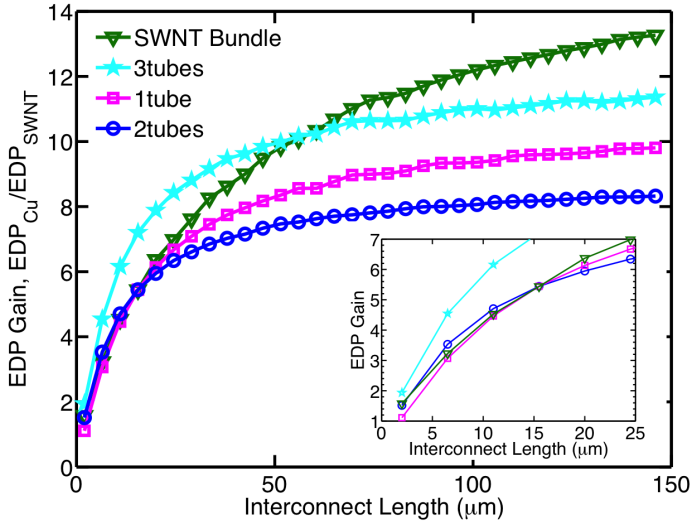

Figure 24: EDP offered by single or a few SWNT interconnect designs with various number of tubes and bundles of SWNTs as a function of interconnect length assuming that drivers and receivers are $5\times$ the minimum size. Kinetic inductance is assumed to be equal to its theoretical value, which is $8nH/\mu m$ per conduction channel.

The plots in Figure 23 and Figure 24 are obtained assuming the theoretical kinetic inductance value of $8nH/\mu m$ per conduction channel [89]. There are two channels that contribute to conduction in a metallic SWNT with a $2nm$ diameter. Therefore, the kinetic inductance component is $4nH/\mu m$. The experimental results for kinetic inductance reported in literature range from $4nH/\mu m$ to $60nH/\mu m$

[93, 94]. To quantify the impact of kinetic inductance on CMOS interconnects, similar simulations are performed with various kinetic inductance values in this range as illustrated in 25.

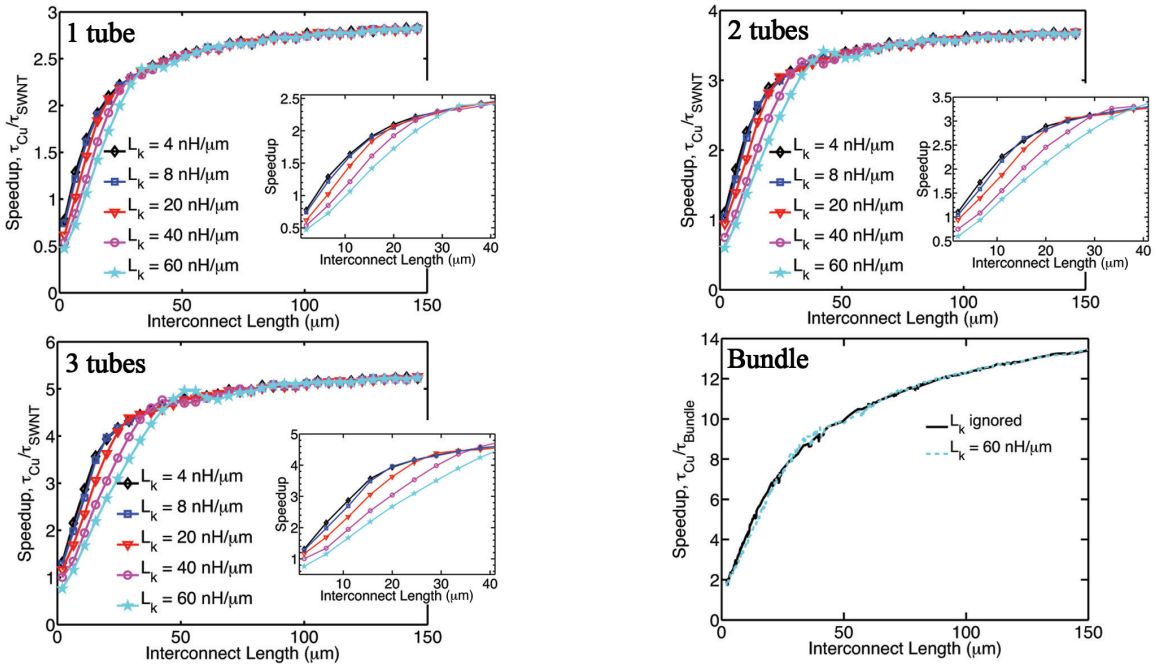

Figure 25: Speedup offered by single or few SWNT interconnect designs with various number of tubes and bundles of SWNTs as a function of interconnect length assuming that drivers and receivers are 5× the minimum size. Kinetic inductance per conduction channel is varied.

It can be seen in this plot that in the range of 4–20$nH/\mu m$ kinetic inductance values, the impact on the interconnect performance is quite small. For the extreme value of 60$nH/\mu m$, however, the performance degradation of SWNT interconnect designs cannot be neglected. For long interconnects, the resistance–capacitance component of the delay, which increases quadratically with interconnect length dominates the total unified delay expression [95]. The time of flight, which depends on the line inductance increases linearly with the interconnect length and its impact reduces for long interconnects. The impact of the kinetic inductance can

be neglected for SWNT bundles since there are many tubes in parallel. The difference between assuming an extreme kinetic inductance value of $60nH/\mu m$ and completely ignoring it is negligible for bundles as illustrated in Figure 25.

Impact of broken tubes on speedup and EDP gain for two– and three–tube SWNT interconnects designs are illustrated in 26 and 27, respectively. Speedup and EDP improvements are only slightly reduced if one of the tubes in a two–tube design is broken close to the receiver. If the tube is broken close to the driver, however, performance of interconnects is significantly degraded. The worst case scenario for broken tubes occurs when all but one of the tubes are broken close to the driver. Major improvements in EDP are achieved even in this worst–case scenario for both two– and three–tube designs. In light of the results obtained from the HSPICE simulations, the requirements needed to overcome Cu interconnect performance using SWNT interconnects are determined and summarized in Table 22.

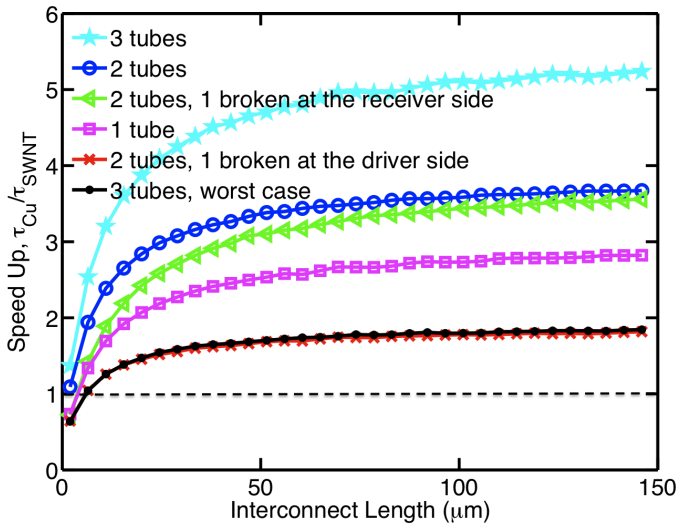

Figure 26: Speedup as calculated in Figure 23 with the impact of broken tubes for two–tube and 3–tube designs in the worst possible case included.

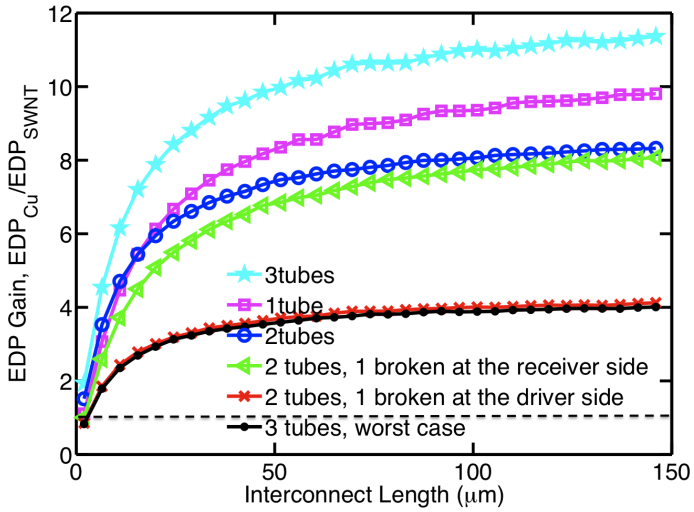

Figure 27: EDP gain as calculated in Figure 24 with the impact of broken tubes for two–tube and 3–tube designs in the worst possible case included.

Table 22: Status update on key metrics.

	Needed	Demonstrated
Tube diameter (nm)	> 2	2.5 ± 0.4 [96] $1,2 \pm 0.3$ [97]
Tube density ($SWNTs/\mu m$)	> 250	$20 - 40$ [96] $10 - 30$ [97]
Conversion to metallic	100%	25% [88]
Contact + quantum resistance ($K\Omega$)	10	10-30 [97]
Alignment	$\sim 100\%$	99.5% [97]

4.4 Conclusions

It is shown that new opportunities arise for emerging interconnect technologies as alternatives for the conventional Cu/low–κ technology at the local and intermediate levels, where wire dimensions are small at future technology nodes.

To avoid interconnects from becoming bottlenecks, the RC delay and EDP of interconnects have to be reduced. Considering the advances in manufacturing highly dense and horizontally aligned individual metallic tubes, this chapter has concentrated on evaluating the potential performances of single– or few–SWNT interconnect designs.

Contrary to previous studies on SWNT interconnect applications, which have all concentrated on bundles of SWNTs for reduced resistance p.u.l., it is shown that individual tubes can be used at extremely small dimensions. It is demonstrated that individual SWNT interconnects can compete with or even outperform minimum–size copper interconnects at the 11–nm minimum wire width and beyond, while lowering the energy–per–bit by more than $3\times$. Using multiple parallel tubes in the design can further improve RC delay of SWNT interconnects with only a slight degradation in EDP. Potentially broken tubes can drastically reduce the improvement in both delay and EDP, but multiple tube designs with broken tubes may still outperform copper interconnects even in the worst case.

CHAPTER 5

SYSTEM–LEVEL DESIGN AND PERFORMANCE BENCHMARKING FOR MULTILEVEL INTERCONNECT NETWORKS FOR CNFETS

In this chapter, the first system–level study on the impact of carbon nanotube field–effect transistors (CNFETs) on multilevel interconnect networks is presented. It is demonstrated that the respective 4.3× and 8× improvements in intrinsic delay and EDP of CNFETs at 16–*nm* technology node over Si–CMOS switches are quickly overshadowed by the delay and EDP of interconnects. For repeater–inserted interconnects, delay and EDP improvements saturate at 2.08×. However, CNFETs offer a major advantage in terms of the required number of metal levels because of the availability of a larger number of repeaters compared to Si–CMOS switches.

5.1 Introduction

Carbon nanotube field–effect transistors, illustrated in Figure 28, are promising candidates to replace CMOS and aid in the extension of Moore's Law due to their much better CV/I intrinsic gate delay [98, 99, 100, 101]. It is reported that CNFETs can offer 6× and 14× improvements over bulk n–type and p–type MOSFET devices, respectively [98, 99] even when device non–idealities are considered. It has been shown in [98, 99, 102] that this improvement is significantly degraded by parasitic capacitances, such as the gate to source and drain extension fringe capacitances, and interconnect capacitances at the device level, such as the capacitance between the gate and source/drain contacts. In a real circuit with interconnects of very different length scales, the improvement offered by CNFETs would be affected by interconnects even further. In addition, as described in chapters 2 through 4 and illustrated in Figure 29, *RC* delay of an average length interconnect increases with technology scaling and quickly becomes comparable to the intrinsic

CMOS device *RC* delay. Figure 30 illustrates the trend for EDP.

In light of these facts, it becomes necessary to consider the impact of interconnects on the potential performances of all emerging devices, especially the ones with intrinsic delays smaller than that of MOSFET since the impact of interconnects will be more pronounced. On the other hand, as designed in Chapter 2, a MIN comprises many metal levels with various interconnect dimensions that accommodate interconnects of very different length scales. CNFET drivers and repeaters offer different output resistance and input capacitances and the ramifications of replacing Si–CMOS switches with CNFETs have to be studied carefully.

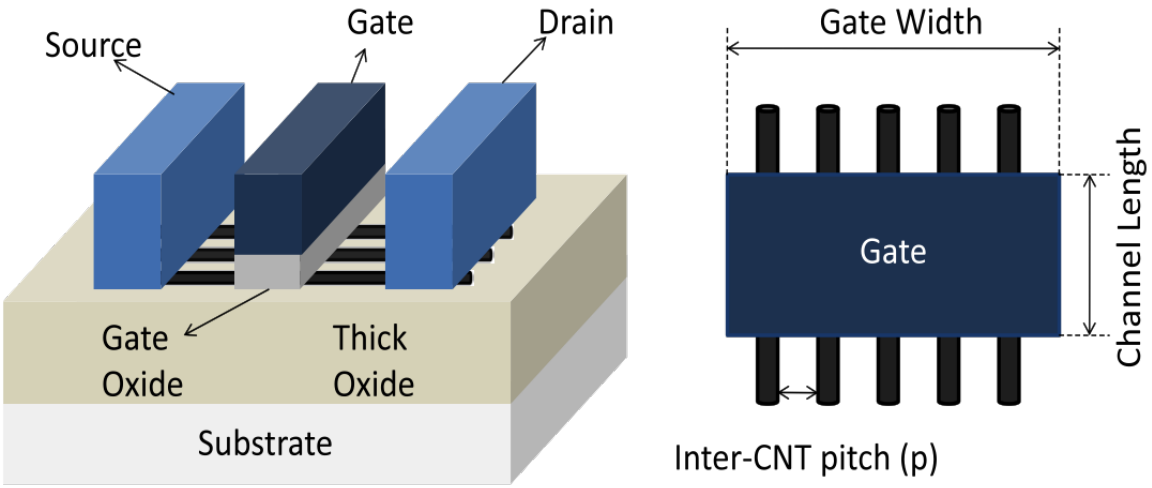

Figure 28: 3–D view of a CNFET (left) and the top view of the gate of a CNFET (right) regenerated from [99].

In this chapter, for the first time, a comprehensive study on the impact of interconnects on the potential performance of CNFET circuits and the impact of CNFETs on the design and performance of MINs is presented. First, the improvement in speed, energy per binary switching operation, and EDP offered by CNFET circuits over CMOS circuits versus interconnect length is quantified at the 16–*nm* technology node. To find the device resistance and capacitance values, HSPICE

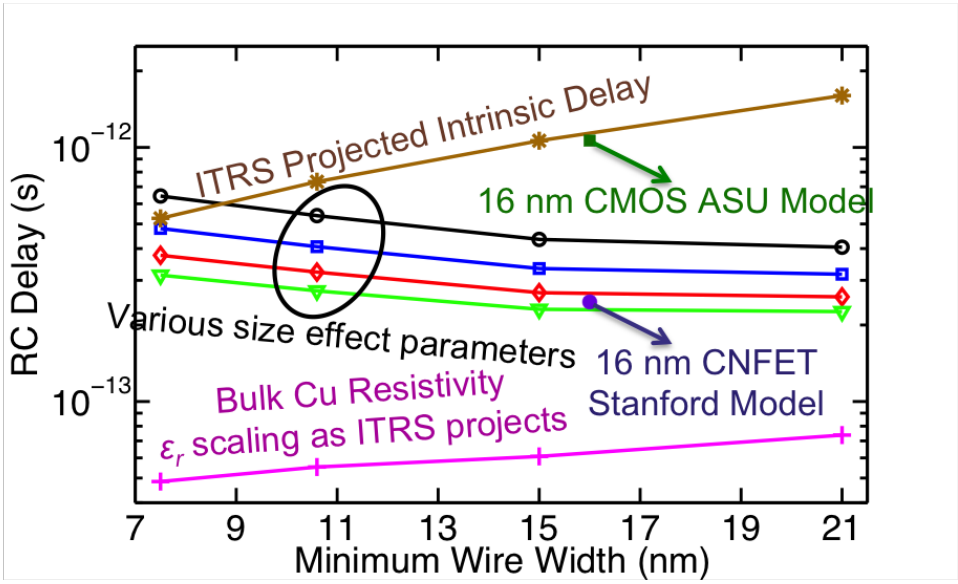

Figure 29: The *RC* delay of a 10–gate–pitch–long interconnect is plotted versus the technology generation for various experimentally reported size effect parameters. For reference, the bulk Cu resistivity scenario and intrinsic delay of CMOS switches based on ITRS projections are also plotted. For the 16–*nm* technology node, intrinsic delays of CMOS and CNFET switches are shown based on ASU predictive models and Stanford University CNFET model, respectively [98, 99, 103].

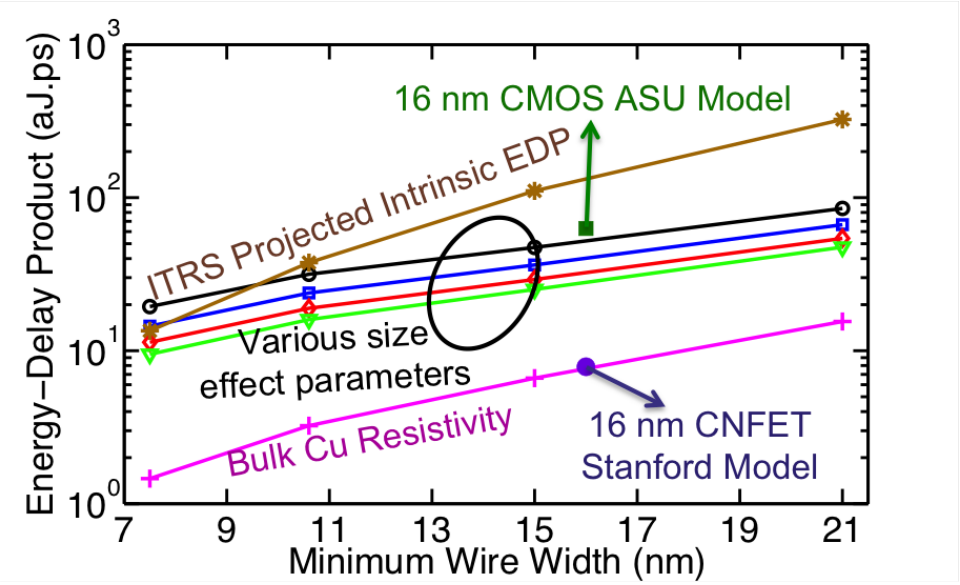

Figure 30: The EDP comparison for the same items in Figure 29.

simulations are performed using the CNFET SPICE model developed at Stanford University [98, 99], and predictive models based on Berkeley short channel IGFET Model (BSIM) 4 developed at the Arizona State University for planar MOSFETs. For long lengths, interconnects with repeaters and various cross–sectional dimensions are considered. Next, the MINs for both CNFET and CMOS circuits are designed and compared in terms of the required number of metal levels and power dissipation with the design methodology that was presented in Chapter 2. The impact of using fast switches on repeater insertion and the consequent effect on via blockage in an MIN are also investigated.

Section 5.2 elaborates on the models and technology parameters used in benchmarking the potential performances of CNFET circuits against their CMOS counterparts. Impact of interconnects on performance is quantified for both short interconnects at the local level and long interconnects where repeaters are inserted. Impact of CNFET circuits on repeater insertion is underlined. Section 5.3 compares the optimal MIN design results for CNFETs and MOSFETs. The results are summarized and concluding remarks are given in Section 5.4.

5.2 CNFET Circuit Performance

5.2.1 Technology Parameters

In this study, the inter–CNT pitch, illustrated in Figure 28 and defined by the sub–lithographic self assembly process of CNT growth [86, 104, 82, 105], is varied from 5 nm (200 $CNTs/\mu m$) to 20 nm (50 $CNTs/\mu m$). A 0.7 V supply voltage is assumed following the 2010 ITRS projections [25] for the 16–nm technology node corresponding to the technology year 2018. The threshold voltage of CNFET devices are adjusted such that the ON current through a CNFET device with a high–density of tubes is adjusted to $\sim 15\mu A/CNT$ while maintaining low OFF currents by setting the required model parameters listed in [98, 99] accordingly. Resistance and capacitance values of a minimum size CNFET inverter with 5 nm,

10 nm and 20 nm inter–CNT pitch at 0.7 V supply voltage are calculated to be 9.493 $K\Omega/26\ aF$, 12.963 $K\Omega/20.5\ aF$ and 24.138 $K\Omega/16\ aF$, respectively. For CMOS inverters, the resistance and capacitance values are 22.047 $K\Omega$ and 48.3 aF, respectively. Assuming a smaller supply voltage of 0.52 V for a CNFET inverter with a 5 nm inter–CNT pitch, the output resistance can be brought to a similar value to CMOS inverters, namely 23.333 $K\Omega$, to save power.

5.2.2 Impact of Unrepeated Interconnects

Under these assumptions, a CNFET inverter with a 5 nm inter–CNT pitch offers speedup and EDP improvement values of 4.3× and 8×, respectively ignoring the impact of interconnects. As Figures 31 and 32 show; however, these intrinsic improvements quickly degrade even at short interconnect lengths. At a 10 gate pitch interconnect length and assuming a fan–out of 3, the improvements have already fallen down to 3× and 4.4×, respectively. CNFET circuits with 5 nm inter–CNT pitch operated at 0.52 V offer only about a 2× improvement in speed, which is due to the smaller device capacitance while offering the same output resistance as Si–CMOS. However, this offers significant savings in energy–per–bit by taking advantage of the quadratic dependence of energy dissipation on the supply voltage. As a consequence, they offer even better EDP gain than high performance CNFET circuits operated at 0.7 V. At long interconnect lengths, the quadratic dependence of delay on interconnect length will be converted to a linear dependence by inserting repeaters as explained in the next subsection.

5.2.3 Impact of Unrepeated Interconnects

The plots obtained in this section are the results of rigorous HSPICE simulations. In order to explain the behavior shown in these plots, compact models for repeater insertion [106] are used. Simulations have shown that the output capacitance of a CMOS inverter is comparable to its input capacitance. The optimal number of

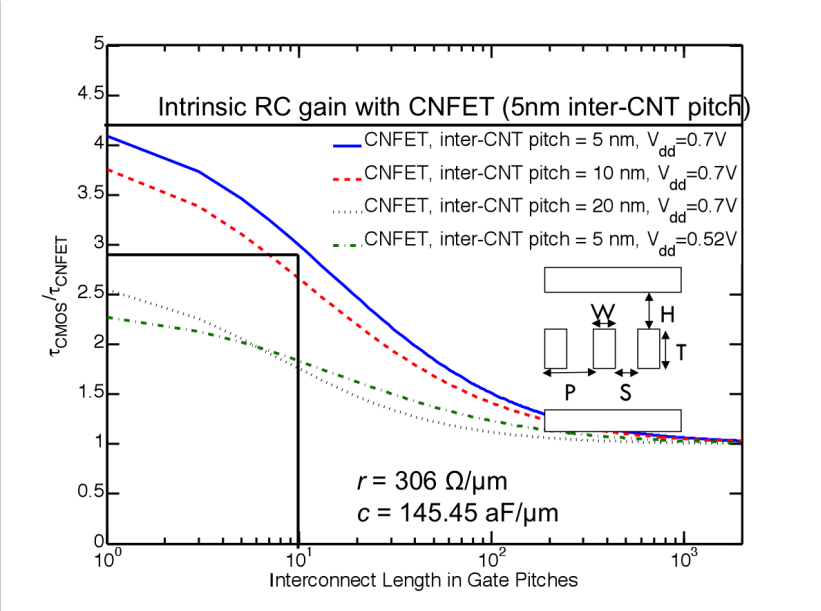

Figure 31: Speedup offered by CNFET circuits over CMOS circuits at various interconnect lengths.

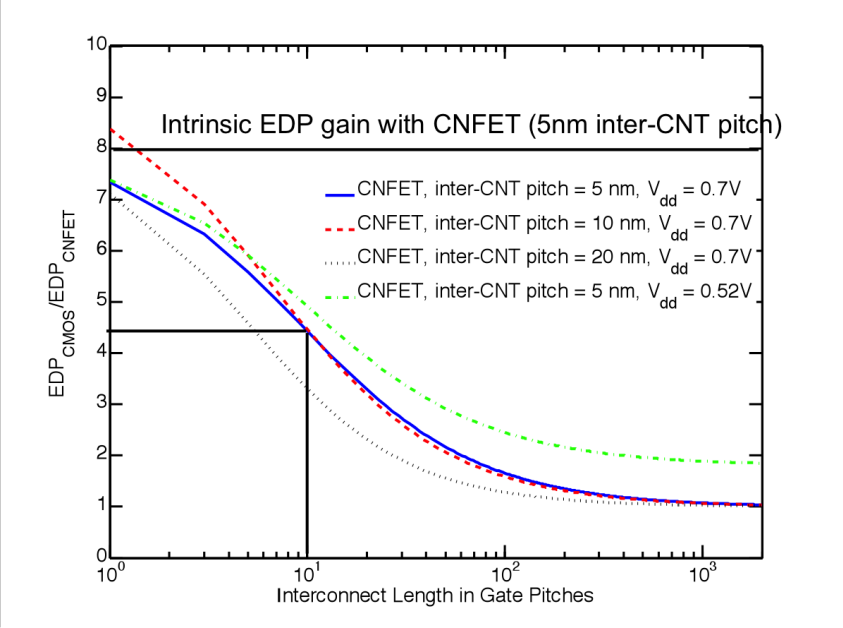

Figure 32: EDP gain offered by CNFET circuits over CMOS circuits at various interconnect lengths.

repeaters is given by,

$$k = \sqrt{\frac{0.4 R_{int} C_{int}}{1.4 R_0 C_0}}, \tag{10}$$

where R_0 and C_0 are the output resistance and input capacitance of a minimum size inverter and R_{int} and C_{int} are the total interconnect resistance and capacitance, respectively. However, often a sub–optimal number of repeaters is used with a minor delay penalty. The delay of a repeated interconnect with a sub–optimal factor of ζ is given as,

$$\tau = \left(1.4 + 0.75\zeta + \frac{0.75}{\zeta}\right) \sqrt{R_0 C_0 R_{int} C_{int}}. \tag{11}$$

The expression for the energy–per–bit for this interconnect is

$$E = \left(\sqrt{\frac{0.4}{1.4}\zeta + 0.5}\right) C_{int} V_{dd}^2, \tag{12}$$

where the first term represents the power dissipation associated with repeaters and the second term represents the power dissipation associated with interconnect segments. Note that the energy–per–bit expression does not depend on device parasitic capacitance parameters.

The expression for the optimal number of repeaters shows that the number of repeaters required will be higher for CNFETs, which have much smaller $R_0 C_0$ delay products compared to CMOS gates. This result is illustrated in Figure 33. The higher number of repeaters may cause problems in a MIN design where a higher number of vias will be required and extra via blockage will reduce the net effective area that can be used for routing wires. As a consequence, this extra via blockage may cause a larger number of metal levels to be required for routing all the wires. On the other hand, using the same number of repeaters for CNFET circuits as CMOS circuits may keep via blockage the same, but it will reduce the speedup advantage offered by CNFETs. Second, it is seen through the delay expression in

equation 11 that for a repeater inserted line, the speedup advantage of any emerging device technology will reduce by the square root. This fact is demonstrated in Figure 34, where only 2.08× improvement in both delay and EDP gain is observed as opposed to the aforementioned respective intrinsic improvement of 4.3× and 8×.

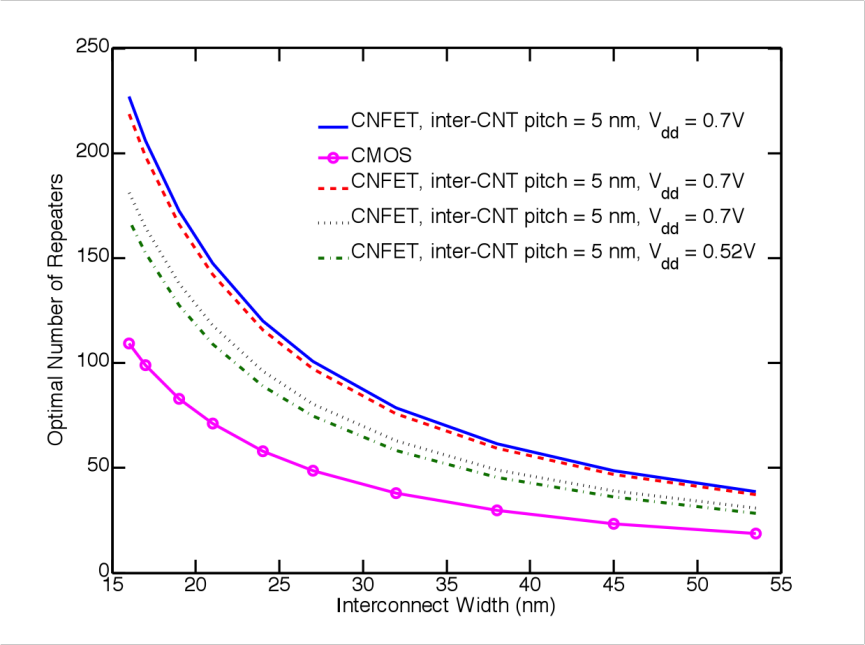

Figure 33: Optimal number of repeaters required for CMOS circuits and CNFET circuits under various conditions.

The discussion so far elaborates on how interconnects may become bottlenecks for any fast switch and limit their speedup and EDP gain in a system. However, to avoid interconnects from becoming bottlenecks, it may be possible to come up with a device which offers the same amount of ON current as a CMOS switch at a smaller supply voltage to take advantage of the quadratic dependence of energy–per–bit on the supply voltage. As illustrated in Figure 34, it is possible to operate a high performance CNFET at 0.52 V to keep the output resistance of an inverter the same as a CMOS inverter, but to increase the EDP gain significantly. This lower supply voltage would lower energy dissipation in both devices and

wires quadratically. Figure 34 also illustrates that even though a larger number of repeaters are required for CNFET circuits, this does not cause extra power dissipation since the power dissipated in repeaters, which is calculated by equation 12, only depends on the total capacitance of the interconnect.

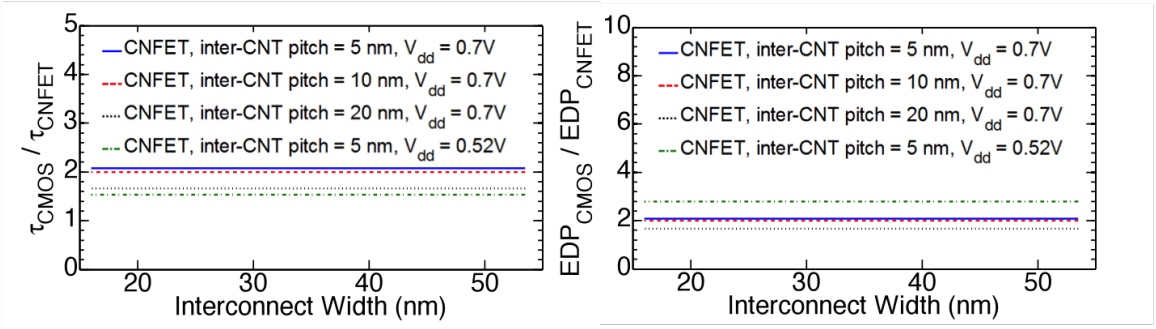

Figure 34: Speedup and EDP gain of an interconnect repeated with CNFET repeaters over CMOS repeaters.

5.3 MIN Design Results for CNFETs

The MINs in this chapter are designed using the methodology that was described in Chapter 2. This is the first time that the design and performance of MINs are investigated for CNFET circuits. To compare the MINs for CNFETs and MOSFETs, the number of cores on the chip and the frequency of operation are varied and the number of metal levels required, the total interconnect power dissipation, and the total chip power dissipation, including interconnects and dynamic and leakage power of logic gates and repeaters, are quantified.

ITRS projections are assumed for the die size and the total number of logic transistors. The area models given in [39] are used for CMOS gates and a simple area model is used for CNFETs, where the width is assumed to be the gate width plus the gate overhead, which is a layout specification and is usually 1.5× the minimum feature size, and the length is assumed to be the sum of gate, contact and source/drain extension lengths. In this work, NAND2 gates are considered.

They typically occupy smaller area when made out of CNFETs as reported in [107], where the area occupied by a minimum size CNFET and a CMOS NAND2 gates are compared. As the gates are made larger, the gap between the areas of CNFET and CMOS NAND2 gates is reduced with the models considered in this work.

The number of metal levels for various core sizes is plotted in Figure 35, where it can be seen that the smallest number of metal levels is required by CNFETs with 5 *nm* inter–CNT pitch. The reason for this is that small sizes of CNFET NAND2 gates with such a high density of tubes can satisfy the frequency constraint and more active area can be reserved for repeater insertion.

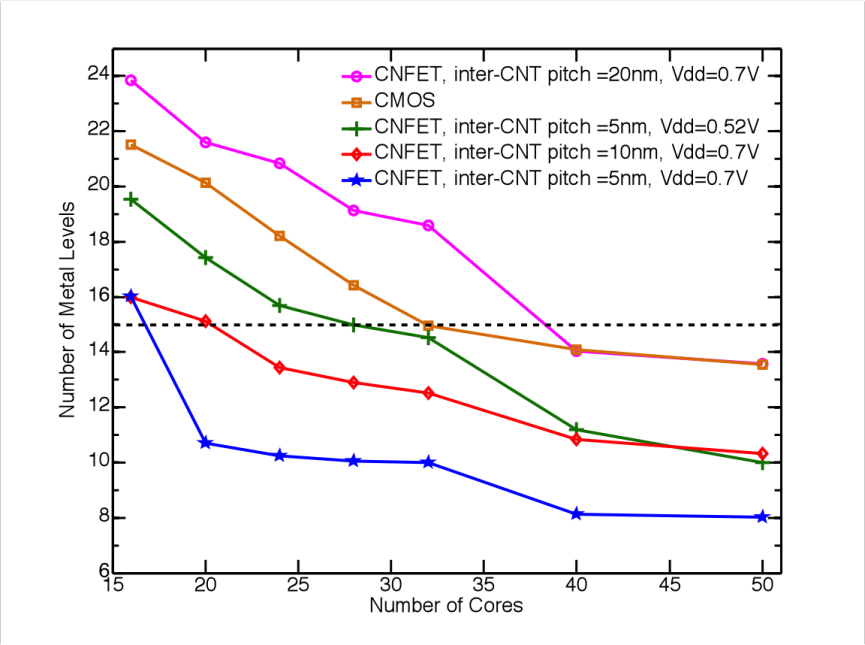

Figure 35: Number of required metal levels for various core sizes assuming different technologies.

Even though a large number of repeaters are inserted in the MIN design for 5 *nm* inter–CNT pitch CNFETs and the number of vias are increased, the wires have smaller dimensions resulting in vias with smaller diameters. As a consequence, via blockage does not cause a significant problem in the MIN design of CNFETs. Figure 35 also shows that the minimum number of cores is limited by the number

of metal levels. MINs for CNFETs with 5 *nm* inter–CNT pitch operated at a smaller supply voltage also require smaller number of metal levels than MOSFETs. As shown in Figures 36 and 37 , due to the quadratic dependence of power dissipation in both interconnects and devices on supply voltage, the best performance can be achieved by using CNFET devices with a high density of tubes operating at a lower supply voltage. Note that the total power dissipation of MOSFET chip is much higher than a CNFET chip even though the interconnect aggregate capacitance of interconnects is the same for both chips. This is due to the fact that CMOS gates have larger capacitances compared to CNFET gates. They also have to be made larger in size to satisfy the clock frequency requirement resulting in higher dynamic and leakage power dissipations.

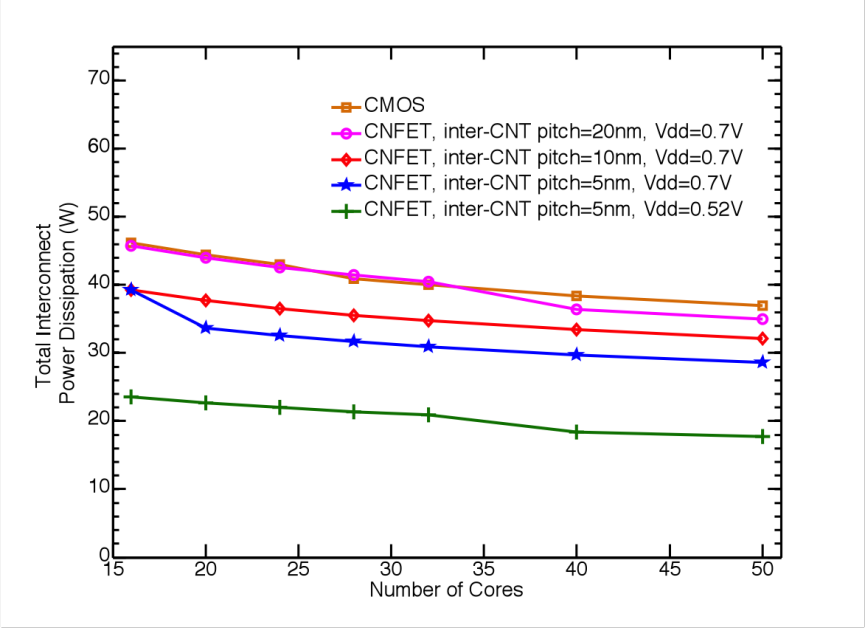

Figure 36: Total interconnect power dissipation of the MIN for various core sizes assuming different technologies.

Figure 38 shows how the required number of metal levels depends on the frequency of operation assuming 32 cores. Due to their high current driving capability and small size, CNFET circuits with 5 *nm* inter–CNT pitch can offer a small

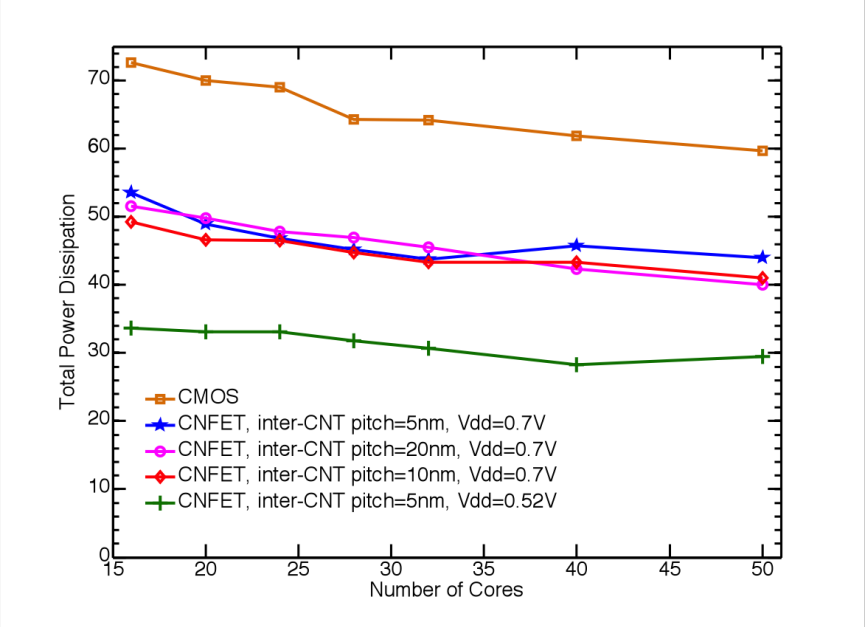

Figure 37: Total power dissipation of the MIN including dynamic and leakage power of logic gates and repeaters.

number of metal levels over a very large frequency range. As the tube density is reduced, the frequency at which the required number of metal levels is below 15 reduces. The steps in this plot correspond to the points where the logic gates have to be upsized in order to meet the given clock frequency constraint. As the frequency is increased, logic gates occupy more and more of the available silicon area leaving smaller areas for repeaters. As a consequence, a smaller number of repeaters can be inserted and the number of metal levels required to route all interconnects increases significantly. In other words, the number of metal levels puts an additional limit on the maximum frequency of operation.

A larger frequency of operation translates into a larger power dissipation at the same technology node. The interconnect power dissipation and the total power dissipation of CNFET MINs are much smaller than that of CMOS if high–density devices are operated at low supply voltages. Figures 38, 39 and 40 illustrate that it is possible to operate at a frequency of 12 GHz with such devices and still keep

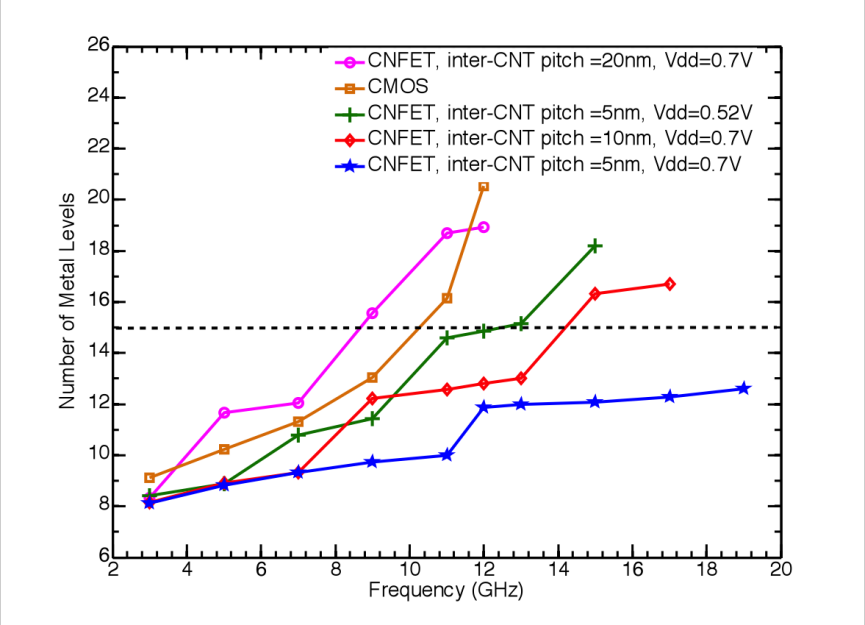

Figure 38: Number of required metal levels for various clock frequencies assuming different technologies.

the number of metal levels below 15 while dissipating as much power as a CMOS chip dissipates at about 6 GHz.

At the ITRS projection of 10.652 GHz, a high–performance CNFET chip operated at a 0.52 V supply voltage requires less than 15 metal levels, dissipates about 1.5$\times$ less interconnect power and dissipate about 2.3$\times$ less total power than a Si–CMOS chip.

5.4 Conclusions

Due to interconnect size effects, the RC delay of even average length interconnects in a MIN become comparable to the intrinsic RC delay of Si–CMOS switches with technology scaling. Therefore, any switch that is faster than Si–CMOS will be slowed down by interconnects even more severely. The degradation in the intrinsic improvements in delay and EDP offered by CNFETs over MOSFETs is quantified by taking interconnect bottleneck into account. For long, repeater–inserted lines,

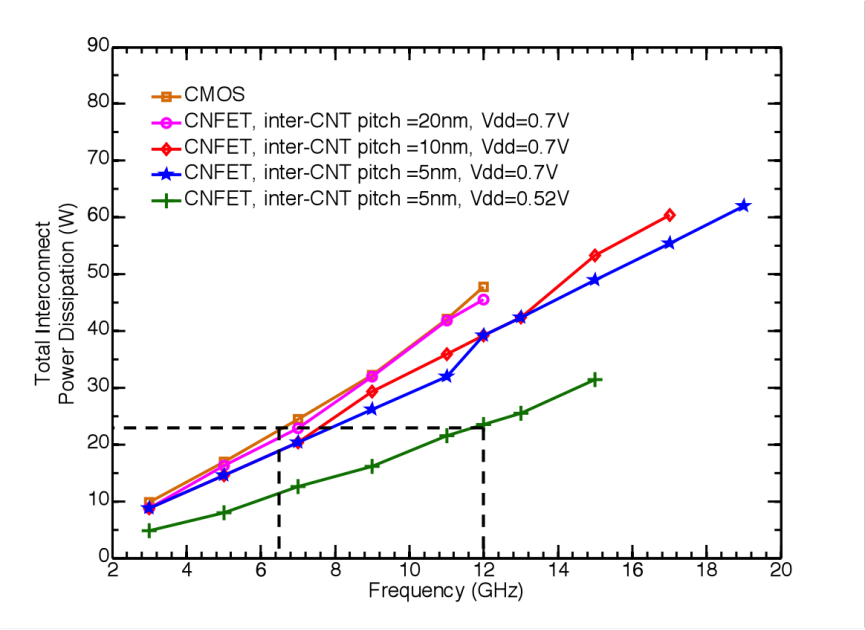

Figure 39: Total interconnect power dissipation of the MIN at various clock frequencies assuming different technologies.

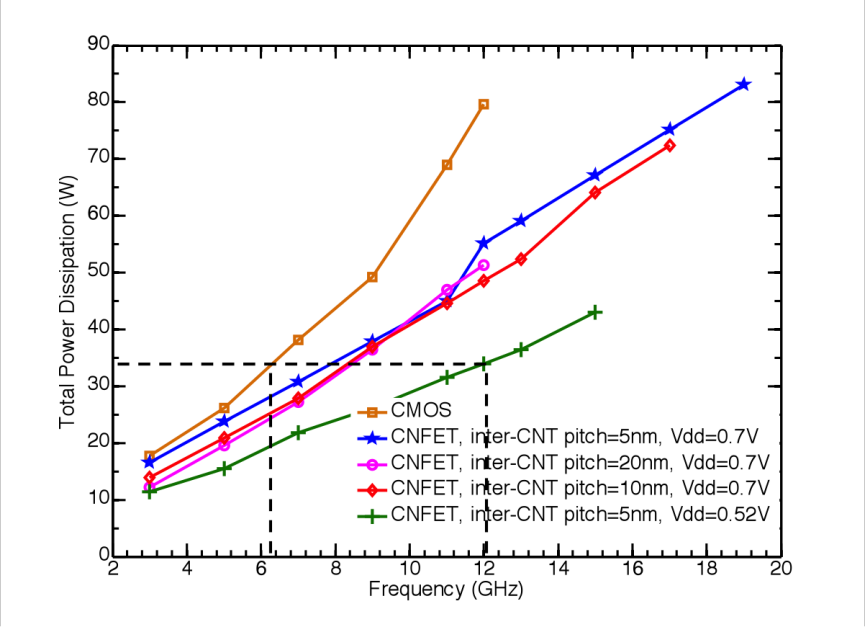

Figure 40: Total power dissipation including dynamic and leakage power of logic gates and repeaters at various clock frequencies.

improvement in delay and EDP saturate at the square root of the gain in the RC delay. MINs are designed for CNFETs and their performance is evaluated in terms of the number of metal levels and power dissipation. CNFETs with high density of tubes offer the smallest number of metal levels, which is mainly because of the larger area that is available for repeater insertion. The higher number of repeaters does not cause extra via blockage even though the number of vias increases significantly. This is because interconnect dimensions at various metal levels are smaller compared to the CMOS interconnect levels and via blockage is directly proportional to wiring pitch of lower metal levels.

CHAPTER 6

CIRCUIT PERFORMANCES OF VARIOUS LOGIC DEVICES WITH CONVENTIONAL AND EMERGING INTERCONNECT TECHNOLOGIES

The trade–offs between the technology parameters of various interconnect technologies are investigated on the basis of their impacts on the circuit performances of emerging post–CMOS devices. In this chapter, FinFETs, sub–threshold CMOS circuits, nanowire–based gate–all–around (GAA) tunneling field–effect transistors (TFETs) and CNFETs are studied. Each of these devices are paired with the conventional Cu/low–κ interconnect, single–wall carbon nanotube (SWNT) interconnect manufactured in horizontal bundles or in a single layer, and multi–layer graphene nanoribbon (GNR) interconnect. The relative performances of all these interconnect technologies with each type of device are evaluated. The interconnect technology option that gives the best performance in terms of circuit delay, energy–per–bit and EDP is reported for each of the device technologies.

6.1 Introduction

Even though there are many strong candidates for the major device technology in the post–CMOS era, each of these candidates will continue to suffer from the limitations caused by interconnects. As the resistance and capacitance values of different device technologies vary significantly, the constraints that they put on interconnects are quite different as well. To obtain the best circuit performance, it is crucial to investigate the interactions of interconnects with all these emerging devices. In previous chapters the interactions of FinFETs and CNFETs with the conventional Cu/low–κ technology have been studied.

Tunneling FETs (TFETs) show promise in overcoming the power wall facing thermionic FETs by allowing for significant reduction in the supply voltage. A

wide range of device architectures and materials has been studied in the realization of TFETs in the previous years including single gate (SG), double gate (DG) and GAA structures [108]. Previous studies have shown that band–to–band–tunneling (BTBT) FETs can potentially offer intrinsic gate delays that are comparable with thermionic FETs at lower supply voltages [108, 109]. In this chapter, the ON/OFF current and input capacitance of InAs nanowire–based GAA tunnel FETs are modeled and their interaction with interconnects is studied based on these device models.

This chapter is divided into five sections. In Section 6.2, assumptions about the interconnect technologies that are considered in this study are explained and the interconnect configurations are illustrated. In Section 6.3, some of the assumptions that are made for simulating the mentioned devices and TFET model details are described. Section 6.4 compares the performances of carbon–based interconnect technologies against Cu/low–κ technology for each device type and tabulates the impact of interconnect resistance and capacitance on the circuit performance of the devices that are investigated here. Section 6.5 underlines the key results and concludes the paper. Results confirm that device resistance and capacitance determine the most appropriate interconnect technology for each device type.

6.2 Interconnect Technology Parameters

The reference structure that is used to compare the relative performances of emerging carbon–based interconnect technologies is the conventional Cu/low–κ interconnect technology configuration shown in Figure 41. The configurations assumed for these emerging technologies are also illustrated in Figure 41.

For both Cu and carbon–based interconnects, the resistance and capacitance values are calculated as described in Section 4.2.1. In a densely packed SWNT bundle, the distance between the nanotubes is assumed to be 0.34 nm [110] due to

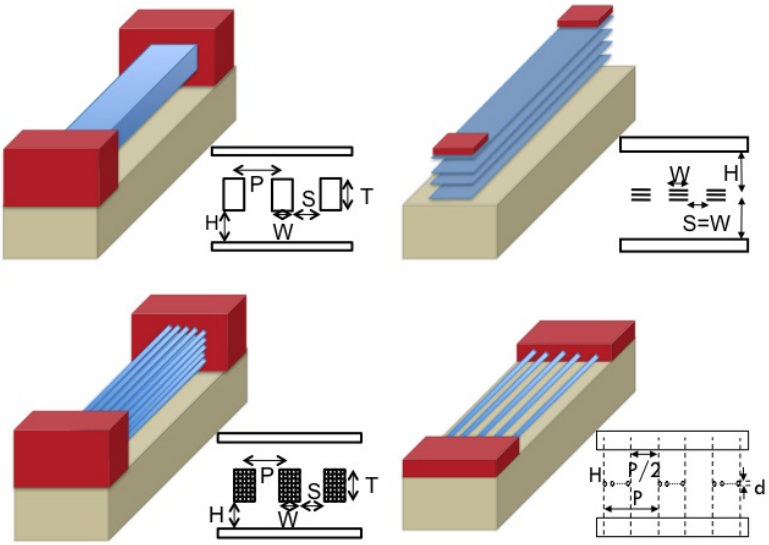

Figure 41: Interconnect configurations for conventional Cu/low–κ technology (top left) assuming $W = S = P/2$ and $H = T = AR \cdot W$, where P, W, S, T, H and AR stand for the wire pitch, wire width, wire spacing, wire thickness, inter–layer dielectric height, and aspect ratio of the wire, respectively, multi–layer GNR interconnect with top contacts (top right), SWNT bundle (bottom left), and mono–layer of well–aligned high density SWNTs (bottom right).

Van der Waals forces. This corresponds to a density of tubes in the cross–section of the bundle of $1/1.5nm^2$. Since, statistically, only $1/3$ of SWNTs are metallic [111] , the density of metallic tubes in the cross–section of a densely packed SWNT bundle is $1/4.5nm^2$. We consider a value of $1/3nm^2$ for the density of metallic tubes in the cross–section of the bundle, which requires that $\sim$45% of the tubes in the bundle should be metallic. For multi–layer GNR interconnects, top contacts are considered, so the contacts couple only to the topmost layer. Appropriate models that consider the effective amount of contribution that each graphene layer provides for current conduction are used [112]. Based on these models, optimum number of layers that minimizes delay and EDP is calculated. A constant $1\mu m$ electron MFP is assumed for GNRs [113].

6.3 Device Technology Parameters

The device architectures that are considered in this chapter are illustrated in Figure 42.

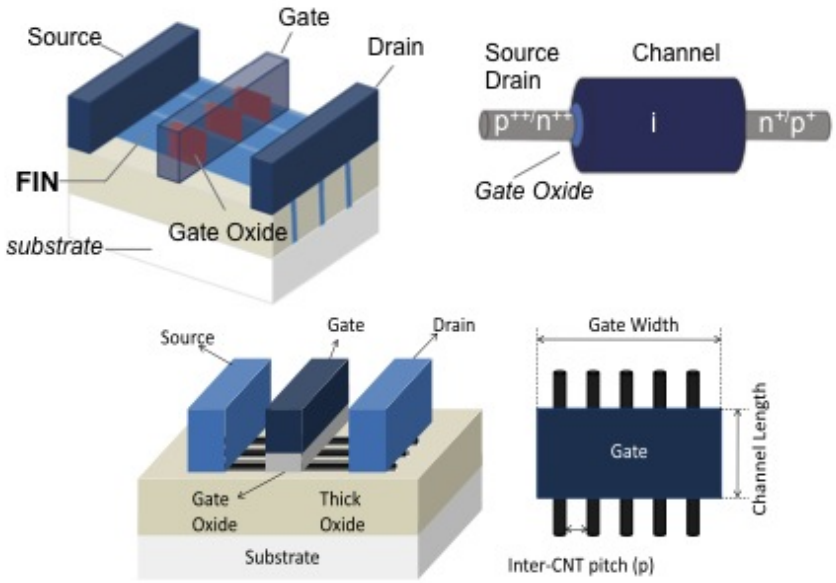

Figure 42: Device architectures for FinFET (top left), nanowire–based GAA TFET (top right), and MOSFET–like CNFET (bottom).

FinFET devices and CNFETs with 5 *nm* inter–CNT pitch are modeled as described in Chapters 2 and 5, respectively. For sub–threshold circuits, complete circuit simulations are performed in HSPICE based on predictive SPICE models [103].

For TFETs, InAs nanowires are considered due to their direct bandgap that eliminates the necessity for phonon assistance in tunneling. InAs is a promising material for realizing TFETs thanks to their small bandgap and light hole and electron effective masses [114], which both increase the ON current of the TFET device. Both p–type and n–type TFETs are realized by assuming n–i–p and p–i–n structures as shown in Figure 43, respectively.

Applying proper bias voltages between the device terminals can modify the band diagram shown in Figure 43. Tunneling occurs between the n/p doped

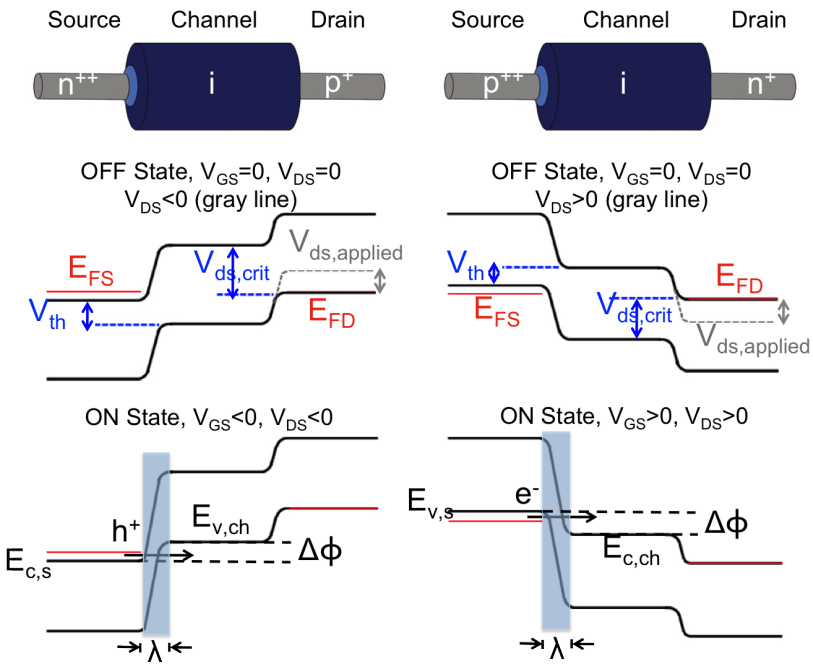

Figure 43: Schematic of an InAs nanowire–based GAA p–type TFET and the corresponding band diagram in the OFF/ON states (left), same information for an n–type TFET (right).

source and the intrinsic channel by introducing a tunnel window ($\Delta\Phi$ in Figure 43) using a negative/positive voltage at the gate in a p–type/n–type TFET. The threshold voltage of the device is defined as the amount of voltage that has to be applied at the gate such that the valence/conduction band in the channel is at the same energy level as the conduction/valence band edge in the source region in a p–type/n–type TFET. In the OFF state, the change in the potential in the channel has a one to one dependence on the applied gate voltage. Applying a gate voltage that is larger in magnitude than this threshold voltage introduces a non–zero energy window where tunneling occurs. From this point on, the impact of the charges inside the channel have to be taken into account in calculating how the position of the valence/conduction band in the channel changes with the applied gate voltage. The drain current flowing through the device due to the tunneling of carriers at the source–channel junction is modeled using the

Wentzel–Kramers–Brillouin (WKB) approximation for calculating tunneling probability [115]. Various TFET modeling publications take the approach to perform atomistic simulations or CAD based simulations for calculating the ON current of TFETs [108, 114]. However, for the purpose of performing circuit–level simulations quickly and with reasonable accuracy, we have used an analytical expression derived by using Landauers formula [109] given by,

$$I_{ON} = \frac{2q}{h} T_{WKB} k_B T \cdot \ln \left[\frac{\left(1+exp\left(\frac{\left(E_{c,s}-E_{FS}\right)}{k_B T}\right)\right)\left(1+exp\left(\frac{\left(E_{v,ch}-E_{FD}\right)}{k_B T}\right)\right)}{\left(1+exp\left(\frac{\left(E_{c,s}-E_{FD}\right)}{k_B T}\right)\right)\left(1+exp\left(\frac{\left(E_{v,ch}-E_{FS}\right)}{k_B T}\right)\right)} \right]. \quad (13)$$

In this equation, q is the electron charge, h is Plancks constant, T_{WKB} is the tunneling probability, k_B is Boltzmann constant and T is the temperature. $E_{c,s}$, $E_{v,ch}$, E_{FS} and E_{FD} represent the conduction band edge at the source, valence band edge at the channel, Fermi level at the source and Fermi level at the drain, respectively. This expression can be used for calculating the ON current through a p–type TFET. However, current through an n–type device can be calculated with a similar approach.

Using the relation in equation 13, current through a p–type device is plotted versus the gate voltage in Figure 44, where various nanowire diameter and carrier effective masses are assumed with a constant oxide thickness of 1 nm. The drain current through the device increases with reduced nanowire diameter and carrier effective mass. In this study, it is assumed that the only impact of scaling down the nanowire diameter on the device characteristics is allowing for a better gate control over the channel, which improves the drain current. The diameter dependences of bandgap and effective carrier mass are ignored for simplicity. However, below 6 nm, these diameter dependencies cause significant reduction in current and cannot be ignored [114]. At this diameter, the results match well with atomistic full–band simulations, which calculate drain currents of $\sim 130 \mu A/\mu m$ normalized to the diameter of the nanowire [114].

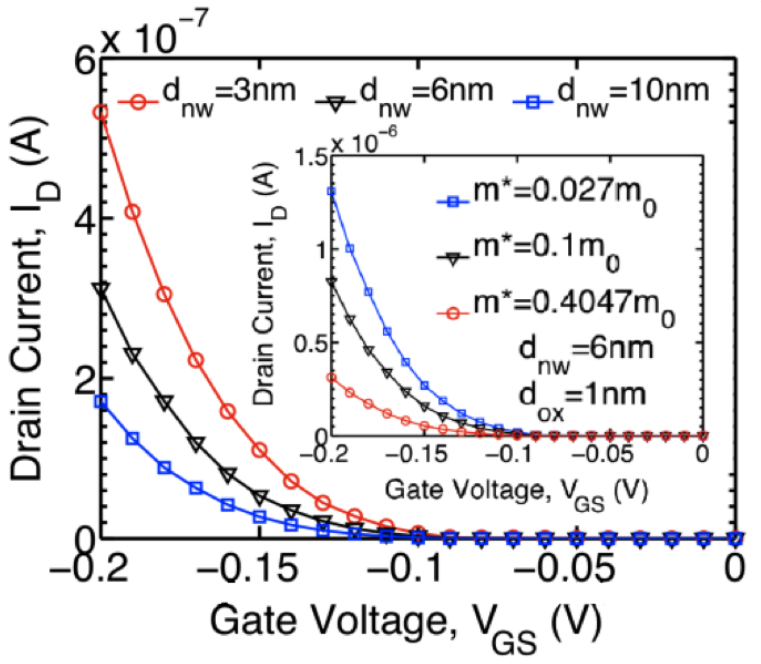

Figure 44: I_D–V_{GS} curve of a p–type TFET for various nanowire diameters and carrier effective masses. Higher currents are achieved at smaller nanowire dimensions due to enhanced gate control. Smaller effective masses increase the tunneling probability; hence offer larger current values.

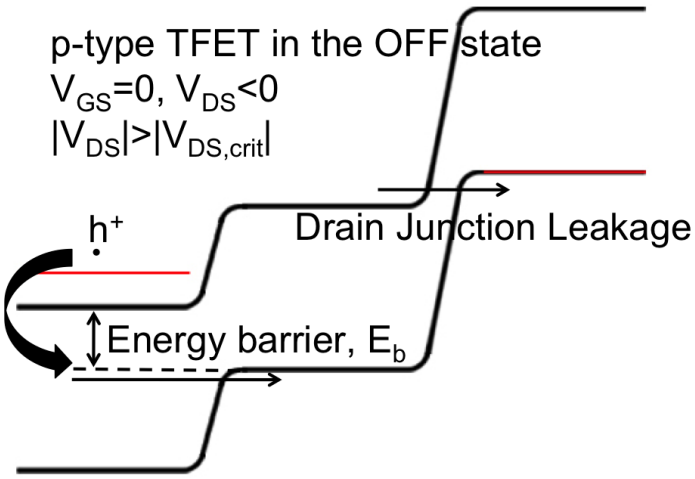

Figure 45: Leakage mechanisms considered in this work shown on a p–type TFET.

Scaling the gate oxide also enhances the gate control and increases the tunneling probability, which in turn increases the current through the device. In this study, we assumed that the channel length is significantly larger than the screening length to suppress short channel effects and avoid junction overlap.

Two leakage mechanisms are considered in this work; namely, the trap assisted tunneling in the source–channel junction [116] and the conduction in the drain–channel junction at high V_{ds} values as illustrated in Figure 45.

To obtain low OFF currents through TFET devices, the current conduction must be primarily through the source–channel junction and not the channel–drain junction. To suppress tunneling current through the channel–drain barrier, the source side is assumed to be highly degenerate such that the Fermi–level (E_{FS}) lies $\sim 4k_BT$ above/below the conduction/valence band and the drain side Fermi level (E_{FD}) is assumed to lie on the valence/conduction band edge for a p/n–type TFET. The tunneling probability through the channel–drain junction of the device that depends on the applied bias between the drain and source also puts a limit on the maximum supply voltage value. Supply voltage values above the critical V_{ds} as illustrated in Figure 43 give rise to significant OFF state current to run through the device. The current through the channel–drain junction can also be reduced by other methods such as increasing the bandgap of the material, which introduces a drive current penalty, using broken–gap heterojunction materials, and having a drain underlap [117].

To ensure that the potential inside the channel changes with respect to the gate voltage only, we assume that the gate oxide thickness (d_{ox}) is very small and the oxide capacitance is much larger than the drain capacitance [108]. In 1–D devices operating in the quantum capacitance limit (Q_{CL}), the oxide capacitance is much larger than the quantum capacitance (C_q) as well, which provides very small gate

capacitances since $C_g = C_{ox} \times C_q/(C_{ox} + C_q) \sim C_q$ [108, 118]. The quantum capacitance can be calculated by

$$C_q = q\frac{\partial Q_{ch}}{\partial E_{v,ch}} = q^2 \frac{\partial}{\partial E_{v,ch}} \int_{E_{c,s}}^{E_{v,ch}} T_{WKB} DOS(E) \left(f_s(E) - f_D(E)\right) dE. \quad (14)$$

However, the total gate capacitance is not as small because of the fringing fields from the gate to the source and drain [108, 119], which have been taken into account in this work.

The supply voltage values for FinFETs, high–performance CNFETs and TFETs at the 16–nm technology node are taken as 0.85 V, 0.7 V and 0.18 V, respectively.

6.4 Circuit Analysis Results

The delay and EDP performances of the device–interconnect pairs in this section are calculated using a driver connected to a receiver through an interconnect of varied length assuming a fan–out of 3 as described before. In order to perform a fair comparison, we assume a CNFET inverter that is 5$\times$ the minimum size as the driver. The number of fins in a FinFET and the number of nanowires in a TFET are calculated such that the total width of the devices is the same as the CNFET. The fin pitch and nanowire pitch is assumed to be equal and as given in [59] for each technology node.

Assuming that only Cu/low–κ interconnects are used, the delay and EDP performances of CNFET and TFET circuits are compared against FinFET circuits in Figure 46 at the 16–nm technology node. It is seen that at very small interconnect lengths, CNFET devices can outperform FinFET devices by $\sim 2\times$ in terms of the intrinsic gate delay metric, CV/I, and $\sim 3\times$ in terms of EDP. On the other hand, TFET devices are significantly more resistive; hence they are $\sim 10\times$ slower than FinFET devices, but they offer significant advantages in energy. TFET circuits offer $\sim 22\times$ gain in energy due to the significant supply voltage reduction.

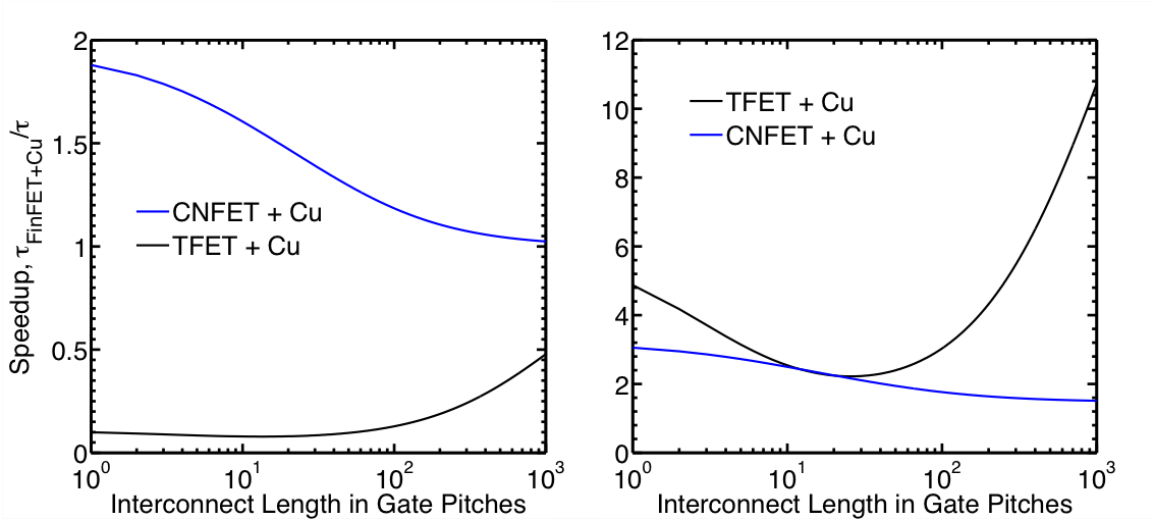

Figure 46: Relative performances of FinFET, CNFET and TFET circuits in terms of circuit delay, τ (left), and EDP (right) using Cu/low–κ interconnects at the 16–nm technology node.

In FinFET circuits, both the resistance p.u.l and capacitance p.u.l of interconnects have a significant impact on the circuit delay. Therefore, to outperform Cu/low–κ interconnects in FinFET circuits, either the resistance p.u.l or the capacitance p.u.l associated with interconnects has to be reduced significantly while avoiding a significant change in the other parameter. Individual SWNT interconnects with 2 nm diameter have much smaller capacitance p.u.l compared to Cu interconnects, but they are too resistive to be used in high–performance circuits at the 16–nm technology node. On the other hand, bundles of SWNT interconnects have significantly lower resistance p.u.l values compared to Cu interconnects at similar capacitance p.u.l values. Therefore, as illustrated in Figure 47, the best interconnect option for a FinFET circuit in terms of circuit delay is SWNTs manufactured in horizontal bundles. Multi–layer GNR interconnects may outperform Cu interconnects if the edges are perfectly smooth, with a probability of electrons

backscattering at the edges equal to 0. Even a moderate 20% edge–scattering probability, which is the best reported value [120], significantly degrades GNR performance as illustrated in Figure 47.

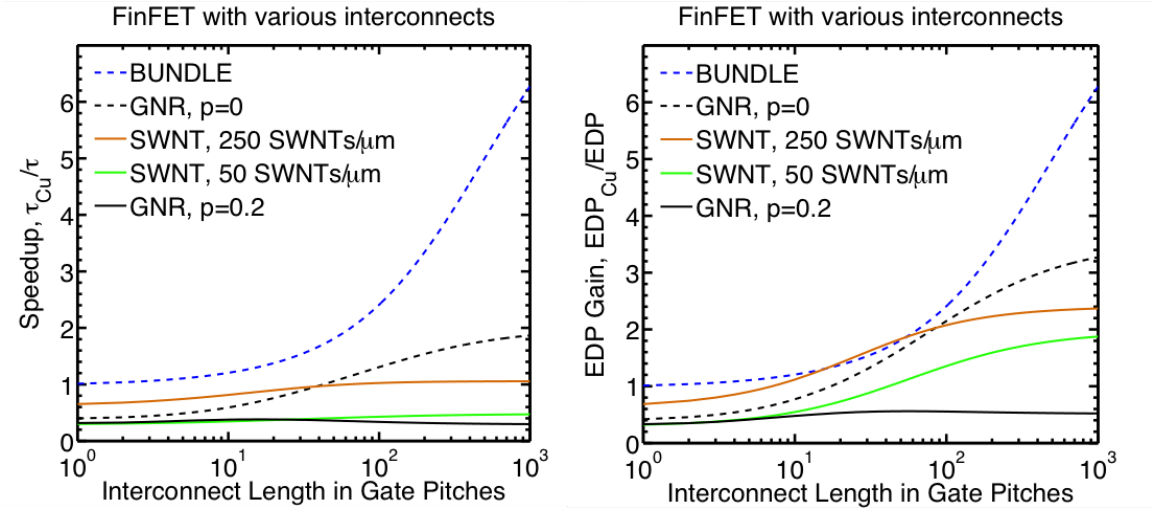

Figure 47: Relative performances of various interconnect technologies in FinFET circuits in terms of circuit delay, τ, (left) and EDP (right) at the 16–nm technology node.

Even though it is not possible to outperform Cu interconnects in terms of circuit delay with a mono–layer of SWNT interconnects at the 16–nm technology node due to their high resistance p.u.l, it is possible to benefit from their smaller capacitance p.u.l compared to Cu interconnects, which translates into a lower power dissipation. As Figure 47 demonstrates, a mono–layer of SWNTs as dense as 250 $SWNTs/\mu m$can offer $\sim 2\times$ better EDP performance than Cu at $\sim$ 100 gate pitches.

Due to the reasons underlined in Chapter 4, more opportunities arise for using these carbon–based interconnect technologies at highly scaled technology nodes. This fact is illustrated in Figure 48, where it can be seen that SWNTs can offer much larger gains in both circuit delay and EDP at the 7–nm technology node. For this to be possible, however, the density of SWNTs has to increase significantly since the minimum interconnect dimensions where tubes have to be placed are much

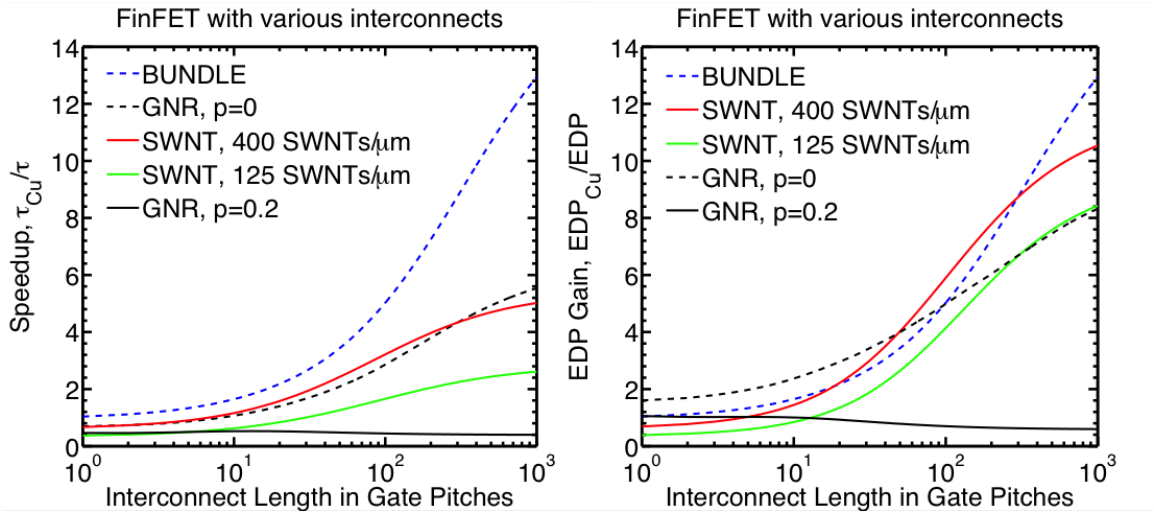

Figure 48: Relative performances of various interconnect technologies in FinFET circuits in terms of circuit delay, τ, (left) and EDP (right) at the 7–nm technology node.

smaller at future technology nodes. Assuming perfectly reliable connections, a density of at least 125 $SWNTs/\mu m$ is required to have a connection between the driver and the receiver at the 7–nm technology node.

The conclusions that can be drawn from simulations using CNFETs are very similar to FinFET circuits. Interconnect resistance and capacitance are equally effective in determining the circuit delay and bundles of SWNTs can offer the best delay performance due to their smaller interconnect resistance p.u.l compared to Cu. CNFET devices offer the smallest output resistance among the device types that are considered in this work. As a consequence, CNFETs are effected more severely from the changes in interconnect resistance p.u.l. The fact that the speedup offered by SWNT bundles over Cu is slightly larger than that in FinFET circuits proves this point. The simulation results obtained for CNFET circuits are plotted in Figure 49.

TFETs have very different requirements for interconnects than FinFETs and CNFETs. Due to the lower ON current offered by TFETs, the output resistance of TFET

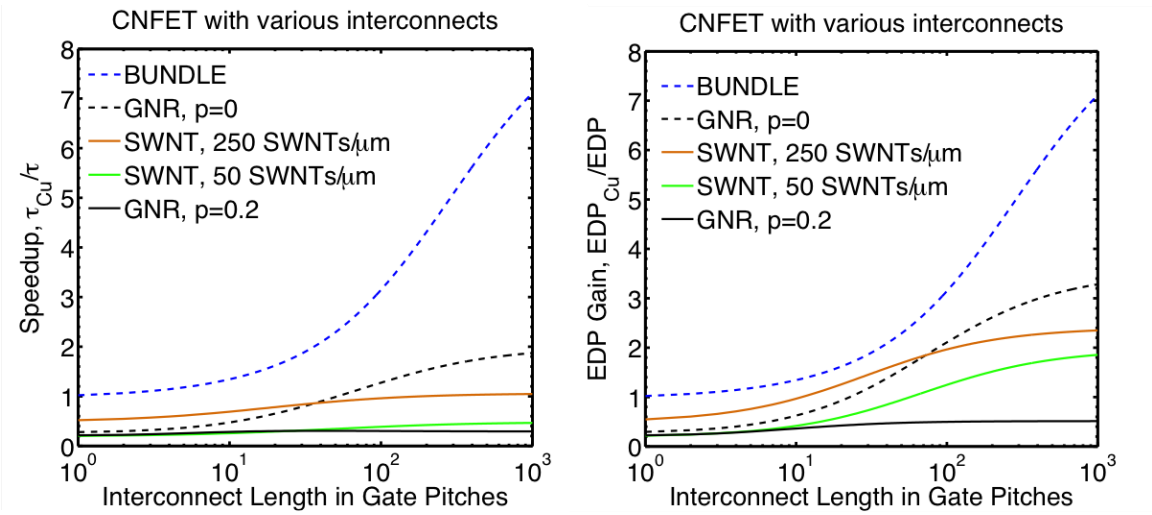

Figure 49: Relative performances of various interconnect technologies in CNFET circuits in terms of circuit delay, τ, (left) and EDP (right) at the 16–*nm* technology node.

devices is much larger than both of these device technologies. As a result, TFETs can tolerate larger interconnect resistance p.u.l. In other words, interconnect resistance is not as crucial in TFET circuits as it is in FinFET and CNFET circuits. Reducing interconnect capacitance p.u.l is more beneficial in reducing the circuit delay in TFET circuits than reducing the interconnect resistance p.u.l. Clearly, reduced interconnect capacitance means lower interconnect power dissipation as well. However, this does not mean that the resistance p.u.l of the interconnect has a negligible impact on the TFET circuit performance. Figures 50 and 51 demonstrate that the best circuit delay can be obtained by using a low–density mono–layer of SWNTs because they offer the smallest interconnect capacitance p.u.l. However, the diameter of the tubes in the mono–layer has a non–negligible impact on the speedup as shown in Figures 50 and 51 due to the different resistance p.u.l values. If tubes with a diameter of 2 *nm* are used, the resistance p.u.l can be reduced compared to 1 *nm* diameter tubes and a better speedup can be achieved. Thus, interconnect resistance p.u.l still has an impact on circuit performance even though it is not as pronounced as it is in the case of CNFET or FinFET circuits. In short, moderately

resistive low–capacitance interconnect technology options must be considered for obtaining the best performance in delay in TFET circuits.

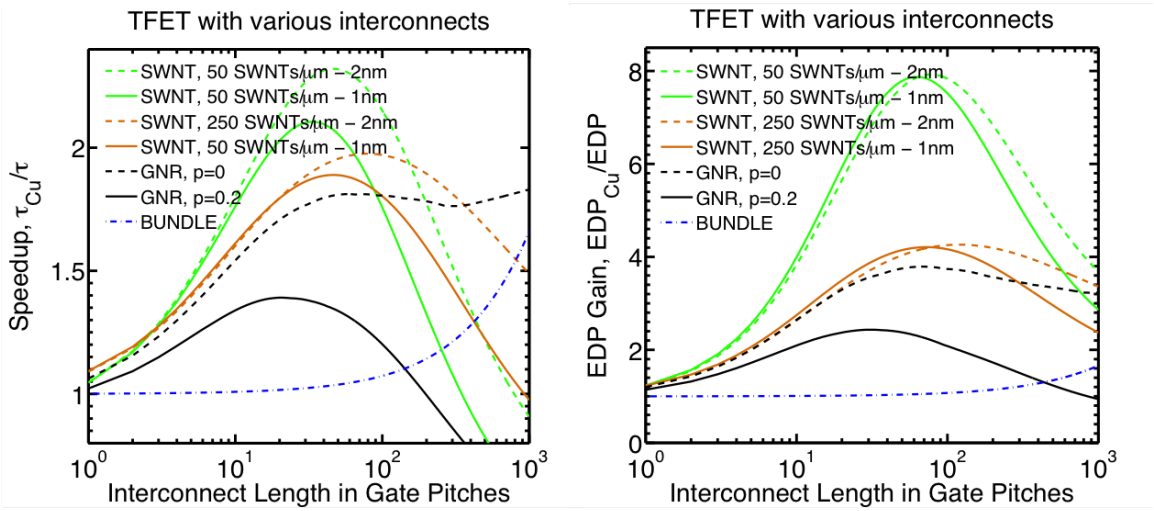

Figure 50: Relative performances of various interconnect technologies in TFET circuits in terms of circuit delay, τ, (left) and EDP (right) at the 16–*nm* technology node.

In sub–threshold CMOS circuits, the interconnect resistance p.u.l is completely dominated by the large driver resistance and has very little impact on circuit performance except at very long interconnect lengths as plotted in Figure 52. Therefore, the important interconnect technology parameter is the p.u.l. capacitance. Low–capacitance carbon–based interconnects all provide better circuit delay than when Cu interconnect is used in sub–threshold circuits. Even with GNR interconnects with an electron backscattering probability of 0.2 at the edges, it is possible to achieve better circuit delay than with Cu/low–κ interconnect. At long interconnect lengths, however, the delay components associated with the resistance of the interconnect are comparable to the components associated with the driver resistance and speedup values shown in Figure 52 drop.

For sub-threshold circuits, only SWNTs with 1 *nm* diameter are considered since wire resistance does not have a significant impact on circuit performance.

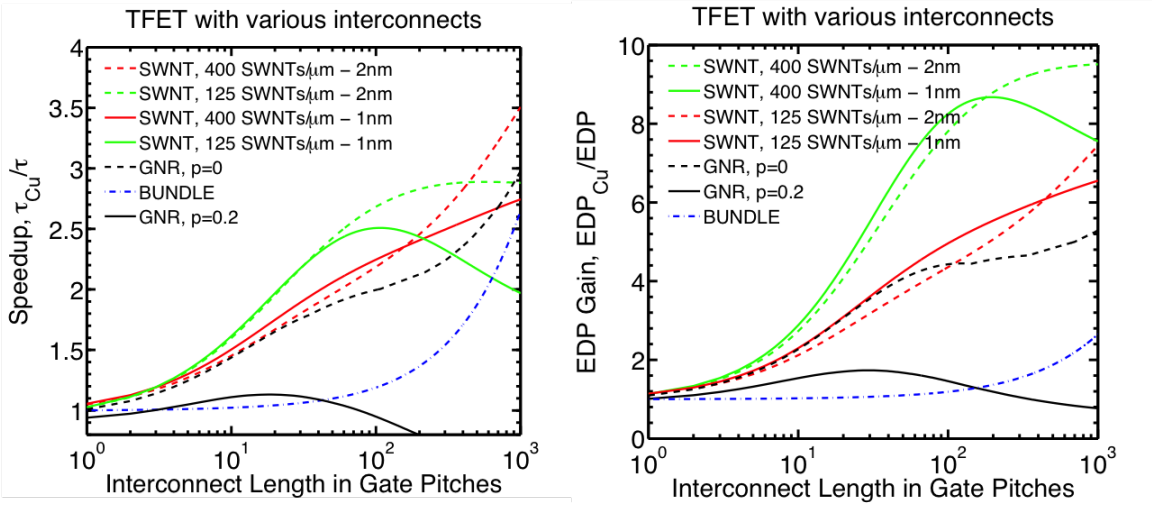

Figure 51: Relative performances of various interconnect technologies in TFET circuits in terms of circuit delay, τ, (left) and EDP (right) at the 7–*nm* technology node.

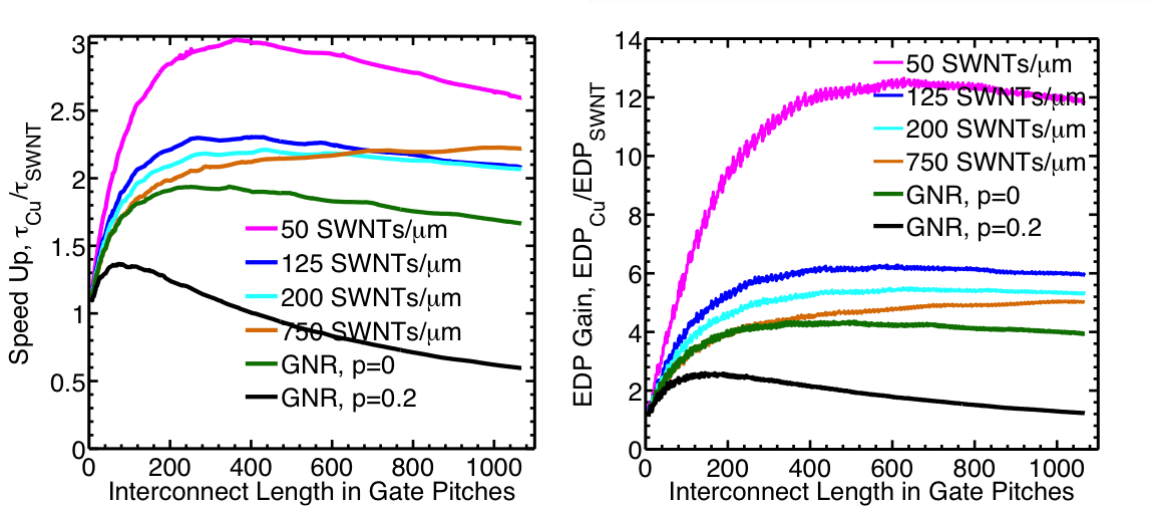

Figure 52: Relative performances of various interconnect technologies in CMOS circuits operated in the sub–threshold regime in terms of circuit delay, τ, (left) and EDP (right) at the 16–*nm* technology node.

Note that as the density of tubes in a mono-layer are increased, the associated capacitance p.u.l increases and the maximum speedup is lowered.

Table 23 summarizes all the results that are explained in this section and quantifies how much interconnect resistance and capacitance p.u.l impact circuit performance for various device types at short and long interconnect lengths. Also, the first three best interconnect options that maximize EDP performance at short and long interconnect lengths for each device are tabulated.

Table 23: Comparison table summarizing the simulation results.

Device Type	Device Resistance	Device Capacitance	Interconnect Resistance Impact		Interconnect Capacitance Impact		Best Interconnect Option (targeting EDP)	
			Short	Long	Short	Long	Short	Short
FinFET	Reference	Reference	✓	✓✓✓	✓✓✓	✓✓✓	Bundle	Bundle GNR* SWNT*
CNFET	Low	Low	✓✓	✓✓✓	✓✓✓	✓✓✓	Bundle	Bundle GNR* SWNT*
TFET	High	Low		✓✓	✓✓✓✓	✓✓✓	SWNT GNR	SWNT SWNT* GNR
Sub–V_{th}	Very High	Reference		✓✓	✓✓✓✓	✓✓✓✓	SWNT SWNT* GNR*	SWNT SWNT* GNR*

6.5 Conclusions

Interconnect requirements of various candidates for the main device technologies in the post–CMOS era are investigated. The driver resistance and receiver capacitance are estimated using predictive SPICE models and running HSPICE simulations for CNFET, FinFET and sub–threshold CMOS circuits. Recently developed analytical models for InAs–based GAA TFET circuits are used to estimate driver output resistance and receiver input capacitance in TFET circuits. The types of interconnects that best suit each of these devices in terms of the circuit delay and EDP are reported.

It is shown that different interconnect technologies can outperform the conventional Cu/low–κ interconnect depending on the type of switches used because the output resistance and input capacitance of these switches can vary quite significantly. Carbon–based interconnects can find use in various device technology options and their use becomes even more beneficial as the technology scaling continues. It is shown for FinFET circuits that SWNT interconnect benefits increase as the technology scales down to 7–nm technology node due to the significant performance degradation of Cu/low–κ interconnect due to size effects.

CHAPTER 7

CONCLUSIONS AND FUTURE DIRECTIONS

Interconnects are an ever growing challenge to continue improving the performances of electronic chips. Both local and global interconnects introduce limitations in both latency and power dissipation not only because of the ever increasing resistive and capacitive load they impose on the system, but also the negative impacts the solutions of these problems have on the system such as repeater power dissipation, repeater area, routing congestion and via blockage. Improving the transistor delay and energy dissipation with every technology generation will escalate the interconnect problem in future technology generations. Furthermore, interconnects impose reliability concerns due to electromigration and time–dependent dielectric breakdown (TDDB) due to larger electric fields.

7.1 Conclusions and Contributions

In this dissertation, first, we designed and benchmarked the conventional Cu/low–κ interconnect technology for future technology nodes. We showed that, contrary to previous publications, the Cu resistivity increase due to miniaturization can cause a significant increase in the required number of metal levels, and we investigated the reasons for this change. We also demonstrated the impacts of various interconnect process parameters, for instance, the interconnect barrier/liner bilayer thickness, and aspect ratio, on the design of a multilevel interconnect network for FinFET devices.

Furthermore, we created a framework to perform an interconnect sensitivity analysis for future FinFET CMOS technology nodes based on actual netlists and timing closed GDSII–level layouts with detailed routing. Multiple standard cell

and interconnect libraries are created to compare the performances of various design options. We showed that the impact of interconnect performance degradation on the circuit speed and power dissipation highly depends on the circuit. By considering three circuits of different sizes and layout structures, we categorized different types of circuits to make generic conclusions on this sensitivity.

Contrary to previous publications, which have indicated that individual single–wall carbon nanotube (SWNT) interconnects are too resistive for high performance CMOS applications and that they must be used in bundles, we demonstrated that they can offer significant delay and energy–per–bit improvements in future high–performance circuits. We compared the performances of various designs comprising one or a few parallel SWNTs against the performance of the conventional Cu/low–κ interconnect considering the impact of potentially broken tubes. Considering that manufacturing horizontal bundles of dense CNT interconnects have turned out to be challenging, and that there have been promising advances in making horizontal arrays of dense SWNTs, we showed that the latter scenario can potentially become an earlier solution to the interconnect problem from the materials perspective.

We presented the first system–level study on the impact of CNFETs on the multilevel interconnect networks. We determined the requirements imposed on supply voltage value and carbon nanotube density for better performance, based on system–level parameters such as the number of metal levels, the maximum clock frequency and the number of logic cores. We showed that any device that is faster than Si–CMOS will be slowed down by interconnects more severely and will need rigorous optimizations for best performance. We also investigated the trade–offs between the technology parameters of various interconnect technologies on the basis of their impacts on the performances of FinFET, CNFET, TFET and sub–threshold circuits.

7.2 Future Work

From the circuit design and reliability point of view, the interconnect metrics studied in this dissertation are not equally important for all wires. For instance, at the local metal levels, the requirements for power/ground wires are very different compared to signal interconnects. The important parameters to optimize are the IR Drop and simultaneous switching noise (SSN) for power/ground interconnects whereas delay, energy–per–bit and crosstalk are the important parameters for signal wires. For power/ground interconnects, targeting low resistance is more important than in short signal interconnects as the resistances of the majority of the short interconnects are dominated by the driver resistance in CMOS chips. For short signal interconnects, targeting a small capacitance is important to reduce both the latency and the power dissipation as the aggregate capacitance accounts for a significant portion of the dynamic power dissipation in electronic chips. Similarly, electromigration is more pronounced for power/ground interconnects since large DC currents run through them. Signal interconnects are less vulnerable to electromigration because they conduct bidirectional AC currents. Therefore, there are many different parameters that need to be co–optimized together and the traditional approach to target smaller resistance and capacitance values while maximizing resistance to electromigration and TDDB can be improved to include these effects.

Based on this discussion, the work in Chapter 2 can be continued such that a variety of design options can be explored based on stochastic wiring distributions to account for the specialization for both devices and interconnects in future technology nodes. Hybrid interconnect structures that can address the different requirements for different connections can be studied with this approach. For instance, based on the results on this dissertation, a hybrid multilevel interconnect network structure with GNR interconnects at the local signal levels can be studied,

which may reduce the total interconnect capacitance at these levels while keeping a comparable performance. Similarly, since via resistance becomes a significant contributor to circuit delay at future technology nodes, the implications of using new via structures that use vertical CNT bundles at future technology nodes can be explored. Furthermore, since different device options have different requirements in terms of the aforementioned interconnect metrics, this study can be extended to include various device options. For instance, those parts on the chip which may not need high speed but require low–power operation can be implemented using TFETs. The ultimate goal in this study can be to create a generic system–level simulator, which can take into account the properties of various device and interconnect options to optimally pair them in terms of user–defined parameters such as speed or power dissipation.

The sensitivity analysis framework that is outlined in Chapter 3 can be extended to multiple device technologies including III-V devices, tunneling FETs and nanowire FETs to investigate the interconnect requirements for these new device technologies based on actual netlists and GDSII–level layouts.

Patterning problems become more pronounced as multiple–lithography techniques become a common method to extend the use of 193–nm lithography tools until EUV lithography is ready. The ramifications of the inherent variation problem with these technologies and the required regularity in layout due to manufacturing limitations can be studied using the framework that is outlined in Chapter 3. To speed the analysis involving variations, the results based on actual netlists and GDSII–level layouts for conventional or emerging devices can be used to calibrate the design methodology based on stochastic wiring distributions. This way, a faster simulation environment can be used to explore more design options accurately.

So far, potential research topics in the conventional charge–based technologies are investigated. The semiconductor industry also encourages researching non–charge–based systems to extend Moore's Law to beyond–2020 technology generations. Any device technology that offers advantages in performance, power dissipation or ease in dimensional scaling will have to be complemented with an interconnect technology that offers similar trades and all this research has to be interconnect–centric. On the other hand, interconnect requirements, noise mechanisms, impact of parasitics on performance can be drastically different for these new technologies.

One of the promising candidates for a new state variable is the electron spin. Spin–based devices communicate through the orientation of the electron rotation. Recently, the all–spin–logic technology is being investigated actively. In this technology, information is transmitted between nanomagnets through a combination of spin diffusion and spin transfer torque. Performance of this system has a greater dependence on the length of the interconnect than the charge–based system. Therefore, correct estimation of the interconnect distribution based on layout becomes critical for accurate performance estimation. It would be very interesting to analyze this issue on the physical layout level.

Furthermore, placement of the cells in this magnetic circuit may put major constraints on circuit performance due to circuit parasitics. Similar to the crosstalk problem in charge–based systems, the stray fields generated by the magnets may impact the energy profiles of the neighboring magnets, which may potentially cause the neighboring magnets to become unstable during switching. Depending on the magnitude of this impact, logic errors may cause the circuit to fail to implement the intended function. Placing the magnets far from each other may alleviate this problem depending on the strength of the fields, but will cause the

wirelength to increase, which contradicts with the requirement to reduce interconnect length. Studying the tradeoffs between these contradicting requirements can be a very interesting research topic. This research can have multiple stages involving (1) creating device and interconnect models for spin–based systems, (2) constructing appropriate standard cell and interconnect libraries for early investigation of performance using commercial placement and routing tools, and (3) creating scripts and patches to attach to these commercial tools, which can take into account correct physical performance models associated with the spin–based system.

One very interesting study would be to investigate the variation mechanisms that may play a role in these emerging systems. During the initial stages of this work, impacts of dimensional variation, variation due to thermal noise and variations in the external field can be analyzed based on physics–based models, which can then be taken into account in the later stages of physical design and layout analysis.

REFERENCES

[1] R. Dennard, F. Gaensslen, V. Rideout, E. Bassous, and A. LeBlanc, "Design of ion–implanted MOSFETs with very small dimensions," *IEEE Journal of Solid-State Circuits*, vol. 9, pp. 256–268, October 1974.

[2] D. Buchanan, "Scaling the gate dielectric: materials, integration and reliability," *IBM Journal of Research and Development*, vol. 43, pp. 245–264, May 1999.

[3] Y. Yeo, Q. Lu, W. Lee, T.-J. King, C. Hu, X. Wang, and T. Ma, "Direct tunneling gate leakage current in transistors with ultrathin silicon nitride gate dielectric," *IEEE Electron Device Letters*, vol. 21, pp. 540–542, November 2000.

[4] P. Bai *et al.*, "A 65nm logic technology featuring 35nm gate lengths, enhanced channel strain, 8 Cu interconnect layers, low-κ ILD and 0.57 μm^2 SRAM cell," in *IEDM Technical Digest*, pp. 657–660, December 2004.

[5] D. Antoniadis, I. Aberg, C. Ni Chleirigh, O. Nayfeh, A. Khakifirooz, and J. Hoyt, "Continuous MOSFET performance increase with device scaling: the role of strain and channel material innovations," *IBM Journal of Research and Development*, vol. 50, pp. 363–376, July/September 2006.

[6] R. Chau, S. Datta, M. Doczy, J. Kavalieros, and M. Metz, "Gate dielectric scaling for high–performance CMOS: from SiO_2 to high–κ," in *Intl. Workshop on Gate Insulator*, pp. 124–126, November 2003.

[7] C. Auth *et al.*, "45nm high-–κ + metal gate strain–enhanced transistors," in *Symposium on VLSI Technology*, pp. 128–129, June 2008.

[8] C. Auth *et al.*, "A 22nm high–performance and low–power CMOS technology featuring fully–depleted tri–gate transistors, self–aligned contacts and high density MIM capacitors," in *Symposium on VLSI Technology*, pp. 131–132, June 2012.

[9] J. Meindl *et al.*, "Interconnecting device opportunities for gigascale integration (GSI)," in *IEDM Technical Digest*, pp. 23.1.1–23.1.4, December 2001.

[10] M. Bohr, "The new era of scaling in an SoC world," in *IEEE International Solid State Circuits Conference*, pp. 23–28, February 2009.

[11] D. Edelstein *et al.*, "Full copper wiring in a sub–0.25 μm CMOS ULSI technology," in *IEDM Technical Digest*, pp. 773–776, December 1997.

[12] Y.-M. Lin *et al.*, "100–GHz Transistors from Wafer Scale Epitaxial Graphene," *Science*, vol. 327, p. 662, February 2010.

[13] A. Bachtold, P. Hadley, T. Nakanishi, and C. Dekker, "Logic circuits with carbon nanotube transistors," *Science*, vol. 294, pp. 1317–1320, October 2001.

[14] A. Naeemi and J. Meindl, "Design and performance modeling for single–walled carbon nanotubes as local, semiglobal, and global interconnects in gigascale integrated systems," *IEEE Transactions on Electron Devices*, vol. 54, pp. 26–37, January 2007.

[15] A. Naeemi and J. Meindl, "Compact physics–based circuit models for graphene nanoribbon interconnects," *IEEE Transactions on Electron Devices*, vol. 56, pp. 1822–1833, September 2009.

[16] R. Nathanael, V. Pott, H. Kam, J. Jeon, and T.-J. Liu, "4–terminal relay technology for complementary logic," in *IEDM Technical Digest*, pp. 1–4, December 2009.

[17] R. Beausoleil *et al.*, "Nanoelectronic and nanophotonic interconnect," in *Proceedings of the IEEE*, vol. 96, pp. 230–246, February 2008.

[18] A. Krishmamoorthy *et al.*, "Computer systems based on silicon photonic interconnects," in *Proceedings of the IEEE*, vol. 97, pp. 1337–1361, July 2009.

[19] B. Behin-Aein, D. Datta, S. Salahuddin, and S. Datta, "Proposal for an all–spin logic device with built–in memory," *Nature Nanotechnology*, pp. 266–270, February 2010.

[20] J. Kilby, "Invention of the integrated circuit," *IEEE Transactions on Electron Devices*, vol. 23, pp. 648–654, July 1976.

[21] S. Danko, "New developments in the auto–sembly technique of circuit fabrication," in *Proceedings of the National Electronics Conference*, pp. 542–550, October 1951.

[22] R. Noyce, "A look at future costs of large integrated arrays," in *Proceedings of the Fall Joint Computer Conference*, pp. 111–114, 1966.

[23] G. Moore, "Cramming more components into integrated circuits," *Electronics Magazine*, vol. 38, pp. 114–117, April 1965.

[24] H. Bakoglu and J. Meindl, "Optimal interconnection networks for ulsi," *IEEE Transactions on Electron Devices*, vol. 32, pp. 903–909, May 1985.

[25] "International technology roadmap for semiconductors." Online, http://www.itrs.net.

[26] D. Miller, "Rationale and challenges for optical interconnects to electronic chips," *Proceedings of the IEEE*, vol. 88, pp. 728–749, June 2000.

[27] D. Miller, "Device requirements for optical interconnects to silicon chips," *Proceedings of the IEEE*, vol. 97, pp. 1166–1185, July 2009.

[28] K. Koo, H. Cho, P. Kapur, and K. Saraswat, "Performance comparisons between carbon nanotubes, optical, and Cu for future high performance on–chip interconnect applications," *IEEE Transactions on Electron Devices*, vol. 54, pp. 3206–3215, December 2007.

[29] G. Lopez, J. Davis, and J. Meindl, "A new physical model and experimental measurements for copper interconnect resistivity considering size effects and line-edge roughness (LER)," in *IEEE International Interconnect Technology Conference*, 2009.

[30] C. Moore, "Data processing in exascale–class computer systems," in *The Salishan Conference on High Speed Computing*, April 2011.

[31] N. Magen, A. Kolodny, U. Weiser, and N. Shamir, "Interconnect power dissipation in a microprocessor," in *International Workshop on System Level Interconnect Prediction*, pp. 7–13, 2004.

[32] J. Davis, V. De, and J. Meindl, "A stochastic wire-length distribution for gigascale integration (GSI) - Part I: derivation and validation," *IEEE Transactions on Electron Devices*, vol. 45, pp. 580–589, March 1998.

[33] A. Naeemi, R. Sarvari, and J. Meindl, "On–chip interconnect networks at the end of the roadmap: limits and nanotechnology opportunities," in *IEEE International Interconnect Technology Conference*, pp. 201–203, June 2006.

[34] J. Gambino, T. Lee, F. Chen, and T. Sullivan, "Reliability challenges for advanced copper interconnects: electromigration and time–dependent dielectric breakdown (TDDB)," in *IEEE International Symposium on the Physical and Failure Analysis of Integrated Circuits*, pp. 677–684, July 2009.

[35] S. Nassif, G.-J. Nam, and S. Banerjee, "Wire delay variability in nanoscale technology and its impact on physical design," in *International Symposium on Quality Electron Design*, pp. 591–596, March 2013.

[36] R. Sarvari, A. Naeemi, R. Venkatesan, and J. Meindl, "Impact of size effects on the resistivity of copper wires and consequently the design and performance of metal interconnect networks," in *IEEE International Interconnect Technology Conference*, 2005.

[37] D. Sekar, A. Naeemi, R. Sarvari, and J. Meindl, "Intsim: A CAD tool for optimization of multilevel interconnect networks," in *IEEE International Interconnect Technology Conference*, 2007.

[38] D. Frank, W. Haensch, G. Shahidi, and O. Dokumaci, "Optimizing CMOS technology for maximum performance," *IBM Journal of Research and Development*, vol. 50, pp. 419–431, July/September 2006.

[39] J. Eble, *A generic system simulator with novel on–chip cache and throughput models for gigascale integration*. Ph.d. thesis, Georgia Institute of Technology, Atlanta, November 1998.

[40] R. Venkatesan, J. Davis, K. Bowman, and J. Meindl, "Optimal n–tier multilevel interconnect architectures for gigascale integration (GSI)," *IEEE Transactions on VLSI Systems*, vol. 9, pp. 899–912, December 2001.

[41] D. Sekar, R. Venkatesan, K. Bowman, A. Joshi, J. Davis, and J. Meindl, "Optimal repeaters for sub–50 nm interconnect networks," in *IEEE International Interconnect Technology Conference*, pp. 199–201, June 2006.

[42] Q. Chen, J. Davis, P. Zarkesh-Ha, and J. Meindl, "A compact physical via blockage model," *IEEE Transactions on VLSI Systems*, vol. 9, pp. 689–692, December 2000.

[43] R. Sarvari, A. Naeemi, P. Zarkesh-Ha, and J. Meindl, "Design and optimization for nanoscale power distribution networks in gigascale systems," in *IEEE International Interconnect Technology Conference*, pp. 190–192, June 2007.

[44] S. Damaraju *et al.*, "A 22nm IA multi–CPU and GPU system–on–chip," in *IEEE International Solid–State Circuits Conference*, pp. 56–57, February 2012.

[45] C. Auth, "22nm fully–depleted tri–gate CMOS transistors," in *IEEE Custom Integrated Circuits Conference*, pp. 1–6, September 2012.

[46] S. Borkar, "Thousand core chips —a technology perspective," in *Design Automation Conference*, pp. 746–749, 2007.

[47] J. Plombon, E. Andideh, V. Dubin, and J. Maiz, "Influence of phonon, geometry, impurity, and grain size on copper line resistivity," *Applied Physics Letters*, vol. 89, no. 11, pp. 113124–1–113124–3, 2006.

[48] D. Nikonov and I. Young, "Overview of beyond–CMOS devices and a uniform methodology for their benchmarking." arXiv preprint arXiv:1302.0244, 2013.

[49] S. Naffziger *et al.*, "The implementation of a 2–core multi–threaded Itanium family processor," *IEEE Journal of Solid-State Circuits*, vol. 41, pp. 197–209, January 2006.

[50] F. Wu, G. Levitin, and W. Hess, "Low-temperature etching of Cu by hyrogen-based plasmas," *ACS Applied Materials and Interfaces*, vol. 2, pp. 2175–2179, July 2010.

[51] W. Steinhoegl, G. Schindler, G. Steinlesberger, M. Traving, and M. Engelhardt, "Comprehensive study of the resistivity of copper wires with lateral dimensions of 100nm and smaller," *Journal of Applied Physics*, vol. 97, pp. 023706–1–023706–7, January 2005.

[52] M. Shimada, M. Moriyama, K. Ito, S. Tsukimoto, and M. Murakami, "Electrical resistivity of polycrystalline Cu interconnects with nano–scale linewidth," *Journal of Vacuum Science and Technology B*, vol. 24, pp. 190–194, January 2006.

[53] W. Steinhoegl, G. Schindler, G. Steinlesberger, M. Traving, and M. Engelhardt, "Impact of line edge roughness on the resistivity of nanometer–scale interconnects," *Microelectronic Engineering*, vol. 76, pp. 126–130, October 2004.

[54] H. Kitada *et al.*, "The influence of the size effect of copper interconnects on *RC* delay variability beyond 45nm technology," in *IEEE International Interconnect Technology Conference*, pp. 10–12, June 2007.

[55] H.-C. Chen, H.-W. Chen, S.-P. Jeng, C.-M. Wu, and J.-C. Sun, "Resistance increase in metal nano–wires," in *International Symposium on VLSI Technology, Systems and Applications*, pp. 1–2, April 2006.

[56] W. Besling, M. Broekaart, V. Arnal, and J. Torres, "Line resistance behavior in narrow lines patterned by a TiN hard mask spacer for 45 nm node interconnects," *Microelectronic Engineering*, vol. 76, pp. 167–174, October 2004.

[57] J. Guillaumond *et al.*, "Analysis of resistivity in nano–interconnect: full range (4.2–300 *K*) temperature characterization," in *IEEE International Interconnect Technology Conference*, pp. 132–134, June 2003.

[58] W. Steinhoegl, G. Schindler, and M. Engelhardt, "Unraveling the mysteries behind size effects in metallization systems," in *Semiconductor International*, vol. 28, pp. 34–38, May 2005.

[59] S. Sinha, G. Yeric, V. Chandra, B. Cline, and Y. Cao, "Exploring sub-20nm FinFET design with predictive technology models," in *Design Automation Conference*, pp. 283–288, 2012.

[60] M. Dunga, C.-H. Lin, A. Niknejad, and C. Hu, *FinFETs and Other Multi–Gate Transistors*, ch. 3: BSIM–CMG: A compact model for multi–gate transistors, pp. 113–153. Springer US, 2008.

[61] A. Kaloyeros, E. E.T., K. Dunn, and O. Van der Straten, "Zero thickness diffusion barriers and metallization liners for nanoscale device applications," *Chemical Engineering Communications*, vol. 198, pp. 1453–1481, June 2011.

[62] Nangate, "Nangate FreePDK45 Open Cell Library." Online http://www.nangate.com.

[63] Synopsys, "Synopsys Design Compiler." Online http://www.synopsys.com.

[64] Cadence Design Systems, "Encounter Digital Implementation System." Online http://www.cadence.com.

[65] J.-C. Chen, T. Standaert, E. Alptekin, T. Spooner, and V. Paruchuri, "Interconnect performance and scaling strategy at 7nm node," in *IEEE International Interconnect Technology Conference*, pp. 93–96, May 2014.

[66] Synopsys, "Synopsys Raphael." Online http://www.synopsys.com.

[67] H. Ren and M. Naik unpublished.

[68] R. Ho, K. Mai, and M. Horowitz, "The future of wires," *Proceedings of the IEEE*, vol. 89, pp. 490–504, April 2001.

[69] J. Davis *et al.*, "Interconnect limits on gigascale integration (GSI) in the 21^{st} century," *Proceedings of the IEEE*, vol. 89, pp. 305–324, March 2001.

[70] A. Nieuwoudt and Y. Massoud, "Evaluating the impact of resistance in carbon nanotube bundles for VLSI interconnect using diameter–dependent modeling techniques," *IEEE Transactions on Electron Devices*, vol. 53, pp. 2460–2466, October 2006.

[71] O. Jamal and A. Naeemi, "Ultra–low power single–wall carbon nanotube interconencts for subthreshold circuits," *IEEE Transactions on Nanotechnology*, vol. 10, pp. 99–101, January 2011.

[72] J. Li, Q. Ye, A. Cassell, H. Ng, R. Stevens, J. Han, and M. Meyyappan, "Bottom–up approach for carbon nanotube interconnects," *Applied Physics Letters*, vol. 82, pp. 2491–2493, April 2003.

[73] F. Kreup, A. Graham, M. Liebau, G. Duesberg, and R. Seidel, "Carbon nanotubes for interconect applications," in *International Electron Devices Meeting*, pp. 683–686, 2004.

[74] M. Nihei *et al.*, "Low–resistance multi–walled carbon nanotube vias with parallel channel conduction of inner shells [IC interconnect applications]," in *IEEE International Interconnect Technology Conference*, pp. 234–236, June 2005.

[75] M. Nihei *et al.*, "Electrical properties of carbon nanotube via interconnects fabricated by novel damascene process," in *IEEE International Interconnect Technology Conference*, pp. 204–206, June 2007.

[76] Y. Awano, "Carbon nanotube technologies for LSI via interconnects," *IEICE Transactions on Electronics*, vol. E89-C, pp. 1499–1503, November 2006.

[77] Y. Choi *et al.*, "Integration and electrical properties of carbon nanotube array for interconnect applications," in *IEEE Conference on Nanotechnology*, pp. 262–265, 2006.

[78] B. Wei *et al.*, "Microfabrication technology: organized assembly of carbon nanotubes," *Nature*, vol. 416, pp. 495–496, April 2002.

[79] D. Futaba *et al.*, "84% catalysy activity of wafer–assisted growth of single walled carbon nanotube forest characterization by a statistical and macroscopic approach," *Journal of Physical Chemistry B*, vol. 110, pp. 8035–8038, April 2006.

[80] Y. Maeda *et al.*, "Large scale separation of metallic and semi–conducting single–walled carbon nanotubes," *Journal of American Chemical Society*, vol. 127, pp. 10287–10290, July 2005.

[81] Y. Chai, X. Z., and P. Chan, "Electron–shading effect on the horizontal aligned growth of carbon nanotubes," *Applied Physics Letters*, vol. 94, p. 043116, January 2009.

[82] C. Kocabas, S. Kang, T. Ozel, M. Shim, and J. Rogers, "Improved synthesis of aligned arrays of single–walled carbon nanotubes and their implementationin thin film type transistors," *The Journal of Physical Chemistry C*, vol. 111, no. 48, pp. 17879–17886, 2007.

[83] L. Ding, D. Yuan, and J. Liu, "Growth of high–density parallel arrays of long single–walled carbon nanotubes on quartz substrates," *Journal of American Chemical Society*, vol. 130, pp. 5428–5429, April 2008.

[84] W. Zhou, R. C., and B. P.J., "Wafer scale synthesis of dense aligned arrays of single–walled carbon nanotubes," *Nano Research*, vol. 1, pp. 158–165, August 2008.

[85] A. Ismach, L. Segev, E. Wachtel, and E. Joselevich, "Atomic–step–templated formation of single wall carbon nanotube patterns," *Angewandte Chemie*, vol. 43, pp. 6140–6143, November 2004.

[86] S. Han, X. Liu, and C. Zhou, "Template free directional growth of single–walled carbon nanotubes on a– and r–plane sapphire," *Journal of American Chemical Society*, vol. 127, pp. 5294–5295, 2005.

[87] H. Ago, K. Nakamura, K.-I. Ikeda, N. Uehara, N. Ishigami, and M. Tsuji, "Aligned growth of isolated single–walled carbon nanotubes programmed by atomic arrangement of substrate surface," *Chemical Physics Letters*, vol. 408, pp. 433–438, June 2005.

[88] Y. Kim *et al.*, "Highly aligned scalable platinum–decorated single wall carbon nanotube arrays for nanoscale electrical interconnects," *ACS Nano*, vol. 3, pp. 2818–2826, September 2009.

[89] A. Naeemi and J. Meindl, "Carbon nanotube interconnects," *Annual Review of Materials Resarch*, vol. 39, pp. 255–275, January 2009.

[90] C. White and T. Todorov, "Carbon nanotubes as long ballistic conductors," *Nature*, vol. 393, pp. 240–242, May 1998.

[91] J. Park *et al.*, "Electron–phonon scattering in metallic single–walled carbon nanotubes," *Nano Letters*, vol. 4, pp. 517–520, March 2004.

[92] Comsol, "Comsol Multiphysics." Online http://www.comsol.com.

[93] J. Plombon, K. OBrien, F. Gstrein, V. Dubin, and Y. Jiao, "High–frequency electrical properties of individual and bundled carbon nanotubes," *Applied Physics Letters*, vol. 90, pp. 063106–1—063106–3, February 2007.

[94] K. Saraswat and H.-S. Wong unpublished.

[95] R. Venkatesan, J. Davis, and J. Meindl, "Compact distributed RLC interconnect models—Part IV: Unified models for time delay, crosstalk, and repeater insertion," *IEEE Transactions on Electron Devices*, vol. 50, pp. 1094–1102, April 2003.

[96] W. Zhou, L. Ding, S. Yang, and J. Liu, "Synthesis of high–density large–diameter, and aligned single–walled carbon nanotubes by multiple cycle growth methods," *ACS Nano*, vol. 5, pp. 3849–3857, May 2011.

[97] J. Zhang *et al.*, "Carbon nanotube electronics—materials, devices, circuits design, modeling, and performance projection," in *International Electron Devices Meeting*, pp. 23.1.1–23.1.4, December 2011.

[98] J. Deng and H.-S. Wong, "A compact SPICE model for carbon nanotube field–effect transistors including non–idealities and its application—part I: model of the intrinsic channel region ," *IEEE Transactions on Electron Devices*, vol. 54, pp. 3186–3194, December 2007.

[99] J. Deng and H.-S. Wong, "A compact SPICE model for carbon nanotube field–effect transistors including non–idealities and its application—part II: full device model and circuit performance benchmarking," *IEEE Transactions on Electron Devices*, vol. 54, pp. 2195–3205, December 2007.

[100] J. Guo, A. Javey, H. Dai, and L. M., "Performance analysis and design optimization of near–ballistic carbon nanotube field–effect transistors," in *Proceedings of the International Electron Devices Meeting*, pp. 703–706, December 2004.

[101] H.-S. Wong, J. Appenzeller, V. Derycke, R. Martel, S. Wind, and P. Avouris, "Carnon nanotube field effect transistors —fabrication, device physics and circuit implications," in *Proceedings of the International Solid State Circuits Conference*, pp. 370–371, 2003.

[102] N. Patil, J. Deng, S. Mitra, and H.-S. Wong, "Circuit–level performance benchmarking and scalability analysis of carbon nanotube transistor circuits," *IEEE Transactions on Nanotechnology*, vol. 8, January 2009.

[103] Arizona State University, "Predictive technology models." url http://www.ptm.asu.edu.

[104] S. Kang, C. Kocabas, T. Ozel, M. Shim, N. Pimparkar, M. Alam, S. Rotkin, and J. Rogers, "High–performance electronics using dense, perfectly aligned arrays of single–walled carbon nanotubes," *Nature Nanotechnology*, vol. 2, pp. 230–236, 2007.

[105] X. Li, L. Zhang, X. Wang, I. Shimoyama, X. Sun, W.-S. Seo, and H. Dai, "Langmuir–Blodgett assembly of densely aligned single–walled carbon nanotubes from bulk materials," *Journal of American Chemical Society*, vol. 129, no. 16, pp. 4890–4891, 2007.

[106] H. Bakoglu, *Circuits, Interconnections and Packaging for VLSI*. Addison-Wesley, 1990.

[107] N. Patil, J. Deng, A. Lin, and H.-S. Wong, "Design methods for misaligned and mispositioned carbon–nanotube immune circuits," *IEEE Transactions on Computer–Aided Design of Integrated Circuits and Systems*, vol. 27, pp. 1725–1736, October 2008.

[108] A. Seabaugh and Q. Zhang, "Low–voltage tunnel transistors for beyond CMOS logic," *Proceedings of the IEEE*, vol. 98, pp. 2095–2110, December 2010.

[109] J. Appenzeller, J. Knoch, T. Bjork, H. Riel, H. Schmidt, and W. Reiss, "Toward nanowire electronics," *IEEE Transactions on Electron Devices*, vol. 55, pp. 2827–2845, November 2008.

[110] A. Maarouf, C. Kane, and E. Mele, "Electronic structure of carbon nanotube ropes," *Physical Review B*, vol. 61, pp. 11156–11165, April 2000.

[111] M. Dresselhaus, G. Dresselhaus, and P. Avouris, *Topics in Applied Physics, Carbon Nanotubes: Synthesis, Structure, Properties and Applications*. New York: Springer-Verlag, 2000.

[112] V. Kumar, S. Rakheja, and A. Naeemi, "Performance and energy–per–bit modeling of multilayer graphene nanoribbon conductors," *IEEE Transactions on Electron Devices*, vol. 59, pp. 2753–2761, October 2012.

[113] K. Bolotin *et al.*, "Ultrahigh electron mobility in suspended graphene," *Solid State Communications*, vol. 146, pp. 351–355, 2008.

[114] M. Luisier and G. Klimeck, "Atomistic full–band design study of inas band–to–band–tunneling field–effect transistors," *IEEE Electron Device Letters*, vol. 30, pp. 602–604, June 2009.

[115] S. Sze and K. Ng, *Physics of semiconductor devices*. Wiley-Interscience, 2 ed., 1981.

[116] N. Mojumder and K. Roy, "Band–to–band tunneling ballistic nanowire fet: circuit–compatible device modeling and design of ultra–low–power digital circuits and memories," *IEEE Transactions on Electron Devices*, vol. 56, pp. 2193–2201, October 2009.

[117] U. Avci, R. Rios, K. Kung, and I. Young, "Comparison of power and performance for the TFET and MOSFET and considerations for p–TFET," in *IEEE Conference on Nanotechnology*, pp. 869–872, August 2011.

[118] S. Datta, *Quantum transport: atom to transistor*. Cambridge University Press, 2005.

[119] S. Xiong, T.-J. King, and J. Bokor, "Study of the extrinsic parasitics in nano–scale transistors," *Semiconductor Science and Technology*, vol. 20, pp. 652–657, May 2005.

[120] X. Wang, Y. Ouyang, X. Li, H. Wang, J. Guo, and H. Dai, "Room–temperature all–semiconducting sub–10 nm graphene nanoribbon field–effect transistors," *Physical Review Letters*, vol. 100, p. 206803–1–206803–4, May 2008.