A Hierarchical Optimization Engine for Nanoelectronic Systems Using Emerging Device and Interconnect Technologies

A Doctoral Dissertation
Presented to
The Academic Faculty

by

Chenyun Pan

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

Georgia Institute of Technology
August, 2015

Copyright © 2015 by Chenyun Pan

A Hierarchical Optimization Engine for Nanoelectronic Systems Using Emerging Device and Interconnect Technologies

Approved by:

Dr. Azad J. Naeemi, Advisor
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Jeffery A. Davis
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Saibal Mukhopadhyay
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Sudhakar Yalamanchili
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Francky Catthoor
Electrical Engineering Department
Katholieke Universiteit Leuven

Date Approved: July 7th, 2015

To my parents

ACKNOWLEDGEMENTS

First and foremost, I want to express my sincere gratitude to my advisor, Professor Azad Naeemi, for his continuous support and guidance during my PhD study and research. I was given a tremendous amount of freedom to explore various research subjects that I am really interested in, and it is my true honor and pleasure to work with him. I also appreciate Prof. Jeff Davis, Prof. Saibal, Prof. Francky, and Prof. Sudha for taking time being my committee members.

Second of all, I would like to thank my colleagues and friends from the Nanoelectronics Research Lab, Shaloo, Ahmet, Vachan, Nick, Sou-Chi, Sourav, Ramy, Victor, Divya, Rouhollah, Javaneh, and Phillip, for the productive research collaboration and constructive comments and feedback during the dry-run of my conference presentations. My sincere thanks also go to my colleagues and friends in IMEC, Praveen, Peter, Dmitry, Rogier, Prof. Francky, Zsolt, Ivan, Antonino, and Victor, for their research discussion and fun time during my two visits. In addition, I want to thank my other collaborators Redwan and Prof. Avik from Virginia Tech. Again and always, I wish to thank the great research and study environment that Georgia Tech has provided and all the research organizations and institutes below that sponsor my five years' research, study, and traveling: Semiconductor Research Collaboration, Global Research Collaboration, Nanoelectronic Research Initiative, INDEX center, National Science Foundation, and IMEC.

Last but not least, I want to thank my parents, who always support, love, and motivate me to seek and enjoy the research that I am fond of. My sincere appreciation also goes to my significant other, Liwei, whose thoughtful care and devotion brings my PhD study into fruition.

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENTS .... **iv**

LIST OF TABLES .... **xi**

LIST OF FIGURES .... **xii**

SUMMARY .... **xxii**

Chapter 1 Introduction and Background **1**

1.1 Si CMOS FETs/Cu Interconnect Scaling Challenges 1

1.2 Emerging Technologies 4

1.2.1 Device-Level Innovations 4

1.2.2 Interconnect Innovations 5

1.2.3 System-Level Innovations 6

1.3 Dissertation Motivation 7

1.4 Conclusions 10

Chapter 2 Device-Level Modeling and Simulation **12**

2.1 Introduction 12

2.2 Graphene PN Junction Devices 14

2.2.1 Resistance Modeling 17

2.2.2 Capacitance Modeling 19

2.2.3 Validation and Comparison with NEGF Simulation 20

2.2.4 Performance Comparison with CMOS FETs 22

2.3 Tunneling Devices.. 24

2.4 Planar and FinFET Si CMOS Devices 28

2.5 Gate-All-Around Nanowire FETs 28

2.6 Conclusions 33

Chapter 3 Interconnect- and Circuit-Level Modeling and Simulation................. 35

3.1 Introduction 35

3.2 Metal Interconnect... 38

3.2.1 Resistance Model 38

3.2.2 Capacitance Model... 41

3.3 Graphene Interconnect 44

3.3.1 Resistance Model 44

3.3.2 Capacitance Model... 46

3.3.3 Simulation Results 47

3.4 Graphene PN Junction Logic 63

3.4.1 Basic Logic Function 64

3.4.2 Reconfigurable Logic Gate 66

3.4.3 SRAM Cell.. 67

3.5 Gate-All-Around Nanowire FETs 70

3.5.1 Ring Oscillator Analysis 70
3.5.2 ARM Core Analysis 73
3.6 Conclusions 81
Chapter 4 System-Level Modeling and Design Methodology Integration 84
4.1 Introduction 84
4.2 Hierarchical Memory Model 87
4.3 Empirical CPI Model 88
4.4 Multi-Level Interconnection Network Models 92
4.5 Power Distribution Network Model 95
4.6 Process Variation Model 99
4.7 Thermal Model 99
4.8 Hierarchical Design Methodology and Validation 100
4.9 Conclusions 104
Chapter 5 Device Technology Optimization 106
5.1 Introduction 106
5.2 Conventional Si CMOS Devices 108
5.2.1 Comparison between Planar FET and FinFET 109
5.2.2 Performance and Area Scaling Trends 111
5.3 Graphene PN Junction Devices 114
5.3.1 Single-Core Optimization 114

5.3.2 Multi-Core Optimization 119

5.4 Tunneling Devices 122

5.5 Optimization under Process Variation 125

5.5.1 Single-Core Optimization 127

5.5.2 Multi-Core Optimization 131

5.6 Conclusions 136

Chapter 6 Interconnect Technology Optimization 139

6.1 Introduction 139

6.2 Device and Interconnect Scaling Trends 142

6.3 Interconnect Implications 146

6.4 A Modified Interconnect Structure 147

6.5 Al-Cu Hybrid Interconnect Architecture 152

6.5.1 Proposed Fabrication Process Flow 153

6.5.2 Simulation Results 154

6.6 Graphene Interconnect 158

6.7 Optimization under Process Variation 162

6.8 Conclusions 168

Chapter 7 System-Level Benchmarking and Optimization 171

7.1 Introduction 171

7.2 3D Technology Optimization 174

7.2.1 Single logic core optimization ... 174
7.2.2 Impact of the Via Diameter and Capacitance ... 177
7.2.3 Comparison between the TSV- and MIV-based systems ... 178
7.3 Heterogeneous Multi-core Integration ... 180
7.3.1 Symmetric Multi-Core Analysis ... 181
7.3.2 Asymmetric Multi-core Analysis ... 183
7.3.3 Heterogeneous Multi-core Analysis ... 185
7.4 Chip/Package Co-Optimization ... 191
7.4.1 Package and Circuit-Level Simulation ... 191
7.4.2 System-Level Optimization Results ... 195
7.5 Conclusions ... 199
Chapter 8 Conclusions and Future Work ... 201
8.1 Conclusions of dissertation ... 201
8.2 Future Work ... 205
8.2.1 Device-Level Modeling and Analyses ... 205
8.2.2 Circuit- and Interconnect-Level Modeling and Analyses ... 206
8.2.3 System-Level Modeling and Analyses ... 206
8.2.4 Hierarchical Optimization Engine Development ... 207
List of Publications ... 208
References ... 210

VITA **227**

LIST OF TABLES

Table 1: Device-Level Comparison of Various Metrics between GPNJ and Si CMOS Devices. 23

Table 2: Technology Assumptions for LFET and VFET Devices 30

Table 3: Interconnect Process Variation Assumptions for N14, N10, and N7 Nodes. 38

Table 4: The Default Configuration of the Simulation. 48

Table 5: Comparison between GPNJ- and CMOS-based SRAM Cell. 70

Table 6: The Raw Data Collected for Various Processor Generations for the Intel Microprocessor. 89

Table 7: Package Configurations for Resistance, Capacitance, and Inductance 98

Table 8: Configurations for the Single-Core Design. 110

Table 9: Simulation Results of a Single Core for Planar CMOS and FinFET Devices. 110

Table 10: The Coefficient and Exponent of Power-Law Relation between Throughput and Core Area. 111

Table 11: System-Level Comparison of Various Optimal Design Parameters and Performance Metrics between Si CMOS and GPNJ Logic Cores. 117

Table 12: The Standard Deviation and Mean Optimal Clock Frequency under Each Individual Source of Interconnect Process Variation Using Three Fabrication Processes 163

Table 13: Comparison between Optimized MIV- and TSV-Based Two-Layer 16-Core Processors. 179

Table 14: Design Parameters and Performance Metrics for an Asymmetric Multi-Core Processor at Various Parallelisms of the Program. 185

LIST OF FIGURES

Figure 1: Emerging technologies on the device, interconnect, and system levels [17]. 4

Figure 2: Hierarchical optimization engine overview. 9

Figure 3: Angular dependence of quasiparticle transmission through the electrostatically generated GPNJ.......... 14

Figure 4: Transmission probability as a function of the incident angle θ for various gap distances between two gates. 15

Figure 5: Schematic of a tilted GPNJ device built on a graphene sheet. 16

Figure 6: Illustration of the operation principle of a GPNJ switches. (a) and (d) show p- and n-type GPNJ switches at ON state, respectively, (b) and (c) show p-type GPNJ switches at OFF state for electrons coming from various angles and parallel direction, respectively, (e) and (f) show n-type GPNJ switches at OFF state for electrons coming from various angles and parallel direction, respectively. 16

Figure 7: The Brillouin zone of graphene. 18

Figure 8: Cross section view of GPNJ devices and interconnects. 20

Figure 9: Comparison of the average transmission probability of a GPNJ device between experimental and simulation data [64]. 21

Figure 10: Conductance versus control voltage of a GPNJ switch based on NEGF results for various width of the graphene sheet. 21

Figure 11: ON and OFF resistances versus controlled voltage at Vdd = 0.5V. 23

Figure 12: Schematic of an InAs nanowire and band diagram in the OFF/ON states for (a) p-type and (b) n-type GAA TFETs. 25

Figure 13: ID-VGS 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 [80]. 26

Figure 14: Leakage mechanisms considered in this work shown on a p-type TFET. 27

Figure 15: The 3D view and layout of VFETs and LFETs with various configurations. 29

Figure 16: The cross-sectional view of the n-type LFETs with (a) 2fin/3stack and (b) 3fin/2stack using wrap contacts at source and drain regions. 30

Figure 17: Intrinsic leakage current versus ON current for LFET devices with three Vth options at two supply voltages. In the legend, LFET- xFyS indicates that the number of fins and stacks used in an LFET are 'x' and 'y', respectively. 31

Figure 18: The ON current versus the supply voltage for LFETs and VFETs using 7 options with low- and high-Vth flavors. In the legend, VFET-xNW indicates that the number of nanowires used in a VFET is 'x'. 33

Figure 19: The impact of CD and overlay variations on the width of interconnects for the (a) LELE double patterning, (b) SADP, and (c) SAQP fabrication techniques. 39

Figure 20: The resistance per unit length normalized to the nominal value versus the deviation in σ_0, which is quantitatively shown in the bottom part of Table 3 for five independent sources of interconnect variations. 40

Figure 21: The cross-section view of multilevel interconnects under (a) no process variation, (b) CD variation, (c) etch variation, (d) CMP variation, and (e) overlay variation. 42

Figure 22: The comparison between the compact model and Raphael simulation for capacitance per unit length versus various types of process variation. 43

Figure 23: Comparison of the resistance per unit length between graphene interconnects and copper interconnects. (a) shows resistance per unit length versus the width of the interconnects at various interconnect length (b) shows the resistance per unit length versus the length of the interconnects for various interconnect widths. 45

Figure 24: Cross-section view of the multi-layer graphene interconnect using (a) side contact (b) top contact. 46

Figure 25: (a) Capacitance per unit length of the multi-layer graphene interconnects versus the width for various numbers of layers. (b) Cross-sectional view of the copper interconnects and graphene interconnects based on the assumption that the interconnect pitch is twice of the interconnect width. 47

Figure 26: The schematic of a 32-bit Kogge-Stone adder. 48

Figure 27: Delay improvements of multilayer graphene interconnects compared with copper interconnects versus the number of graphene layers at three MFP values. 49

Figure 28: Optimal delay improvement and number of graphene layers versus interconnect length for various MFP values. 50

Figure 29: Optimal delay improvement and number of graphene layers versus interconnect length for various MFP values. Here, the contact resistance is 1000 $\Omega\cdot\mu m$. 51

Figure 30: Optimal delay improvement and number of graphene layers versus interconnect length for various MFP values. Here, the width of the interconnect is 40nm. 52

Figure 31: Optimal delay improvement and number of graphene layers versus interconnect length for various MFP values. Here, the width of the interconnect is 14nm; the technology node is 7nm. 52

Figure 32: Optimal delay improvement versus interconnect length for various MFP values at a low supply voltage Vdd = 0.5V with (a) normal threshold voltage and (b) high threshold voltage devices. 53

Figure 33: Optimal number of graphene layers versus interconnect length for various MFP values at a low supply voltage V_{dd} = 0.5V with (a) normal threshold voltage and (b) high threshold voltage devices. 54

Figure 34: Optimal delay improvement versus interconnect length for various MFP values. Here, the number of graphene layers is fixed to be 10. 55

Figure 35: Energy consumption and EDP improvement versus the number of graphene layers. 55

Figure 36: Optimal EDP improvement and number of graphene layers versus interconnect length for various MFP values. 56

Figure 37: Optimal delay improvement versus interconnect length at various MFP values with edge roughness of (a) 0.2 and (b) 1.0. 57

Figure 38: Delay improvement versus contact resistance of graphene at various MFP values. The total number of cells on the bit line is (a) 64 and (b) 256. The bit-line length and width between two nearby cells are 0.26 μm and 40 nm, respectively. 58

Figure 39: Delay improvement versus contact resistance of graphene at various MFP values. The total number of cells on the word line is (a) 64 and (b)

256. The word-line length and width between two nearby cells are 0.43 μm and 40 nm, respectively. 59

Figure 40: Flowchart of the multi-layer graphene interconnects benchmarking based on commercial tools. 60

Figure 41: The layout of a placed and routed ARM Cortex-M0. 61

Figure 42: Delay improvement and relative energy consumption of an ARM core versus the number of graphene layers. 61

Figure 43: EDP improvement versus the number of graphene layers with a contact resistance of (a) 100 $\Omega \cdot \mu m$ and (b) 1000 $\Omega \cdot \mu m$. 62

Figure 44: EDP improvement versus the number of graphene layers with edge roughness coefficient p of (a) 0.2 and (b) 1. Here, the contact resistance is assumed to be 100 $\Omega \cdot \mu m$. 63

Figure 45: 3D plot of a GPNJ inverter. 64

Figure 46: Layout of an GPNJ inverter (a) using chemical doping control (b) using electrostatic gate control. 65

Figure 47: Layout of a GPNJ NAND2 gate (a) using electrostatic gate control, (b) using chemical doping control, and (c) using chemical doping control that takes into account the curved corner during lithography process. 66

Figure 48: Reconfigurable logic gates based on GPNJ devices. 67

Figure 49: Read and write operation of a 6-transistor SRAM cell. 68

Figure 50: Various performance metrics of an SRAM using GPNJ devices. (a) write margin, (b) access time, and (c) read margin. 69

Figure 51: Leakage current versus the operational frequency of a 15-stage ring oscillator using LFETs and VFETs with 7 options at the supply voltage of 0.6V. 71

Figure 52: Energy per switch versus the operational frequency of a 15-stage ring oscillator using LFETs and VFETs with 7 options at the supply voltage of 0.4 and 0.6V. 72

Figure 53: Actual frequency versus the target frequency of a synthesized ARM core using seven configurations of LFETs and VFETs at Vdd = 0.6V. 75

Figure 54: Comparison of the (a) maximum clock frequency and (b) percentage of interconnect delay in the critical path at three supply voltages for an ARM core using seven configurations of LFETs and VFETs. 75

Figure 55: Energy per clock cycle versus the clock frequency of a synthesized ARM core using seven configurations of LFETs and VFETs at Vdd = 0.6V. 76

Figure 56: Energy dissipation per clock cycle versus the frequency at three supply voltages for an ARM core (a) optimized at each supply voltage and (b) optimized only at Vdd = 0.6V using seven configurations of LFETs and VFETs. 77

Figure 57: Cell counts of different Vth cell usage versus the frequency target of an ARM core using core using seven configurations of LFETs and VFETs at supply voltage of (a) 0.6V and (b) 0.4V. LVT, SVT, and HVT represent low-Vth, standard-Vth, and high-Vth devices. 78

Figure 58: Dynamic energy dissipation versus the leakage energy dissipation for an ARM core using seven configurations of LFETs and VFETs at timing target of (a) 5 ns and (b) 1.4 ns.. 80

Figure 59: Total core area versus supply voltage of an ARM core using seven configurations of LFETs and VFETs at the timing target of (a) 2.2 ns and (b) 0.8 ns. 81

Figure 60: $CPIlogic$ versus the number of logic transistors of a single core for (a) Intel microprocessor family (b) IBM POWER family. 91

Figure 61: Interconnect density function and the cumulative interconnect distribution for inter-layer and all interconnects. 95

Figure 62: Equivalent circuit model for the PDN, including both chip and package. 96

Figure 63: Delay and leakage distribution generation of variation-aware system-level design. 98

Figure 64: Flowchart of the system-level design methodology. 101

Figure 65: Comparison of simulation results with actual data in terms of (a) throughput, (b) clock frequency, (c) logic core area, and (d) number of logic transistors.. 103

Figure 66: Optimization results for a single core implemented by the conventional planar CMOS and FinFET devices. (a) and (b) show the throughput versus the supply voltage and the number of logic gates. (c) and (d) show the pie chart of power components for each part. 109

Figure 67: (a) Throughput versus core area for a single-core FinFET processor at various technology nodes. (b) Optimal percentage of the area that is occupied by logic. 112

Figure 68: Relative throughput versus core area of FinFET single core at the 16 nm technology node for various Eble's exponents. 114

Figure 69: Throughput versus multiple design parameters. (a) Throughput versus Vdd and number of logic gates at $V_g = 0.6V$ and $D_{gap} = 40nm$, and the optimal point is shown as the purple star in (b). (b) Throughput versus gap distance and control voltage, where each point is obtained from optimal Vdd and number of logic gates. 116

Figure 70: The pie chart of the power consumption of each component for (a) CMOS core and (b) GPNJ core. 117

Figure 71: Various optimal design parameters and performance metrics at various technology nodes for a single-core GPNJ processors. 118

Figure 72: Various optimal design parameters and performance metrics at various technology nodes for a multi-core GPNJ processors. 120

Figure 73: System-level throughput optimization for a TFET single core, given a fixed core area as 5 mm^2. (a) and (b) show the throughput versus the supply voltage and the number of logic gates under 2 W/cm^2 and 10 W/cm^2 power density constraints, respectively. (c) and (d) show the power and delay components at the optimal throughput design point with 2 W/cm^2 power density constraint. 121

Figure 74: The comparison between the optimized system using TFETs and CMOS low-power devices for various optimized parameters versus the core area with 2 W/cm^2 power density constraint. 123

Figure 75: The optimal throughput comparison among TFETs, CMOS high performance, and CMOS low power devices under various power density budgets, given a 5 mm^2 core area. (a) low power density range, (b) high performance range. 124

Figure 76: The optimal throughput of a single core using tunneling devices versus core area at various power density budgets. 124

Figure 77: The optimal throughput of a single core using TFETs versus core area at various total power consumption budgets. 125

Figure 78: (a) and (d) show the optimal throughput versus the supply voltage and the number of logic gates at the nominal state, where the red points are the design points being investigated in the bottom four figures. (b) and

(c) show the frequency and throughput distributions for various number of logic gates at vdd = 0.5V. (e) and (f) show the frequency and throughput distributions for various supply voltages at N_g = 22M. 126

Figure 79: Throughput versus supply voltage and the number of logic gates for a 5 mm2 core with 100 W/cm2 power density constraints based on 80% yield implemented by frequency tuning. 128

Figure 80: Throughput versus supply voltage and the number of logic gates for a 5 mm2 core with 100 W/cm2 power density constraints based on 80% yield implemented by adaptive supply voltage. 129

Figure 81: Leakage power versus the frequency of a core. The red dots are the samples applied with adaptive supply voltage with 100 W/cm^2 power density constraint and the blue dots are raw data. 130

Figure 82: Comparison of cumulative throughput distributions for the core implemented with two post tuning methods. 130

Figure 83: Cumulative distribution function for frequency and effective throughput per core of a multi-core processor with the assumption that the maximum power density for each core is 100 W/cm^2. 132

Figure 84: Comparison of the frequency between two multi-core processors with and without applying the power reallocating technique for two samples (a) and (b). 134

Figure 85: Comparison of the probability density function for the average frequency between two asynchronous processors. 135

Figure 86: Overall performance improvement versus the number of cores in the multi-core processor. 136

Figure 87: Various device and interconnect metrics for three device structures using copper interconnects at various technology nodes. 143

Figure 88: Normalized single-stage inverter delay for three device structures using copper interconnects at various technology nodes. 145

Figure 89: Normalized energy consumption for three device structures using copper interconnects at various technology nodes. 145

Figure 90: The cross-sectional view of the interconnect structure with a constant pitch and a width of (a) half pitch and (b) 0.6×pitch. 147

Figure 91: Single-stage delay breakdown versus the relative width of the interconnect for the circuits at the 7nm technology node using (a) VFET

and (b) FinFET devices. Here, the length of the interconnect is 10 gate pitches. 148

Figure 92: The contour figure of the delay improvement at various relative widths and aspect ratios of the interconnects. 149

Figure 93: Optimal (a) clock frequency improvement and (b) relative width for three device structures at various technology nodes, assuming the aspect ratio is 3. 150

Figure 94: Optimal (a) EDP improvement and (b) relative width for three device structures at various technology nodes, assuming the aspect ratio is 3. 151

Figure 95: Flow chart of subtractive Al-Cu hybrid interconnect process. 153

Figure 96: Resistivity versus the width of Cu and Al interconnects with two grain boundary reflectivity co-efficient R and surface specularity parameter p. Here, the aspect ratio is 2. 154

Figure 97: Chip clock frequency versus aspect ratio for six interconnect configurations at the 16nm technology node. Here, $R = p = 0.5$. 155

Figure 98: (a) Optimal clock frequency and (b) aspect ratio versus technology node for six different interconnect configurations that are defined in Figure 97. 157

Figure 99: Throughput versus the number of graphene layers at various MFP using different materials as the substrates. (a) Smooth edge (b) Edge scattering probability P = 0.2. 160

Figure 100: (a) Optimal throughput versus core area for GPNJ processor using graphene and copper interconnects. The red curves show the improvement of using multi-layer graphene interconnects with perfect edge and edge scattering probability p = 0.2. (b) Optimal number of graphene layers versus core area for a single core using multi-layer graphene interconnects. 161

Figure 101: The histograms of the optimal clock frequency under each individual source of interconnect variation, including CD core/spacer, etch, CMP, and overlay variations, with various σ values using three fabrication processes such LELE, SADP, and SAQP. 162

Figure 102: The histogram of the delay for the longest interconnects on various metal levels. 163

Figure 103: The relative frequency at 90% yield versus σ/σ_0 under five sources of interconnect variations at three technology nodes using LELE double

patterning, SADP, and SAQP interconnect fabrication processes. Here, σ_0 represent the default interconnect variability values shown in Table 3.. 166

Figure 104: The relative clock frequency versus various σ values with the combination of five types of interconnect variations at the 14 nm, 10 nm, and 7 nm technology nodes using LELE, SADP, and SAQP fabrication processes. σ_0 represent the default interconnect variability values shown in Table 3. 167

Figure 105: Optimization results for a single core implemented by the conventional 2D integration and 3D integration with MIVs. (a) and (b) show the throughput versus the supply voltage and the number of logic gates. (c) and (d) show the power and delay components for each part. 175

Figure 106: The optimal throughput and power consumption versus the number of logic layers for the system implemented by MIV-based 3D ICs, given D_{via} = 100 nm, C_{via} = 0.4fF. 176

Figure 107: Optimal throughput of a single core versus the diameter of the vias for various numbers of logic layers. 177

Figure 108: (a) Basic configuration of MIV- and TSV- based systems. (b) Optimal throughput versus the logic percentage for both MIV- and TSV-based two-layer 16-core processors, assuming the program is fully parallelized. 179

Figure 109: Comparison of the optimal throughput between MIV- and TSV-based processors for various number of logic layers. 180

Figure 110: Various design parameters and performance metrics versus number of cores for a symmetric multi-core processor. (a) four bullets in the legend correspond to the optimal chip throughput, logic area percentage, link utilization during serial and parallel part of the program, respectively. Here, the parallelism is assumed to be 0.9; the die are is 200 mm^2; and the power budget is 200W. 182

Figure 111: Diagram for the symmetric versus asymmetric multi-core processor. 183

Figure 112: Contour figure of the chip throughput versus big-to-small core ratio and number of cores in an asymmetric multi-core processor. Here, the parallelism of the program is 0.8, and the total power budget is 150W. 184

Figure 113: Chip throughput versus power consumption of heterogeneous multi-core processors with four configurations running a program with various parallelisms. In the legend, three acronyms CMOSHP, CMOSLP, and TFET represent three technologies: CMOS high-performance devices, CMOS low-power devices and TFETs, respectively. CMOSHP-TFET

represents a heterogeneous multi-core processor with a big core implemented with CMOS high-performance devices and small cores implemented with TFETs. 186

Figure 114: Optimal chip throughput versus power consumption for various multi-core scenarios, including switching big or small cores alternately and simultaneously. 188

Figure 115: Energy consumption per instruction versus power consumption of heterogeneous multi-core processors with four configurations running a program with parallelism of 0.5. 190

Figure 116: Three packaging configurations. (a) without off-chip decap (b) with side decap (c) with center decap. 192

Figure 117: The envelop of the transient response of the power supply noise for three package configurations without inserting sleep transistors. 193

Figure 118: Transient response of the virtual ground during active and idle states. 194

Figure 119: DC drop and sleep transistor area as a percentage of the chip area versus width of the transistor under various supply voltages. 195

Figure 120: Relative throughput, energy, and EDP versus width of sleep transistors. The solid and dash lines represent processors implemented with power-gating and clock-gating techniques, respectively. 197

Figure 121: Optimal relative EDP versus average number of sequential instructions for three package configurations. 198

SUMMARY

The objective of this dissertation is to develop a fast and efficient hierarchical optimization engine to benchmark and optimize various emerging device and interconnect technologies and system-level innovations at the early design stage. As the semiconductor industry approaches sub-20nm technology nodes, both devices and interconnects are facing severe physical challenges. Many novel device and interconnect concepts and system integration techniques are proposed in the past decade to reinforce or even replace the conventional Si CMOS technology and Cu interconnects. To efficiently benchmark and optimize these emerging technologies, a validated system-level design methodology is developed based on the compact models from all hierarchies, starting from the bottom material-level, to the device- and interconnect-level, and to the top system-level models. Multiple design parameters across all hierarchies are co-optimized simultaneously to maximize the overall chip throughput instead of just the intrinsic delay or energy dissipation of the device or interconnect itself. This optimization is performed under various constraints such as the power dissipation, maximum temperature, die size area, power delivery noise, and yield. For the device benchmarking, novel graphen PN junction devices and InAs nanowire FETs are investigated for both high-performance and low-power applications. For the interconnect benchmarking, a novel local interconnect structure and hybrid Al-Cu interconnect architecture are proposed, and emerging multi-layer graphene interconnects are also investigated, and compared with the conventional Cu interconnects. For the system-level analyses, the benefits of the systems implemented with 3D integration and heterogeneous integration are analyzed. In addition, the impact of the power delivery noise and process variation for both devices and interconnects are quantified on the overall chip throughput.

CHAPTER 1 INTRODUCTION AND BACKGROUND

1.1 Si CMOS FETs/Cu Interconnect Scaling Challenges

Decades of relentless complementary metal–oxide–semiconductor (CMOS) technology scaling has brought significant performance improvement to the electronic community. The transistor density doubles every 18-24 months, and the cost per transistor drops exponentially over the past 40 years, inciting a blooming semiconductor industry [1]. Behind this extraordinary scaling history are huge amounts of research efforts and investments to overcome numerous engineering challenges associated with the scaling, especially in the last two decades [2].

From the device perspective, as the technology nodes approach sub-20nm [3], the semiconductor industry has faced great challenges such as severe short-channel effects, large parasitic capacitance and resistance, and random doping variations [2]. A shorter channel length leads to a faster switching time and a smaller device footprint area, but it causes the drawback of a larger leakage current due to the drain induced barrier lowering (DIBL) effects [4]. The leakage power dissipation is a major component of the total power consumption in modern VLSI systems. Since the power consumption sets limitations for both high-performance and low-power portable devices, supply voltage scaling is the most effective way to suppress the power dissipation [5]. In order to maintain the driving current, a reduced supply voltage requires a lower threshold voltage. This again increases the static power dissipation when the field-effect transistors (FETs) are at the off state. To suppress the leakage current without compromising the driving current, the electrostatic gate control of the FETs is improved and novel device structures have evolved from the conventional planar FETs to the multi-gate devices, such as

FinFETs, Pi-gate, and Omega-FET devices, and even to the ultimate gate-all-around (GAA) devices [3, 6]. These novel device structures, however, bring larger parasitics, especially at small dimensions because parasitics are inversely proportional to the distance between two surfaces. The parasitics mainly come from the inversion, centroid and quantum capacitances associated with the gate and channel, the source and drain extension, fringe capacitance, and gate to source and drain capacitances [2]. These parasitic capacitances have an ever-increasingly large impact on the delay and the dynamic power dissipation [2]. To reduce the parasitic capacitance, low-k dielectrics and air-gaps are developed with challenges of the fabrication difficulties and reliability issue [7, 8]. At small nodes, the resistivity also increases significantly due to the extra scatterings at regions of the source and drain contact, extension, and the interface. It is of critical importance to reduce the Schottky barrier height at the contact interface for future CMOS devices. Moreover, the scaling of the oxide thickness has already pushed the CMOS technology to its reliability limits, such as the gate oxide leakage, time-dependent dielectric breakdown (TDDB), and negative-bias temperature instability (NBTI) [9]. Besides the aforementioned traditional scaling challenges, the current CMOS scaling encounters new issues, including atomic spacing limiting critical dimensions, interface and support layers dominating the physical structures, and quantum confinement and scattering effects [2].

From the interconnect perspective, ideally, the resistance per unit length increases by $1/\alpha^2$, where α is the scaling factor, since the cross-sectional area is proportional to α^2. In reality, at sub-20nm nodes, the interconnect resistivity faces an ever-increasing size effect. It has been demonstrated that the surface and grain boundary scatterings largely increase the effective resistivity of copper interconnects [10]. Diffusion barriers, which are very poor conductors, will take an ever-increasing fraction of the interconnect volume. These aforementioned reasons cause the resistance per gate pitch being more than $1/\alpha$ larger

compared with the previous technology generation. For the interconnect capacitance, since the capacitance per unit length remains the same based on the assumption that the aspect ratio and interlayer dielectric (ILD) permittivity stay unchanged, the capacitance per gate pitch reduces by α. Therefore, the overall intrinsic interconnect RC delay per gate pitch keeps increasing. From the manufacture perspective, the minimum pitch is already beyond the single-exposure limit of the photolithography process [11]. Multiple patterning techniques are required to fabricate interconnects with small dimensions, which inevitably aggravates the variability, including the overlay and critical dimension (CD) core/spacer variations [12]. Other sources of interconnect variation arising from etching, polishing, and orientation also have large impacts on the chip performance [13]. Moreover, the interconnect scaling also faces severe reliability challenges such as the electromigration (EM), low-k dielectric TDDB, and mechanical weaknesses [14].

To overcome the aforementioned challenges, many advanced fabrication and process techniques have been invented and applied to the conventional Si CMOS FETs with copper interconnects, such as strained silicon, high-K metal gate, low-K inter-layer dielectric material, multiple patterning techniques, and etc. However, these fabrication and process techniques are only temporary remedies for the scaling problems. Continuous innovations and inventions of emerging technologies are required to sustain the performance growth. Furthermore, improvements in device or interconnect technologies alone are neither sufficient nor cost-effective to provide enough gain in the computing power due to the economic limitation. Note that the fab cost associated with the lithographic equipment, clean room facilities, and process complexity has grown exponentially [15, 16]. Therefore, circuit- and system-level innovations are of crucial importance to further improve the power and energy efficiencies to sustain the Moore's Law scaling and overall computational advancement.

1.2 Emerging Technologies

To overcome challenges in the conventional Si CMOS with the copper interconnect technology, novel post-CMOS technologies are proposed from the bottom material level, to the device and interconnect levels, and to the top system level. Figure 1 shows several selected emerging technologies that have been proposed recently.

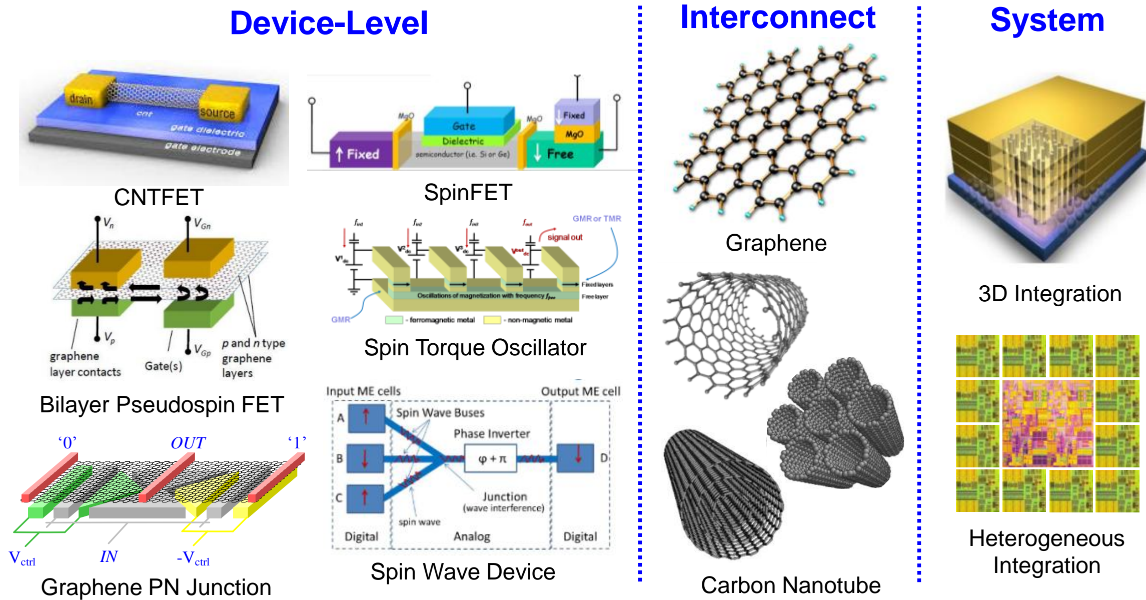


Figure 1: Emerging technologies on the device, interconnect, and system levels [17].

1.2.1 Device-Level Innovations

The most popular device innovations are divided into two categories: transistor circuits and majority-gate circuits. The first category contains electronic devices, such as graphene PN junction (GPNJ) and carbon nanotube FET (CNTFET) devices, and other devices including ferroelectric, piezoelectric and orbitronic devices. Spintronic devices mainly belong to the second category. For the first category, carbon-based electronics are proposed after the discoveries of graphene and carbon nanotubes (CNT). A GPNJ device uses the unique angular dependent transmission probability of several GPNJs to create a switch and achieve a large ON ratio [18, 19]. CNTFET is also a promising alternative to

the bulk silicon transistor in light of its low-power and high-performance design due to the ballistic transport and low OFF-current properties [20, 21]. A bilayer pseudospin FET (BisFET) is another graphene device, employing tunneling between two graphene monolayers [22, 23]. A spinFET combines a MOSFET and a switchable magnetic element [24]. The parallel or anti-parallel magnetization state determines whether the resistance of a spinFET is low or high. A spin wave device (SWD) is another device concept that contains nanomagnets connected by ferromagnetic interconnects [25, 26]. Short pulses of spin waves are excited in and propagate along the interconnects. Computation are performed by the interactions among the phases of spin waves, which determine the magnetization directions of the output nanomagnet. Other spin-based devices include the nanomagnetic logic (NML) [27], all spin logic (ASL) [28, 29], and spin torque oscillator (STO) logic [30] devices.

1.2.2 Interconnect Innovations

To further reduce the interconnect capacitance per unit length, the ultra-low-k material is pursued, and fabrication techniques are proposed for porous material and air-gap interconnect structures. A misalignment-free multilevel air-gap interconnect with a via-base structure was fabricated by self-aligned gap formation and etch back [31]. The goal is to overcome the reliability issues and achieve robust interconnects. Since lowering the ILD permittivity without compromising reliability is increasingly hard, carbon-based interconnect material is proposed in light of its high capability in conducting current and small capacitance. Potential candidates are the multi-layer graphene interconnects [32, 33], single- or multi-wall carbon nanotube interconnects [34, 35], and carbon nanotube bundles [36]. The optical interconnect is another potential candidate to solve the communication bottleneck in high-performance integrated circuits. It has been proposed for both on-chip and chip-to-chip interconnects [37], which have the advantages of

immunity to electrical noise, long distances communication, and ultra-high bandwidth. The major challenges of this technology is the energy and area overhead associated with the light generation and detection circuits [38]. From the manufacturing point of view, with the technology scaling, double patterning appears to be one of the cost-effective advanced patterning options [39]. One conventional double patterning technique is the litho-etch-litho-etch (LELE) process [40]. In this approach, two sequential steps of lithography and etch processes are performed, which make the LELE process suffer from the overlay variation, especially at a small technology node [40]. Recently, a more advanced double patterning technique, self-aligned double patterning (SADP), is developed to bypass the overlay variation by using a pattern-mask and a block-mask [41, 42]. As the technology scales down to sub-10nm nodes, triple or even quadruple patterning schemes are required due to the single-exposure resolution limitation [43]. Recently, the self-aligned quadruple patterning (SAQP) is proposed and demonstrated [43]. This process needs two steps of sidewall spacer processes, bringing more sensitivity to the CD spacer variation.

1.2.3 System-Level Innovations

To further improve the power or energy efficiency of the system and sustain Moore's law scaling, novel system-level integration techniques are proposed. One solution is the three-dimensional (3D) integration. 3D integrated circuits (ICs) have intrigued a lot of research in the past decade due to their potential benefits, including extending Moore's Law by increasing the device density, overcoming the barriers in the interconnect scaling, and providing further performance improvement with less power consumption [44]. A monolithic CMOS 3D chip can potentially provide the equivalent transistor count of a biological system, such as human brain, to enable many applications, such as man-machine interface and cognitive or neuromorphic computing [45]. Another system-level

innovation is the heterogeneous integration. It uses different device technologies for logic cores of different sizes to improve the energy efficiency. One potential application has been proposed to use both the conventional CMOS FETs and tunneling FETs (TFETs) to achieve a significant power-saving without compromising the performance of a multi-core processor [46]. Since the process variation is a major challenge for the processors at small nodes, the system-level variation-aware design are also proposed and implemented to suppress the impact due to various device and interconnect variations. For a multi-core processor, the systematic within-die (WID) variation induces the performance and power asymmetry [47]. Some cores may operate faster and violate the power density constraint while other cores may operate slower and be unable to fully utilize their allocated power budget. For a synchronous system, this asymmetry leads to significant performance degradation since the clock frequency is limited by the slowest core. One technique is to disable the slowest core so that the overall clock frequency can be increased [48]. Another technique is to use asynchronous systems, where each core runs at its own maximum frequency. To further take advantage of the asynchronous operation and better utilize the total power budget, a power reallocation technique is proposed [49]. By transferring the allocated power from the cores at worse process corners to more power efficient cores at better process corners, this method can provide better performance and yield compared with the conventional asynchronous system based on the same total power consumption budget.

1.3 Dissertation Motivation

As Si CMOS devices with Cu interconnects approach their scaling limits, there is a global search for novel devices and interconnects that can augment or even replace their conventional counterparts [50]. The high cost of developing novel technologies makes it crucial to develop a methodology to effectively evaluate these emerging technologies and

quantify their impact on the overall chip performance at the early design stage, and not rely solely on the intrinsic speed and energy consumption of emerging devices themselves. New devices may offer fundamentally different energy-delay trade-offs, and one needs to co-optimize device- and system-level parameters for maximum chip throughput under a given power/area budget. A smaller device may shorten the aggregate length of interconnects, but it may have higher ON resistance and leakage current. Advantages of emerging technologies can be fully evaluated only if the characteristics of devices and the overall performance of the system are simultaneously considered and co-designed.

Among research on evaluating various novel technologies [46, 50-55], focuses have been placed on the device- and circuit-level benchmarking [50-52], in which only delay and energy-delay-product (EDP) are considered for an intrinsic device/gate or a simple circuit such as an inverter chain, an SRAM, or an adder. Some work has performed system-level analyses and optimization for emerging devices [54, 55], but architecture-level information is neglected, where transistors are unfairly treated equally and the chip performance is assumed to be proportional to the number of binary switching operations per unit time. In [46], a more complicated cycle-accurate simulator is used to evaluate the performance of TFET technology that can provide detailed and accurate performance estimation, but it is based on one specific architecture and system configuration, and the design space exploration is rather limited due to the time-consuming simulation. To efficiently optimize the system-level parameters of a multi-core processor such as the number of cores, previous studies [56-58] have shown optimal numbers of cores to achieve various performance metrics based on Amdahl's Law [59], which states that the parallel speedup of a multi-core processor is limited by the serial code in a program. In these studies, as a rule of thumb, it is assumed that the performance of a logic core is proportional to the square-root of the complexity or the area of the core. Under this rule

of thumb, optimization can be efficiently performed for a multi-core processor. However, it lacks some detailed information, including the device-level characteristics, the interconnection network, and the hierarchical memory system, causing the loss of insight regarding the device-circuit-system interactions. Since memory bandwidth is one of the significant bottlenecks in multi-core processors, it would be incomplete to perform the throughput analysis and optimization without including the limits imposed by a finite memory bandwidth. In this dissertation, the aforementioned square-root rule of thumb used in [56-58] is revisited by using a combination of a memory model, a thermal model, device-level models, interconnection network models, and an empirical cycle-per-instruction (CPI) model obtained from two major processor families.

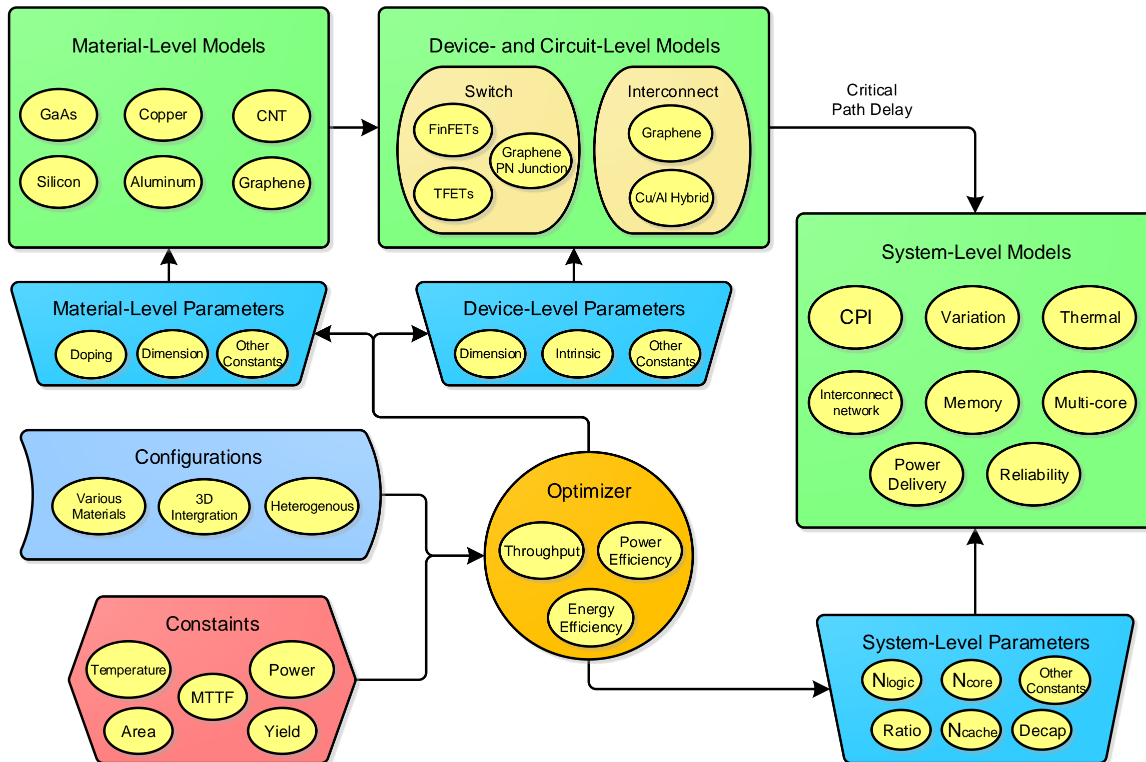


Figure 2: Hierarchical optimization engine overview.

Figure 2 shows an overview of the proposed hierarchical design engine. It includes a compressive set of models at the material, device/interconnect, circuit, and system levels

that allow exhaustive fast cross-layer optimizations for post-CMOS nanoelectronic systems. At the material level, there are many optimizations and choices to be made such as the kind of material to be used (e.g, copper, aluminum, graphene, or Si), or the doping level (e.g. in GPNJ devices and multi-layer graphene interconnects). At the device level, there are many parameters to be optimized such as the gap distance, oxide thickness, and supply voltage, which affect the device density, delay, power dissipation, and reliability. None of these optimizations can be done purely at the device level and one has to account for the circuit- and system-level implications.

Finally, at the system level, parameters such as core size, number of cores, logic-to-cache ratio, number of stacked dice in 3D chips, big-to-small core ratio in a heterogeneous chip must be optimized. The proposed research will explore cross-layer optimizations for various sets of constraints and for various applications and to show case the most promising materials, devices, circuits and architectures for each case. For such system-level analysis, it is highly inefficient to first fully design cores of various sizes and complexities with complete logic circuits, floorplanning, and timing analyses to quantify the throughput for the multi-core system. Instead, decisions on the complexity and size of the cores must be made at a higher level, which requires simplified system-level design models that can provide reasonably accurate results at the early design stage without the detailed knowledge of the individual logic circuits. The methods should be fast, efficient, and preferably analytical to enable exploring a large multi-parameter design space, including device, circuit, and system level parameters.

1.4 Conclusions

This chapter describes the scaling challenges in the conventional Si CMOS technology with the copper interconnects as the technology nodes approach sub-20nm. Several emerging beyond-CMOS technologies are introduced across various levels from

the device- and interconnect-level to the top system-level. The motivation of this dissertation is illustrated that a fast and efficient hierarchical optimization methodology is desirable to provide reasonably accurate benchmarking and optimization for various emerging technologies at the early development stage. The overview of the proposed design methodology is also depicted at the high level.

The rest of this dissertation is organized as follows. Chapter 2 illustrates the bottom device-level models that are used in the hierarchical engine, including the conventional planar FETs, FinFETs, nanowire GAAFETs, TFETs, and GPNJ devices. In Chapter 3, interconnect-level models are derived, including the metal and graphene interconnects. The circuit-level benchmarking is also performed for the multi-layer graphene interconnects and GAAFETs. Chapter 4 introduces the system-level models, such as a hierarchical memory model, an empirical CPI model, multi-level interconnection network models, a thermal model, and a process variation model. At the end of Chapter 4, all models are integrated into a validated hierarchical design methodology. In Chapters 5-7, based on the models and hierarchical optimization engine developed in Chapters 2-4, optimizations and benchmarkings are performed for various emerging device and interconnect technologies and novel system-level configurations, including 3D and heterogeneous integration. Chapter 8 concludes and highlights the major findings from this dissertation and provides potential extensions and future work for this research.

CHAPTER 2 DEVICE-LEVEL MODELING AND SIMULATION

2.1 Introduction

This chapter demonstrates the modeling approaches and performs simulations on the device level. Ever since the discovery of graphene with excellent physical and electrical properties, and compatibility with top-down lithographical patterning, much interest has been drawn in the electron-device community and graphene-based FETs have been introduced in [60]. However, graphene is a gapless material and creating a bandgap, which is essential for a conventional FET, would require patterning graphene nanoribbons as narrow as a few nanometers. The technological challenges with small dimensions have hindered fabricating graphene FETs with sufficient Ion/Ioff ratios. Klein tunneling through a potential barrier and angular dependent transmission probability of electrons in GPNJ can open a new path towards manufacturable graphene devices for digital logic circuits [61]. The original device concept is proposed by using the MUX-based logic gate [62]. It has shown significant advantages over CMOS devices in terms of delay-power product and signal restoration for a similar device footprint. However, the device-level resistance model is derived only at the equilibrium state for devices with symmetric pn junctions, which may overestimate the device performance.

In this chapter, an improved device structure and a more elaborate device model is derived to account for the transmission probability for asymmetric junctions and non-idealities, such as contact resistance and rounded corners. The physical analytical transmission model used in this chapter has also been compared with experimental results and non-equilibrium Green's Function (NEGF) simulation results obtained for a single tilted graphene junction [63, 64]. Since GPNJ devices will not be a "drop in" alternative

for Si CMOS switches, such a device-level modeling is key in understanding the potential advantages of GPNJ devices.

For the low-power applications, TFETs show promise in overcoming the power wall faced by thermionic FETs by allowing significant reduction in the supply voltage [65]. A wide range of device architectures and materials have been studied in the realization of TFETs in the previous years, including single gate (SG), double gate (DG) and GAA structures [65]. 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 [65, 66]. In this Chapter, the InAs nanowire based TFET is chosen for the modeling in light of direct bandgaps that eliminate the necessity for phonon assistance in tunneling. For the comparison, the conventional CMOS FinFET model is taken from ASU Predictive Technology Model (PTM) [67] down to the 7 nm technology node.

Beyond the 7 nm technology node, the FinFET device structure was unable to provide further improvement due to the poor electrostatic gate control, whereas the GAA structure was a strong candidate to keep the scaling at sub-5nm technology nodes [68]. However, even for an advanced channel structure, the cell width of the conventional lateral device is still severely limited by the gate length, the spacer thickness, and the source/drain contact size [69]. Hence, to improve the scalability, the vertical nanowire FET (VFET) was proposed and developed by fabricating the channel in the vertical direction, and the source and drain are on the top and bottom of the gate [70-72]. In doing so, the SCE is further alleviated, and the gate length and spacer thickness are relaxed without compromising the parasitic resistance from the source and the drain contacts. The deficiency is that there is no known solution yet to induce stress into the channel during the fabrication of a VFET, and consequently, the effective channel mobility is reduced, leading to a lower ON current than that in a lateral nanowire FET (LFET).

The rest of this chapter is organized as follow. Section 2 introduces the GPNJ device modeling approaches for the ON resistance, leakage current, and input capacitance as well as the NEGF and experimental validation. The intrinsic device-level comparison is performed between CMOS and GPNJ. In Section 3, InAs nanowire TFETs are modeled based on the compact analytical equations. Section 4 describes the SPICE models that are used for the conventional planar and multi-gate CMOS FETs. At the end, two types of CMOS GAA FETs are modeled and compared with various configurations in Section 5.

2.2 Graphene PN Junction Devices

A GPNJ is formed by laying a graphene sheet on an insulating layer with two split gates underneath them. These gates are biased with opposite voltage polarities, as shown in Figure 3. A positive voltage would raise the Fermi energy level of the graphene region above the gate and lead to an n-type electrostatic doping. Likewise, a negative voltage would make the graphene region above the gate become p-doped. Once the GPNJ is created, a unique angular dependent transmission for an electron beam with an incidence angle θ passing through a GPNJ is observed, as shown in Figure 3.

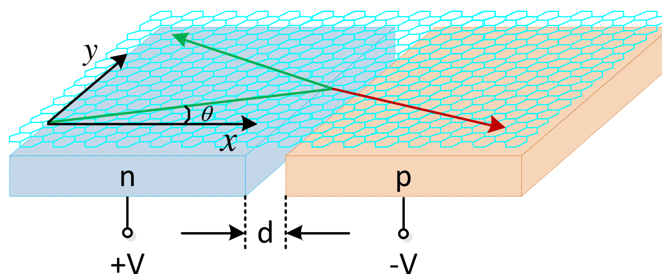


Figure 3: Angular dependence of quasiparticle transmission through the electrostatically generated GPNJ.

The transmission probability is given as [73]

$$T(\theta) = cos^2(\theta)\, e^{-\pi d k_F \, sin^2(\theta)} \tag{1}$$

where $k_F = E_F/\hbar v_F$ is the Fermi wavevector in the graphene band diagram [74], E_F is the Fermi energy, v_F is the Fermi velocity, and $\hbar$ is Planck's constant divided by 2π. The Fermi energy E_F in the graphene region above each gate depends on the applied gate voltage V_g and the oxide properties of the device [75]

$$E_F = \frac{1}{\gamma t_{ox}}\left(\sqrt{\varepsilon^2 + 2\gamma\varepsilon q V_g t_{ox}} - \varepsilon\right) \tag{2}$$

where $\gamma = (4\pi q^2)/(h^2 v_F^2)$ is a constant depending on the graphene properties, $v_F \approx 8 \times 10^5 m/s$ is the Fermi velocity, t_{ox} is the oxide thickness between the metal gate and graphene sheet, q is the elementary charge, and h is the Planck's constant.

Figure 4 shows relations between the transmission probability T and θ in polar coordinates. The transmission probability decreases with an increase in the junction gap and a decrease in the incident angle. Electrons with a normal incident angle pass through the junction directly. Thus, a GPNJ can be considered as an electron collimator.

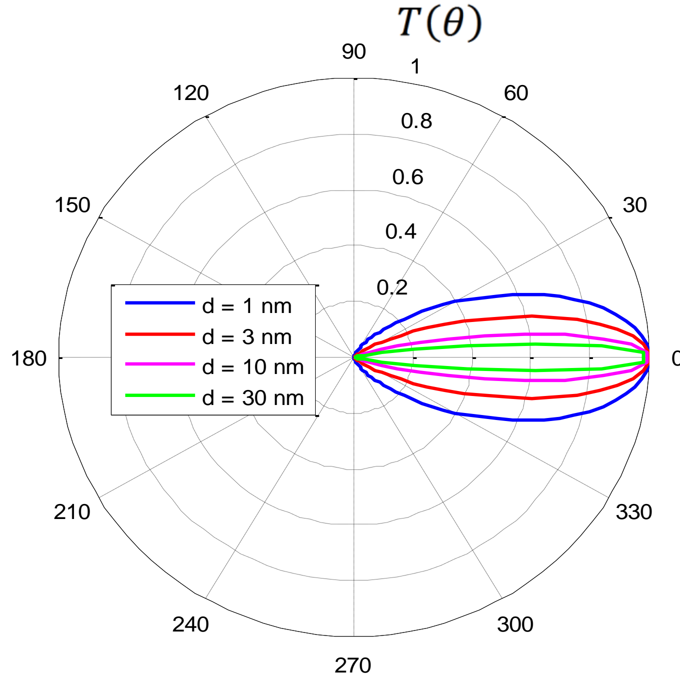


Figure 4: Transmission probability as a function of the incident angle θ for various gap distances between two gates.

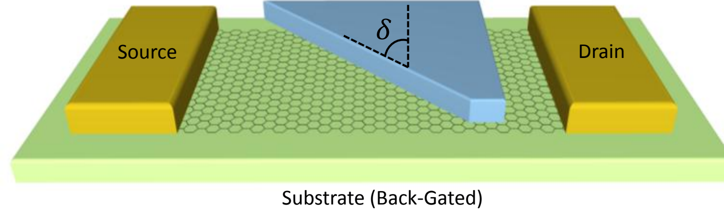


Figure 5: Schematic of a tilted GPNJ device built on a graphene sheet.

Figure 5 shows another type of GPNJ configuration. The gap is geometrically tilted at an angle δ rather than vertical. The electron transmission probability is given as [76]

$$\begin{aligned} T_\delta(\theta) &= \frac{1}{2}[T(\theta+\delta) + T(\theta-\delta)] \\ &= 0.5cos^2(\theta+\delta)\, e^{-\pi k_F d\, sin^2(\theta+\delta)} + 0.5cos^2(\theta-\delta)\, e^{-\pi k_F d\, sin^2(\theta-\delta)} \end{aligned} \quad (3)$$

where θ is the incident angle, and δ is the tilted angle of the GPNJ.

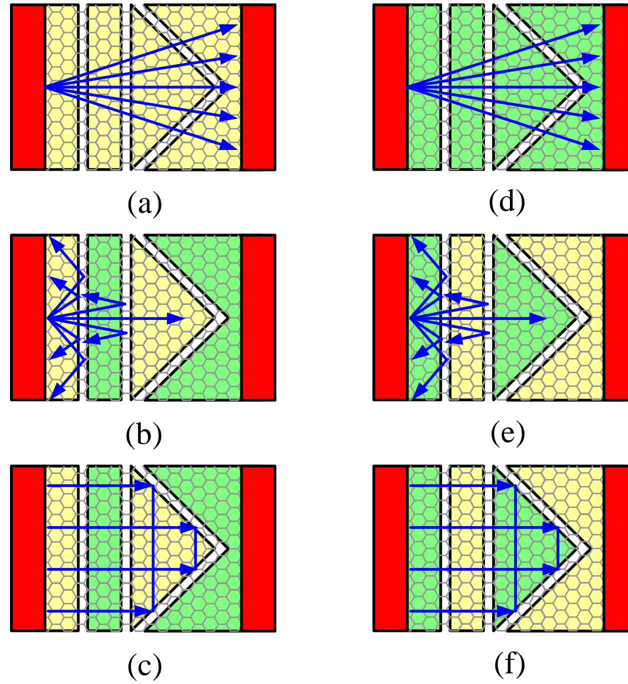


Figure 6: Illustration of the operation principle of a GPNJ switches. (a) and (d) show p- and n-type GPNJ switches at ON state, respectively, (b) and (c) show p-type GPNJ switches at OFF state for electrons coming from various angles and parallel direction, respectively, (e) and (f) show n-type GPNJ switches at OFF state for electrons coming from various angles and parallel direction, respectively.

Based on these two types of configurations in Figure 3 and Figure 5, a basic GPNJ switch is created. Figure 6 (a) and (d) show the p- and n-type switches at ON state, respectively. The yellow and green gates are applied with positive and negative voltages, respectively. Since all gates in both switches are applied with the same voltage, no junction is created. Therefore, the electrons with all the incident angles can pass through the switches. Otherwise, if two adjacent gates are applied with opposite voltage polarities, two vertical and one tilted junctions will be created. In Figure 6 (b) and (e), two vertical GPNJs work as collimators to filter those electrons with large incident angles. Most of the electrons with normal incident angles that pass through the first two junctions are blocked and reflected back by the third tilted junction, shown in Figure 6 (c) and (f),. As a result, the switches are turned off, and those remaining electrons passing through the entire switches contribute to the leakage current.

2.2.1 Resistance Modeling

Both ON and OFF resistances depend on the transmission probabilities of the electrons passing through the junction. Previous work in [62] used the transmission in a symmetric GPNJ to evaluate the resistance. However, this resistance is accurate only at a small bias and a low temperature. To obtain a realistic value at room temperature, Landauer formula is used herein as [74]

$$I = \frac{2q}{h} \int_{-\infty}^{+\infty} dE M_{eff}(E)[f(E-\mu_1) - f(E-\mu_2)] \tag{4}$$

where $f(x)$ is the Fermi distribution function, μ_1 and μ_2 are the electrochemical potentials on two sides of the graphene, and E is the energy level. M_{eff} is the effective number of modes at a given energy level E, which can be written as

$$M_{eff}(E) = \sum_{n} T(\theta_n). \tag{5}$$

The transmission probability for an asymmetric junction is needed and written as [64]

$$T(\theta) = \begin{cases} u(\theta_C - \theta_1) \dfrac{\cos\theta_1 \cos\theta_2}{\cos^2\left(\dfrac{\theta_1 + \theta_2}{2}\right)} e^{-\pi d \frac{2k_{F1}k_{F2}}{k_{F1}+k_{F2}} sin(\theta_1) sin(\theta_2)} & \text{, for PN junctions} \\ u(\theta_C - \theta_1) \dfrac{\cos\theta_1 \cos\theta_2}{\cos^2\left(\dfrac{\theta_1 + \theta_2}{2}\right)} & \text{, for } p^+p \text{ or } n^+n \text{ junctions} \end{cases} \tag{6}$$

where $u(x)$ is a step function, k_{F1} and k_{F2} are the Fermi wavevectors at the incident and emergent sides, respectively, θ_1 and θ_2 are the incident and emergent angles of the electrons passing through the junction, respectively, and $\theta_c = \sin^{-1}(-k_{F2}/k_{F1})$ is the critical angle according to the Snell's Law as

$$k_{F1}\, sin\,\theta_1 = k_{F2}\, sin\,\theta_2. \tag{7}$$

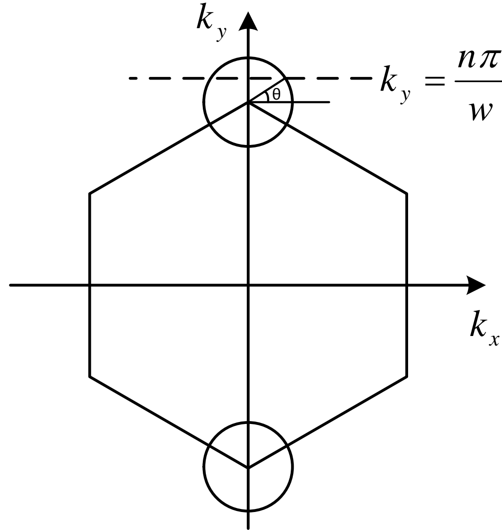


Figure 7: The Brillouin zone of graphene.

The incident angle of an electron beam is obtained from the Brillouin zone of the graphene sheet as shown in Figure 7. $k_y = n\pi/W$ is the quantized transverse wavevector

of electrons, and W is the width of the graphene sheet. The number of conduction modes depends on the number of quantized values of k_y passing through a circle with a radius of Fermi wavevector k_F, which is modulated by applying a voltage to the backgate.

For a given mode, the incident angle is written as

$$\theta_n = arcsin\left(\frac{\left|k_y - \frac{2\pi}{3b}\right|}{k_F}\right). \tag{8}$$

Transmission probabilities for an electron to pass three junctions are calculated from (6) and denoted as T_1, T_2, and T_3, respectively. The overall transmission probability, T, is then written as [77]

$$\frac{1-T}{T} = \frac{1-T_1}{T_1} + \frac{1-T_2}{T_2} + \frac{1-T_3}{T_3}. \tag{9}$$

Once the current is obtained from (4), the total resistance of the device is obtained as

$$R_{device} = \frac{V_{dd}}{I} + R_c \tag{10}$$

where $R_c = \rho_c/W$ is the additional non-ideal contact resistance, and ρ_c is the contact resistivity. The experiment work in [78] reported that the total contact resistance including both the ideal quantum resistance and the non-ideal contact resistance is 185 ± 20 Ω·μm at room temperature and 120 ± 20 Ω·μm at low temperatures. In current work, ρ_c is taken as 100 Ω·μm.

2.2.2 Capacitance Modeling

The input gate capacitance C_g is the equivalent capacitance of the electrostatic and quantum capacitances that are in series

$$c = \left(c_{ox}^{-1} + c_Q^{-1}\right)^{-1} \tag{11}$$

where $c_Q = \gamma E_F$ and $c_{ox} = \varepsilon / t_{ox}$. t_{ox} is the gate oxide thickness, and $\gamma = (4\pi q^2)/(h^2 v_F^2)$ is a constant depending on the graphene properties. To evaluate the capacitance of a graphene interconnect, t_{ox} is treated as the interconnect pitch based on the typical dielectric thicknesses in conventional Cu/low k interconnect technologies as shown in Figure 8.

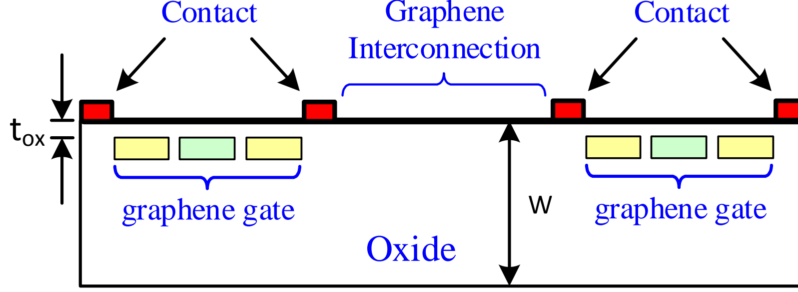


Figure 8: Cross section view of GPNJ devices and interconnects.

2.2.3 Validation and Comparison with NEGF Simulation

To validate the analytical approach, the experimental and simulation data for a tilted split GPNJ are compared in Figure 9.

The average transmission per mode through a GPNJ is written as [64]

$$T_{av}(E_F) = \frac{1}{2} \int T(\theta) cos\theta d\theta \tag{12}$$

where $T(\theta)$ is the angle dependent transmission in (6). In the case of a junction with tilted angle δ, the transmission is modified as $T(\theta + \delta)$. For a symmetric abrupt junction with no gap distance, the transmission is

$$T_{av}(E_F) = \frac{1}{2} \int T(\theta)\, cos^2(\theta + \delta)\, d\theta = \frac{2}{3} cos^4 \frac{\delta}{4}. \tag{13}$$

The average transmission reduces with δ. The reduction originates from the rotation of the transmission lobe into a low mode density region [63]. For a split junction, the impact of a tilt is calculated numerically and is shown in Figure 9 along with the T_{av} from

experiments [79]. The experimental data match well with the simulation and confirm the rotation of the transmission lobe and effectively modifying the incident angles.

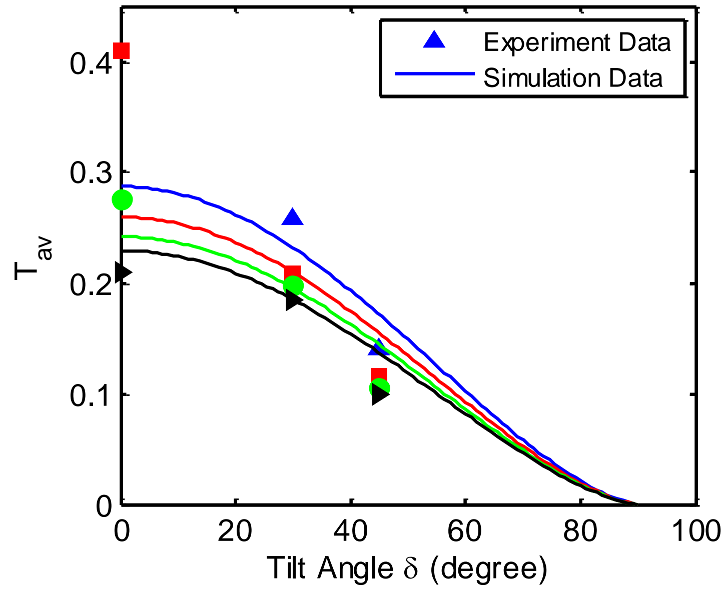


Figure 9: Comparison of the average transmission probability of a GPNJ device between experimental and simulation data [64].

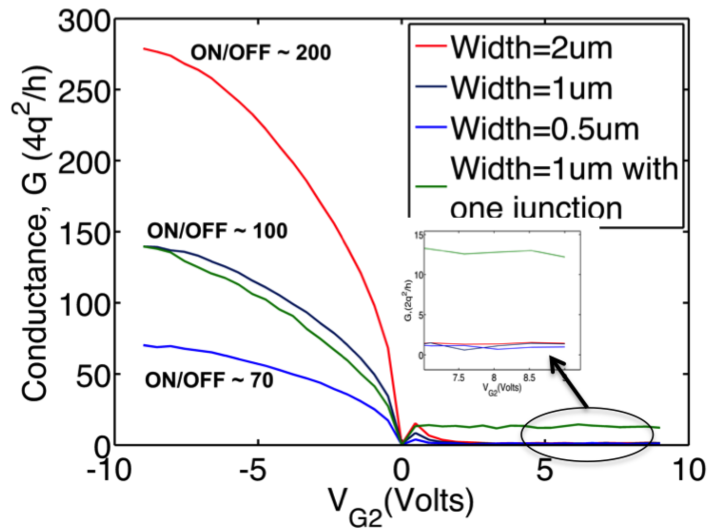


Figure 10: Conductance versus control voltage of a GPNJ switch based on NEGF results for various width of the graphene sheet.

The analytical models in this chapter do not consider the effects of charge impurity and electron phonon scatterings and edge effect, which may introduce non-ideal electron transmission probabilities and randomize electron trajectories. For a single junction, results herein match well with the experimental results in Figure 9. However, for multi-junction devices where non-ideal effects may play more prominent roles, currently there are no experimental data available. The preliminary NEGF simulation results from Virginia Tech are shown in Figure 10. One can observe that the ON/OFF ratio drops as the channel width decreases. This mismatch between the analytical expression and NEGF results are not being understood yet. Therefore, the performance projections obtained in this chapter represent the upper bound on the potential performance of GPNJ circuits.

2.2.4 Performance Comparison with CMOS FETs

The basic assumption to calculate the performance metrics is that the minimum feature size F is 16 nm. The gap between two vertical split gates d is equal to the minimum feature size. The gap distance for the 45° tilted junction is taken as $1.5F$ to further reduce the leakage current. The minimum width of the graphene sheet is set as twice of the gap distance to increase the triangular gate area and to make better electrostatic coupling between the gates and graphene. The supply voltage V_{dd} is set as $0.5V$. The control voltage is applied on the leftmost rectangular gate and the third triangle gate, as shown in Figure 7; the input voltage is applied on the other two gates. From the results shown in Figure 11, the ON and leakage resistances increase as the control voltage increases. This is because electrons coming from the graphene sheet with a large k_F have a small critical angle according to the Snell's Law, which collimate the electrons and block electrons with incident angles larger than the critical value. Thus, decreases in both ON and OFF currents lead to higher resistances. The increase of the leakage resistance is more significant than that of the ON resistance, suggesting that the ON/OFF ratio

improves as the control voltage increases. Therefore, the control voltage is one design parameter that needs to be optimized in the later system-level analysis.

Compared with previous MUX-based graphene logic gate [62], an additional vertical junction are added as collimators to further block electrons with large incident angles before they reach the 45° tilted junction to further reduce the leakage current. The absolute value of the control voltages applied on the green and yellow gates are increased to further suppress the leakage current and improve the ON/OFF ratio in Figure 11.

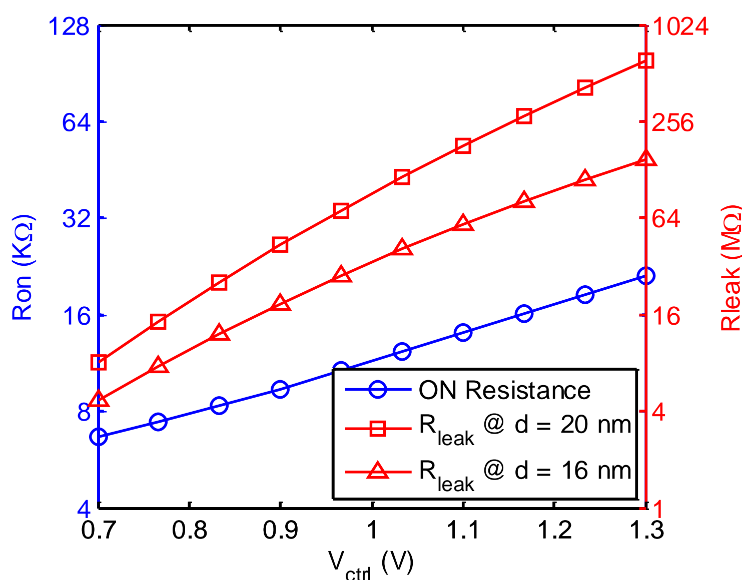


Figure 11: ON and OFF resistances versus controlled voltage at Vdd = 0.5V.

Table 1: Device-Level Comparison of Various Metrics between GPNJ and Si CMOS Devices.

Metrics	CMOS	GPNJ
I_{on} (μA/μm) @ Vdd = 0.5V	416.5	841.8
I_{off} (nA/μm) @ Vdd = 0.5V	231.6	210.8
I_{on} (μA/μm) @ Vdd = 0.7V	1136.2	2351.2
I_{off} (nA/μm) @ Vdd = 0.7V	865.0	838.2
Contact Resistance (Ω*μm)	-	100
C_g @ W = 4F (aF)	36	33.77

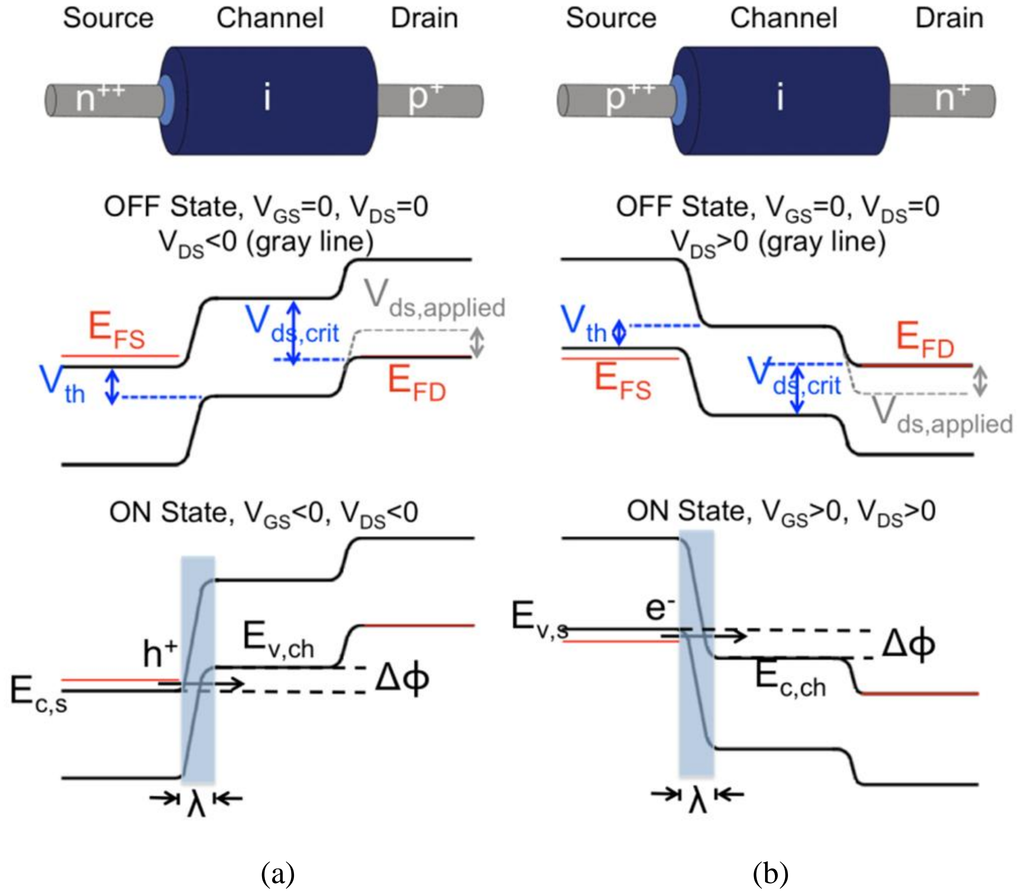
For comparison purpose, device models for Si CMOS technology are taken from Predictive Technology Model (PTM) at the 16 nm technology node [67]. The channel width and the supply voltage of CMOS switches are the same as those of GPNJ switches. The control voltage of GPNJ switches is chosen as 1.2V to have a comparable leakage resistance with a CMOS transistor. Table 1 shows comparisons among various metrics. The major advantage of a GPNJ switch is that it provides a higher driving current at a comparable leakage performance compared to its CMOS counterpart.

2.3 Tunneling Devices

TFETs show potentials in overcoming the power wall issues in terms of a significant reduction in the supply voltage. In this chapter, InAs nanowires are considered in light of direct bandgaps that eliminate the necessity for phonon assistance in tunneling. In Figure 12, p- and n-type TFETs are realized by assuming n-i-p and p-i-n structures, respectively. In Figure 12, the band diagram is modified by applying proper bias voltages between the device terminals [80].

Tunneling occurs between the n/p doped source and the intrinsic channel by introducing a tunnel window ($\Delta\Phi$ in Figure 12) 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 larger than the threshold voltage introduces a non-zero energy window where tunneling occurs. From this point on, the impact of the charges inside the channel is taken

into account to calculate how the position of the valence/conduction band in the channel changes with the applied gate voltage.



(a) (b)

Figure 12: Schematic of an InAs nanowire and band diagram in the OFF/ON states for (a) p-type and (b) n-type GAA TFETs.

The drain current flowing through the device due to the tunneling of carriers at the source-channel junction is modeled using the WKB approximation to calculate tunneling probability [81] and Landauer's formula [74]

$$I_{ON} = \frac{2q}{h} T_{WKB} k_B T \cdot \ln\left[\frac{\left(1+e^{(E_{c,s}-E_{FS})/k_BT}\right)\left(1+e^{(E_{v,ch}-E_{FD})/k_BT}\right)}{\left(1+e^{(E_{c,s}-E_{FD})/k_BT}\right)\left(1+e^{(E_{v,ch}-E_{FS})/k_BT}\right)}\right], \tag{14}$$

where $T_{WKB} = \exp\left(-4\lambda\sqrt{2m^*}E_g^{3/2}/3q\hbar(E_g + \Delta\Phi)\right)$ is the tunneling probability, and $\lambda = (\varepsilon_{nw} d_{nw}^2 \ln(1 + 2d_{ox}/d_{nw})/8\varepsilon_{ox})^{1/2}$ is the screening length. Using the relation in (14), current through a p-type device is plotted versus the gate voltage in Figure 13 with various nanowire diameter and carrier effective masses at a constant oxide thickness of 1 nm. The drain current through the device increases with reduced nanowire diameter and carrier effective mass. In this work, 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. Below 6 nm, however, diameter dependencies cause significant reduction in current and are no longer ignored [82]. Therefore, system-level optimizations are performed for a 6 nm diameter nanowire device. The results match well with those in atomistic full-band simulations, and the drain current normalized to the nanowire diameter is about 130 μA/μm [82].

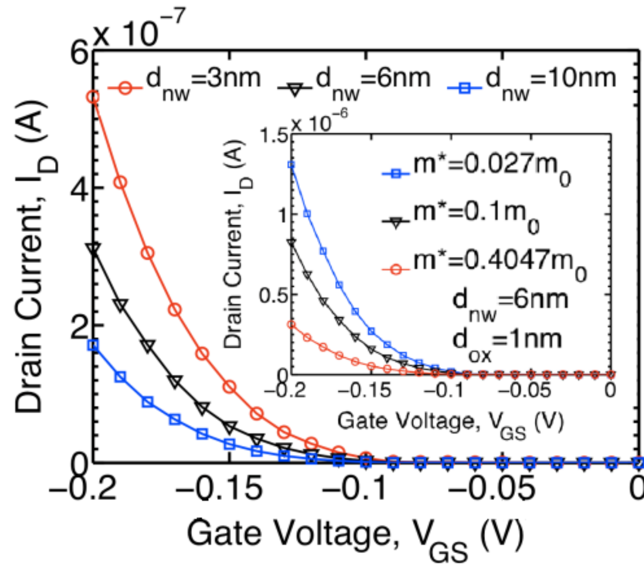


Figure 13: ID-VGS curve of a p-type TFET for various nanowire diameters and carrier effective masses. Higher currents are achieved at smaller nanowire dimensions due to enhanced gate control. Smaller effective masses increase the tunneling probability; hence offer larger current values [80].

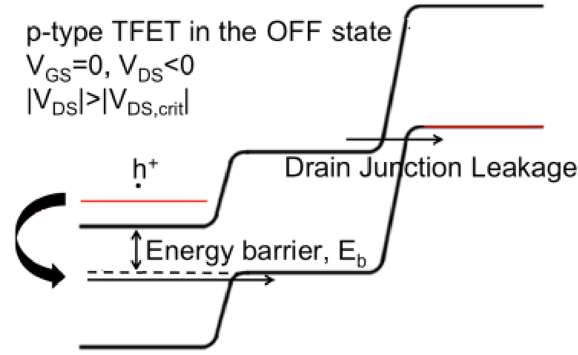


Figure 14: Leakage mechanisms considered in this work shown on a p-type TFET.

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 work, 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: the trap assisted tunneling in the source-channel junction [83] and the conduction in the drain-channel junction at high V_{ds} values as shown in Figure 14.

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 ~4kT 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, which 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} 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 [84].

To ensure that the potential inside the channel only changes with respect to the gate voltage, we assume that the gate oxide thickness (d_{ox}) is small and the oxide capacitance

is much larger than the drain capacitance [65]. In 1-D devices operating in the quantum capacitance limit (QCL), the oxide capacitance is much larger than the quantum capacitance (C_q), which provides small gate capacitances since $C_g=C_{ox}\times C_q/(C_{ox}+C_q)\sim C_q$ [65, 74]. The quantum capacitance is calculated by

$$C_q = q\frac{\partial Q_{ch}}{\partial E_{v,ch}} = q^2\frac{\partial}{\partial E_{v,ch}}\int_{E_{c,s}}^{E_{v,ch}} T_{WKB}DOS(E)\big(f_s(E)-f_D(E)\big)dE \qquad (15)$$

Note that the total gate capacitance is not small because of the fringing fields from the gate to the source and drain [65, 85], which have been taken into account in this work.

2.4 Planar and FinFET Si CMOS Devices

The SPICE model are taken from ASU Predictive Technology Model (PTM) [67], which provides accurate, customizable, and predictive model files for future CMOS transistors. These predictive model files are compatible with standard circuit simulators, such as SPICE, and scalable with a wide range of process variations. Recently, a new set of models has been released for multi-gate transistors (PTM-MG), such as bulk FinFET, from 20nm to 7nm nodes have been released. Two versions, high-performance (HP) and low-standby power (LSTP), are offered based on the BSIM-CMG which is a dedicated model for multi-gate devices.

2.5 Gate-All-Around Nanowire FETs

The device-level compact models for VFETs and LFETs are adopted from the previous work [68]. These models are developed based on the compact BSIM-CMG model that has been calibrated to TCAD simulations to account for the quasi-ballistic transport [86, 87]. The analytical model is applied in the calculations of parasitic resistance and capacitance for various device configurations up to Metal 1 level.

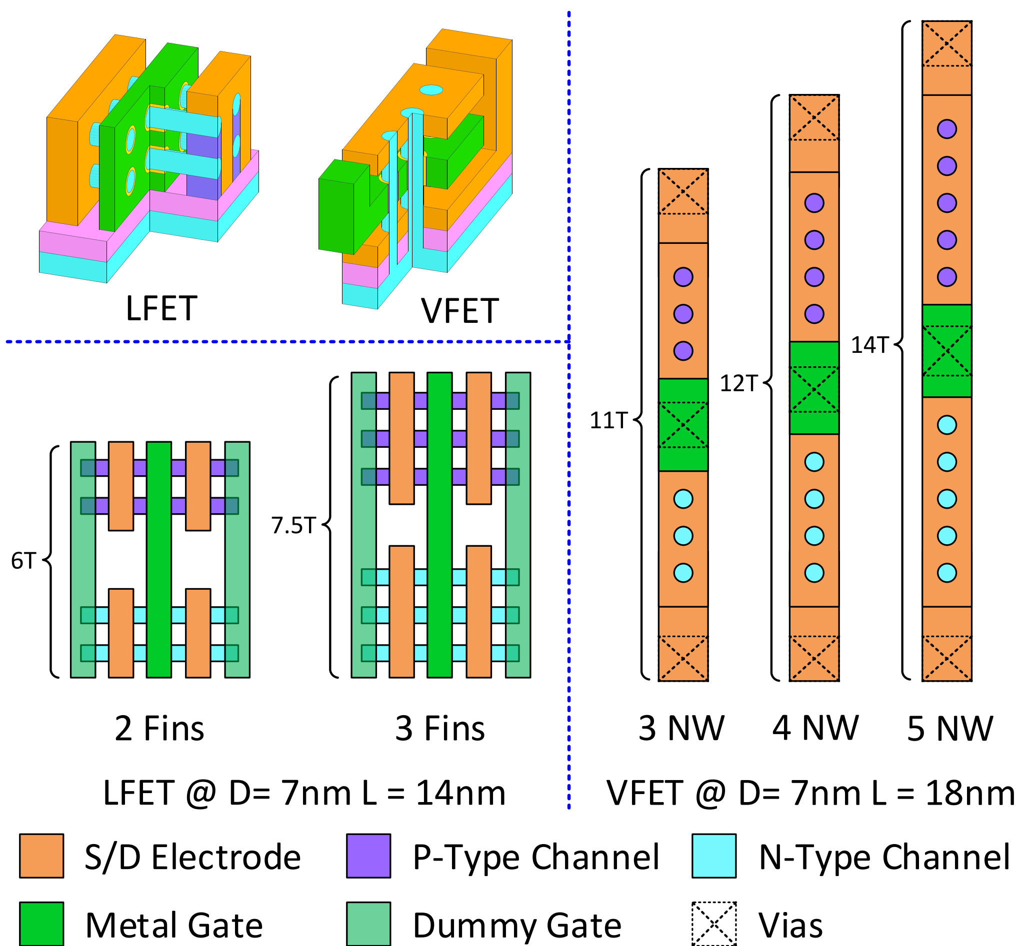


Figure 15: The 3D view and layout of VFETs and LFETs with various configurations.

Figure 15 shows the 3D and layout views of VFETs and LFETs at various configurations. For VFETs, the source and drain are on the top and bottom of the gate, therefore adjacent transistors are completely independent from each other. For the LFET devices, however, a dummy gate is required on each side of the cell to achieve the source and drain isolation between the nearby cells, which significantly increases the area overhead [88]. The advantage of VFETs is that the channel length is in the vertical direction. Increasing the channel length does not enlarge the footprint area or compromise the contact area. All devices are assumed to be fabricated on (001) wafer,

using Si/$Si_{0.5}Ge_{0.5}$ as the channel material for lateral nFET/pFET and Si for VFETs. Table 2 lists the technology assumptions for LFET and VFET devices.

Table 2: Technology Assumptions for LFET and VFET Devices

Parameters	LFET	VFET
Gate Length (nm)	14	18
Spacer Thickness (nm)	4	5
Gate Pitch (nm)	32	32
Metal 1 Pitch (nm)	24	24
Fin/NW Pitch (nm)	18	18
NW Diameter (nm)	7	7
Number of Fins/NWs	2, 3	3, 4, 5
Cell Height (Tracks)	6, 7.5	11, 12, 14
Nominal Vdd (V)	0.6	0.6
Channel Stress (GPa)	1	0
Channel Orientation	[110]	[001]

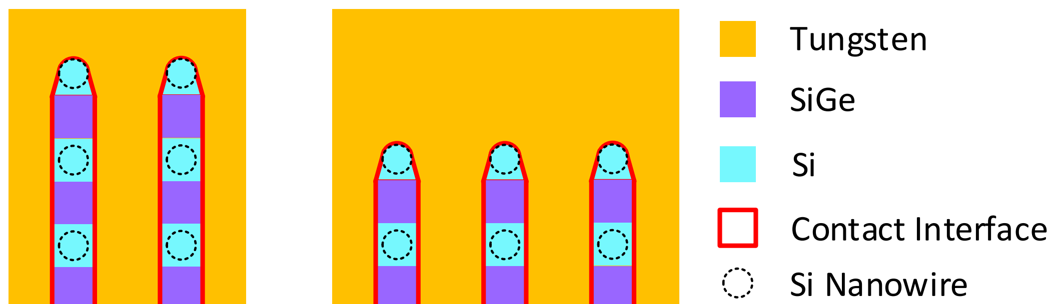


Figure 16: The cross-sectional view of the n-type LFETs with (a) 2fin/3stack and (b) 3fin/2stack using wrap contacts at source and drain regions.

In the previous device-level modeling [68], it is assumed that the source and drain contacts of LFETs are based on the commonly used epitaxially grown structures, which provides an additional channel stress on top of the initial strain level of 1.5 GPa from the strain relaxed buffer (SRB) [89]. Recently, a new contact model is developed for the wrap contact [90], which surrounds the whole fin structure and creates a larger contact interface, as shown in Figure 16. Although it may downgrade the strain from the initial SRB by 0.5 GPa, it reduces the parasitic source and drain access resistance by about 30%

compared with the contact built on the source and drain epitaxy, especially for devices with a small fin pitch [90]. This leads to up to 15% of the improvement in the ON current based on the same leakage. If the embed epitaxy can provide an optimistic stress boosting of 0.5 GPa, the wrap contact still has a comparable ON current. Therefore, in this dissertation, the wrap contact model is used in the source and drain of LFETs.

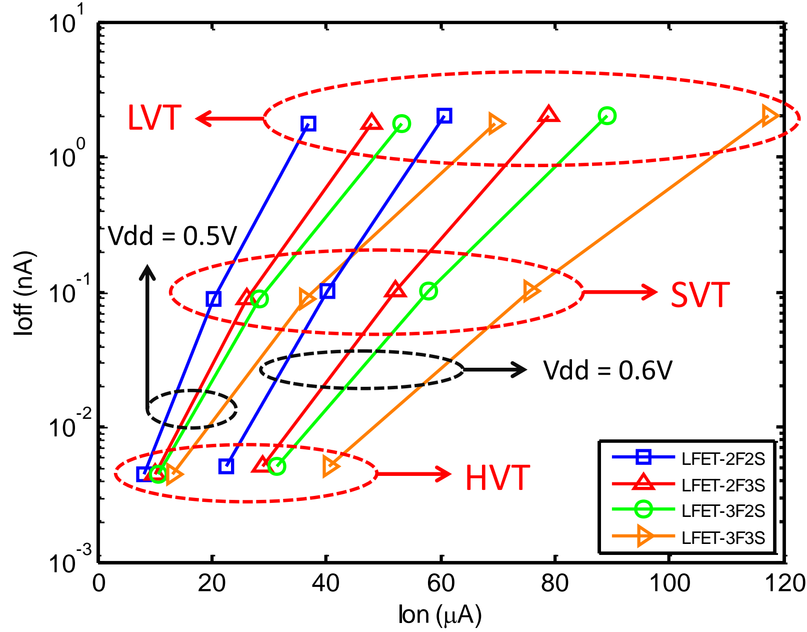


Figure 17: Intrinsic leakage current versus ON current for LFET devices with three Vth options at two supply voltages. In the legend, LFET- xFyS indicates that the number of fins and stacks used in an LFET are ‘x’ and ‘y’, respectively.

To compare various device options, the work functions of n- and p-type LFETs and VFETs are adjusted individually at the nominal supply voltage of 0.6V. The leakage currents for the high-, standard-, and low-Vth devices are assumed to be 5pA, 0.1nA, and 2nA, respectively. After the work function is set, Figure 17 shows the intrinsic leakage and average ON current of n- and p-type LFETs with four configurations at two supply voltages. In general, the ON current drops as the supply voltage decreases, especially for the high-Vth devices. Since they operate at near-threshold region, more than 50% of the

current reductions are observed. The leakage current, however, is insensitive to the supply voltage, thus the leakage power decreases almost linearly as the supply voltage scales. Another observation is that the ON current of the LFET with 3fin/2stack is higher than the one with 2fin/3stack even though the total number of nanowires is the same. The reason is that the contact area is smaller for a 2-fin LFET, as shown in Figure 16. In addition, the current needs to pass through the narrow spacings of the tungsten between the top two layers of nanowires to reach the bottom nanowires. Therefore, the parasitic source and drain access resistances of 2-fin devices are larger than those of 3-fin devices, leading to a smaller ON current. However, this ON current advantage for the 3-fin LFET diminishes as the device threshold increases. This is because the device is more dominated by the channel resistance instead of the parasitic resistance, especially at a small supply voltage. Moreover, the 3-fin LFET uses a 7.5-track layout as shown in Figure 15 and thus consumes a large footprint area. This area/performance trade-off will be investigated in Section IV.

To compare the DC performance of LFETs and VFETs, Figure 18 shows the average ON current of n- and p-type FETs among 7 device configurations with two Vth flavors at three supply voltages. In most cases, the ON current of VFETs is relatively low compared with that of LFETs mainly because of the lack of stress in the channel during the fabrication process. At a low supply voltage, however, the difference between the ON currents of VFETs and LFETs decreases, and VFETs even offer larger driving currents at the supply voltage of 0.4V with the high-Vth flavor. This is because a longer channel of VFET provides a steeper subthreshold slope and offers a larger ON current based on the fixed leakage target when the device operates at near-threshold region.

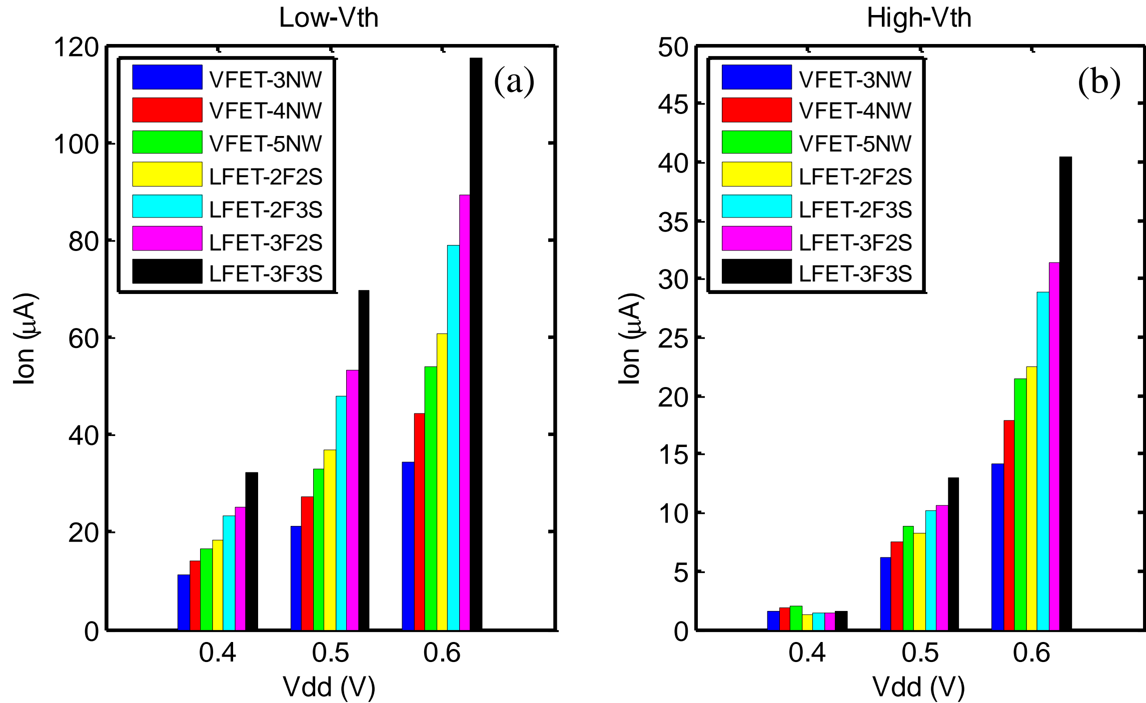


Figure 18: The ON current versus the supply voltage for LFETs and VFETs using 7 options with low- and high-Vth flavors. In the legend, VFET-xNW indicates that the number of nanowires used in a VFET is 'x'.

2.6 Conclusions

This chapter illustrates and develops the detailed modeling approaches, assumptions, and simulations at the device level. Based on the property of the angular dependent transmission probability of electrons observed in GPNJs, a modified device structure is presented. More elaborate physical models, including ON resistance, leakage current, contact resistance, and footprint area, are also developed based on the ideal edges and perfect ballistic transport to better evaluate the upper limit of delay and power consumptions of GPNJ circuits. To take into account the lithography patterning limitations, GPNJ circuits with curved corners are investigated as well. Compared with Si CMOS devices, GPNJ devices can potentially offer a larger driving current with comparable leakage current and footprint area.

The analytical device-level models for TFETs have also been developed to efficiently evaluate the overall system-level metrics in Chapter 5. It is demonstrated that the IV characteristics of a TFET behaves like a Si CMOS switch, but it provides much lower leakage current and ultra-low supply voltage at the cost of low ON current.

For the comparison, the conventional Si CMOS planar FET and FinFET device models are taken from the ASU PTM, which are based on calibrated TCAD simulations. For the ultimate CMOS GAA FETs, two device structures with various configurations are modeled based on the previous modeling work from IMEC. LFETs in general provide larger ON currents than VFETs as a result of the channel stress. At a low supply voltage with a high-Vth flavor, however, the ON current for LFETs downgrades more significantly because of the poor SCE control.

CHAPTER 3 INTERCONNECT- AND CIRCUIT-LEVEL MODELING AND SIMULATION

3.1 Introduction

This chapter demonstrates the modeling approaches and performs simulations on the interconnect and circuit levels. With the scaling of the CMOS technology, the delay and power consumption of the interconnects become ever more important issues because the die size does not scale proportionally and the aggregate length of interconnects increases with technology scaling. Graphene is a promising candidate to reduce the ever-increasing performance gap between devices and interconnects [91]. However, graphene is a two-dimensional structure and suffers from the large quantum resistance and the limited number of populated conduction channels. In addition, the graphene quality, such as the contact resistance, edge roughness, and mean-free-path (MFP) have significant influences on the delay [91]. Therefore, the performance improvement brought by graphene interconnects needs to be thoroughly investigated at both material and circuit levels to fully understand their potential advantages. In this chapter, for the first time, large realistic circuit designs are used to evaluate this potential. Both a 32-bit adder and an SRAM have been fully simulated in SPICE using accurate experimentally-calibrated FinFET and copper interconnect models for 10 and 7 nm technology nodes. It is found that the potential performance improvements offered by graphene interconnects highly depend on the intrinsic and parasitic capacitances and resistances of the devices and interconnects. Hence, it is critical to use realistic models for FinFET devices and Cu interconnects at any specific technology node. Multiple material-, device-, and circuit-

level parameters, including width, length, MFP, edge roughness, contact resistance, number of layers, supply/threshold voltages of the devices, and layout and placement efficiency have been analyzed to minimize the delay and EDP.

In this Chapter, the circuit-level modeling is illustrated for both conventional metal interconnects and novel multi-layer graphene interconnects. The SPICE simulation is also performed to quantify the potential improvement of using graphene interconnects for a 32-bit adder and an SRAM. In addition, an ARM Cortex-M0 processor is used to further benchmark the graphene interconnect at a larger and more realistic scale. Several logic gates based on GPNJ devices are presented as well, and the GPNJ-based SRAM is simulated and compared with its CMOS counterpart.

For the performance of the ultimate CMOS GAA FETs, recent studies [68, 69] have compared LFETs and VFETs. It was observed that VFETs provide smaller energy and footprint area than their LFET counterparts. These studies, however, were performed by only using a ring oscillator, a multiplier, and an SRAM as the representative benchmarking circuits. In [69], the multiplier is only studied at the placement and routing level to demonstrate the area saving without investigating the electrical performance. In [68], numbers of fins in an LFET and nanowires in a VFET were assumed to be four and seven, respectively, which may not be a fair comparison between the potential performances of these two devices. In this chapter, for the first time, the performance comparison is established with the implementation of an ARM core. 50 cells are characterized based on compact device-level models adopted from [68], which are calibrated by TCAD simulations. Compared with the previous work [68], more realistic physical dimensions, such as the gate extension and nanowire pitch, are used. In addition, we have also improved the LFET with a wrap-around contact scheme, which reduces the contact resistance for the LFET, improving the device ON current [90]. During the synthesis, multiple-Vth optimization is performed to develop a comprehensive

understanding of the ultimate performance advantages of different device structures. A variety of relevant device configurations are investigated, including different number of fins, nanowires, and nanowire stacks. This optimization flow is crucial to account for the trade-offs that otherwise could not be captured in the intrinsic device-level ON/OFF current and footprint area or a simple ring oscillator circuit. For instance, a VFET usually consumes less energy because of smaller parasitic capacitances, but its ON current is lower due to the lack of stress during the fabrication process. For a given frequency target, this means that more low-Vth and stronger cells and buffers are needed, which subsequently leads to an increased energy and leakage dissipation. These trade-offs also vary for different performance targets and device configurations, such as the number of fins, nanowires, and nanowire stacks. Therefore, a larger scale analyses are required to fully understand the advantages of different device options.

The rest of this chapter is organized as follow. Section 2 illustrates the resistance and capacitance per unit length of the conventional metal interconnects, where impacts of process variations on three interconnect fabrication processes are analyzed. In Section 3, multi-layer graphene interconnects are modeled and circuit-level comparison are performed. The performance improvement of using graphene interconnects are observed for a 32-bit adder and an SRAM. More detailed analyses are also made for an ARM core by performing placement and routing with commercial tools. Section 4 presents the GPNJ-based logic gates, including the conventional complementary logic and the reconfigurable logic gates. The GPNJ-based SRAM cell is also simulated and compared with its CMOS counterpart. At the end, two types of GAA FETs structures are benchmarked and compared in Section 5 by using SPICE simulation for a ring oscillator and multi-Vth optimization flow for an ARM core.

3.2 Metal Interconnect

In this section, the compact analytical models for the capacitance and resistance of the metal interconnect are introduced. For the interconnect geometries and material properties, Table 3 lists the process assumptions under both the nominal state and process variations for three technology nodes, where five sources of variations are investigated.

Table 3: Interconnect Process Variation Assumptions for N14, N10, and N7 Nodes.

	Process Feature	N14	N10	N7
Nominal Case	CD Line/Core (nm)	32	24	18
	CD Spacer (nm)	32	24	18
	Aspect Ratio	2	2	2
	Height Trench (nm)	64	48	36
	Height Stack (nm)	128	96	72
	K-Value low-k Dielectric	2.5	2.3	2.2
	Metallic Barrier Thickness (nm)	3	3	1
	Dielectric Barrier Thickness (nm)	15	10	7
	K-Value Dielectric Barrier	5.5	5.5	5.5
Variations (3σ)	CD Line/Core (nm)	3	3	3
	CD Spacer (nm)	1.5	1.5	1.5
	Etch (nm)	7	5	3.75
	CMP (nm)	9	9	9
	Overlay (nm)	6	5	4

3.2.1 Resistance Model

Copper interconnects are widely used in the semiconductor industry due to their low resistivity. As technology nodes go down, the size effects, including the surface and grain boundary scatterings, drastically increase the effective resistivity which can be modeled as

$$\rho_{eff} = \rho_0 \left\{ \frac{1}{3} \left[\frac{1}{3} - \frac{\alpha}{2} + \alpha^2 - \alpha^3 \ln(1+\alpha) \right]^{-1} + 0.45(1-p) \frac{\lambda}{w_0} \left(\frac{1+AR}{AR} \right) \right\} \quad (16)$$

$$\alpha = \frac{\lambda R}{d(1-R)} \tag{17}$$

where λ is the bulk MFP of an electron, d is the average separation of the grain boundaries, R is the fraction of electrons scattered by the potential barrier at the grain boundary, p is the fraction of electrons elastically scattered at the surface, w_0 is the actual width of the interconnect that excludes the barrier thickness, and AR is the aspect ratio, which is the height divided by the width of the interconnect.

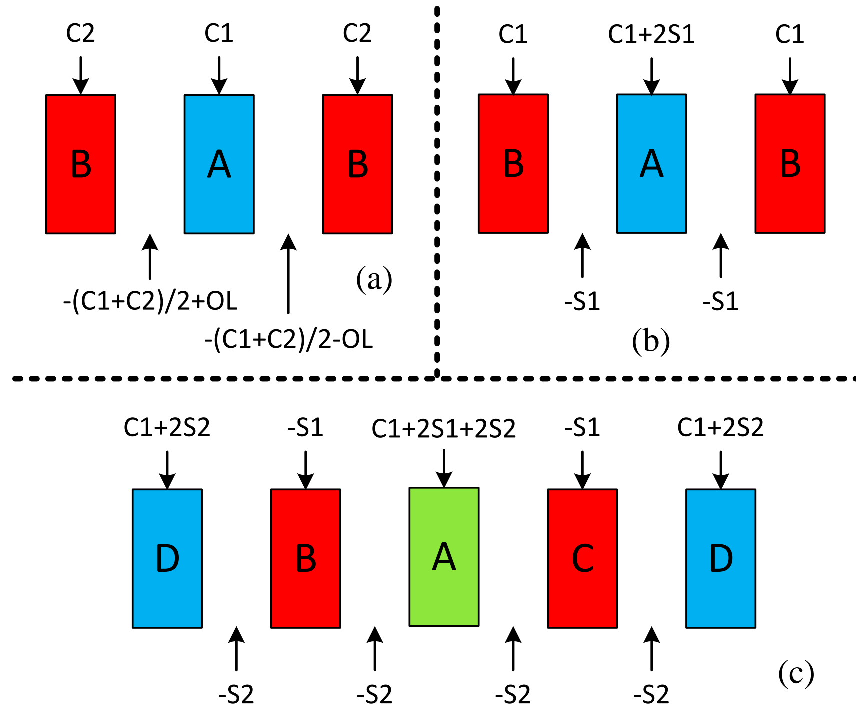


Figure 19: The impact of CD and overlay variations on the width of interconnects for the (a) LELE double patterning, (b) SADP, and (c) SAQP fabrication techniques.

Figure 19 shows the cross-sectional view and the impact of CD and overlay variations on the interconnect width for three patterning techniques [92]. Those interconnects, whose resistance are the most sensitive to the CD variations, are marked as 'A'. C1/C2 and S1/S2 are two independent CD line/core and spacer variations, respectively, that need to be added or subtracted on the nominal widths or spacings of the

interconnects, and OL represents the overlay variation. For the LELE double patterning process, there are two independent CD line/core variations C1 and C2 for two nearby interconnects; for the SAQP process, two independent CD spacer variations S1 and S2 are involved.

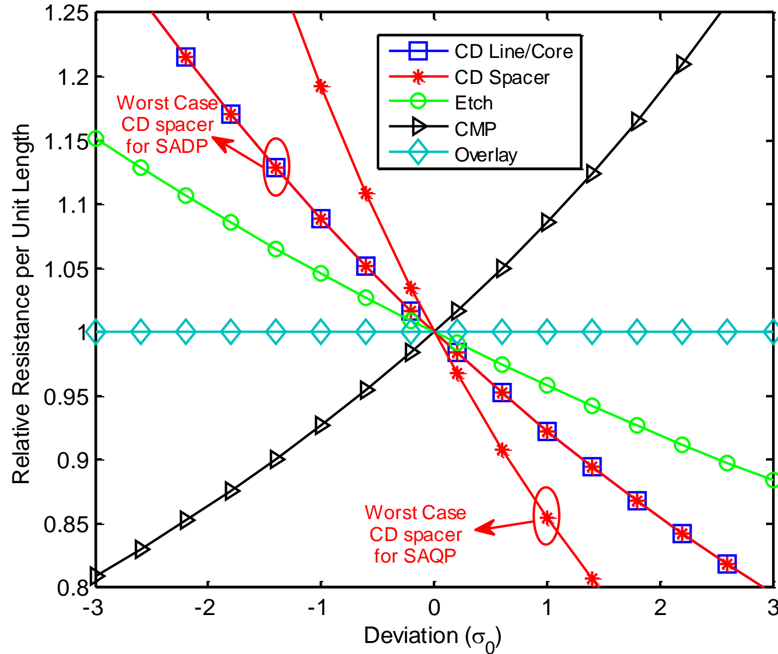


Figure 20: The resistance per unit length normalized to the nominal value versus the deviation in σ_0, which is quantitatively shown in the bottom part of Table 3 for five independent sources of interconnect variations.

Based on the assumption of the metallic barrier thickness and tapering trench angle shown in Table 3 and the line edge roughness of 0.5, the resistance per unit length relative to the nominal case is depicted in Figure 20, where five sources of interconnect variations are investigated, including CD line/core, CD spacer, etch, chemical mechanical polishing (CMP), and overlay variations. For the CD spacer variation, two curves are plotted, taking into account the worst case scenarios for the SADP and SAQP processes. Among five sources of interconnect variations, the CD spacer causes the largest impact on the interconnect resistance for the SAQP process at the worst case scenario. The

overlay variation for the LELE process does not affect the interconnect resistance. Compared with the interconnect capacitance in Figure 20, the interconnect resistance has the opposite trends under various sources of variations, including CD, etch, and CMP variations. Therefore, the impact of the interconnect variation on the overall performance depends on the circuits and systems. If the device has a large output resistance and a small input capacitance, the delay is dominated by the interconnect capacitance variation; if the driver resistance is small, the resistance variation has a larger impact.

3.2.2 Capacitance Model

The interconnect capacitance model without process variation, shown in Figure 21 (a), follows previous work [93]. The total capacitance is estimated as the summation of the line-to-line and line-to-ground capacitances, which can be written as the following equations [93]:

$$C_{total} = 2 \cdot C_{L2L}(W,T,S,H,\varepsilon) + 2 \cdot C_{L2G}(W,T,S,H,\varepsilon) \tag{18}$$

where

$$\frac{c_{L2G}}{\varepsilon} = \frac{W}{H} + 1.086\left(1 + 0.685e^{-\frac{T}{1.343S}} - 0.9964e^{-\frac{S}{1.343H}}\right)\left(\frac{S}{S+2H}\right)^{0.0476}\left(\frac{H_\rho}{H}\right)^{0.337} \tag{19}$$

$$\frac{c_{L2L}}{\varepsilon} = \frac{T}{S}\left(1 - 1.897e^{\left(-\frac{H}{0.31S}+\frac{T}{2.474S}\right)} + 1.302e^{-\frac{H}{0.082S}} - 0.13e^{-\frac{T}{01.326S}}\right) + 1.72\left(1 - 0.6548e^{-\frac{W}{0.3477H}}\right)e^{-\frac{S}{0.651H}} \tag{20}$$

where C_{L2L} and C_{L2G} are the line-to-line and line-to-ground capacitances of the interconnects, respectively; $W, T, S,$ and H are the interconnect cross-sectional dimensions that are shown in Figure 21 (a), and ε is the dielectric permittivity.

For the asymmetric interconnect structures caused by etch, CMP, and overlay variations (Figure 21 (c), (d), and (e)), the equation (2) is modified as

$$\begin{aligned} C_{total} = & C_{L2L}(W, T, S_1, (H_1 + H_2)/2, \varepsilon) \\ & + C_{L2L}(W, T, S_2, (H_1 + H_2)/2, \varepsilon) \\ & + C_{L2G}(W, T, (S_1 + S_2)/2, H_1, \varepsilon) \\ & + C_{L2G}(W, T, (S_1 + S_2)/2, H_2, \varepsilon) \end{aligned} \tag{21}$$

where H_1 and H_2 are the dielectric thicknesses on the top and bottom of the interconnects, respectively, and S_1 and S_2 are the spacings on the left and right sides of the interconnects, respectively.

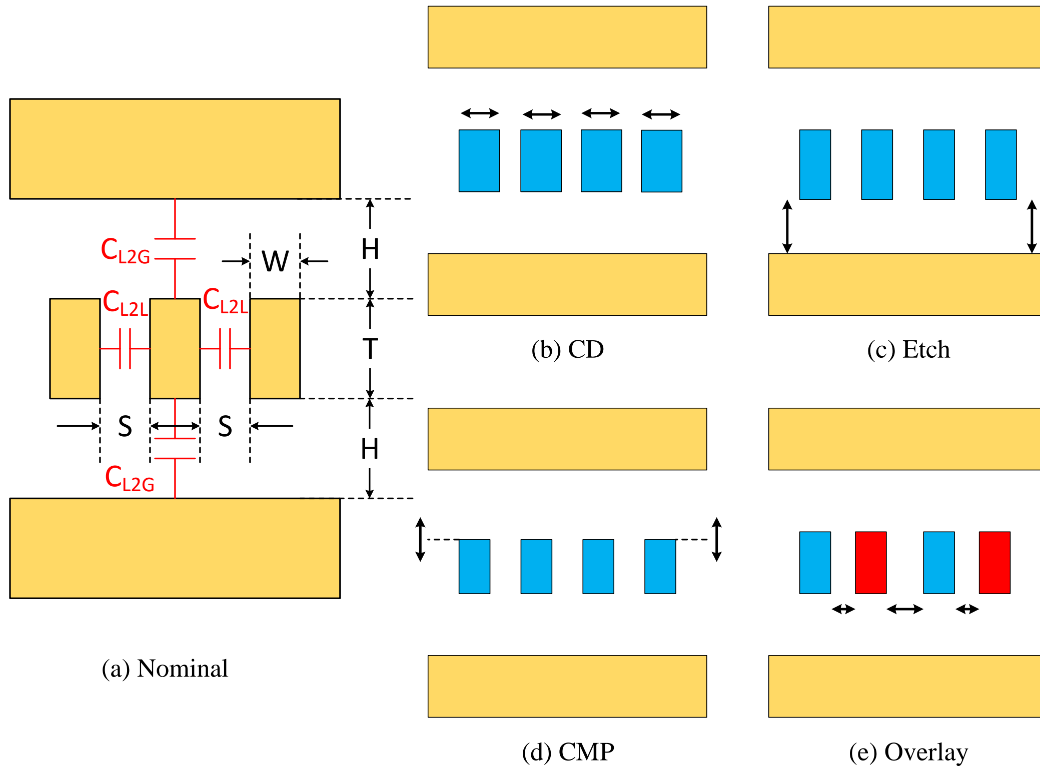


Figure 21: The cross-section view of multilevel interconnects under (a) no process variation, (b) CD variation, (c) etch variation, (d) CMP variation, and (e) overlay variation.

To validate this approach, the field solver Raphael has been used [94], and the comparisons for the interconnect under various sources of interconnect variations are shown in Figure 22. Since there is a thick dielectric barrier at the top of the interconnect on each metal layer as shown in Table 3, an effective dielectric constant 1.17 is used in the Raphael simulation. The results indicate that the maximum errors between the results obtained from the compact model and Raphael simulations are less than 2%. Therefore, by using these compact physical models, reasonably accurate estimations can be achieved, making a fast and efficient system-level benchmarking possible.

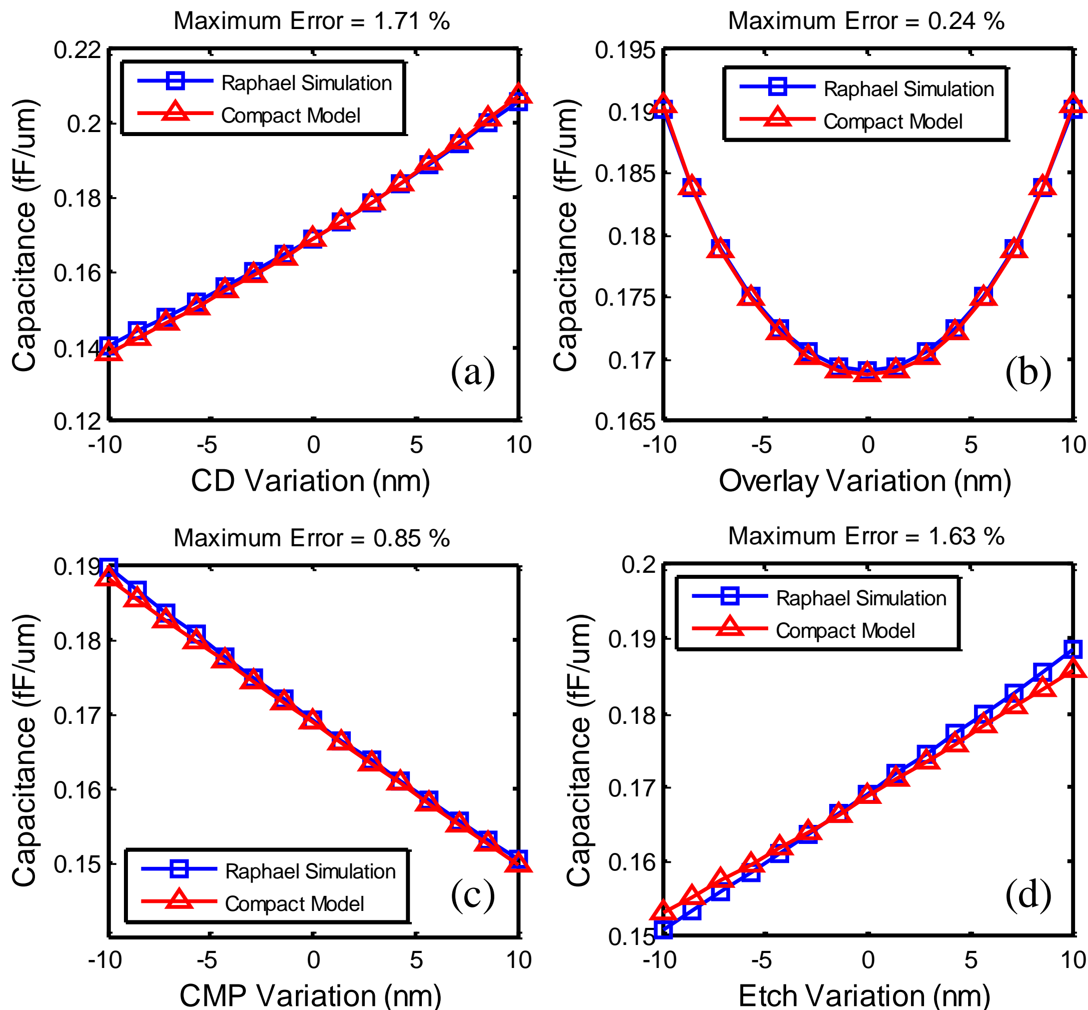


Figure 22: The comparison between the compact model and Raphael simulation for capacitance per unit length versus various types of process variation.

3.3 Graphene Interconnect

In this subsection, the compact resistance and capacitance models are described to take into account various design and intrinsic parameters of graphene interconnects, such as the electron MFP, edge roughness, dimensions, and the number of graphene layers. These models are used for the circuit-level simulation to quantify the potential benefits over the conventional copper interconnects in terms of the delay, energy, and EDP.

3.3.1 Resistance Model

To obtain the current in a single-layer graphene interconnect whose length is comparable or longer than the electron MFP, (4) needs be updated as

$$I = \frac{q}{h} \int_{-\infty}^{+\infty} 2 \sum_{n} \frac{l_{eff_n}}{L + l_{eff_n}} [f(E - \mu_1) - f(E - \mu_2)] dE \tag{22}$$

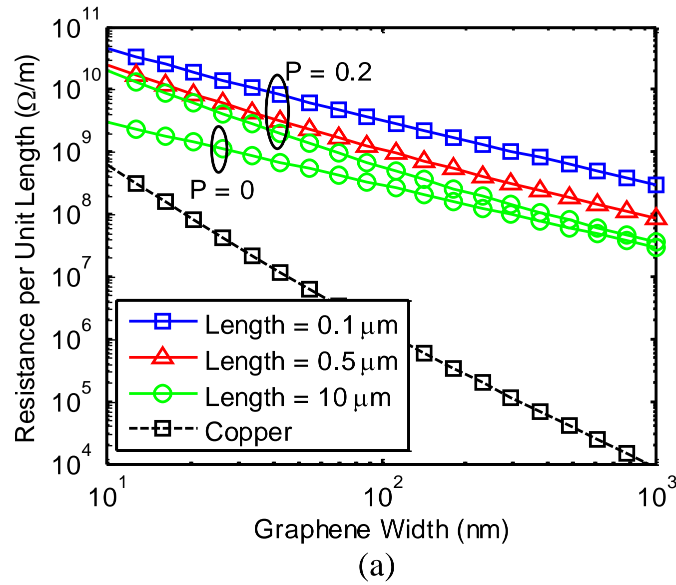
where L is the length of the interconnect, and l_{eff_n} is the effective MFP at the nth subband in graphene, which is given by

$$\frac{1}{l_{eff_n}} = \frac{1}{l_D} + \frac{1}{l_n} \tag{23}$$

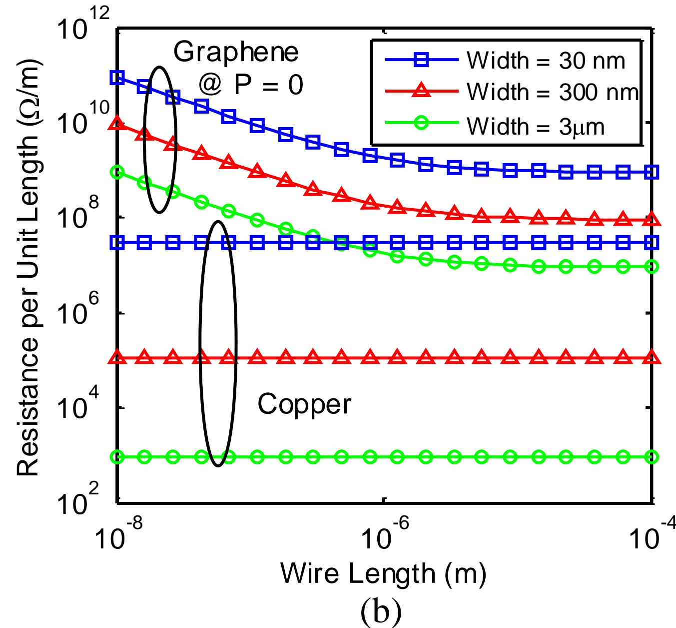
where l_D is the defect-induced MFP, and l_n is the MFP due to the electron scatterings at the graphene edges. The edge-scattered MFP is expressed as

$$l_n = \frac{W}{P} \cdot \frac{k_x}{k_y} = \frac{W}{P} tan^{-1}(\theta_n) \tag{24}$$

where k_x and k_y are the transverse and longitudinal wavevectors with respect to the K points in the Brillouin zone, P is the backscattering probability at the edges, θ_n is the incident angle of the electron beam at the nth subband in the zigzag graphene, and the Fermi energy E_F of the graphene is assumed to be $0.4eV$ in this work.



(a)



(b)

Figure 23: Comparison of the resistance per unit length between graphene interconnects and copper interconnects. (a) shows resistance per unit length versus the width of the interconnects at various interconnect length (b) shows the resistance per unit length versus the length of the interconnects for various interconnect widths.

Figure 23 shows the resistance per unit length for mono-layer graphene and copper interconnects at various lengths, widths, and edge scattering probabilities. Copper has a

smaller intrinsic interconnect resistance compared with graphene. To further reduce the resistance per unit length, multi-layer graphene interconnects are used because a larger current flow through multiple graphene sheets. For multi-layer graphene interconnects, two types of contact are depicted in Figure 24, including side and top contacts. The top contact is made on the top of the topmost layer of the graphene sheet, while the side contact is made with the assumption that every graphene layer is electrically connected with the contact. Using the side contact can provide a smaller resistance, but it is more difficult to manufacture compared with the top contact. The resistance model for the multi-layer graphene interconnects using the top contact follows the work [95].

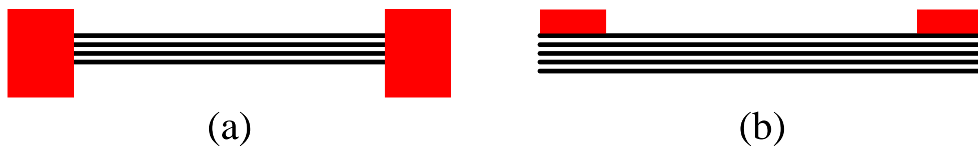


Figure 24: Cross-section view of the multi-layer graphene interconnect using (a) side contact (b) top contact.

3.3.2 Capacitance Model

The value of the electrostatic capacitance per unit length of a multi-layer graphene interconnect is obtained from the simulation results of RAPHAEL [94]. Due to the small quantum capacitance and graphene sheet thickness, it provides a smaller capacitance per unit length as shown in Figure 25. Thus, it reduces both the interconnect delay and the circuit dynamic power dissipation, which are major advantages of graphene interconnects.

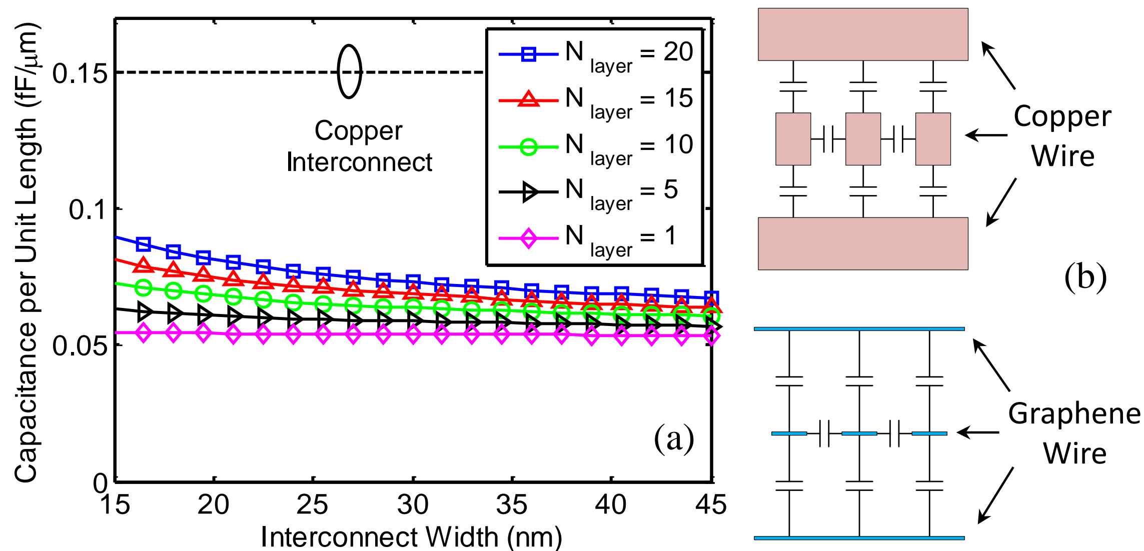


Figure 25: (a) Capacitance per unit length of the multi-layer graphene interconnects versus the width for various numbers of layers. (b) Cross-sectional view of the copper interconnects and graphene interconnects based on the assumption that the interconnect pitch is twice of the interconnect width.

3.3.3 Simulation Results

Based on the interconnect resistance and capacitance models described in the previous subsections, two widely used circuits are chosen as the representative designs to perform the circuit/technology co-optimization and benchmarking for graphene interconnects [96, 97]. The device-level models are taken from IMEC, which incorporates TCAD simulations, silicon measurements for certain parts of the device (contact resistance etc.) for ground rules at both 10nm and 7nm technology node. The device parasitics are obtained from detailed Raphael simulations with an appropriate metal stack assumption [98]. The interconnect resistance model follows the one by IMEC, which is based on detailed simulations and calibrations with silicon measurements of interconnects at different trench widths.

3.3.3.1 32-bit Adder

An adder is a common execution unit in processors. It is the building block for more complex logic circuits such as multipliers and dividers. Therefore, fast circuit simulations are desirable to make multi-parameter optimization feasible. The adder analyzed in this chapter is the 32-bit Kogge-Stone adder, which is the fastest configuration in the family of carry look-ahead adders [99]. For other types of carry look-ahead adders, the relative performance improvement is similar because repeaters are inserted for long interconnects to keep the same segment length for all drivers. Figure 26 shows the schematic of the 32-bit Kogge-Stone adder with a total of seven stages exist in the critical path.

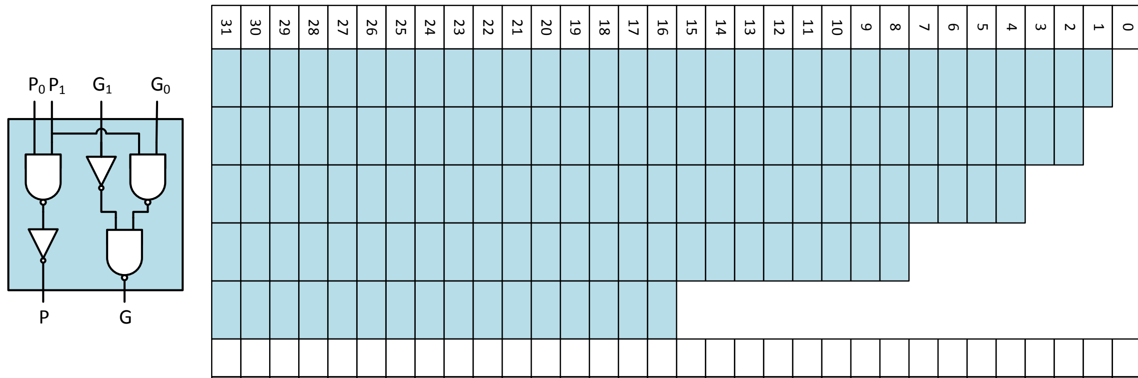


Figure 26: The schematic of a 32-bit Kogge-Stone adder.

Table 4: The Default Configuration of the Simulation.

Parameters	Value
Technology Node (nm)	10
Interconnect Pitch (nm)	40
Aspect Ratio	2
Block Pitch (um)	5
Contact Resistance ($\Omega\cdot\mu m$)	100

To measure the performance advantages of multilayer graphene interconnects over copper interconnects, improvement in the critical path delay is plotted versus the number

of graphene layers in Figure 27. Three different MFP values, 0.1, 0.3, and 1.2 μm, are used in the evaluation, corresponding to three substrates, SiO_2, BN, and air, respectively [100, 101]. The building block includes carry propagation and generation circuits, as shown by the blue blocks in Figure 26. The block pitch is defined as the center distance between two basic circuit blocks in the adder. For long interconnects that connect two distant blocks, repeaters are inserted. The default configuration is shown in Table 4.

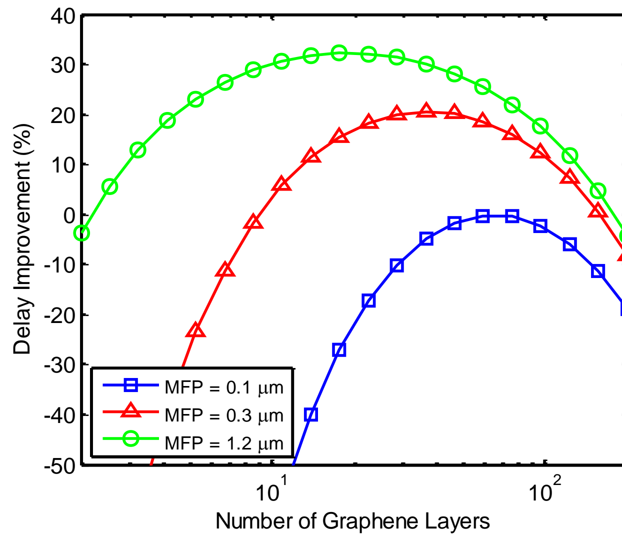


Figure 27: Delay improvements of multilayer graphene interconnects compared with copper interconnects versus the number of graphene layers at three MFP values.

Optimal numbers of graphene layers are observed to achieve the maximum delay improvement at various MFP values. This is because with small numbers of graphene layers, the large resistance of the graphene interconnect dominates the delay; increasing the number of graphene layers significantly reduces the interconnect resistance. However, if there are too many graphene layers, the large capacitance of the graphene overshadows the benefits of the resistance saving. Therefore, the improvement starts to decrease when the number of layers exceeds a certain value. Another observation is that the optimal number of graphene layers decreases as the MFP increases. The reason is that the interconnect resistance is smaller for a higher quality graphene, and the circuit

performance is more dominated by the capacitance. Further research is needed to enable a large number of graphene layers, especially for the SiO_2 case. The maximum delay improvement of suspended graphene is about 30% higher than that of graphene on BN.

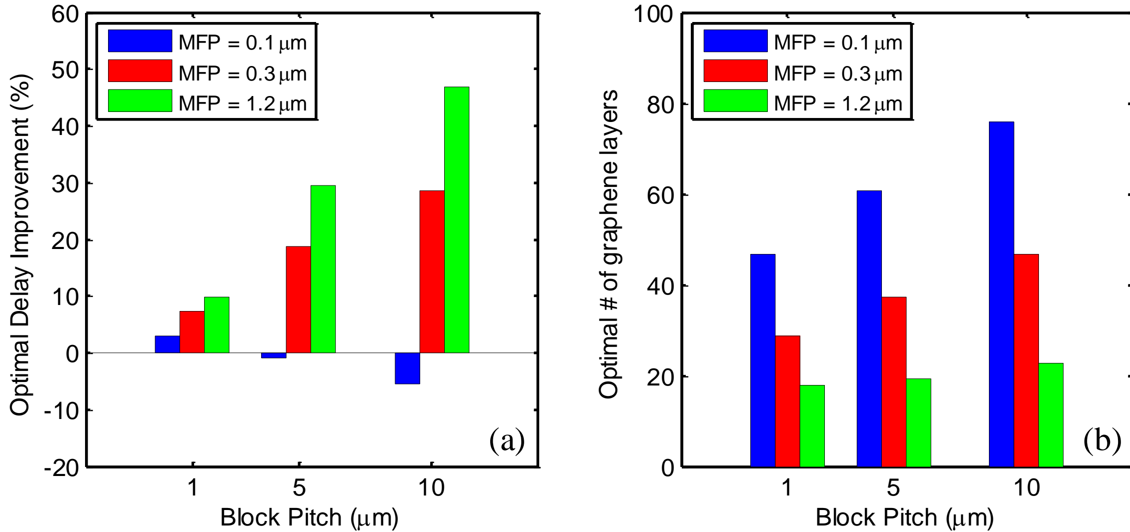


Figure 28: Optimal delay improvement and number of graphene layers versus interconnect length for various MFP values.

The optimal points in Figure 27 are shown in the middle columns in Figure 28, where the interconnect length is 5 μm. Since the block pitch varies upon different router and layout efficiency, two more block pitches are investigated in Figure 28. Depending on the MFP, the advantage of graphene interconnects varies even for the same increase in interconnect length. If the MFP is short (blue bars), the delay is limited by the large resistivity of the graphene. The delay improvement drops as the interconnect length increases. If the MFP is large (red and green bars), the graphene resistance is not as high, and the capacitance advantage of graphene interconnects increases with length, leading to an improved delay improvement. Since graphene provides a smaller capacitance than copper, the circuit benefits more if interconnects are longer. For all the MFP values, the

optimal numbers of graphene layers keep increasing as the interconnect length increases because of the larger interconnect resistance.

For a pessimistic contact resistance value of 1000 $\Omega{\cdot}\mu m$, the delay improvement drops by about 5% in general as shown in Figure 29 compared with Figure 28. Moreover, about 5 to 15 more layers are required to achieve the maximum delay improvement, especially for the circuits with short interconnects because the resistance of a short interconnect is dominated by the contact resistance.

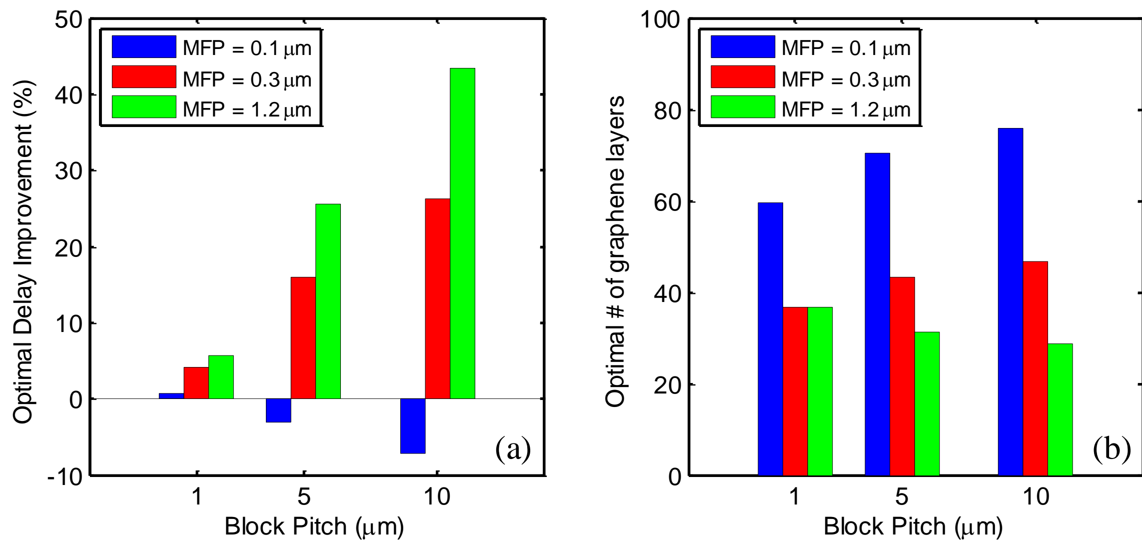


Figure 29: Optimal delay improvement and number of graphene layers versus interconnect length for various MFP values. Here, the contact resistance is 1000 $\Omega{\cdot}\mu m$.

The results presented so far are based on minimum width interconnects, the improvement in delay and the optimal number of graphene layers for a wider graphene are plotted in Figure 30. It can be seen that the graphene interconnects with a small MFP can clearly benefit from the larger width because the resistance of graphene interconnects dominates the delay. However, for the high quality graphene, the delay improvement drops about 5% ~ 10% because the resistance of copper interconnects decreases more significantly than that of graphene with wider interconnects. The optimal number of

graphene layers slightly decreases because the resistance of the interconnect is less dominant.

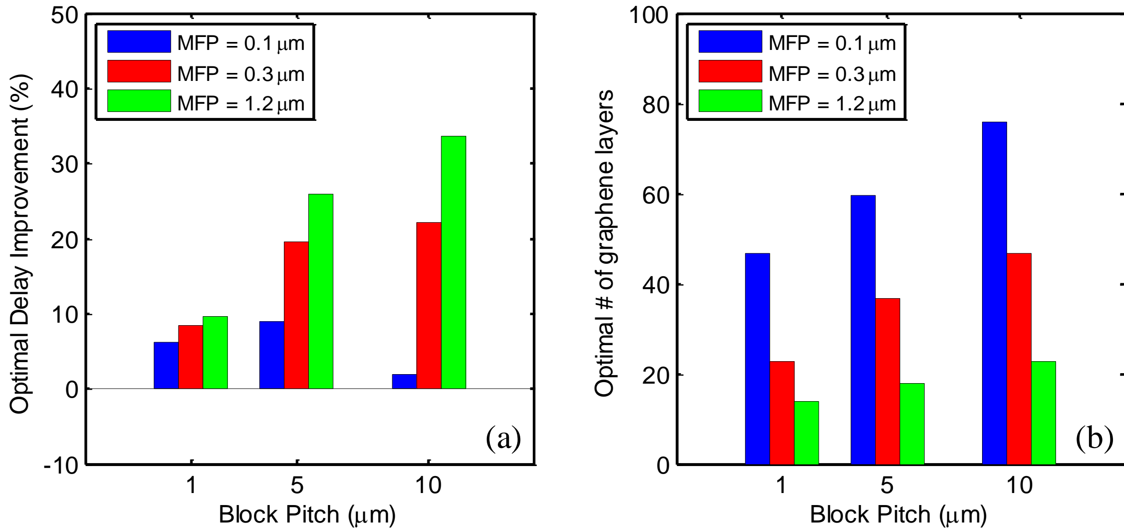


Figure 30: Optimal delay improvement and number of graphene layers versus interconnect length for various MFP values. Here, the width of the interconnect is 40nm.

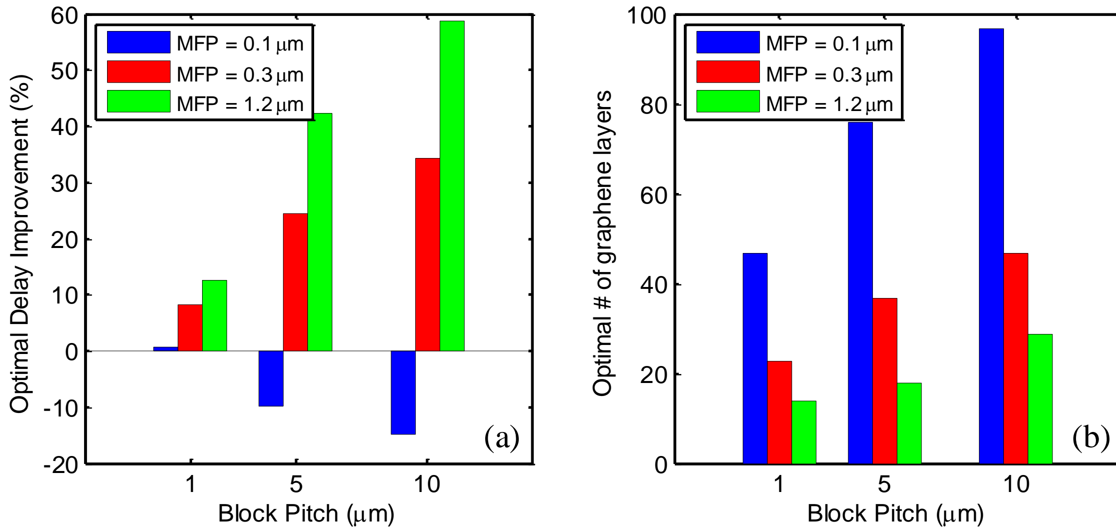


Figure 31: Optimal delay improvement and number of graphene layers versus interconnect length for various MFP values. Here, the width of the interconnect is 14nm; the technology node is 7nm.

For a smaller 7nm technology node, less improvement is observed for a short MFP value, shown in Figure 31, because the narrower interconnect makes the resistance of the graphene even more dominant. However, for a long MFP, as the technology scales, the input capacitance of devices is smaller and the driver resistance becomes larger, which makes the circuit less sensitive to the interconnect resistance and more sensitive to the interconnect capacitance. Therefore, the overall benefit of using graphene interconnect is more significant.

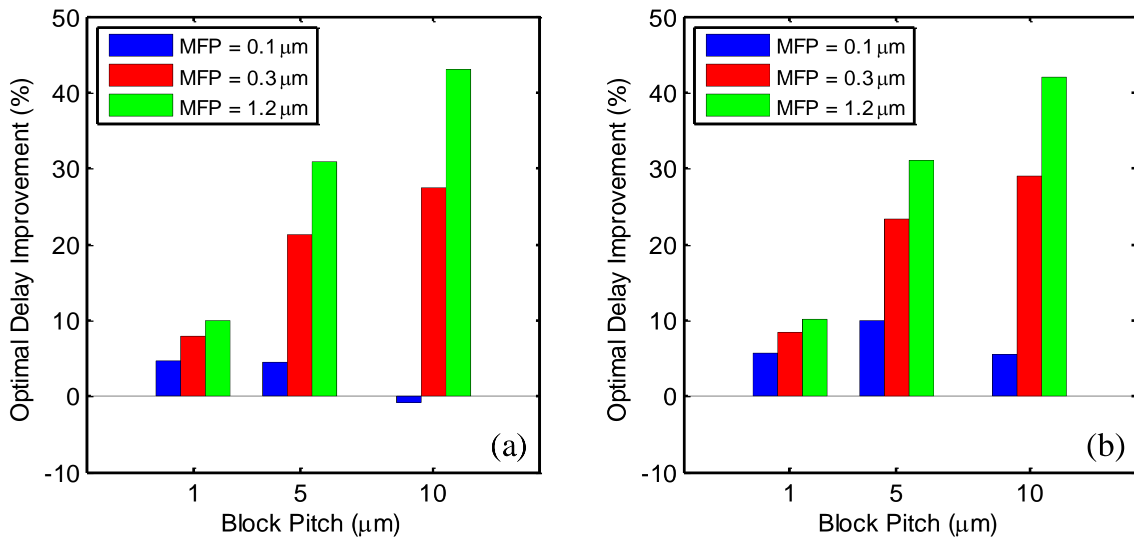


Figure 32: Optimal delay improvement versus interconnect length for various MFP values at a low supply voltage Vdd = 0.5V with (a) normal threshold voltage and (b) high threshold voltage devices.

If the supply voltage decreases from 0.7V to 0.5V, the delay improvement offered by graphene interconnects with short MFP increases. However, there is negligible impact on those with a large MFP as evident from Figure 32 (a) and Figure 28 (a). This is because lowering the supply voltage increases the output resistances of the devices and therefore makes interconnect resistance less important. Likewise, using high threshold voltage devices, shown in Figure 32 (b), offers an additional delay improvement for graphene interconnects with short MFP. In addition, up to 10 layers of graphene can be saved by

using a low supply voltage, shown in Figure 33, especially for short interconnect lengths, because the high driver resistance makes the interconnect resistance less dominant. These results indicate that low quality graphene interconnects are suitable only for the low-power application domains.

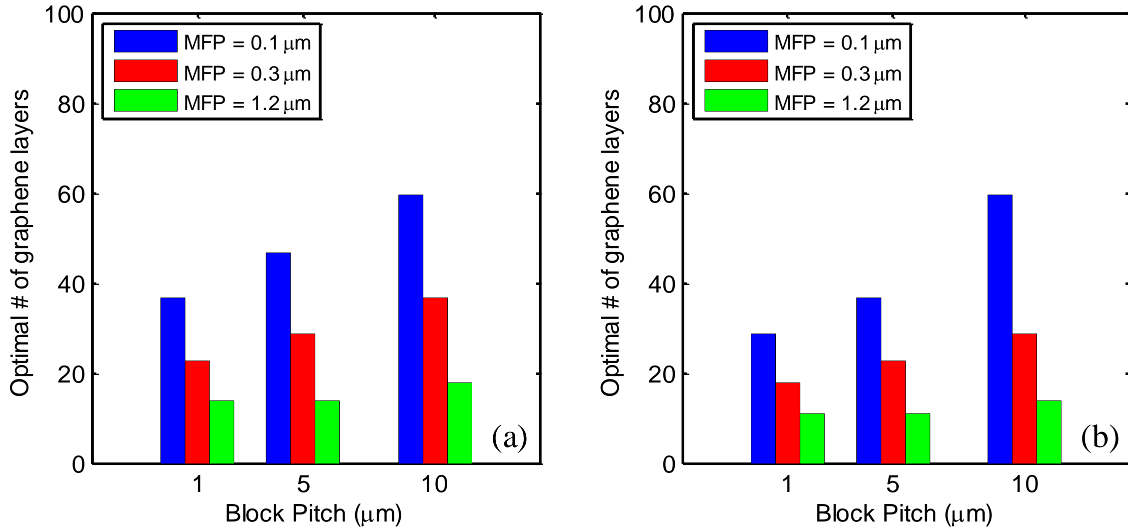


Figure 33: Optimal number of graphene layers versus interconnect length for various MFP values at a low supply voltage V_{dd} = 0.5V with (a) normal threshold voltage and (b) high threshold voltage devices.

As can be seen from previous results regarding the optimal number of graphene layers, up to 80 layers are needed to achieve the maximum delay improvement, especially for the long interconnect with a low MFP value. However, it may not be practical to have such large numbers of layers. For instance, in [32], the maximum number of graphene layers is 10. Therefore, the simulation is performed to evaluate the case when the number of graphene layers is set to be 10, shown in Figure 34. The results indicate that the MFP needs to be longer than 0.3 μm to have a gain over copper interconnects. For a low MFP graphene, the delay increases significantly for a long interconnect length due to the large resistivity.

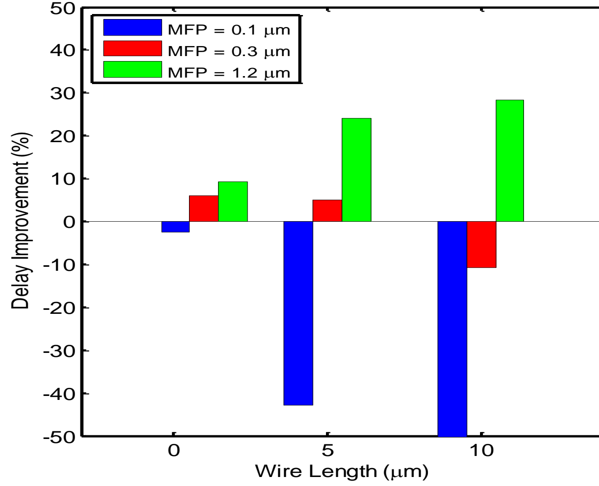


Figure 34: Optimal delay improvement versus interconnect length for various MFP values. Here, the number of graphene layers is fixed to be 10.

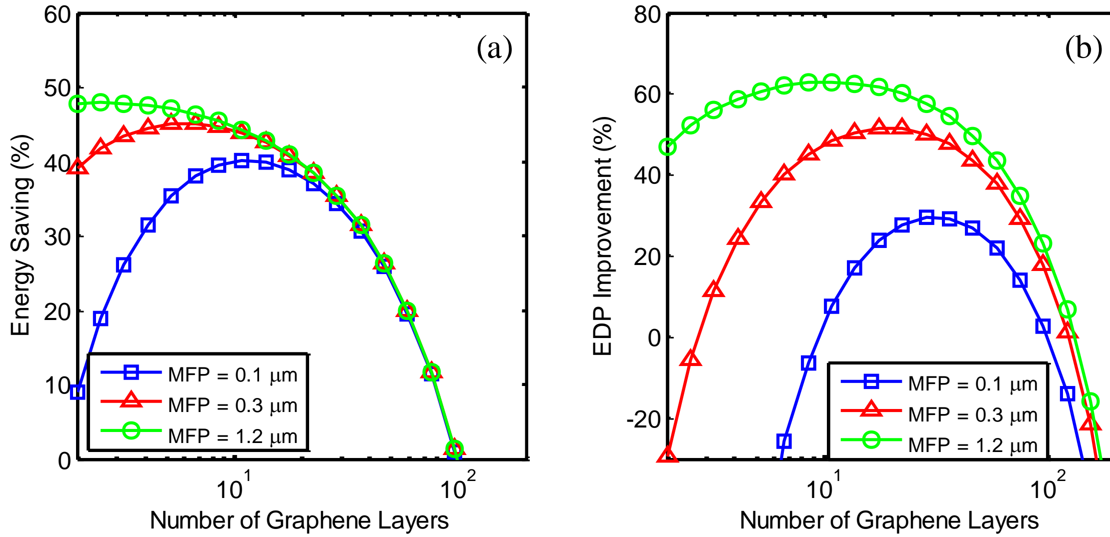


Figure 35: Energy consumption and EDP improvement versus the number of graphene layers.

The performance metric in the previous analyses is delay. The energy consumption is also another important metric that needs to be addressed. In Figure 35 (a), based on the default configuration in Table 4, for a high-quality graphene, the energy improvement keeps decreasing as the number of graphene layers increases due to a larger line-to-line

capacitance. However, if the MFP is small, the energy saving increases at the beginning as the number of graphene layers increases. This is because the delay of the adder using a small number of graphene layers is dominated by the large resistance of the graphene interconnect, which significantly increases the transition time and dynamic power consumption. Moreover, the constant leakage power also contributes to the energy consumption. Therefore, an optimal number of layers can be observed to achieve the maximum energy saving.

Figure 35 (b) shows the EDP improvement, focusing on the trade-off between the energy and the delay. Optimal numbers of graphene layers exist to achieve the maximum EDP improvement for graphene interconnects with various MFPs. Compared with Figure 27 (b), fewer graphene layers are required to achieve the optimal design point if the energy consumption is taken into account.

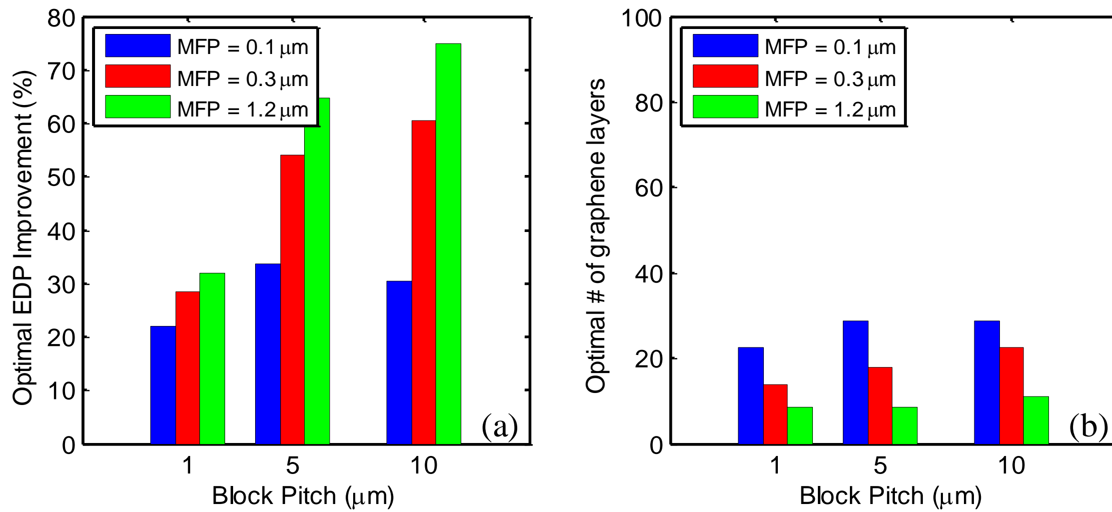


Figure 36: Optimal EDP improvement and number of graphene layers versus interconnect length for various MFP values.

For various interconnect lengths and MFP values, the optimal EDP and number of graphene layers are shown in Figure 36, where one can observe that even for short MFP graphene, more than 20% improvements in the EDP can be achieved. Moreover, the

optimal number of graphene layers required to minimize the EDP is much smaller than that to minimize the delay as shown in Figure 28 (b).

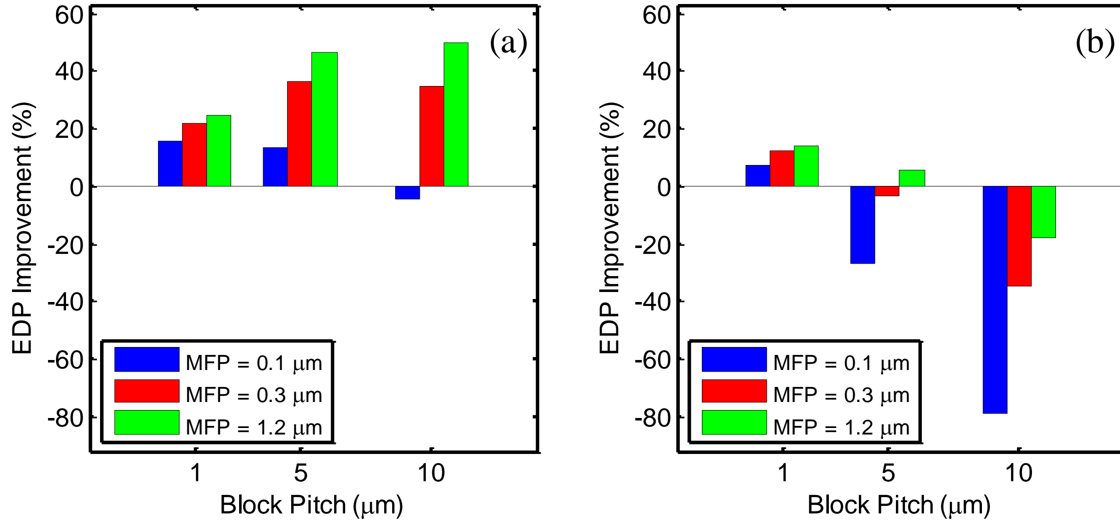


Figure 37: Optimal delay improvement versus interconnect length at various MFP values with edge roughness of (a) 0.2 and (b) 1.0.

All the results presented up to this point assume atomistically smooth edges and therefore no edge scatterings. If the edges are rough, the improvements in the EDP decrease depending on the probability of edge scatterings for electrons, p. From Figure 37, the circuits with short interconnects are not sensitive to the edge roughness because the interconnect resistance is dominated by the contact resistance and quantum resistance, which are independent of the interconnect length. However, for the circuits with long interconnects, the EDP drops significantly, indicating the importance of the smooth edges for long graphene interconnects.

3.3.3.2 64-by-64 and 256-by-256 bit SRAM Circuits

Static random-access memory (SRAM) is commonly used as the on-chip memory due to its fast access time and small latency. The specific design used is a 6-transistor SRAM, which is based on lithography assumptions and electrical considerations to be a

112 cell where it can give a large enough SNM [98, 102]. Two main interconnect components in the SRAM design are investigated, including the bit line and the word line. For other interconnects such as the ones in the internal decoder or the sense amplifier connections, they are relatively short and less critical compared to the bit and word lines.

Figure 38 shows the delay improvement of the graphene bit line in the SRAM. A contact resistance is assumed at each connection point between the access transistor and the bit line. The delay improvement is smaller for larger contact resistances. As the cell number increases, the maximum contact resistance allowed for graphene interconnects to outperform copper interconnects drops because of large aggregate values of lump resistance and capacitance.

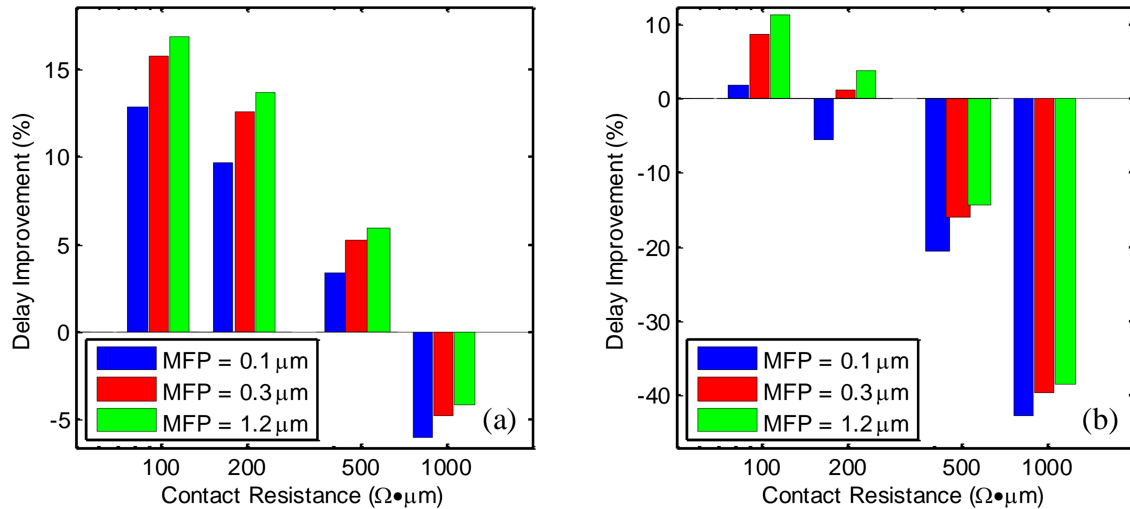


Figure 38: Delay improvement versus contact resistance of graphene at various MFP values. The total number of cells on the bit line is (a) 64 and (b) 256. The bit-line length and width between two nearby cells are 0.26 μm and 40 nm, respectively.

Compared with the bit line, the delay improvement for the word line is more sensitive to the contact resistance of the graphene interconnect as shown in Figure 39. This is because the interconnect segment between two cells for the word line is longer in

this SRAM design. For a 256-bit SRAM word line, the contact resistance needs to be less than 100 Ω·μm to have a delay gain over copper interconnects.

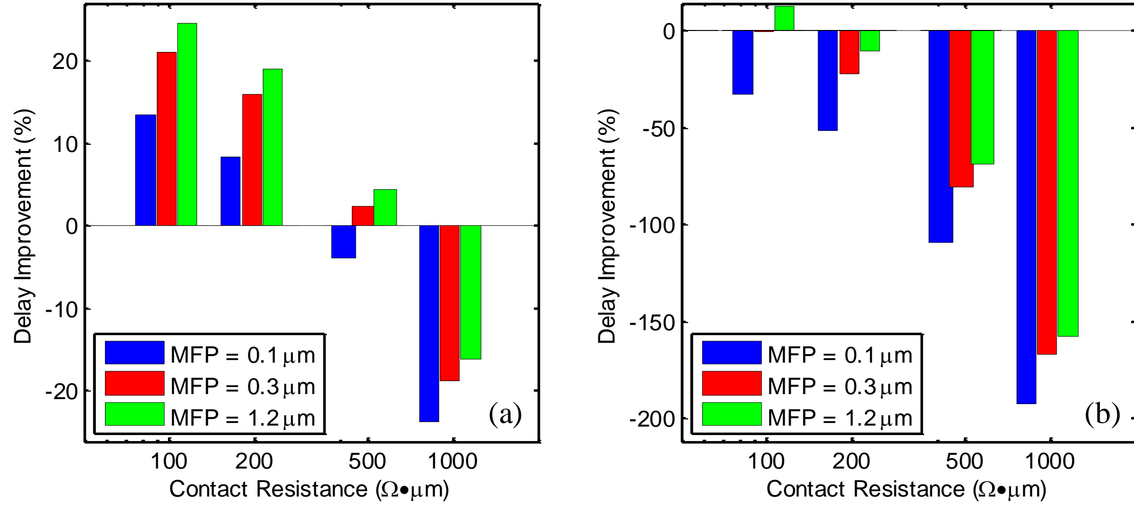


Figure 39: Delay improvement versus contact resistance of graphene at various MFP values. The total number of cells on the word line is (a) 64 and (b) 256. The word-line length and width between two nearby cells are 0.43 μm and 40 nm, respectively.

3.3.3.3 ARM Cortex-M0

In this subsection, the multilayer graphene interconnects are benchmarked against their copper counterparts for an ARM Cortex-M0 processor implemented with the IMEC N10 library [98]. It is the smallest cost-effective 32-bit ARM processor. The overall flowchart is shown in Figure 40. First, the ARM core is synthesized at various clock frequencies using Synopsys Design Compiler. The design that provides the minimum EDP is picked, which is 1.5ns. Second, the placement and routing is performed by using Cadence Encounter. The number of interconnection network levels are set to be four, where the width and spacing of the copper interconnects are 24nm with an aspect ratio of 2 for all levels. After the placement and routing, shown in Figure 41, the parasitics of copper interconnects are extracted by Cadence Encounter, including both resistance and capacitance of each copper interconnect segment at all nodes. Third, all the parasitics are

updated based on the graphene resistance and capacitance. The via resistance is modified to include the non-ideal contact resistance and the ideal quantum resistance of multilayer graphene interconnects, which are obtained based on the graphene interconnect model described in Chapter 3. The sheet resistance of the interconnects is also updated to take into account the resistivity of the graphene due to the finite MFP and edge roughness. The capacitance is modified based on the values from [103]. Finally, the critical path delay and the power consumption of the ARM core are estimated by using Synopsys PrimeTime.

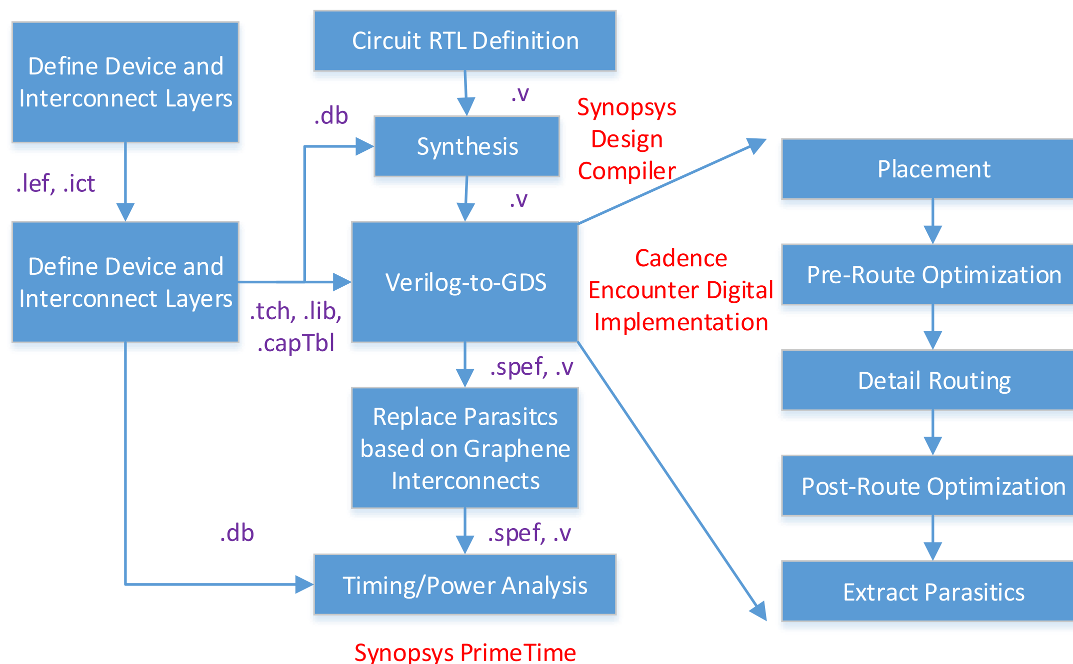


Figure 40: Flowchart of the multi-layer graphene interconnects benchmarking based on commercial tools.

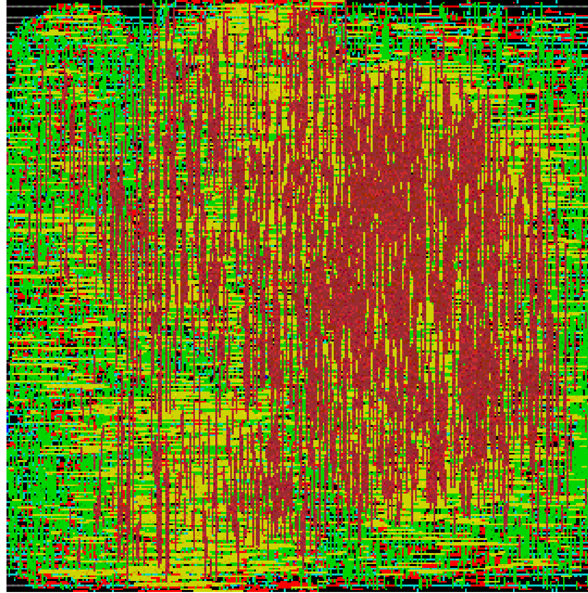


Figure 41: The layout of a placed and routed ARM Cortex-M0.

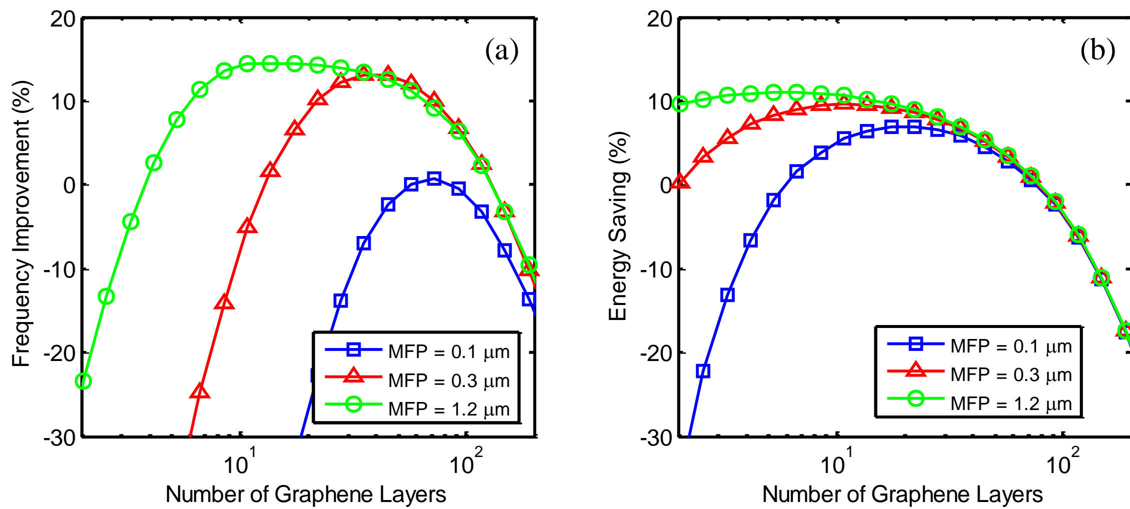


Figure 42. Delay improvement and relative energy consumption of an ARM core versus the number of graphene layers.

Figure 42 (a) shows the critical path delay of the ARM core versus the number of graphene layers. Once again, an optimal number of graphene layers exists to achieve the minimum delay, whose behavior and explanation have been illustrated for Figure 27 (a). In Figure 42 (b), optimal numbers of graphene layers exist that minimize the energy

consumption per cycle. The reason is that if the number of layers is too small, the clock frequency significantly drops due to the large sheet resistance and contact resistance of graphene. This causes a larger transition time and increases the power consumption of the logic cells. In addition, the constant leakage power also contributes to the energy consumption when the clock cycle is too long; if the number of graphene layers is too large, the interconnects become too thick, which again increases the critical path delay and the dynamic power consumption associated with the large line-to-line capacitance. Moreover, fewer numbers of graphene layers are needed to achieve the minimum points for high quality graphene interconnects with longer MFPs. Compared with the core using copper interconnects, up to 15% and 10% of the improvements in the delay and energy consumption are observed, respectively.

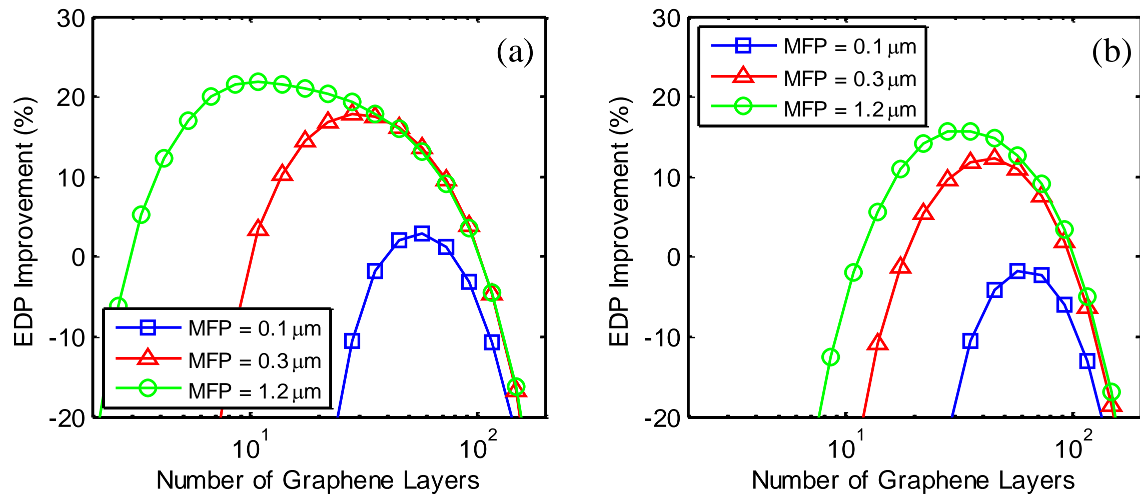


Figure 43. EDP improvement versus the number of graphene layers with a contact resistance of (a) 100 Ω·μm and (b) 1000 Ω·μm.

To balance the delay and energy consumption of the ARM core, the EDP improvement is investigated and plotted in Figure 43 for both aggressive and more realistic contact resistance values. The EDP improvement drops from 22% to 16% if a larger contact resistance is assumed. Furthermore, if the graphene edge is not smooth,

Figure 44 shows the EDP improvements for the graphene with edge roughnesses of 0.2 and 1, respectively. One can observe that a smooth edge is critical for the graphene interconnects to offer a gain over copper interconnects, especially for the graphene with small MFPs. Note that the size of the processor being analyzed in this chapter is the simplest ARM core where the length of the interconnect is relatively short. If a more advanced processor is used, even more benefits can be observed by using graphene interconnects.

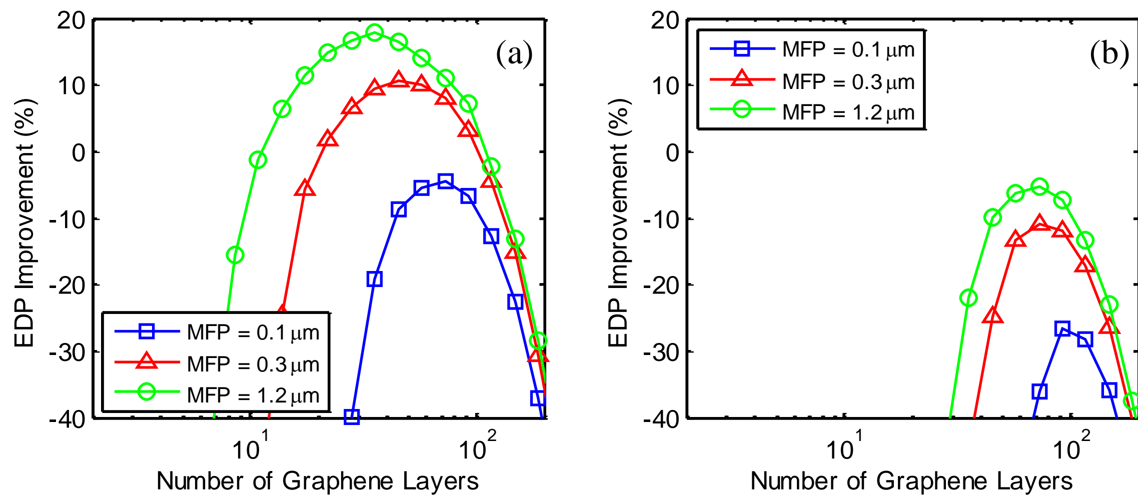


Figure 44. EDP improvement versus the number of graphene layers with edge roughness coefficient *p* of (a) 0.2 and (b) 1. Here, the contact resistance is assumed to be 100 $\Omega\cdot\mu m$.

3.4 Graphene PN Junction Logic

Based on the device-level GPNJ modeling developed in Chapter 2, the basic logic function can be achieved by using the combination of the p-type and n-type GPNJ switches as well as the complex reconfigurable GPNJ logic gate. In this subsection, a gate-level modeling is presented that takes into account various doping mechanisms and a realistic layout. In addition, a circuit-level simulation is performed for an SRAM cell based on GPNJ switches.

3.4.1 Basic Logic Function

A GPNJ inverter is composed of two GPNJ switches that share one contact in the middle of the inverter as the output, shown in Figure 45. Green and yellow gates are applied with positive and negative control voltages. The input is connected to three gray gates that can turn on either side of the GPNJ switch depending on the input. If the input is logic '1', the output will be connected with the left logic '0'; if the input is logic '0', the output will be connected with the right logic '1'.

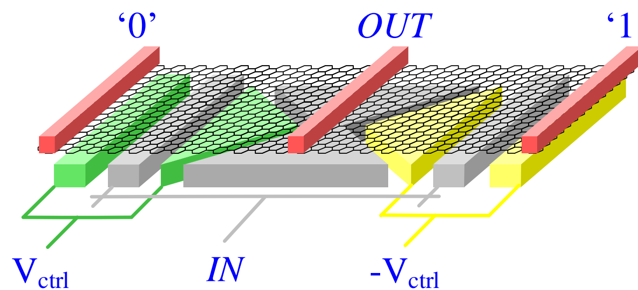


Figure 45: 3D plot of a GPNJ inverter.

In Figure 46, the footprint area model for a GPNJ inverter is depicted that takes into account two control mechanisms. The left figure shows the usage of chemical doping [104] to shift the Fermi energy level of the graphene, whereas the right one uses the electrostatic gate control. One can observe that by using the chemical doping method, the footprint area for a minimum size GPNJ inverter can be reduced by 33%. Therefore, in this work, the chemical doping is assumed for both circuit- and system-level analyses. To be compatible with the standard CMOS process, the design rules such as the minimum spacing between two nearby vias follow the work [50].

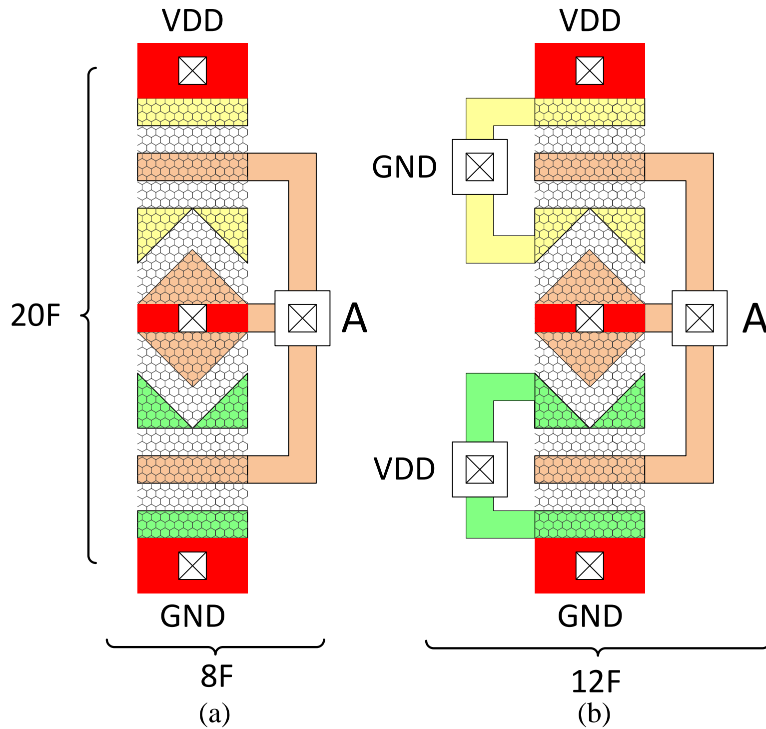


Figure 46: Layout of an GPNJ inverter (a) using chemical doping control (b) using electrostatic gate control.

The layout for a NAND2 gate is shown in Figure 47, where a 20% of the footprint area reduction can be achieved by using the chemical doping control method. The resistance and capacitance values for a GPNJ NAND2 gate can also be obtained based on the width of the graphene sheet in the pull-up and pull-down network. Here, the GPNJ devices with curved corner are considered to take into account the non-ideal lithography patterning, as shown in Figure 47 (c). The footprint area and the input capacitance are 25% and 50% larger, respectively. The impact of the penalties will be quantized in the later chapter.

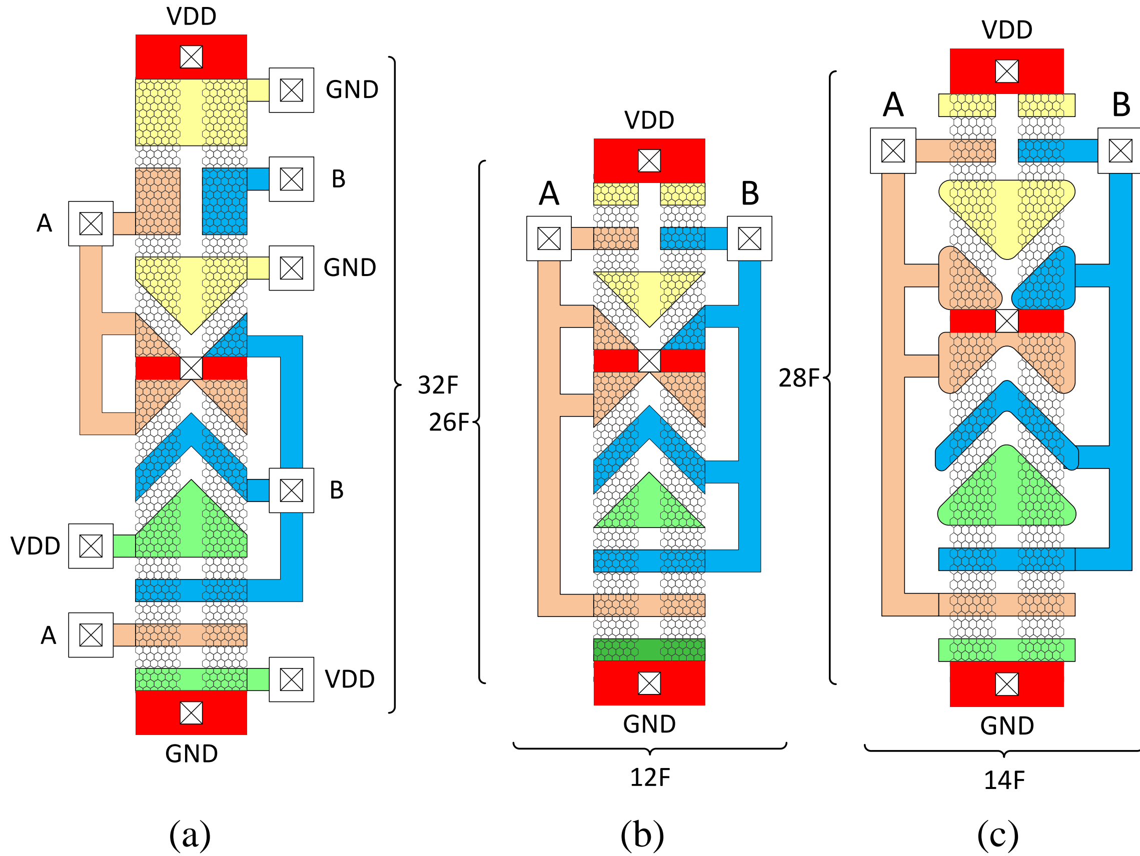


Figure 47: Layout of a GPNJ NAND2 gate (a) using electrostatic gate control, (b) using chemical doping control, and (c) using chemical doping control that takes into account the curved corner during lithography process.

3.4.2 Reconfigurable Logic Gate

To create a more complex logic function, the reconfigurable GPNJ-based logic gates is originally proposed in [62]. It takes advantage of the fact that the pull-up and pull-down network of a GPNJ-based logic gate is reconfigurable during the real-time operation. However, since the original modeling only considers the current flow of symmetric GPNJs at the equilibrium state, it causes driving current inefficiency issues for the realistic rail-to-rail operation. In this chapter, a more elaborate reconfigurable logic is developed, as shown in Figure 48. This compound logic gate can achieve a multiplexer and an XOR gate in a very efficient way compared with its CMOS counterpart.

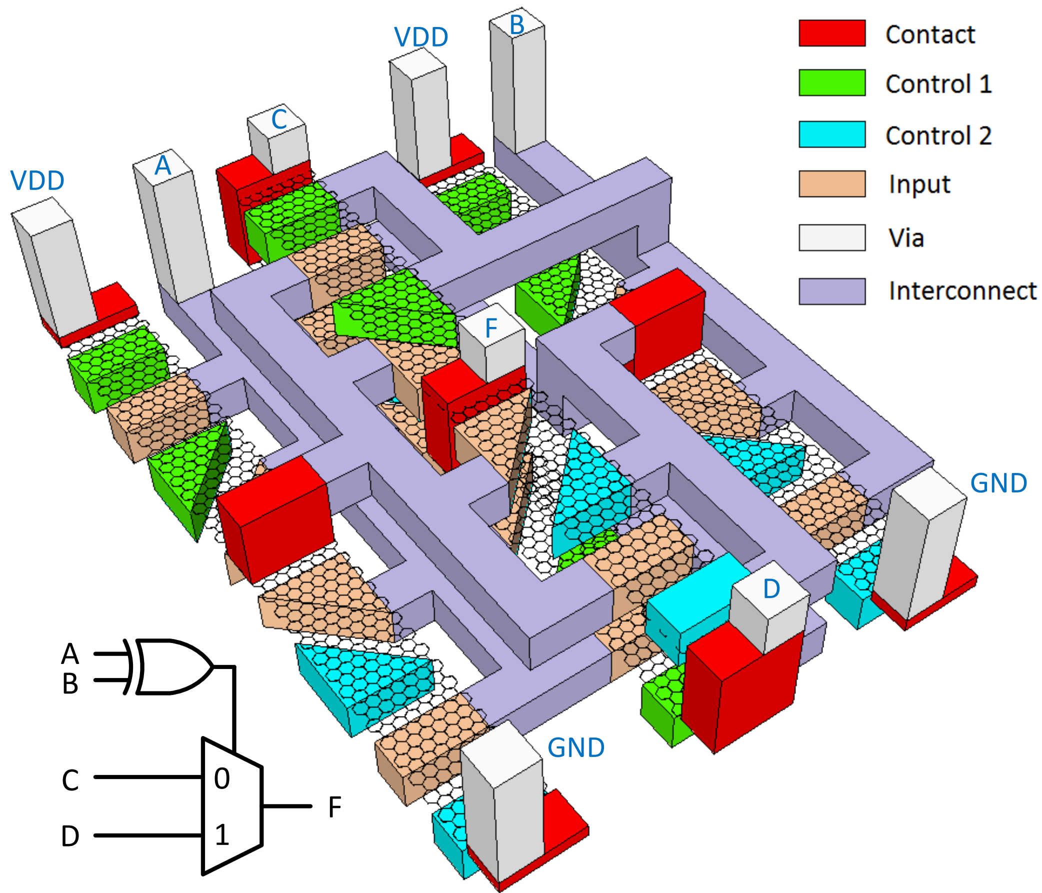


Figure 48: Reconfigurable logic gates based on GPNJ devices.

3.4.3 SRAM Cell

The most commonly used SRAM cell is made of six transistors, and the basic read and write operations are illustrated in Figure 49. The key performance metrics of an SRAM is the access time and read and write noise margins. The access time can be measured by the time needed to discharge the equivalent bit line capacitor to the amount of voltage difference that can be sensed by a sense amplifier at the end of the bit line. The read noise margin is measured by fixing bitlines to Vdd and sweeping the internal Q and $\bar{Q}$ from 0 to Vdd. The window between two curves after the sweep is the static noise

margin. The write noise margin is measured by increasing the BL from 0 to Vdd and finding the maximum value of the voltage that can flip the stored data.

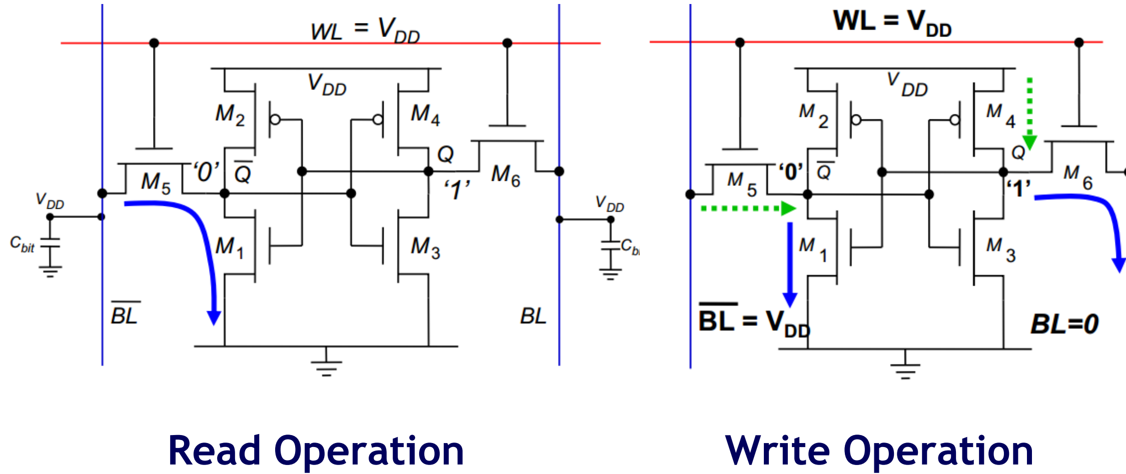


Figure 49: Read and write operation of a 6-transistor SRAM cell.

Figure 50 show the SPICE simulation results of the read and write margins and the access time for a GPNJ-based 6-transistor SRAM cell. The supply voltage of the SRAM is set at 0.5V using GPNJ devices at the 16nm technology node. A Verilog-a model is developed to perform the SPICE simulations, and it is based on the compact device-level model built in Chapter 2. Due to the supreme electrostatic of the 2D graphene material and the steep subthreshold slope, it provides a larger read current without compromising the read and write noise margins. The comparison of various performance metrics are shown in Table 5. For the CMOS SRAM, it is implemented with the 16nm planar CMOS technology with the supply voltage of 0.7V. The read and write noise margins of a GPNJ-based SRAM improves about 59% and 44%, respectively. The leakage power of the GPNJ-based SRAM cell is 57% less, and it provides 6.8X faster access time compared with its CMOS counterpart.

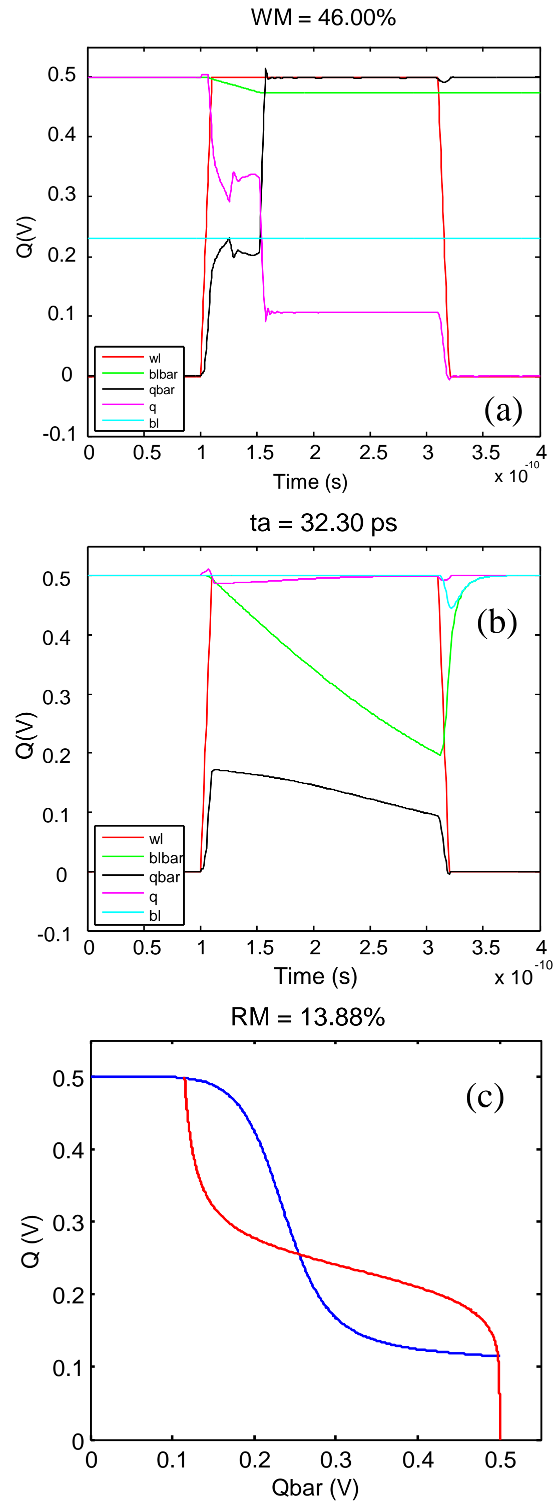


Figure 50: Various performance metrics of an SRAM using GPNJ devices. (a) write margin, (b) access time, and (c) read margin.

Table 5: Comparison between GPNJ- and CMOS-based SRAM Cell.

Parameters	**GPNJ**	**CMOS**
Write Margin relative to Vdd (%)	46	32
Access Time (ps)	32.3	218.34
Read Margin relative to Vdd (%)	8.74	13.88
Leakage Power (nW)	6.4	14.8

3.5 Gate-All-Around Nanowire FETs

In this section, VFETs and LFETs with various configurations are benchmarked and compared for a ring oscillator and an ARM core based on the device-level compact models described in Chapter 2.

3.5.1 Ring Oscillator Analysis

To investigate the impacts of the device input capacitance and the back-end-of-line (BEOL) resistance and capacitance, a 15-stage inverter-based ring oscillator with a fan-out of 3 and a wire load of 300 contacted gate pitch (CGP) is used as a circuit representative to perform the circuit-level performance analysis and comparison between LFETs and VFETs. The models for the resistance and capacitance of the interconnect follow previous work [105]. The copper grain boundary reflectivity co-efficient R, surface specularity parameter p, effective dielectric contact, and barrier thickness are assumed to be 0.15, 0, 2.3, and 1nm, respectively.

Figure 51 shows the comparison of the leakage currents versus the frequency. Three Vth flavors and 7 device configurations are investigated at the nominal supply voltage of 0.6V. LFETs in general provide a higher frequency because of a larger intrinsic device ON current. The performance gap between LFETs and VFETs, however, shrinks at the high-Vth flavor because LFETs suffer more from the SCE.

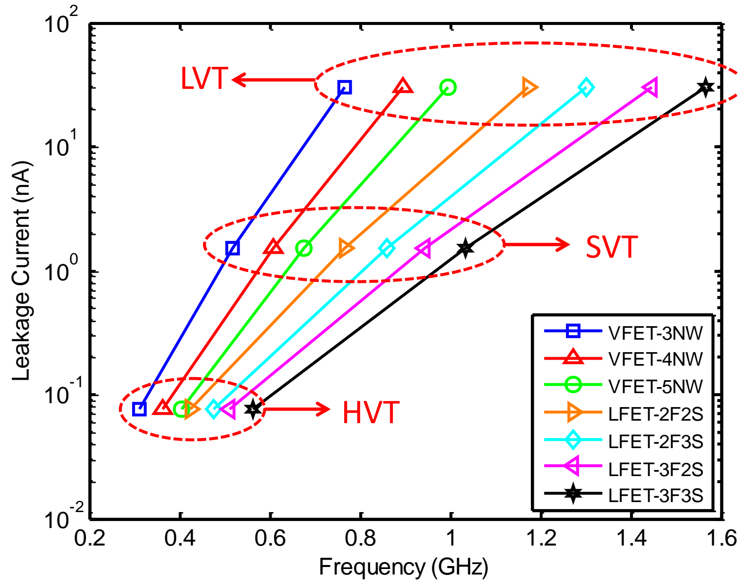


Figure 51: Leakage current versus the operational frequency of a 15-stage ring oscillator using LFETs and VFETs with 7 options at the supply voltage of 0.6V.

To compare the energy dissipation, Figure 52 shows the energy per switch versus the frequency at two supply voltages. One observation is that the energy per switch slightly increases as the device threshold voltage increases. The additional energy mainly comes from the larger short circuit current for the low-Vth devices during the switching even though the switching time is shorter than that of high-Vth devices. Compared with the ring oscillator using LFETs with 2fin/3stack, the one using 3-fin and 2-stack LFETs consumes about the same energy dissipation because the parasitic capacitances are close. Since the 3-fin and 2-stack LFET has a smaller parasitic resistance, it provides a higher frequency.

One advantage of VFETs is the small parasitic capacitance, which offers the ring oscillator using VFETs with 3 nanowires 15-35% energy saving compared with the LFET counterparts. If the supply voltage scales to 0.4V, both the frequency and the switching energy drop for VFETs and LFETs. In addition, Figure 52 shows that the frequency of a ring oscillator made of VFETs with the standard-Vth flavor at supply voltage of 0.4V is

comparable with the one using LFETs. For the high-Vth case, VFETs offer even larger frequency than LFETs due to the better SCE control, indicating that VFETs are more suitable for the low-power application with an ultra-low supply voltage.

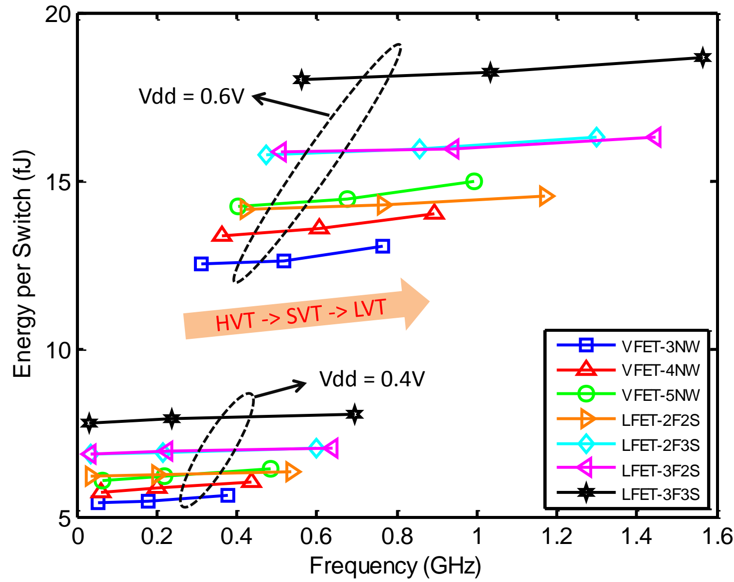


Figure 52: Energy per switch versus the operational frequency of a 15-stage ring oscillator using LFETs and VFETs with 7 options at the supply voltage of 0.4 and 0.6V.

The device- and circuit-level simulations show that both VFETs and LFETs have their own favorable operation conditions, such as the threshold voltage and supply voltage. Even within the same device structure, there are trade-offs among the energy, leakage, and frequency. One can either use a low supply voltage with high-Vth devices to achieve a small energy and leakage, or use a large supply voltage with low-Vth devices to boost the frequency. Therefore, to fully understand the ultimate potential performances and to obtain a fair comparison between these two device structures with various configurations, it is crucial to analyze and benchmark at the higher system-level dynamic energy/area/leakage at a given frequency target, where multiple-Vth devices are used simultaneously.

3.5.2 ARM Core Analysis

A standard cell characterization tool, Cadence Liberate, is used for the library generation [106]. For each device configuration and supply voltage, three libraries are created for the low-, standard-, and high-Vth devices, respectively. A total number of 50 standard cells, including two driver strengths for logic cells and four driver strengths for buffer cells, are characterized. Cell areas for various configurations of LFETs are updated based on the scaled FinFET-based N7 cell library [88], where all fins are terminated with a dummy gate to achieve the isolation between two nearby cells. Within each cell, an additional dummy gate is inserted for every two active gates to avoid short circuits between any two nearby gate contacts. For VFETs, dummy gates are not necessary thanks to the fully source and drain isolation. Based on real layouts of some representative cells, including the register, buffer, and combinational logic cells, the result shows that the empirical average widths of VFET cells are about half of their LFET counterparts [88]. Thus, the cell area of 4-nanowire VFETs with a 12-track layout is equivalent to the one of 2-fin LFETs with a 6-track layout. Since there are no actual physical layouts designed for the 5nm technology node, within cell parasitics are not accounted for. For an ARM core using 2-fin LFETs, the wire load model uses the scaled average wire length as the FO3 net length, which is extracted from the same ARM core at the 7nm node after performing the placement and routing [97]. For other device configurations, the average wire lengths are assumed to be proportional to the square root of cell areas. The interconnect resistance and capacitance models are the same as those described in Section 2. Once all libraries are generated, an ARM Cortex-M0 processor is synthesized using the Synopsys Design Compiler based on the multi-Vth flow with the leakage optimization enabled. Finally, power, timing, and area analyses are performed for each device configuration.

3.5.2.1 Frequency Comparison

Figure 53 shows the actual clock frequency of an ARM core at each frequency target after the synthesis. Seven configurations of LFETs and VFETs are optimized at the nominal supply voltage of 0.6V. For each device configuration, the operating frequency saturates when the timing target is beyond a certain point. The maximum frequency increases as the numbers of fins, nanowires, and nanowire stacks increase. In general, the LFET cores provide higher frequencies than VFET cores due to the larger driving capability. One can also observe that the VFET core with 5 nanowires has almost the same maximum frequency as the one using 4 nanowires. This is because the area overhead of 14-track layout causes a larger wire load, which counteracts the larger driving current of the VFETs with 5 nanowires.

From the comparison of maximum clock frequencies among seven device configurations in Figure 54 (a), the frequency advantage of LFET cores decreases as the supply voltage scales. If one only targets for the performance without considering the energy dissipation, LFETs with a larger number of fins and nanowire stacks are better choices. Figure 54 (b) shows the percentage of the interconnect delay in the overall critical path delay. The interconnect delay is obtained by subtracting the critical path delay without adding the wire load from the total delay with the wire load. One can observe that the interconnect has a larger delay impact on a VFET core, because the rise and fall transition times are more affected by the wire load capacitance due to a smaller input capacitance of VFETs. Compared with an LFET core with 2fin/3stack, the one with 3fin/2stack suffers more from the interconnect due to a longer wire length from a larger footprint area. As the supply voltage increases, the percentage contribution of the interconnect delay decreases. This is mainly caused by the fact that more buffers, inverters, and combinational logic cells are inserted, leading to a more device capacitance dominated situation.

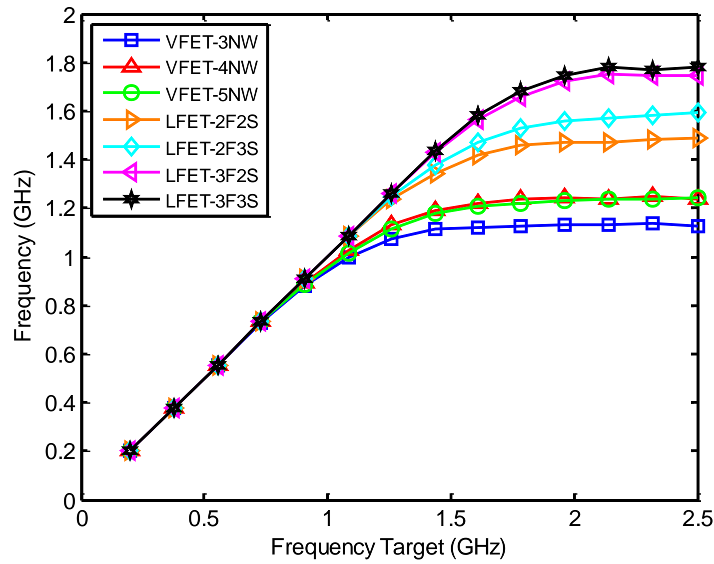


Figure 53: Actual frequency versus the target frequency of a synthesized ARM core using seven configurations of LFETs and VFETs at Vdd = 0.6V.

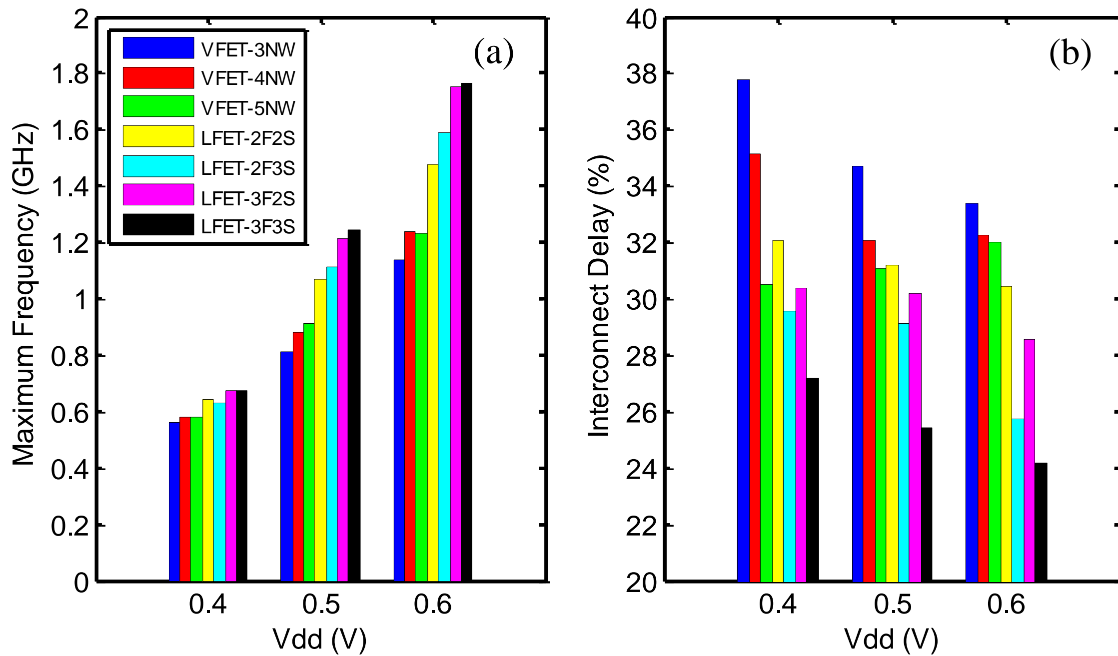


Figure 54: Comparison of the (a) maximum clock frequency and (b) percentage of interconnect delay in the critical path at three supply voltages for an ARM core using seven configurations of LFETs and VFETs.

3.5.2.2 Energy-Frequency Trade-offs

Since the energy and power efficiencies become one of the major concerns in the modern microprocessor design, Figure 55 addresses the energy dissipation per clock cycle for the ARM core using various device configurations. One can observe that the energy per clock cycle rises as the frequency increases, especially when it is approaching the saturation point. This is because the optimizer tries to insert more buffers and increases the driver strengths to push the frequency to its limit. For better energy efficiency, the bottom right of the figure is preferable.

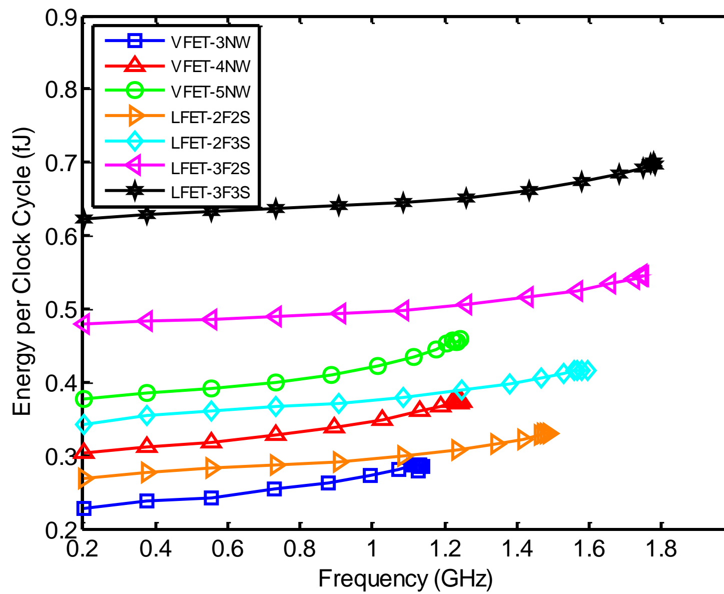


Figure 55: Energy per clock cycle versus the clock frequency of a synthesized ARM core using seven configurations of LFETs and VFETs at Vdd = 0.6V.

To account for both the energy and clock frequency, the saturation points at the tip of the curves in Figure 55 are put together and shown in Figure 56 (a), and two more supply voltages of 0.4V and 0.5V are included. Since the bottom right of the figure is a better design corner, LFET cores in general provide better performance-energy trade-offs. However, at the low frequency range, such as 0.5GHz, the VFET core with 3 nanowires

offer about 15% energy saving compared with the LFET core with 2fin/2stack at the supply voltage of 0.4V.

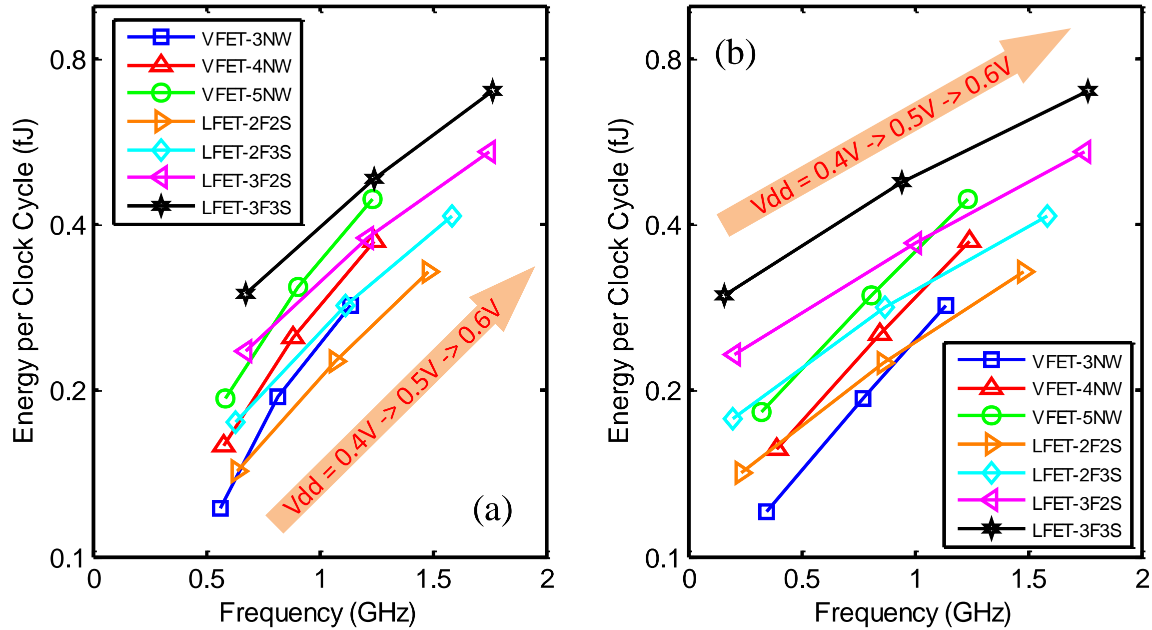


Figure 56: Energy dissipation per clock cycle versus the frequency at three supply voltages for an ARM core (a) optimized at each supply voltage and (b) optimized only at Vdd = 0.6V using seven configurations of LFETs and VFETs.

Analyses are also performed for an ARM core that is synthesized at a nominal supply of 0.6V, and the supply voltage is dynamically scaled down during runtime without re-optimizing the system. The updated energies and frequencies at 0.4V and 0.5V are shown in Figure 56 (b). The clock frequencies for all device configurations shift to the left compared with Figure 56 (a), especially for LFETs, whose ON current downgrades more significantly as shown in Figure 18. At the supply voltage of 0.4V, the VFET core with 3 nanowires offers about 25% improvement in both energy and clock frequency compared with an LFET core with 2 fin/2stack. Figure 57 shows the comparison of cell counts of ARM cores optimized at two supply voltages using three

Vth devices. This clearly illustrates that the best choice strongly depends on the performance and energy target and the application context for the device.

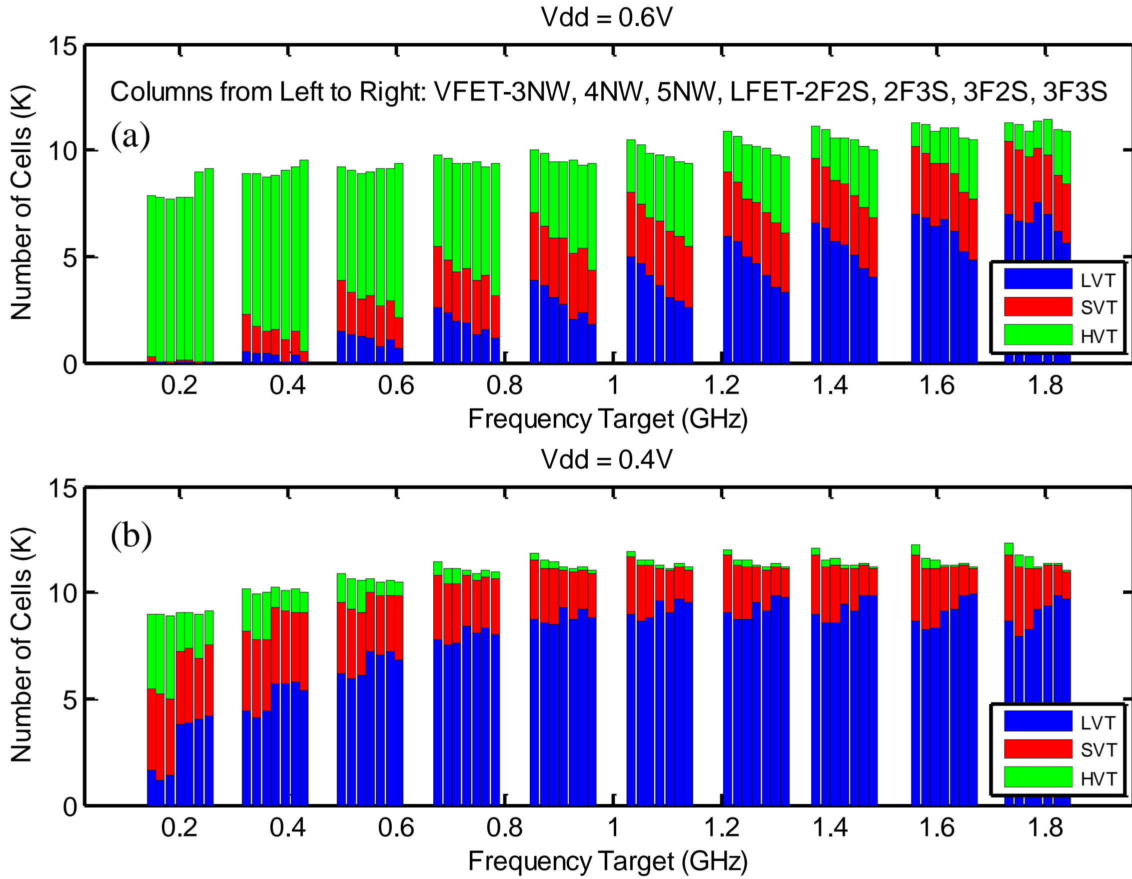


Figure 57: Cell counts of different Vth cell usage versus the frequency target of an ARM core using core using seven configurations of LFETs and VFETs at supply voltage of (a) 0.6V and (b) 0.4V. LVT, SVT, and HVT represent low-Vth, standard-Vth, and high-Vth devices.

At a fixed supply voltage, the total number of cells increases as the frequency target increases, because more inverter, buffer, and logic cells are inserted to achieve a smaller timing target. Meanwhile, more low-Vth and fewer high-Vth cells are used because a larger driving current is required to charge or discharge the load from the input gate capacitance and interconnect capacitance. Compared with the core optimized at 0.4V, the one at the nominal voltage uses more high-Vth cells and fewer low-Vth cells. This is

because the optimization tool is set to minimize the leakage if the given frequency target is reached. Since the ON current rises as the supply voltage increases, more high-Vth cells can be used to suppress the leakage without violating the timing constraint. Comparing the cell usage for different Vth devices of LFETs and VFETs, one observation is that the optimized VFET cores at 0.4V prefers to use fewer low-Vth and more high-Vth cells than their LFET counterparts. The reason is that the high-Vth VFET at a low supply voltage has the advantage of ON current from a better electrostatic gate control as illustrated in Section II. Another observation is that for both supply voltages, the total number of cells in VFET cores is larger than those in LFET cores at a higher frequency target. These additional cell usages mainly come from the buffer and inverter cells, which are required in VFET cores to reach the same timing target due to lower driving capabilities of logic cells.

3.5.2.3 Leakage/Dynamic Energy/Area Trade-offs

In this subsection, the trade-offs among the leakage and dynamic energy as well as the overall area of the ARM core are investigated for given performance targets. Figure 58 illustrates the dynamic versus leakage energy dissipations per clock cycle at two timing targets, and the bottom left of the figure is the preferred corner. As the supply voltage decreases, the data points for each device configuration shift to the bottom right. This is because more low-Vth and fewer high-Vth devices are used to reach the timing as shown in Figure 57. Note that the leakage current is insensitive to the supply voltage as shown in Figure 51, but using a low-Vth cell instead of a high-Vth cell increases the leakage current dramatically. Therefore, for a given performance target, an ARM core designed at a higher supply voltage saves the leakage energy dissipation at the cost of higher dynamic energy consumption. At slow timing targets (e.g. at 5ns), a VFET core with 3 nanowires provides the best dynamic and leakage energy trade-offs. However, as

the core frequency increases, an LFET core with 2fin/2stack surpasses the VFET core due to the better driving capability, allowing the LFET core to use more high-Vth cells and save leakage energy dissipation.

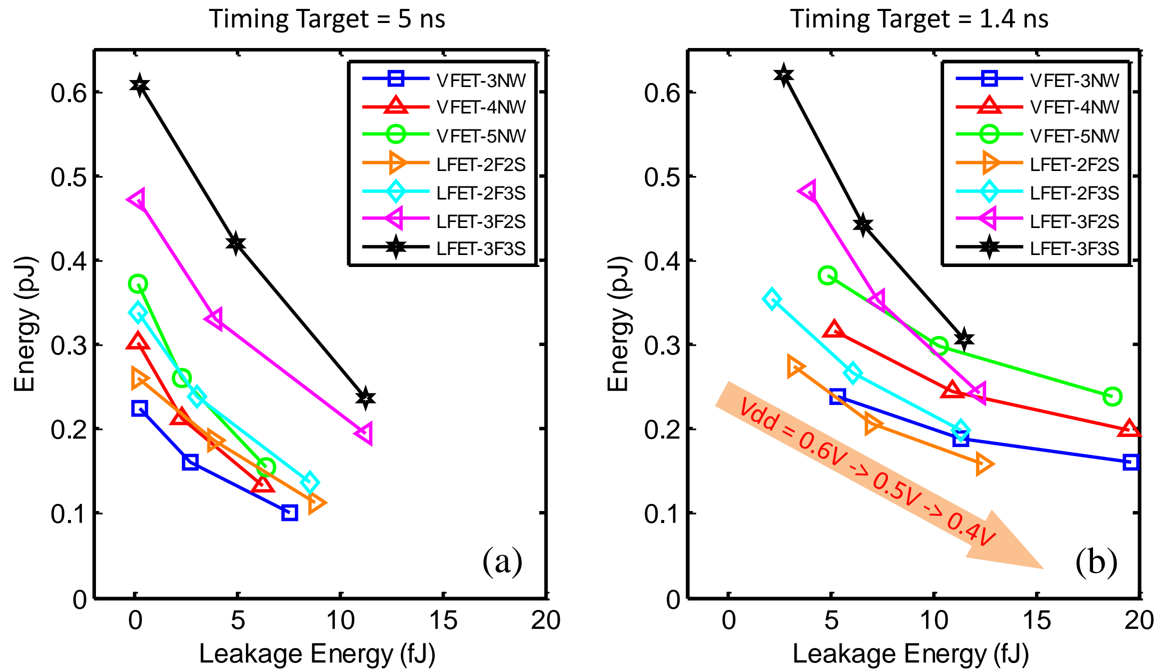


Figure 58: Dynamic energy dissipation versus the leakage energy dissipation for an ARM core using seven configurations of LFETs and VFETs at timing target of (a) 5 ns and (b) 1.4 ns.

One last important metric that needs to be addressed is the core area, and Figure 59 shows the results for seven configurations of LFETs and VFETs at two different clock period targets. In general, the total core area increases as the supply voltage decreases, especially at a shorter clock period. This is because more inverter, buffer, and combinational cells are inserted to meet the timing requirement. Due to the low current driving capability, VFET cores require more buffers to meet a slow timing target, and consequently cause a larger area increase than LFET cores. However, this buffer overhead disappears for slow timing targets. For example at 1.4ns, the VFET cores with 3

nanowires offer 7% and 25% area saving compared with LFET cores with 2- and 3-fin devices, respectively.

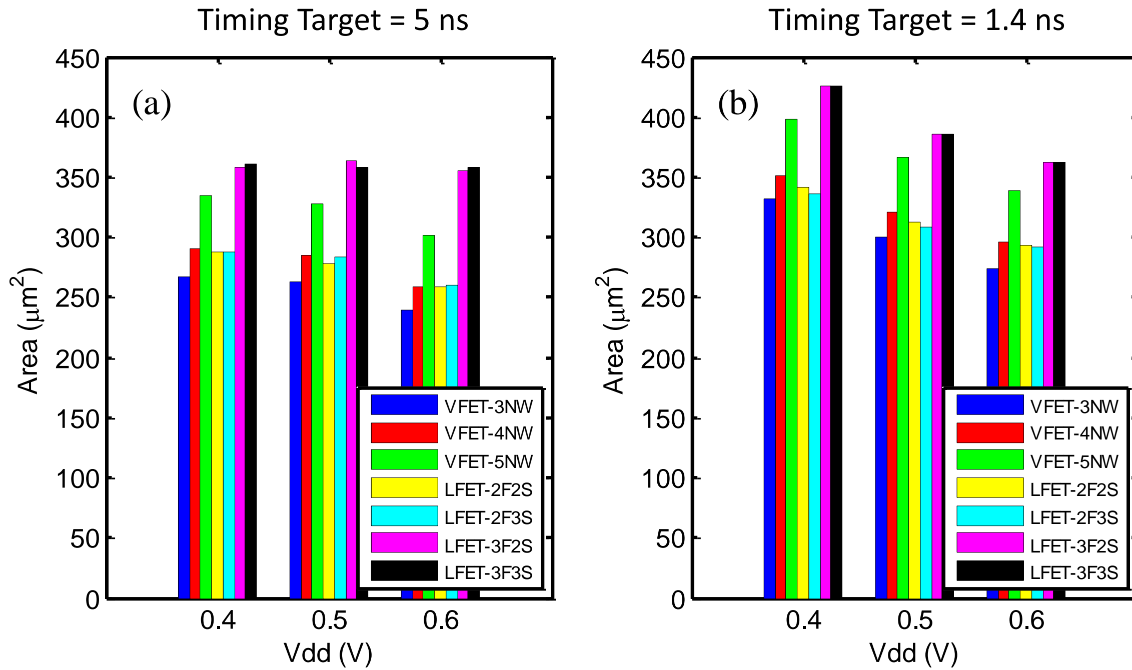


Figure 59: Total core area versus supply voltage of an ARM core using seven configurations of LFETs and VFETs at the timing target of (a) 2.2 ns and (b) 0.8 ns.

3.6 Conclusions

This chapter illustrates and develops the detailed modeling approaches, assumptions, and simulations at the circuit level.

The resistance and capacitance models of multi-layer graphene interconnects are developed based on quantum transport theories. It is shown that the major benefit of using graphene interconnects comes from the capacitance saving. In addition, the resistance and capacitance per unit length of the conventional copper interconnects are presented both with and without process variations.

To quantify the performance benefits, multilayer graphene interconnects are benchmarked against their copper counterparts for a 32-bit adder, an SRAM, and an

ARM core processor. Improvements are observed in terms of delay and EDP. Circuit-level analyses on the adder and SRAM show that these improvements are largely influenced by the contact resistance, MFP, and edge roughness. The major findings are: 1) An optimal number of graphene layers exists to achieve the best performance. To minimize the EDP, 5~50 fewer layers are required than those needed to minimize the delay. A wider graphene needs fewer layers, especially at a block pitch longer than 5 μm and a MFP of 100 nm. 2) As the contact resistance degrades from 100 $\Omega\cdot\mu m$ to 1000 $\Omega\cdot\mu m$, the delay improvement drops by around 5%, and the optimal number of graphene layers increases by about 5-15. 3) In the SRAM application, a low contact resistance is required to ensure gains for both bit and word line delays. 4) At a MFP longer than 300 nm, an additional 10% of the delay improvement is observed for the 7 nm technology node. At a short MFP of 100 nm, graphene offers a negligible improvement in the delay, but it offers 25% saving in the EDP. Furthermore, as the supply voltage decreases to 0.5V, up to 15% of the additional improvement is observed in delay, suggesting a potential applicability of the low-quality graphene in low-power applications. 5) Depending on the MFP, edge scatterings at the rough edges may undermine the delay improvement by 5%~20%. The results of an ARM core processor follow the aforementioned behaviors, where up to 15% and 22% of improvements in clock frequency and EDP, respectively, are observed.

The GPNJ-based logic gates are also presented, including a modified reconfigurable logic function to efficiently achieve both multiplexer and XOR functions. A 6-transistor SRAM implemented by GPNJ devices is also investigated, and the results indicate that the GPNJ-based SRAM provides 59% and 44% better read and write noise margins, respectively, and 57% less leakage power dissipation with a 6.8X faster access time compared with its CMOS counterpart.

At the end, various configurations of LFETs and VFETs are benchmarked and compared for a ring oscillator and an ARM core. LFETs in general provide larger ON currents than VFETs as a result of the channel stress. At a low supply voltage with a high-Vth flavor, however, the ON current for LFETs downgrades more significantly because of the poor SCE. The energy advantage is observed for VFETs due to the small parasitic capacitance. Compared with a 2fin/3stack LFET, the one with 3fin/2stack has a smaller parasitic contact resistance, providing a ring oscillator with a comparable energy and a higher frequency at the cost of 25% larger footprint area. For the ARM core analysis, more low-Vth and fewer high-Vth cells are used at a high frequency target, especially for devices with a low driving capability and operating at a low supply voltage. LFETs are preferable for the high-performance applications because of their larger ON currents. At slow timing targets (e.g. 5ns), VFET cores with 3 nanowires provide a 7% improvement in area and a 20% more in energy compared with LFET cores with 2fin/2stack at the same leakage power.

CHAPTER 4 SYSTEM-LEVEL MODELING AND DESIGN METHODOLOGY INTEGRATION

4.1 Introduction

This chapter describes six compact system-level models for the proposed optimization engine. To accurately obtain the architecture-level information, such as the CPI of the logic core and the cache miss rate, a cycle-accurate simulator is commonly used [107, 108]. However, it usually takes a few hours to generate all the detailed information for each benchmarking program. For a multi-parameter optimization problem, it is not feasible to perform such analyses. To improve the run-time efficiency, an empirical CPI model is developed to fast estimate the CPI of a general microprocessor without the detailed design of the whole processor. Previous work on the empirical CPI model can be traced back to Eble's work in the late 1990s [109]. However, his work was based on old Intel processors. Since then, processors have gone through major changes, and it is essential to revisit this model. In addition, the hierarchical memory model in his work is only applicable to single-core processors. A more detailed validated throughput model that accounts for finite memory bandwidth is required for multi-core processors.

The interconnection network has a large impact on the total power dissipation as well as the maximum operation frequency of a microprocessor. To accurately capture the power and delay associated with the interconnects, placement and routing algorithms are commonly applied for every VLSI design. However, at the early design stage, for a large system that constitutes hundreds of millions of random logic gates, it is extremely time-consuming for each evaluation at each clock frequency. It makes it even harder if one

wants to explore the combinations of various scenarios, such as the interconnect dimensions and material properties. Therefore, to efficiently estimate the interconnection network of a general processor, a rigorous derivation of a complete stochastic wire-length distribution for on-chip random logic networks was developed in 1990s [110]. It was based on Rent's Rule, a well-established empirical relationship between the number of module I/O terminals and the number of gates per module. Later, a modified wire length distribution model was developed to take into account the random arrangement of logic gates in a circuit block. This model provided a better match with the actual commercial products. Based on the new wire length distribution model, multi-level interconnection network design algorithms are developed and validated to fast estimate the total power consumption based on a certain clock frequency. In this dissertation, this interconnection network model is modified and adopted into the hierarchical optimization engine to provide reasonably accurate interconnect network design and estimation.

With integration of billions of transistors in a single chip and increased power consumption and density, the power delivery integrity, thermal, and process variations have become critical concerns in the design of high-performance microprocessors [111-113]. Previous work [114] has developed a compact and accurate physical models for the power delivery model. In this dissertation, this model is modified, updated, and incorporated for the power-gating analyses and chip/package co-optimization. For the thermal model, a Hotspot simulator is adopted into the design methodology to obtain the temperature based on the power dissipation of each individual block. For the process variation analyses, the existing methodologies analyze performance variations using a bottom-up approach, where detailed logic designs of blocks/cores are required for the variation analysis [115-117]. While such an approach is highly accurate at the individual block levels, going forward we face new challenges. With the scaling of the technology node, the number of cores and logic units are expected to increase dramatically [56]. The

design objective is shifting from controlling frequency distribution of individual block/cores to controlling throughput distribution of the multi-core processor for a given power budget and power density constraint. For the system-level variation analysis, it is highly inefficient, if not impractical, to first fully design cores of various sizes and complexities with complete logic circuits, floorplanning, and timing analyses and then apply process variation analysis to quantify the throughput distribution for each core and the multi-core system. Instead, decisions on the complexity and size of the cores must be made at a higher level, which requires simplified system-level design models that can provide reasonably accurate results at the early design stage without the detailed knowledge of the individual logic circuits. The methods should be fast, efficient, and preferably analytical to enable exploring a large multi-parameter design space, including device-, circuit-, and system-level parameters.

The rest of this chapter is organized as follow: in Section 2, a hierarchical memory model is illustrated to addresses the impact of memory bandwidth by including the miss penalty of the on-chip cache and the penalty associated with the link and access latencies of the off-chip main memory. Section 3, an empirical CPI model is extended based on the updated memory model for the latest Intel and IBM processor families. In Section 4, multi-level interconnection network design models and algorithms are introduced. Several modifications are performed to better utilize these models for the proposed hierarchical optimization engine. Section 5 describes a compact power distribution network model to efficiently estimate the power delivery noise for various chip and package configurations, allowing a fast chip and package co-optimization feasible. In Section 6, a process variation model is described, which is used to investigate the impact of various sources of process variation from both devices and interconnects on the overall chip throughput. Section 7 describes the thermal model that is used in the design methodology. At the end, the proposed hierarchical design methodology is developed,

integrated, and validated in Section 8 based on the models from all the device, circuit, and system levels.

4.2 Hierarchical Memory Model

For the memory hierarchy system, instead of using a cycle-accurate architecture simulator, an analytical modeling approach is desired because of the runtime efficiency and computation cost at the early design stage. This model addresses the impact of memory bandwidth by including the miss penalty of the on-chip cache and the penalty associated with the link and access latencies of the off-chip main memory [118]. The overall throughput can be written as

$$TP(n) = \frac{f}{CPI} = \frac{n}{\frac{CPI_{logic}}{f} + M(n)\left(\frac{t_{dram}}{N_{pr}} + t_{link}(n)\right)} \quad (25)$$

where f is the clock frequency, n is the number of cores, M is the miss rate that depends on the size of the cache, t_{link} is the link latency that depends on the bandwidth, data package size, and link utilization, and N_{pr} is an empirical number representing the number of parallel memory requests per cycle, which depends on the complexity of the logic core. It is obtained according to the extracted value from the cycle-accurate simulator for two Intel processors [118]. The miss rate for caches of different sizes is modeled as $M = M_0/\sqrt{S}$, where S is the size of cache in MB, and M_0 is the miss rate for 1 MB cache, which is assumed to be 0.28% according to [109]. This relation is found to provide the most accurate approximation of the average miss rate [118]. For the performance of the on-chip cache, CACTI, an open source cache simulator is used [119] to estimate the cache density and power dissipation. CPI_{logic} is the CPI of a logic core with perfect cache, and it is determined by using an empirical CPI model, which is presented in the next subsection.

4.3 Empirical CPI Model

Previous work about the empirical CPI model can be traced back to Eble's work in the late 1990s [109]. He first observed a power-law relationship existing between the number of logic transistors and the instructions per cycle (IPC) for three different microprocessor families, including both RISC and CISC machines [109]. After that, he presented a logic-memory model that differentiated between logic and cache transistors and weighed the contributions of each appropriately. However, his work is based on old Intel processors. Since then, processors have gone through major changes, and it is essential to revisit this model. In addition, the hierarchical memory model in his work can only be applied to a single-core processor. A more detailed validated throughput model is required to analyze multi-core processors accounting for the finite memory bandwidth. In the last decade, some changes have been made in the architecture of processors, including the shifts from single- to multi-core systems, from 32- to 64-bit machines, and from two- to three-level cache hierarchies. The number of transistors in the processor is also constantly increasing. Thus, it is necessary to revisit this CPI model and extend it for recent processors.

First, since the memory bandwidth becomes one of the significant impediments to the performance, especially for a multi-core processor designed for high-parallel data-intensive applications, a more detailed hierarchical memory model is used as shown in (25).

Second, as the only available benchmark results for the recent processors are SPECint2000 and SPECint2006, the k_{95} in [109] is substituted by k_{2000}, which is estimated by using the computational CPI simulated from a cycle-accurate simulator for two specific Intel microprocessors [118]. For the latest Nehalem and Sandy Bridge processors, SPECint2006 results are also normalized to the SPECint2000 results based on the observation that several processors have the results for both SPEC2000 and

SPEC2006, and the ratios of the scores are close for those processors. Since the SPECint2006 can take advantage of the multi-core computing, Amdahl's law [59] is used to obtain the relationship between the throughput of a single core and the chip throughput of a multi-core processor as

$$TP = \frac{1}{\frac{1-p}{TP(1)} + \frac{p}{TP(n)}} \tag{26}$$

where $TP(1)$ and $TP(n)$ are the throughputs of a multi-core processor running serial and parallel part of the program, respectively, and p is the fraction of the program that can be executed in parallel. Here, p is estimated as 0.35 according to the SPECint2006 results for the same dual-core Itanium processors with two different configurations (one and two active cores).

Table 6: The Raw Data Collected for Various Processor Generations for the Intel Microprocessor.

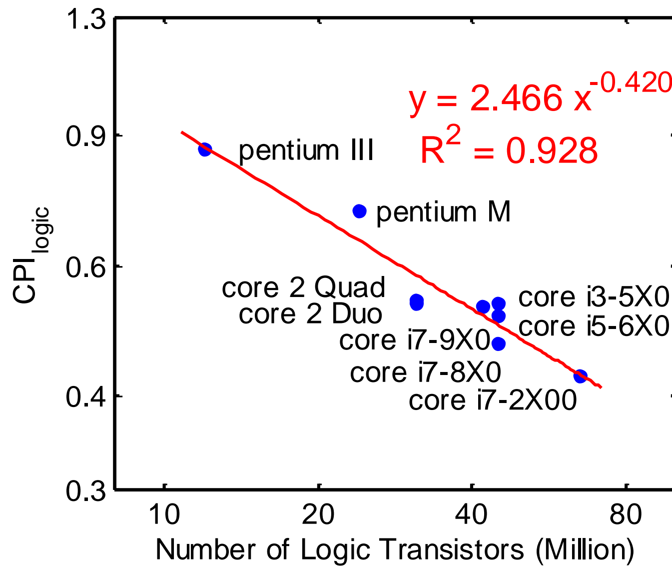
Processors	Node (nm)	Freq (GHz)	Cache (MB)	N_{tran} (Million)	SPECint 2006
Pentium III	180	1.0	0.25	28	2.73*
Pentium M 780	90	2.26	2	144	11.4
Core 2 E8200	45	2.66	6	410	22.65
Core 2 Q9550	45	2.83	12	820	24.4
Core i3-540	32	3.06	4	382	28.43
Core i5-650	32	3.2	4	382	30.93
Core i7-940	32	2.93	8	731	30.8
Core i7-870	32	2.93	8	774	36.8
Core i7-2600	32	3.4	8	995[120]	46.2

* normalized to SPECint2006 from SPECint2000

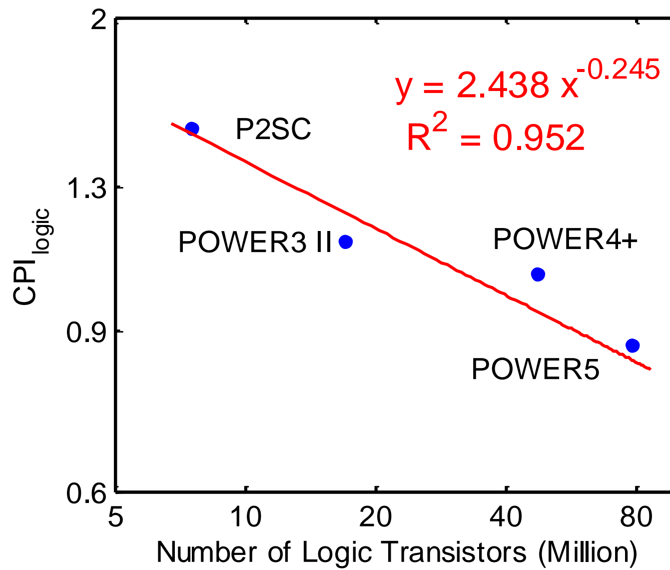
To link the chip throughput to the SPEC benchmarking results, two specific Intel microprocessors [118] are taken as references, which provide the computational CPI based on a cycle-accurate simulator. Once the link is obtained, the throughput TP are evaluated for each Intel processor based on its SPECint 2006 results. To obtain the CPI_{logic}, first, an initial CPI_{logic} is guessed to calculate the overall chip throughput using

(20) and (26). Then, the CPI_{logic} will be increased or reduced if the calculated chip throughput is smaller or larger than TP. After certain number of iterations, the converged CPI_{logic} values are generated.

For the number of logic transistors, most of the total number of transistors in Table 6 can be directly obtained from the Intel website [121], and the last one is obtained based on reference [120]. The number of logic transistors is extracted by using the average value of the following two calculations. First, since each processor generation has several configurations in terms of the cache size, the average transistors per MB cache can be obtained based on the raw total number of transistors for each configuration. Then, the number of logic transistors is equal to the total number of transistors minus the number of transistors in cache. Second, the numbers of logic transistors are estimated by the die area of the logic core based on the reported number of logic transistors in Core 2 processor from Intel [122] as a reference point. For the same technology node, the number of logic transistors is assumed to be proportional to the die area, which is measured from the die photos [122-125]. Based on to the Intel tick-tock development model [126], when each new technology is introduced, the same design is applied to a smaller technology node, which provides the transistor density for the new technology generation. Using this method, the number of logic transistors for all processor generations can be estimated. Since both these two methods may have some deviation from the actual values, the average value of these two methods is used as the final extracted number of logic transistors.



(a)



(b)

Figure 60: CPI_{logic} versus the number of logic transistors of a single core for (a) Intel microprocessor family (b) IBM POWER family.

In the previous subsection, it is shown that a simple empirical CPI model very accurately relates the CPI of a processor core to its number of logic transistors for two different processor families, including both CISC and RISC machines. However, with an

empirical model, predicting future trends needs to be done carefully. First, the coefficient and exponent may vary for different architectures, as depicted in Figure 60. Hence, to utilize the empirical CPI model, future processors are required to use the same building blocks and designs, including the instruction fetch, the dynamic scheduling, execution units, and branch predictions, etc. For other processor families, one can extract the power-law model or develop a simulation tool to quantify the power-law model for existing and emerging architectures. Second, the observed trend may not be applicable if the number of transistors continues to grow, and there may be a diminishing return point beyond which the CPI of the processor may not follow the empirical model. However, the number of transistors per core is not expected to grow significantly as technology advances because of the power constraints and the variation issues [127]. The optimization results in this dissertation also confirm that bigger cores are not desirable due to the low power-efficiency. Finally, a new device technology may allow performing the same function with fewer devices. For such cases, one would need to obtain a CMOS equivalent number of switches.

4.4 Multi-Level Interconnection Network Models

To efficiently generate the multi-level interconnect network, a modified wire length distribution model was developed back in late 1990s [110]. Later, it was modified to provide a better match with the commercial products [128]. This distribution can be written as

$$i(l) = \begin{cases} \frac{\alpha k}{2}\Gamma\left(\frac{l^3}{3} - 2l^2\sqrt{N_{sockets}} + 2lN_{sockets}\right)l^{2p-4}, & 1 \le l < \sqrt{N_{sockets}} \\ \frac{\alpha k}{2}\Gamma\left(2\sqrt{N_{sockets}} - l\right)l^{2p-4}, & \sqrt{N_{sockets}} \le l < 2\sqrt{N_{sockets}} \end{cases} \quad (27)$$

where l is the length of the interconnect in socket pitch, p and k are the Rent's constants, and Γ is written as

$$\Gamma = \frac{2N_{gates}\left(1 - N_{gates}^{p-1}\right)}{-N_{sockets}^{p}\frac{1+2p-2^{2p-1}}{p(p-1)(2p-1)(2p-3)} - \frac{1}{6p} + \frac{2\sqrt{N_{sockets}}}{2p-1} - \frac{N_{sockets}}{p-1}} \tag{28}$$

The average interconnect length of this interconnect distribution for a system with a Rent's constant p that is larger than 0.5 can be simplified and written as

$$L_{avg} = p_{gates}^{1-p} N_{gates}^{p-0.5}\left(\frac{p+1-4^{p-0.5}}{2(p-0.5)(p+0.5)p}\right) \tag{29}$$

Based on this stochastic interconnect distribution model, the multi-level interconnection network can be designed in a very efficient way. For a given clock frequency, first, from the bottom-up direction, the local power distribution network is optimized based on the estimated power dissipation of the devices and interconnects. It generates the number of power and ground vias, which affects the effective wiring efficiency on each metal level. Second, the local signal interconnects are placed based on the minimum interconnect width and pitch. The longest interconnect length on the local level is used for the shortest interconnect on the next intermediate signal interconnects, and the iteration continues until it reaches the maximum number of metal levels or all the interconnects are routed. Third, from the top-down direction, the global clock and power distribution networks are optimized, and the optimal global interconnect pitch at the highest metal level is obtained. Fourth, the repeater insertion algorithm is performed to achieve the minimum EDP. The shortest interconnect length is determined, which is used as the longest interconnect length on the lower intermediate metal level. This iteration also continues until one of the following conditions is satisfied: 1) the iteration count reaches the maximum number of metal levels; 2) no more die area is left for the repeater insertion; or 3) the shortest interconnect length is shorter than one of the longest interconnect lengths obtained from the previous bottom-up routing, and meanwhile, the total number of metal levels with and without the repeater insertion is less or equal than

the maximum number of metal levels. After the repeater insertion algorithm is completed, if the total number of metal levels are still larger than the maximum number of metal levels, the clock frequency needs to be reduced, and the multi-level interconnect network design starts from the beginning. If the total number of metal levels meet the constraint, the total power dissipation is updated according to the number and size of the repeaters, and the simulation goes back to the first step to obtain the updated number of local power vias. This outer iteration keeps running until the total power dissipation converges, and the final design parameters are obtained at a certain given clock frequency.

In the past decades, three-dimensional (3D) integrated circuits (ICs) have intrigued a lot of research in the past decade because of its potential benefits, including extending Moore's Law by increasing the device density, overcoming the barriers in the interconnect scaling, and providing further performance improvement with less power consumption. One commonly used technique to connect different processor layers is the through-silicon via (TSV) based technology [44], while another technology, which has been proposed recently, is the monolithic inter-tier via (MIV) based technology. For the latter, the alignment between layers can be as high as the lithographic alignment, providing about an order of magnitude smaller via diameter and two orders of magnitude smaller via capacitance compared with TSV technology, enabling within-core fine-grained 3D integration [129]. To accommodate the 3D ICs, the wire-length distribution model was extended and derived [130] to estimate both the wire-length distributions within tier and between any two tiers based on a certain distance and the number of tiers.

In this work, the spacing between two nearby tiers is assumed to be the equivalent length of the copper interconnect that has the same capacitance as the inter-tier via. Depending on the diameter and capacitance of the vertical via, optimal number of tiers can be optimized because increasing the number of tiers can reduce the average and the longest interconnect lengths due to the shortcuts in the vertical direction. However, too

many tiers will introduce a large number of vias between tiers, leading to a significant increase of via blockage and a drop in the wiring efficiency.

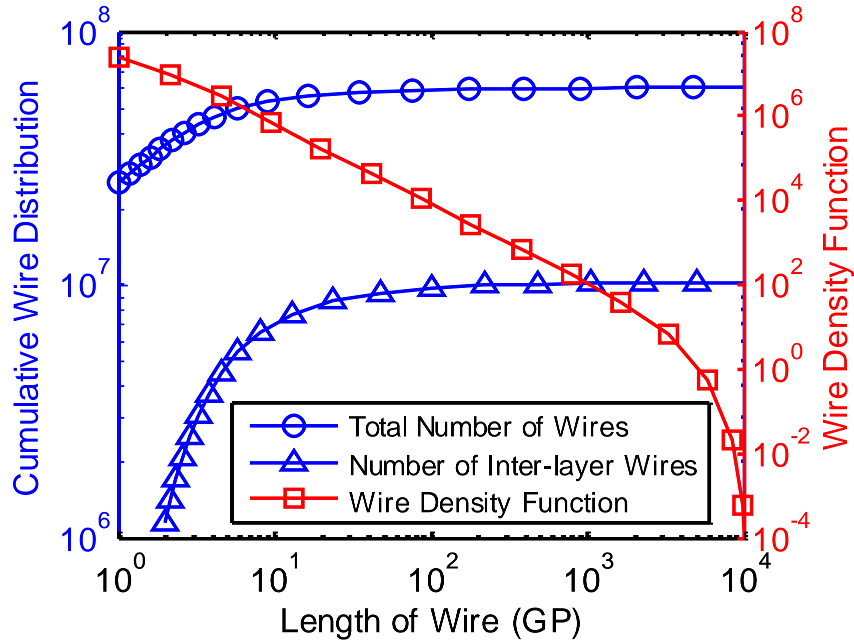


Figure 61: Interconnect density function and the cumulative interconnect distribution for inter-layer and all interconnects.

For a logic core with 20 million logic gates, the cumulative interconnect distributions are shown in Figure 61, indicating the number of inter-tier interconnects reaches 10 million. If the TSV-based 3D technology is utilized, the area occupied by the vias becomes larger than the total die area due to the large diameter of the TSV. However, if a MIV with a 100 nm diameter is available, the fine-grained within-core 3D integration is approachable.

4.5 Power Distribution Network Model

Following the previous work [114], the equivalent circuit model is modified to take into account the sleep transistors, as shown in Figure 62 [131].

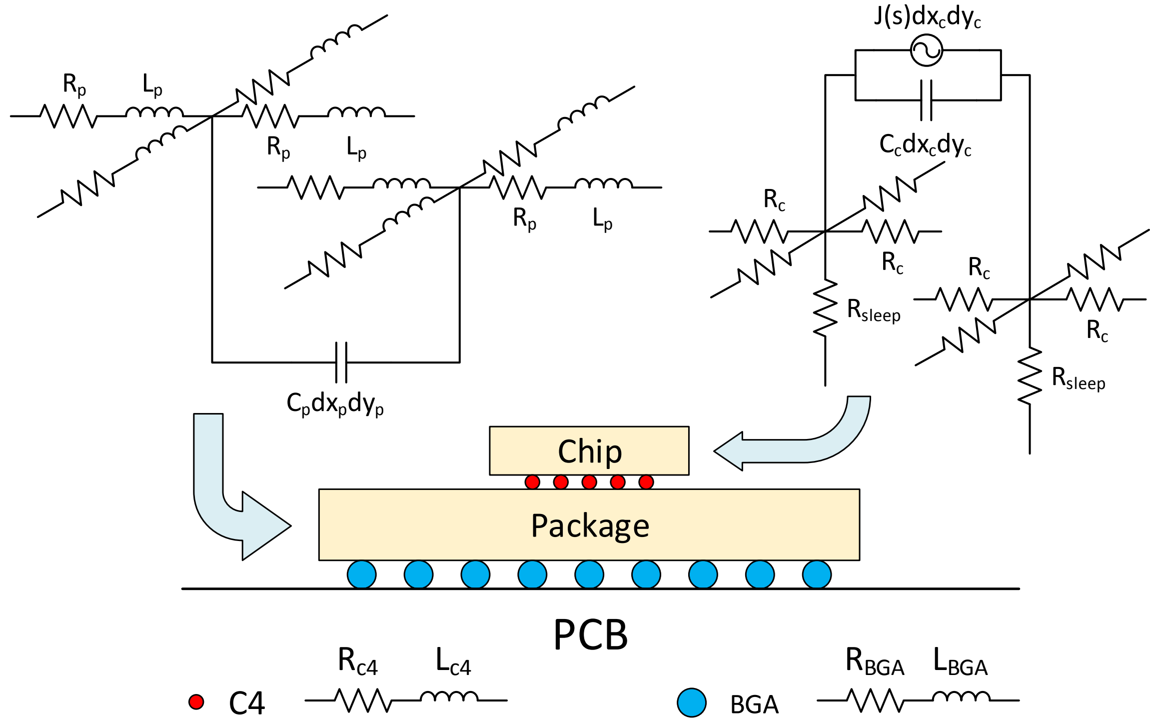


Figure 62: Equivalent circuit model for the PDN, including both chip and package.

To obtain the transient response of this circuit, two sets of partial differential equations in the Laplace domain are built to include both chip and package, which can be written as

$$\nabla^2 V_c(x, y, s) = 2V_c(x, y, s)R_c s C_c + \Phi_c(x, y, s) \tag{30}$$

$$\nabla^2 V_p(x, y, s) = 2V_p(x, y, s)(R_p + sL_p)sC_p + \Phi_p(x, y, s) \tag{31}$$

where R_c and R_p are the lumped resistances between two adjacent nodes on chip and package, respectively, C_c and C_p are the capacitances per unit area for chip and package, respectively, L_p is the package inductance between two adjacent nodes, and Φ_c and Φ_p are the source functions for chip and package, respectively, which are used to make mathematical connections between the chip and the package and between the package and the printed circuit board (PCB) [114]. They can be expressed as

$$\Phi_c(x,y,s) = -\sum_{i=1}^{N_{sp}} R_c J_c(s) \cdot \Delta x \cdot \Delta y \cdot \delta(x - x_{spi})\delta(y - y_{spi}) \\ - \sum_{j=1}^{N_{C4}} \frac{R_c}{sL_{C4j} + R_{C4j}} \cdot \left(V_{c-C4j}(s) - V_{p-C4j}(s)\right) \tag{32}$$

and

$$\Phi_p(x,y,s) = \sum_{i=1}^{N_{C4}} \frac{R_p + sL_p}{sL_{C4j} + R_{C4j}} \cdot \left(V_{c-C4j}(s) - V_{p-C4j}(s)\right) \cdot \delta(x - x_{C4i})\delta(y - y_{C4i}) \\ - \sum_{j=1}^{N_{C4}} \frac{R_p + sL_p}{sL_{BGAj} + R_{BGAj}} \cdot V_{p-BGAj}(s) \cdot \delta(x - x_{BGAj})\delta(y - y_{BGAj}) \\ - \sum_{k=1}^{N_{C4}} \frac{R_p + sL_p}{sL_{ESLk} + R_{ESLk} + \frac{1}{sC_{Decap}}} V_{p-Decapk}(s) \cdot \delta(x - x_{Decapk})\delta(y - y_{Decapk}). \tag{33}$$

where J_c is the current density of the circuit, L_{BGA} and R_{BGA} are the inductance and resistance of the BGA, respectively, C_{Decap} is the capacitance of the decap, L_{ESLk} and R_{ESLk} are the equivalent series inductance and resistance of each decap, respectively, and (x_{spi}, y_{spi}), (x_{C4i}, y_{C4i}), (x_{BGAj}, y_{BGAj}), and (x_{Decapk}, y_{Decapk}) are the location of each switching node, C4 bump, BGA bump, and decap, respectively. The basic parameters, including resistance, inductance, and capacitance for chip, package, C4s, and BGAs, are obtained based on the typical values reported in [114] and the ITRS projections [132]. Once the solutions are obtained for each node, the inverse Laplace transform is performed to obtain the transient response.

Table 7: Package Configurations for Resistance, Capacitance, and Inductance.

Parameters	Value
R_c (Ω)	0.3
R_p (Ω per unit area)	0.05
L_p (nH per unit area)	0.2
C_p (nF/cm2)	2
C_c (nF/cm2)	530
L_{c4} (nH)	0.01
R_{c4} (Ω)	0.002
L_{BGA} (nH)	0.4
R_{BGA} (Ω)	0.01
J_c (A/cm2)	10

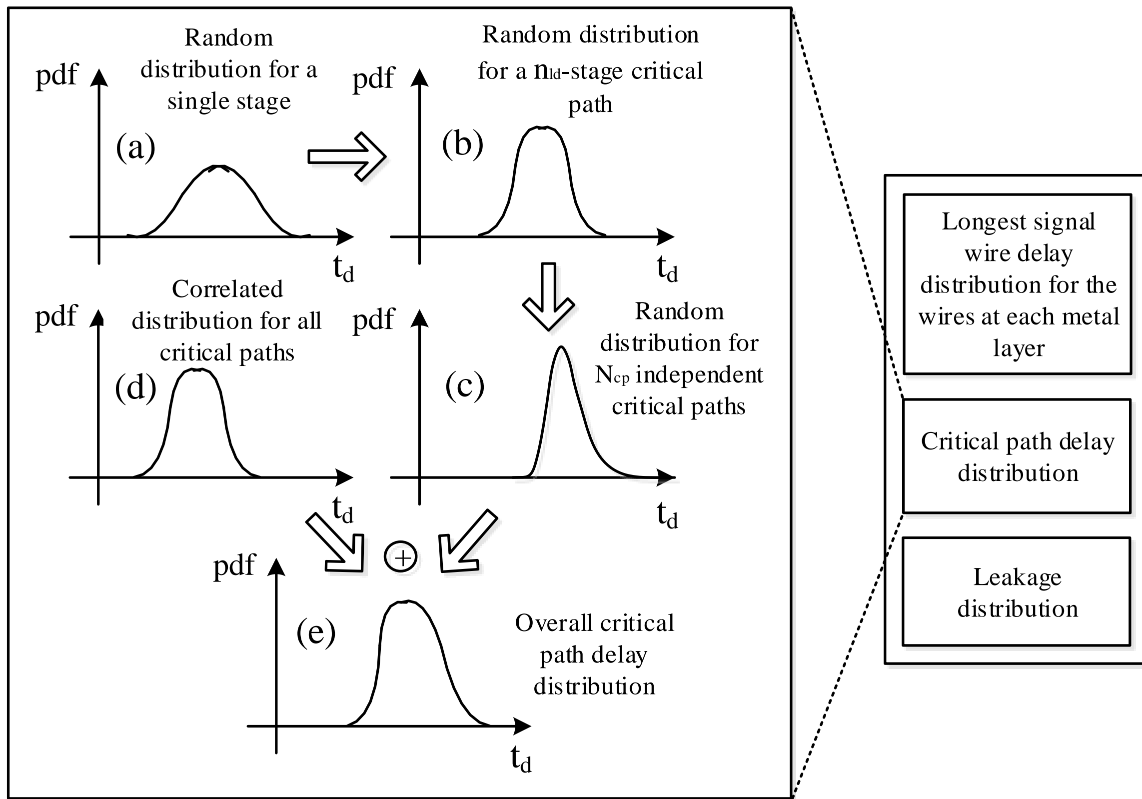


Figure 63: Delay and leakage distribution generation of variation-aware system-level design.

4.6 Process Variation Model

To include the impact due to the process variation, the clock frequency needs to be updated based on the samples taken from the critical path and the longest interconnects delay distributions on each metal level. The device- and circuit-level delay and leakage power distributions are obtained by using HSPICE Monte Carlo simulations. The impact of random and systematic variations on a critical path delay of a logic core is analyzed by following the methodology in the previous work [118].

First, a single-stage NAND2 gate or inverter delay distribution is obtained, as shown in Figure 63 (a). Based on the logic depth and the optimal number of repeaters obtained from the multi-level interconnection network optimization algorithm, Figure 63 (b) shows the delay distribution for a single critical path or signal interconnect with/without repeater insertion. Based on the number of independent critical path, Figure 63 (c) shows the maximum delay distribution of those independent interconnects. Compared with Figure 63 (a), the standard deviation decreases while the mean of delay increases. After taking into account the correlated systematic variation, depicted in Figure 63 (d), the overall critical path delay distribution is obtained by convoluting the random and systematic delay distributions, as shown in Figure 63 (e). Once the samples are taken from the overall delay distributions, the clock frequency is limited by the slowest path from either the critical path or the longest interconnect on each metal level.

4.7 Thermal Model

Power dissipation and thermal issues are increasingly significant in modern processors, which affect the overall energy consumption, maximum frequency, and product reliability. Since a sizable number of device-level parameters are very sensitive to the temperature, especially the leakage current for the CMOS technology, the system-level analyses of power/performance trade-offs would be incomplete without including

the thermal model. To take into account the relation between the power consumption, material property and the temperature of the processor, HotSpot, an open source thermal simulator [133], is utilized to optimize the performance of the processor based on a certain thermal constraint. It is built based on the well-known duality between thermal and electrical phenomena. In this duality, the heat flow passing through a thermal resistance is analogous to the electrical current; the temperature difference is analogous to the voltage; and the thermal capacitance, defining the heat absorbing capability of the material, is analogous to the electrical capacitance which accumulates electrical charge. During simulation, the compact thermal model is represented by a lumped thermal RC network, whose scale is typically relatively small, which can be solved very efficiently and introduce little computational overhead.

4.8 Hierarchical Design Methodology and Validation

The overview of the system-level design methodology is shown in Figure 64 [134]. First, an initial design point is guessed, including the number of cores, the logic-to-cache ratio, the number of logic gates per core and the supply voltage. For given thermal/power and die size area budgets, the device-level characteristics, including channel width, ON/OFF resistance, and the input capacitance, are estimated and simulated by using HSPICE based on Predictive Technology Models for various technology nodes [67].

These device-level parameters are taken as the input for the multi-level interconnection network models. The maximum operation frequency of the logic core is estimated by using a generic critical path model in IntSim [128], which determines the clock frequency based on a given logic depth. It assumes that the logic gates are made of 2-input NAND gates with an average fan-out of 3, and they are sized based on the average interconnect length provided by a modified stochastic wiring distribution model [128]. This model also accounts for the delay of interconnects in a multi-level

interconnect network, which relies on the number and size of repeaters, global clock, and local and global power distribution network models. To obtain the maximum clock frequency under a power density budget, iterations are required to find out the optimal number and size of repeaters on the higher metal levels, the optimal local and global power and clock interconnect pitches and widths, and the optimal lengths, wiring efficiencies and pitches of signal interconnects on each metal level.

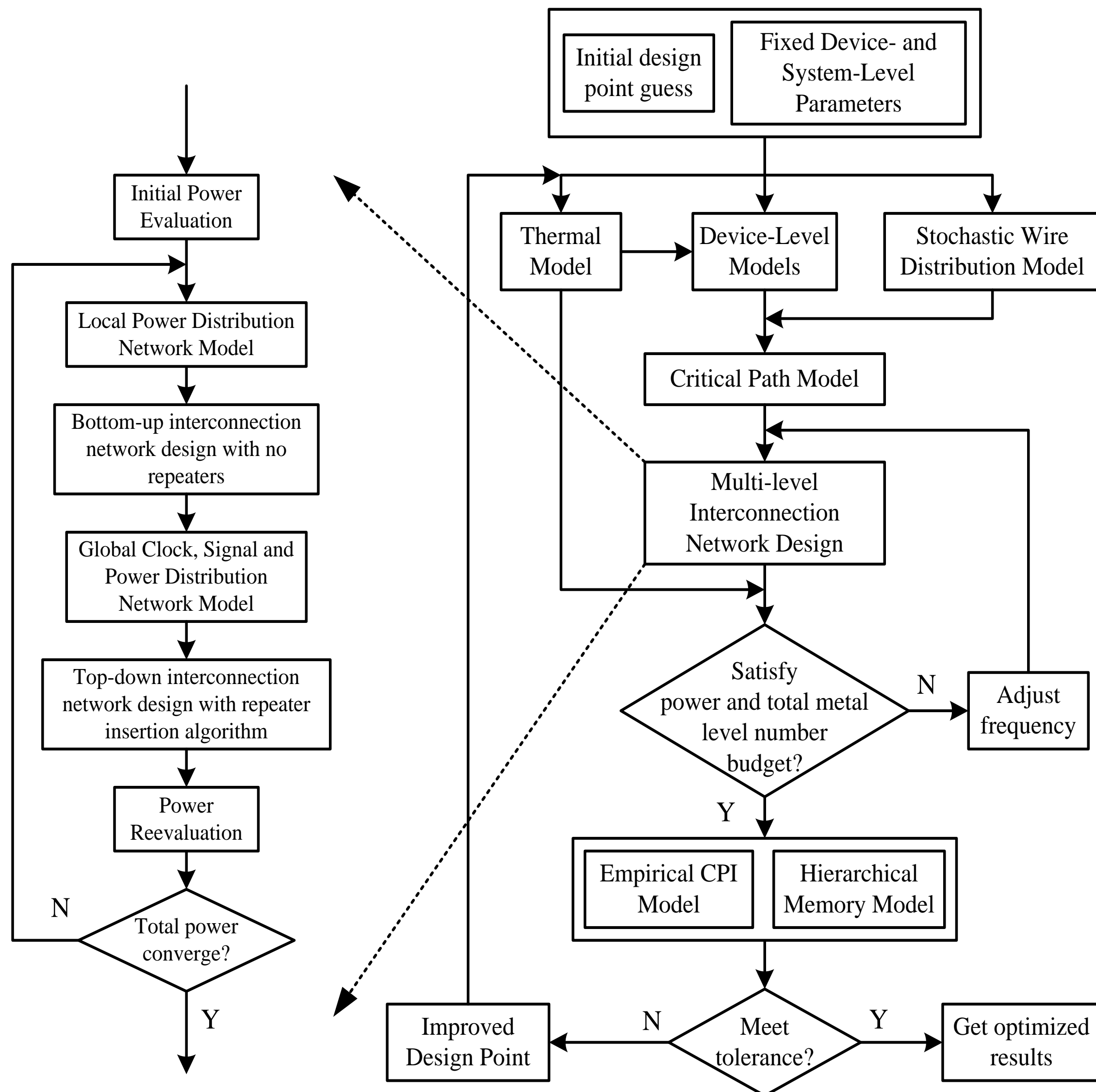


Figure 64: Flowchart of the system-level design methodology.

Due to the high cost of masks in the fabrication process, the maximum number of metal levels is assumed according to the ITRS projection [132]. To relate the temperature with power density and floorplanning of a multi-core processor with on-chip cache, HotSpot, an open source thermal simulator, is used [133], which is built based on the well-known duality between thermal and electrical phenomena. During simulation, the compact thermal model is represented by a lumped thermal RC network, whose scale is typically relatively small, which can be solved very efficiently and introduce little computational overhead. In this work, the maximum thermal budget for the chip is 80 °C.

To evaluate and optimize the throughput, the empirical CPI model and the hierarchical memory model are used as a bridge to connect the circuit-level parameters to the system-level throughput. The empirical CPI model indicates that a power-law relationship exists between the number of logic transistors and the CPI of the logic core as shown in Figure 60. One trade-off that can be made is that for a given logic core size area, one can design a processor by either using fewer logic gates with a larger channel width-to-length ratio to operate at a higher frequency or using a larger number of smaller logic gates to achieve a lower CPI. The hierarchical memory model takes into account the throughput loss due to the cache miss and the off-chip memory access time and link latency based on the memory bandwidth, data transfer size and the link utilization as shown in (20). Another trade-off here is that for a given die area, one can either use more area for the logic function such as branch predictor and execution unit to reduce the computational CPI or assign more area for cache to reduce the miss rate and the miss penalty. The optimizer will explore all the possible combinations of various design parameters such as supply voltage, number of logic transistors, and logic-to-cache ratio to find the optimal design point to achieve the maximum throughput. Once all the initial design points are evaluated, the throughput at the optimal point is compared with those design points that are close to the optimal point. If the difference is larger than the

tolerance (e.g. 1%), more design points will be inserted among the previous evaluated points; if the tolerance is met and no significant improvement is observed, the optimal design parameters as well as the performance metrics will be read. Since most of the models are analytical and compact, it allows us to perform the exhaustive searching very efficiently.

In this section, the system-level design methodology is applied to several commercially available Intel multi-core processors across three technology generations from 65nm to 32nm technology node from three architectures, which are Core 2, Nehalem, and Sandy Bridge microarchitectures. The predictions for the overall throughput, clock frequency, logic core area, and the number of logic transistors are compared with actual values extracted from various resources and publications [121, 122, 135], [123, 136, 137].

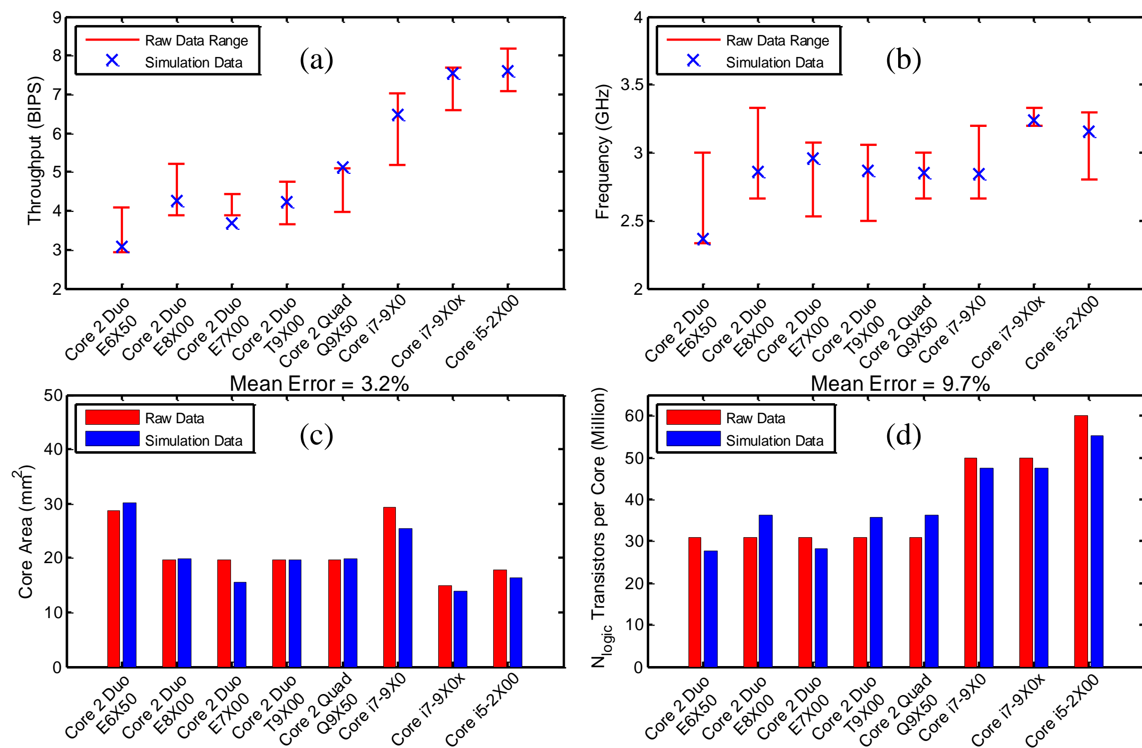


Figure 65: Comparison of simulation results with actual data in terms of (a) throughput, (b) clock frequency, (c) logic core area, and (d) number of logic transistors.

At the device-level, the ON/OFF resistance and capacitance of the logic gates are obtained by HSPICE based on Predictive Technology Model at the 65nm, 45nm, and 32nm technology nodes [67]. The CMOS logic gate footprint model follows the models presented in [50]. At the circuit-level, parameters related with the interconnection network are taken from [128], including Rent's constants and wiring efficiency. At the system-level, Intel [121] provides some processor information, including total die area, number of transistors, supply voltage, clock frequency, memory bandwidth, and power consumption, which can be used to estimate the raw data of the number of logic transistors and the logic core area according to the die photo [122-124]. The die size area budget for the optimization is the summation of the logic core and cache area. The logic depth is estimated by calculating the number of FO4 stages based on the clock period of the processors and the HSPICE simulations based on the supply voltage of the processors. The empirical CPI model is using the Intel trendline shown in Figure 60 (a). The raw throughput of various processors are estimated based on SPEC benchmark results [135].

From the comparison in Figure 65, both the raw data throughput and clock frequency have certain ranges for each type of processor, which are mainly caused by the process variation. In general, the simulation data for the chip throughput and clock frequency are in the range of the published raw data. The geometric mean errors for the logic core area and number of logic transistors compared with the raw data are less than 10%.

4.9 Conclusions

This chapter illustrates and develops the detailed modeling approaches at the system level, where six sets of compact models are introduced, including a hierarchical memory model, an empirical CPI model, multi-level interconnection network design models, a power distribution network model, a process variation model, and a thermal model. In particular, the empirical CPI model is developed based on the Intel and IBM

microprocessors released in the recent decade so that the system-level throughput of the novel device technologies can be effectively evaluated and optimized without designing a full processor. An excellent power-law relationship has been observed between the CPI and the number of logic transistors on chip. The multi-level interconnection network model is also modified to fast obtain the maximum clock frequency based on a fixed number of metal levels and total power dissipation constraints.

At the end of this chapter, a hierarchical design methodology is presented based on all the aforementioned models, which can be used for evaluating and optimizing various device-, interconnect-, and system-level technologies and innovations in the following chapters. This design methodology has also been validated that the simulation data for the chip throughput and clock frequency are in the range of the published raw data. The geometric mean errors for the throughput and frequency compared with the average value of the raw data are less than 5%; and the geometric mean errors for the logic core area and number of logic transistors compared with the raw data are less than 10%.This shows the validity of this methodology so that reasonably accurate benchmarking and optimization can be achieved in a very efficient way.

CHAPTER 5 DEVICE TECHNOLOGY OPTIMIZATION

5.1 Introduction

Many novel device concepts were proposed in the past decade to overcome the limits of the conventional Si CMOS technology, such as carbon-based and spin-based devices. In this chapter, based on the hierarchical optimization engine developed in Chapter 4, two emerging device technologies, GPNJ devices and TFETs, are benchmarked and optimized against their CMOS counterparts in terms of the overall throughput.

The previous work for benchmarking GPNJ devices has only focused on the device-level performance [62]. Since new devices may offer fundamentally different energy-delay trade-offs, it is essential to co-optimize device- and system-level parameters to maximize the chip throughput at a given power or thermal budget. A smaller device may shorten the aggregate length of interconnects, but it may have higher ON resistance and leakage current. The advantage of those emerging technologies can be fully evaluated only if the characteristics of devices and the overall performance of the system are simultaneously considered and co-designed. The high cost of developing novel technologies also makes it vital to gain an early understanding of the potential benefits so that development with little benefit can be avoided.

For the low-power applications, TFETs show promise in overcoming the power wall of thermionic FETs by allowing for significant reduction in the supply voltage [65]. A wide range of device architectures and materials has been studied, and it is shown that TFETs can potentially offer intrinsic gate delays that are comparable with the conventional CMOS technology at lower supply voltages [65, 66]. However, TFETs offer fundamentally different energy-delay trade-offs, and their interaction with interconnects

is substantially different because of a much larger output resistance and a smaller input capacitance compared to thermionic devices. Therefore, device- and system-level co-optimization and benchmarking are crucial to evaluate the potential benefits of the TFETs.

As the technology scales down to the sub-100nm technology nodes, the process variation results in significant spread in the maximum operating frequency and leakage of logic circuits. The increasing power density and on-chip temperature due to the large transistor density and the high clock frequency have also emerged as major concerns for high performance processors. It is critical to ensure that all design decisions satisfy the power density constraint. For example, a higher supply voltage normally implies the ability to operate at a higher frequency. However, under a power density constraint, a higher supply voltage may force the processor to reduce the frequency because of the large leakage and dynamic power dissipations. Since the leakage power is very sensitive to the process variation, the frequency deviation will be significantly affected by the leakage distribution of a processor under a given power density constraint. Therefore, understanding performance and leakage variability under power density constraint is important for current and future processor designs. Significant work has been performed to quantify and minimize the impact of process variation on circuit performance. The statistical static timing analysis and statistical leakage analysis for yield estimation have received major attention [115-117]. These papers focus on the circuit-level analyses, which require specific logic circuits. At a higher level, the maximum frequency distribution for a single core is analyzed [138, 139] by taking into account the impact of die-to-die (D2D) and within-die (WID) variations. In [47], the impact of core-to-core (C2C) variations on the frequency of a multi-core processor is analyzed, which considers the spatially correlated WID variations. In [118], the maximum frequency model presented in [138] has been extended to the throughput distribution for a multi-core processor considering a few given architecture designs. The throughput distribution is

obtained by performing the statistical analysis of the extracted critical paths and running cycle-accurate simulations. Although [118] provides an accurate estimation of the overall performance of the processor, the interaction between the power and frequency is ignored and the design space is not fully explored. In this chapter, the impact due to various device-level random and systematic variations on the overall chip throughput is analyzed based on the variation-aware system-level design methodology developed in Chapter 3. Compared with other works that use cycle-accurate simulators, this methodology significantly reduces the simulation time, making the circuit- and system-level co-optimization possible.

The rest of this chapter is organized as follow. In Section 2, optimization is performed for the Si CMOS devices, such as the planar FETs and FinFETs. Multiple device-level design parameters are optimized and the optimal throughput is compared between planar FET and FinFETs for a single-core processor. In addition, the performance and area scaling analyses are performed for the FinFET devices down to the 7nm technology node. Section 3 optimizes the GPNJ devices for both single-core and multi-core processors. Promising results are shown compared with the conventional CMOS devices. In Section 4, III-V material-based TFETs single-core processors are optimized, targeting for low-power applications with ultra-low power constraints. Section 5 quantifies the impact due to the device-level process variation for both CMOS single- and multi-core processors.

5.2 Conventional Si CMOS Devices

To quantify the performance at the system-level throughput instead of just the intrinsic device-level delay or energy dissipation, both the planar FETs and FinFETs are optimized and compared in terms of the overall chip throughput based on the models from various hierarchies and the design methodology developed in Chapters 2-4.

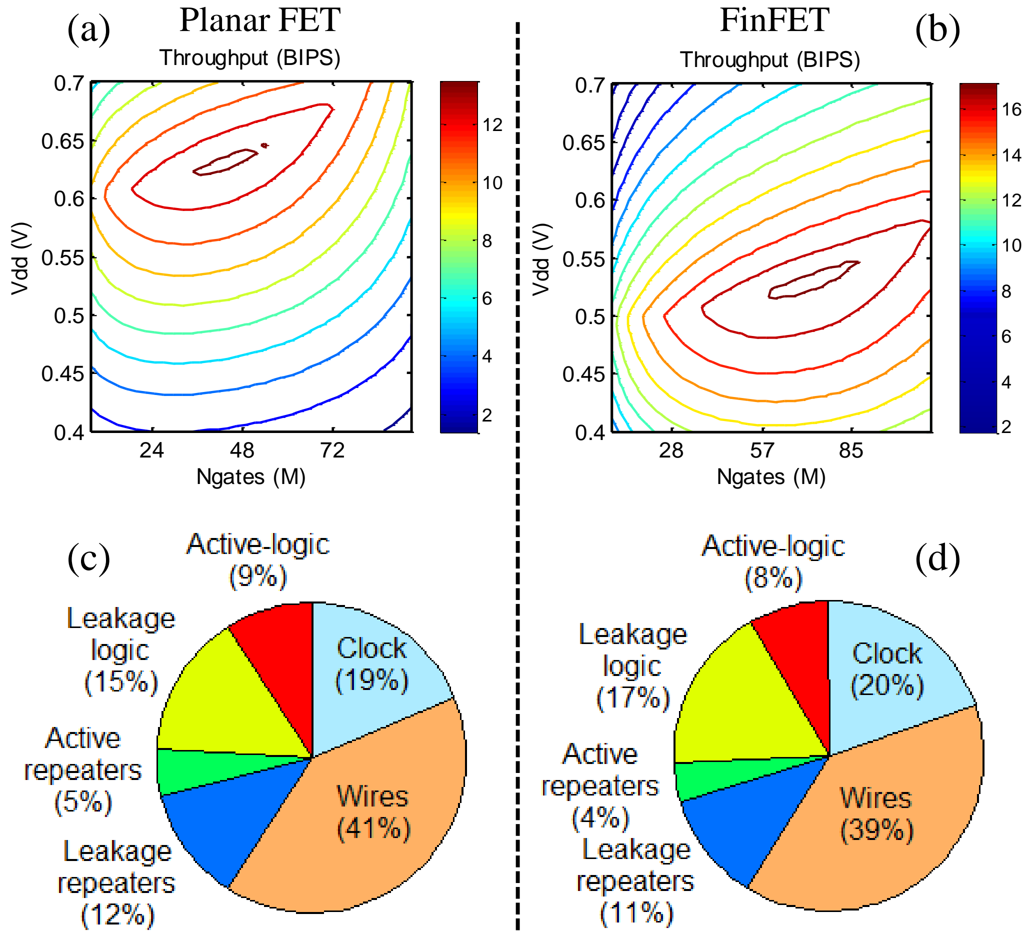


Figure 66: Optimization results for a single core implemented by the conventional planar CMOS and FinFET devices. (a) and (b) show the throughput versus the supply voltage and the number of logic gates. (c) and (d) show the pie chart of power components for each part.

5.2.1 Comparison between Planar FET and FinFET

Based on the assumptions and configurations shown in Table 8, Figure 66 shows the optimization results for a single core using planar CMOS and FinFET devices at the 16nm technology node. From Figure 66 (a) and (b), there exist optimal numbers of gates and supply voltages that maximize the throughput because processors with few transistors suffer from large CPI values as the empirical CPI model indicates; the ones with too many transistors suffer from low clock frequencies constrained by complicated

interconnection networks and large switching energy of the logic gates. Likewise, processors with large supply voltages suffer from the large power dissipation; and the ones with very low supply voltages suffer from the high ON resistance of the devices, which limits the clock frequency.

Table 8: Configurations for the Single-Core Design.

Parameter	**Value**
Technology Node (nm)	16
Total Die Area (mm^2)	25
Thermal Constraint (°C)	80
Miss Rate for 1MB Cache	0.0028
Logic Depth	12
DRAM Access Time (ns)	50
Bandwidth (GB/s)	25
Total Metal Levels	12

Due to the 3D structure of FinFETs, they can provide a larger effective width compared with the planar CMOS transistor. Thus, to achieve the same driving current, FinFET devices need a smaller footprint area so that a larger number of logic gates can fit into the fixed core area without sacrificing the clock frequency as shown in Table 9. Another advantage of FinFETs is the superior electrostatics that provides a steeper subthreshold slope and less drain-induced barrier lowering (DIBL) effect compared to planar FETs [140]. This helps to further reduce the supply voltage and save both dynamic and leakage power dissipations. One can also observe that the optimal average number of fins to reach the maximum throughput is two. In addition, the optimized FinFET core needs a larger cache because of its larger computational capability compared with its planar CMOS counterpart. Due to all of the advantages and behaviors above, the overall throughput improvement for this 25 mm^2 single-core processor is about 26 % if planar CMOS transistors are replaced by FinFETs.

Table 9: Simulation Results of a Single Core for Planar CMOS and FinFET Devices.

Metric	Planar	FinFET
Throughput (BIPS)	13.81	17.38
Frequency (GHz)	4.21	4.40
Supply Voltage (V)	0.63	0.53
Width(F)/Number of Fins	8.64	2
Logic Area (mm^2)	16.16	15.49
Cache Size (MB)	9.01	9.35
N_{gates} (M)	46.13	75.65
Power (W)	17.06	16.42

5.2.2 Performance and Area Scaling Trends

With the same design methodology, the optimal number of logic gates, the clock frequency, and the throughput of a single core with various core areas are obtained as shown in Figure 67. When the area of the logic core is small, a very accurate power-law relationship exists between the throughput and the core area, which can be expressed as

$$T = T_0 \cdot A^{\alpha} \tag{34}$$

The values of the exponent α and the coefficient T_0 in various technology generations are given in Table 10. It can be seen that these values vary very little in various technology generations. However, the values are different as compared to the 0.5 exponent commonly used as a rule of thumb [56].

Table 10: The Coefficient and Exponent of Power-Law Relation between Throughput and Core Area.

Technology Node	20nm	16nm	14nm	10nm	7nm
Coefficient T_0	2.356	3.468	4.421	6.502	9.561
Exponent α	0.471	0.475	0.467	0.456	0.421
T_0 Improvement @ Fixed Area	-	47.2%	27.5%	47.1%	47.0%
Area Reduction @ Fixed Throughput	-	57.1%	41.1%	57.5%	56.2%

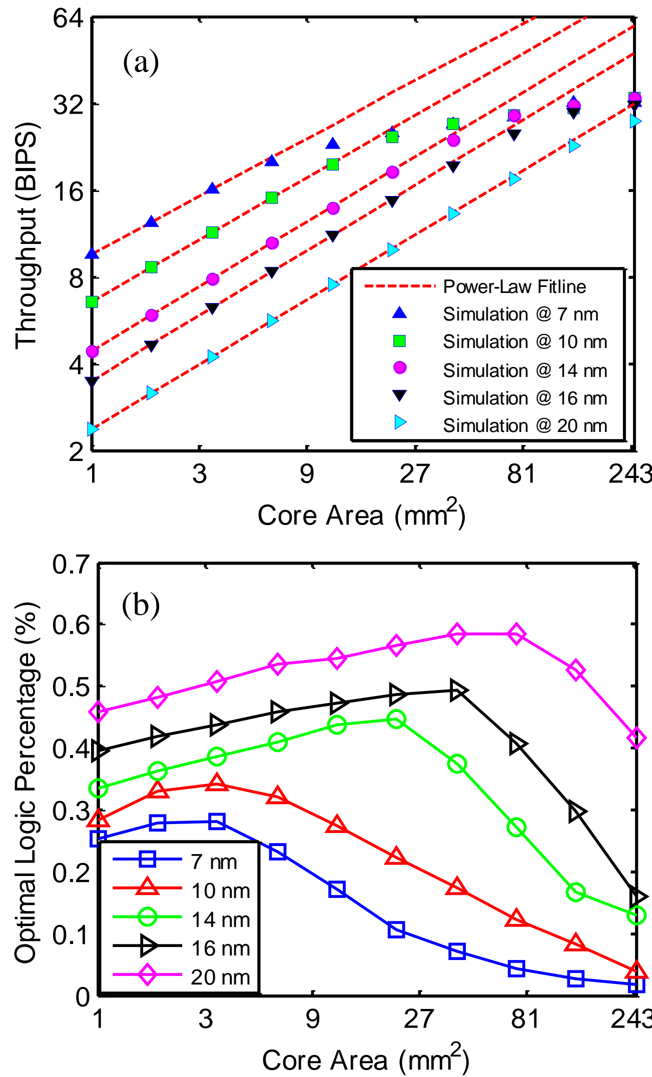


Figure 67: (a) Throughput versus core area for a single-core FinFET processor at various technology nodes. (b) Optimal percentage of the area that is occupied by logic.

Also, as Figure 67 shows, the power law behavior becomes invalid in large core areas. For instance, the power-law relation can hold up to 120 mm^2 at the 20nm technology node. As the technology goes down to the 7nm node, the throughput starts to saturate at 10mm^2. For each new technology generation, if the core based on the Intel architecture is well designed using the predictive FinFET technology, about 27-47% of the throughput improvement is observed at the same area budget; about 41-57% of area

reduction is also observed at the same performance target. However, the optimal throughput saturates if the core area is beyond a certain point. The reason of the saturation is that both the average and the longest interconnection lengths increase as the core size increases, which leads to a longer critical path delay and reduces the clock frequency. Moreover, with the scaling of the technology, the turning points shift to the left, indicating that a large core is not desirable, especially for a processor at a smaller technology node. This is because the interconnect pitch keeps decreasing as the technology scales, and if the size of the core does not decrease proportionally, the power and delay due to interconnects impose increasingly more severe problems. Therefore, even if the transistor density and the intrinsic delay and energy of the devices keep improving thanks to the scaling of the technology, the frequency penalty caused by the interconnection network diminishes the overall throughput improvement, as can be observed in Figure 67 (a) at a large area range.

The optimal percentage of die area that is occupied by the logic is depicted in Figure 67 (b). The processor at a smaller technology node prefers to use a smaller area for logic due to the aforementioned interconnect issue. At a larger die area budget, the optimal logic area drops because using a larger cache to reduce the miss rate and the miss penalty can bring more benefit than using a larger area for logic to reduce the computational CPI. For those large cores with small logic percentage, they are in the saturation region and probably not well designed, especially in a multi-core processor. Later, in the multi-core analysis, it is also confirmed that a large core is not desired, and the optimizer prefers to use a larger number of smaller cores to achieve the maximum chip throughput.

The results above are based on the empirical CPI model for the Intel microprocessor family. If one changes the CPI trendline exponent E_e, Figure 68 shows the relative throughput versus the core area at the 16nm technology node for various E_e. Once again, excellent power-law relations can fit the data when the core area is small for various

processor families. For the same aforementioned reason, the throughput saturates when the area is beyond a certain point. For the rule of thumb square-root relation between the throughput and core area, it may fit very well to a specific processor family. However, for other processor families, the power exponent needs to be modified accordingly.

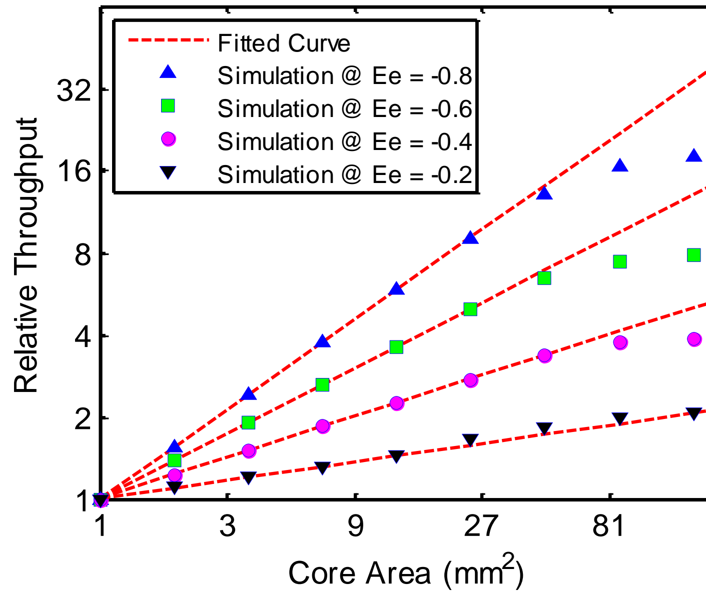


Figure 68: Relative throughput versus core area of FinFET single core at the 16 nm technology node for various Eble's exponents.

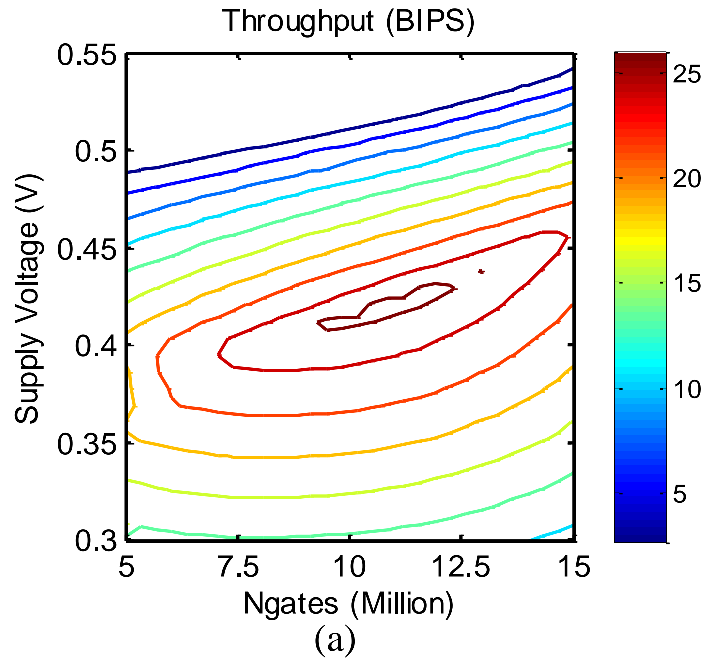
5.3 Graphene PN Junction Devices

In this section, device- and system-level co-optimization is performed for the GPNJ-based devices for both single- and multi-core processors based on the device- and system-level models described in Chapters 2-4.

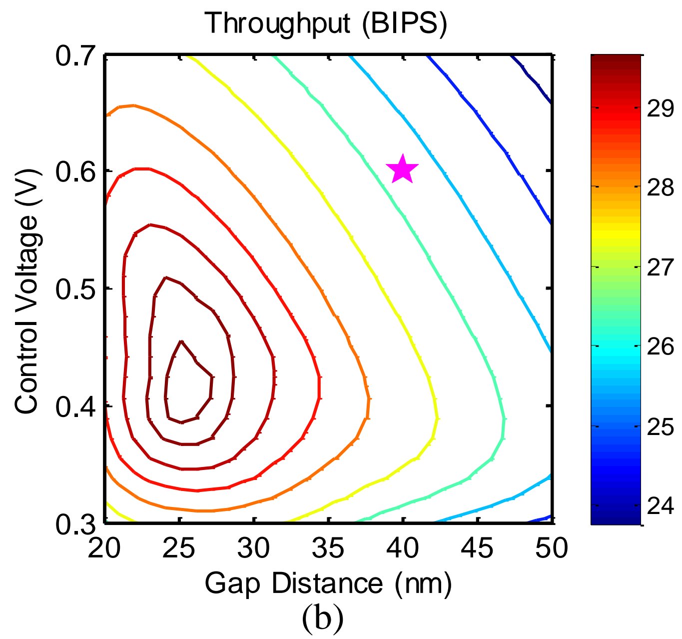
5.3.1 Single-Core Optimization

In this subsection, the optimization is first performed for a 5 mm^2 single-core processor with 80 °C thermal constraint. Various device-level parameters, including supply voltage, control voltage, gap distance, and oxide thickness, are optimized to maximize the overall throughput. For a given gap distance and the control voltage, Figure

69 (a) shows that there exist optimal numbers of logic gate and supply voltage that maximize the throughput because processors with few transistors suffer from large CPI values as the empirical CPI model indicates; the ones with too many transistors suffer from low clock frequencies constrained by complicated interconnection networks and large switching energy of the logic gates. Likewise, processors with large supply voltages suffer from the large power dissipation; and the ones with very low supply voltage suffer from the high ON resistance of the devices, which limits the clock frequency. However, the optimal design point in Figure 69 (a) is only for given gap distance and the control voltage, which is the purple star point in Figure 69 (b). For various gap distances and control voltages, Figure 69 (b) shows the optimal throughput, where each point is obtained based on the optimal supply voltage and the number of logic gates. One can observe that there exist optimal control voltage and gap distance that maximize the throughput because processors with low control voltage suffer from large leakage current as Figure 11 indicates; the ones with large control voltage suffer from the high ON resistance of the devices, which limits the clock frequency. Likewise, processors with large gap distance suffer from the large footprint area, causing a smaller device density; and the ones with very small gap distance have large leakage power dissipation based on Figure 11.



(a)



(b)

Figure 69: Throughput versus multiple design parameters. (a) Throughput versus Vdd and number of logic gates at $V_g = 0.6V$ and $D_{gap} = 40nm$, and the optimal point is shown as the purple star in (b). (b) Throughput versus gap distance and control voltage, where each point is obtained from optimal Vdd and number of logic gates.

The comparison of the optimal device-level parameters and performance metrics between CMOS- and GPNJ-based cores are shown in Table 11. The GPNJ core with sharp edges can provide a 2.1X throughput improvement compared with its CMOS counterpart because GPNJ devices can offer a larger driving current without sacrificing too much leakage performance. At the optimal design point, the power dissipation for each component is shown in Figure 70, where one can observe that the leakage power percentage for the GPNJ core is more significant than the CMOS core. For the GPNJ core with curved corners, the improvement drops to 66% due to the larger input capacitance and footprint area.

Table 11: System-Level Comparison of Various Optimal Design Parameters and Performance Metrics between Si CMOS and GPNJ Logic Cores.

Metrics	**CMOS**	**GpnJ**	**GPNJ Curved**	**Comparison**
V_{dd} (V)	0.59	0.33	0.34	
V_g (V)	-	0.44	0.41	
D_{gap} (nm)	-	25.20	23.17	
R_{nand} (kOhm)	11.67	4.62	3.98	0.34 X
R_{leak} (MOhm)	6.54	2.47	1.37	0.2 X
Freq (GHz)	4.35	11.09	9.20	2.1 X
N_g (Million)	20.23	15.47	13.31	0.64 X
Throughput (BIPS)	13.86	29.70	23.27	1.66 X

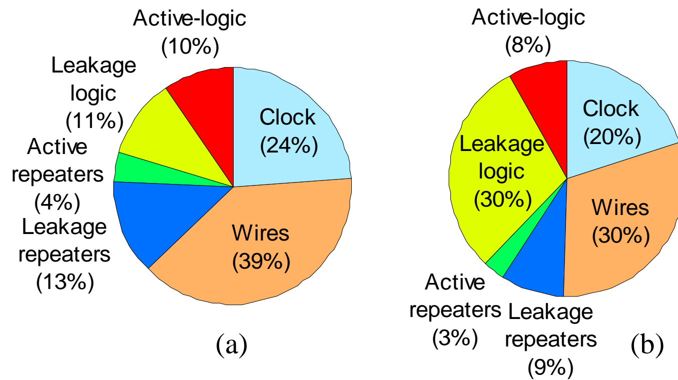


Figure 70: The pie chart of the power consumption of each component for (a) CMOS core and (b) GPNJ core.

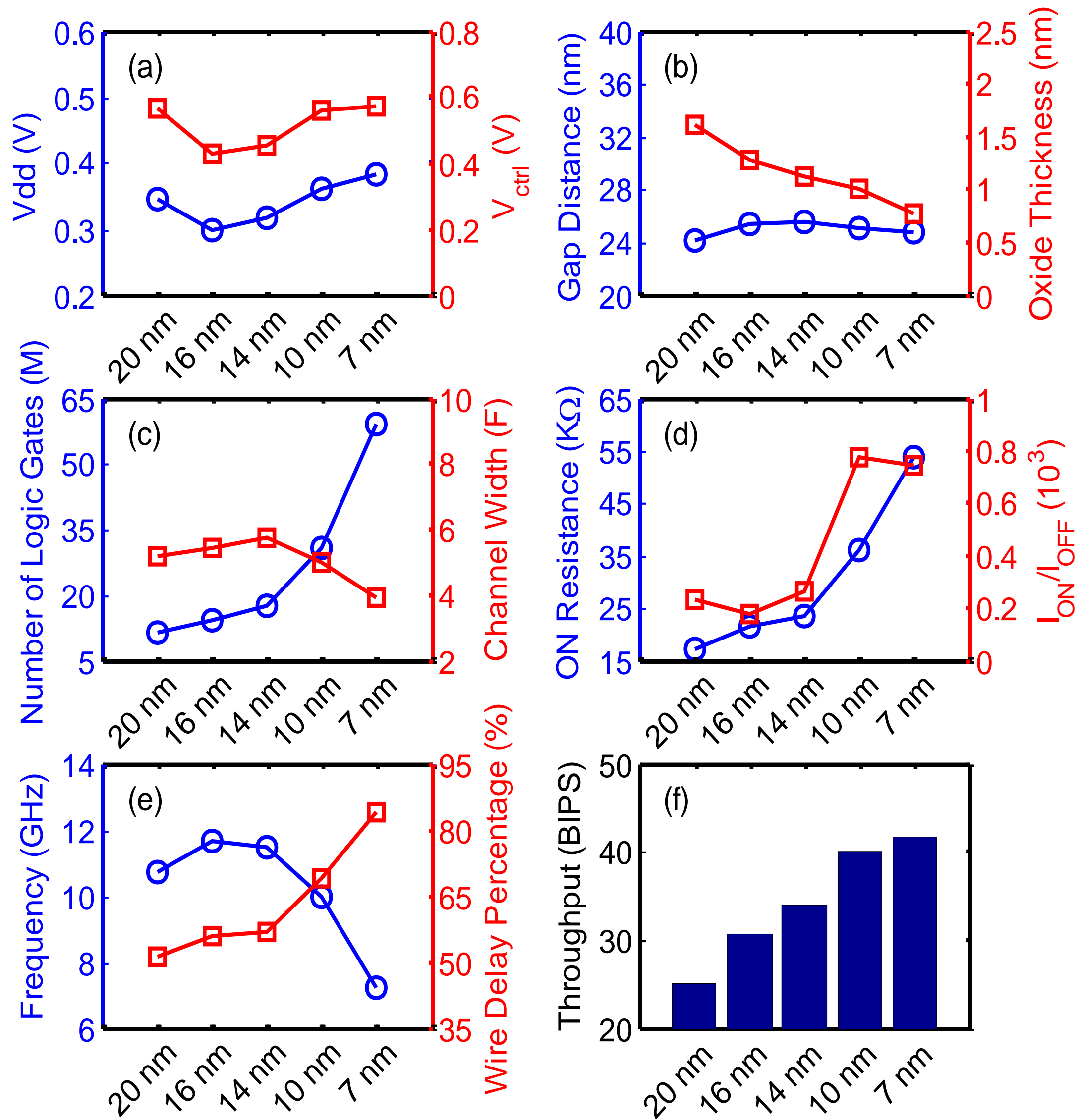


Figure 71: Various optimal design parameters and performance metrics at various technology nodes for a single-core GPNJ processors.

In Figure 71, various design parameters and performance metrics are shown down to the 7 nm technology node for a single GPNJ logic core with a 5 mm^2 die area. The supply voltage starts increasing when the technology node is below 16 nm, which is counterintuitive. The reason is for a fixed core area, the interconnection plays a more important role at a smaller technology node. Therefore, a higher supply voltage is

preferable to reduce the interconnect delay. Due to the fact that the leakage increases exponentially as the gap distance decreases, the optimal gap distance remains at around 25 nm as the technology scales.

At small technology nodes, a lot more logic gates can fit into the same core area so that the complexity of the interconnection network increases dramatically, causing the increase of the interconnect delay in the critical path and the drop of the optimal clock frequency. The ON resistance keeps increasing due to the smaller width of the GPNJ devices, but the ON/OFF ratio improves. Another observation is that there is only 3.7% of the throughput improvement from 10nm to 7 nm technology node. This is because the core area is too big that the throughput enters the saturation region, which is similar to the behavior observed for a FinFET core, shown in Figure 67 (a).

5.3.2 Multi-Core Optimization

For a multi-core processor, the updated results are shown in Figure 72, where two more parameters are optimized, including the number of cores and the logic-to-cache ratio. The memory bandwidth for the 20 nm technology is assumed to be 25 GB/s, and it improves 30% for each new technology generation. Compared with the results shown in the previous single-core analyses, the optimal values of several parameters have different trends. The optimal number of cores keeps increasing as the technology node scales because a large core is not desirable due to the complex interconnection network. In such situation, the supply voltage keeps decreasing instead of going up compared against the previous single-core analysis. The optimal device width does not decrease significantly as the technology scales, indicating that the edge roughness would not cause a significant impact on the GPNJ devices at sub-10 nm technology node. Similar to the single-core analysis, the optimal gap distance does decrease as the technology node goes down because of the severe leakage current at a short gap distance. For each new technology

generation, about 40% of the throughput improvement can be observed. The major improvement comes from the decrease of the area overhead of the GPNJ devices and interconnect pitches, and the increase of the number of metal layers and bandwidth.

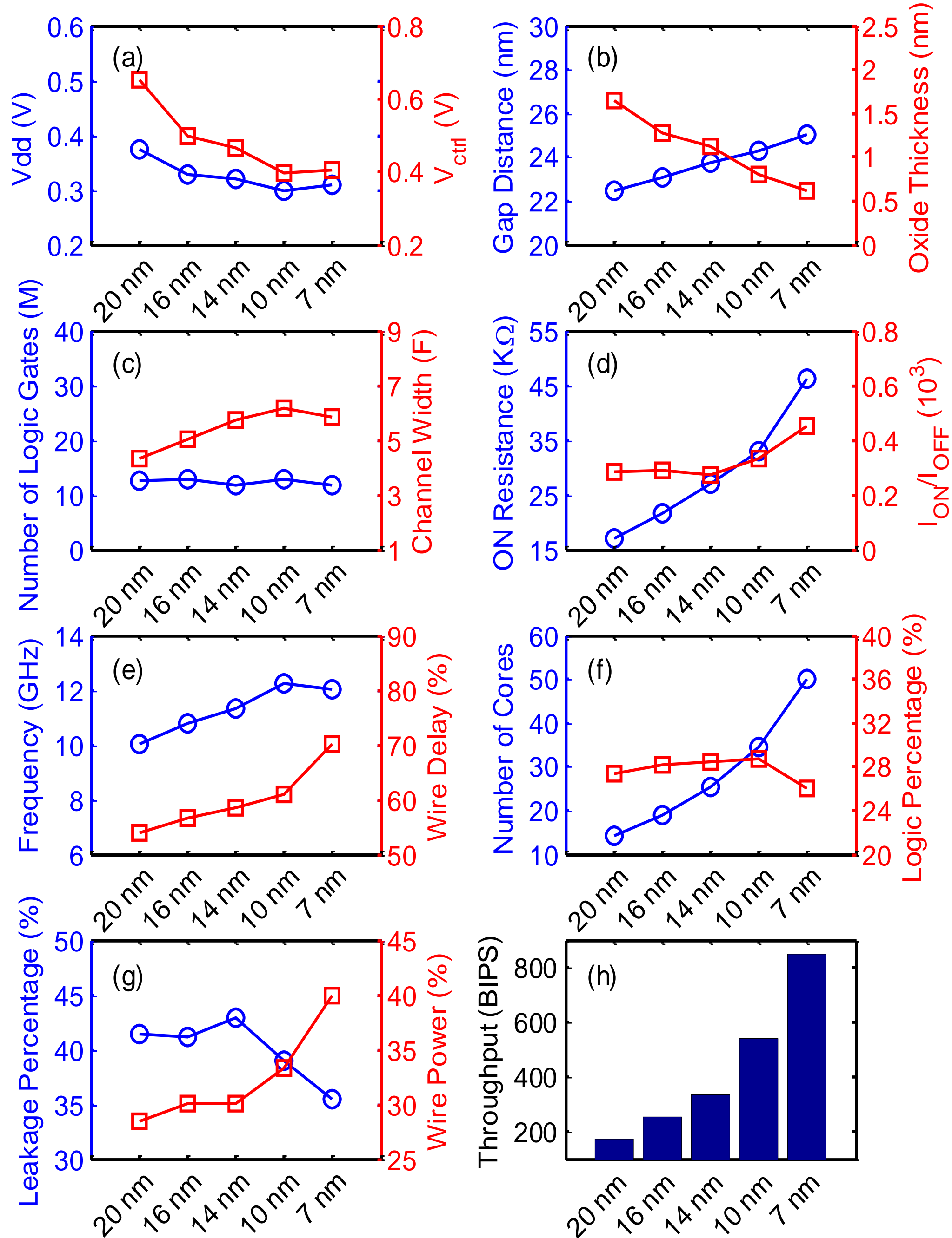


Figure 72: Various optimal design parameters and performance metrics at various technology nodes for a multi-core GPNJ processors.

Another observation is that the optimal width, shown in Figure 72 (c), is several times larger than the minimum feature size based on the power/area/frequency trade-offs in the proposed design methodology. At the 20nm technology node, the width is about 5×20nm = 100nm. Even at 7 nm technology node, the width is about 6×7nm = 42nm. This finding is quite important because transverse quantization and edge scatterings that become prominent in very narrow ribbons have not been considered and can adversely affect both on and off currents.

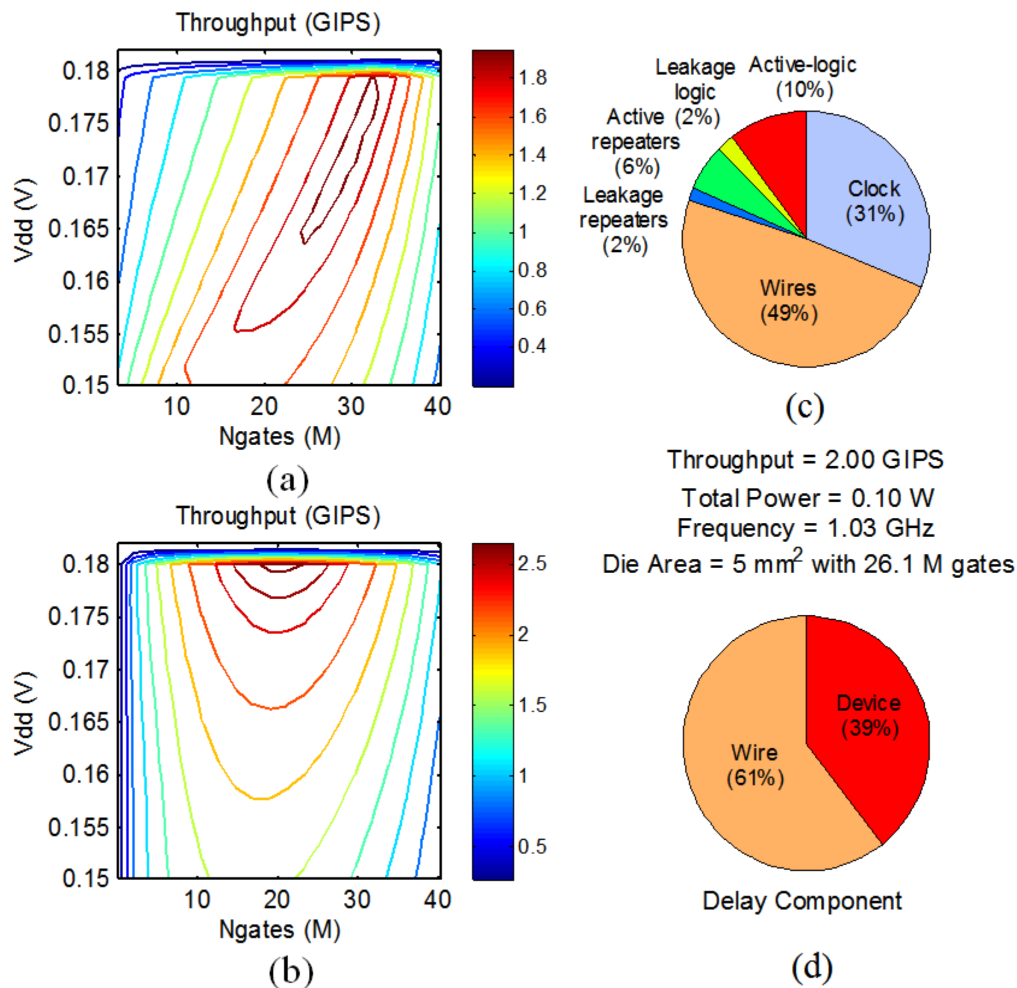


Figure 73: System-level throughput optimization for a TFET single core, given a fixed core area as 5 mm^2. (a) and (b) show the throughput versus the supply voltage and the number of logic gates under 2 W/cm^2 and 10 W/cm^2 power density constraints, respectively. (c) and (d) show the power and delay components at the optimal throughput design point with 2 W/cm^2 power density constraint.

5.4 Tunneling Devices

Based on the hierarchical optimization engine and device-level models developed in Chapters 2-4, a 5 mm^2 TFET single core is optimized. The optimal throughput of the single core versus the supply voltage and the number of logic gates under two power density budgets are shown in Figure 73 (a) and (b). The optimal throughput points can be observed on both plots, indicating that the optimal V_{dd} increases as the power density budget increases. However, beyond 0.18V, the leakage current passing through the drain-channel junction increases drastically. With a 2 W/cm^2 power density budget, about half of the power and 61% of the delay at the optimal point are contributed by the interconnects, as shown in the pie charts in Figure 73.

If one varies the core size, Figure 74 shows that the optimal frequency for a low-power CMOS core is much lower than that of a TFET core because under the same power density constraint, the CMOS core consumes a larger dynamic power due to its higher supply voltage even though the leakage is very small compared with the TFET core. Due to the low I_{ON} of the TFET, a larger number of repeaters are required to satisfy the timing constraint at higher metal levels compared with the CMOS system. One can observe a significant throughput advantage of using TFETs at the given 2 W/cm^2 power density budget.

For various power density budgets, Figure 75 shows the optimal throughput comparison among CMOS high-performance, CMOS low-power, and TFET systems at both low and high power budget ranges, which indicates that a 5mm^2 TFET core can achieve higher throughput below 25 W/cm^2 power density budget compared with a CMOS high-performance core. In addition, the throughput of the TFETs core saturates when the power density is beyond a certain point due to the larger impacts of the critical paths and delays of long interconnects.

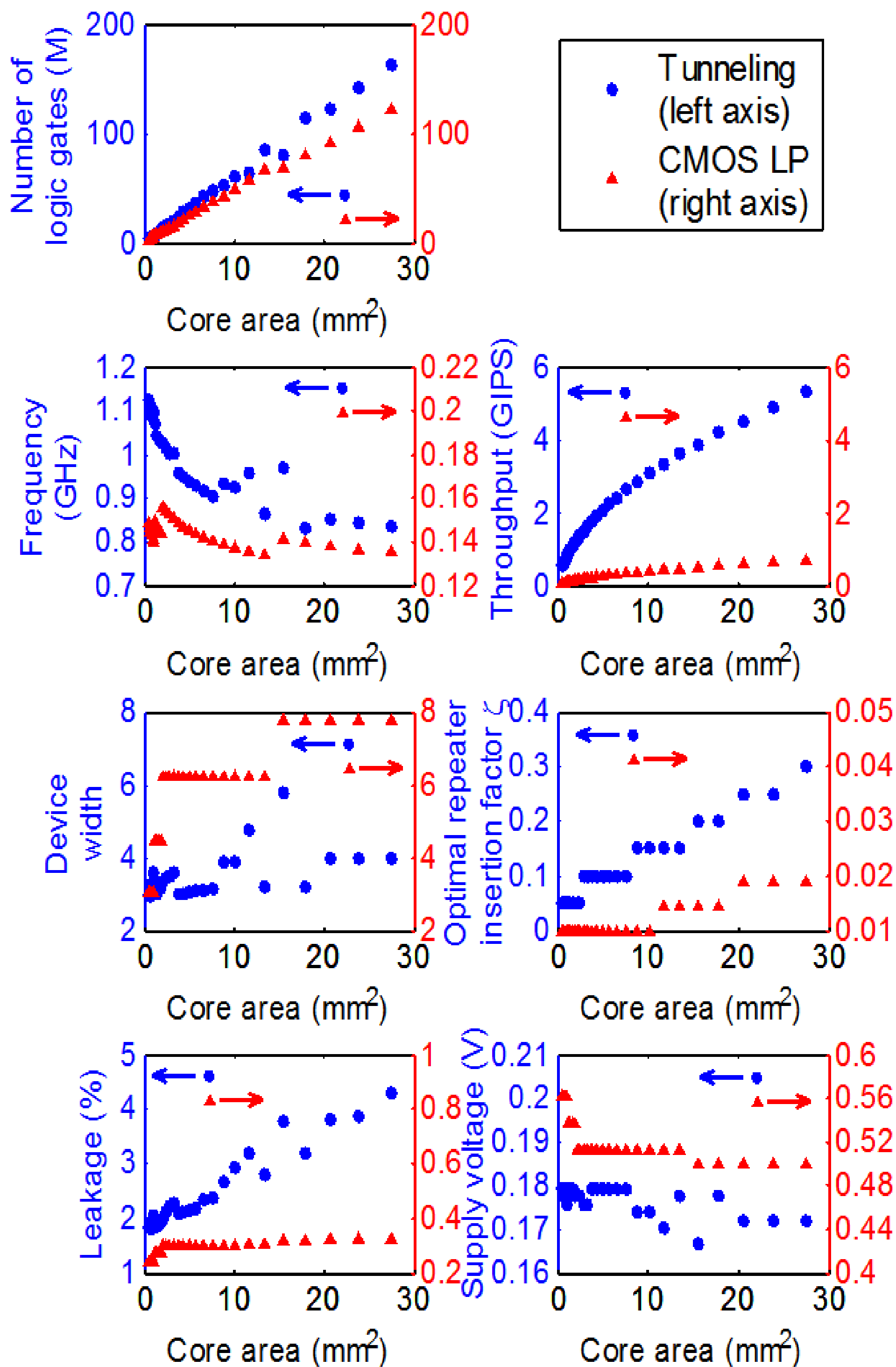


Figure 74: The comparison between the optimized system using TFETs and CMOS low-power devices for various optimized parameters versus the core area with 2 W/cm^2 power density constraint.

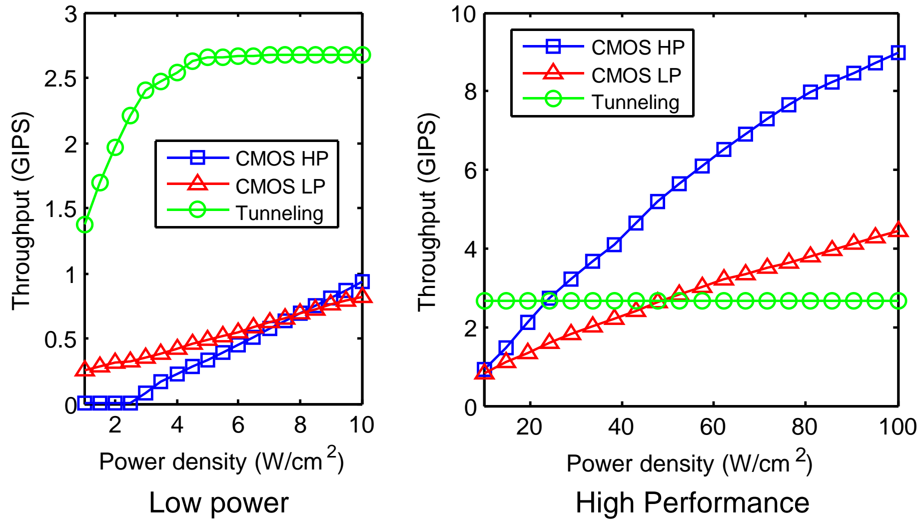


Figure 75: The optimal throughput comparison among TFETs, CMOS high performance, and CMOS low power devices under various power density budgets, given a 5 mm^2 core area. (a) low power density range, (b) high performance range.

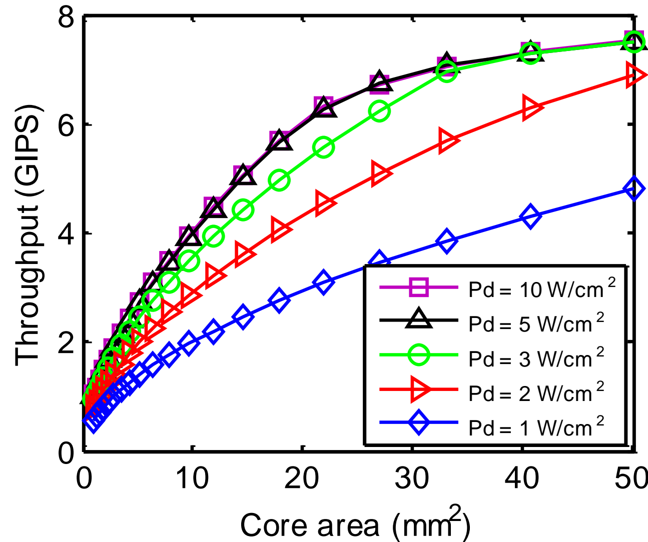


Figure 76: The optimal throughput of a single core using tunneling devices versus core area at various power density budgets.

For various core areas, Figure 76 shows no further improvement if the power density is larger than 5 W/cm^2 because the additional power budget is wasted in the large leakage power as the supply voltage goes above 0.18V. Moreover, the saturation point comes

early for a larger core. If one sets a total power budget, optimal core areas can be observed in Figure 77, because a smaller core cannot fully utilize the given power and a larger core requires longer interconnects.

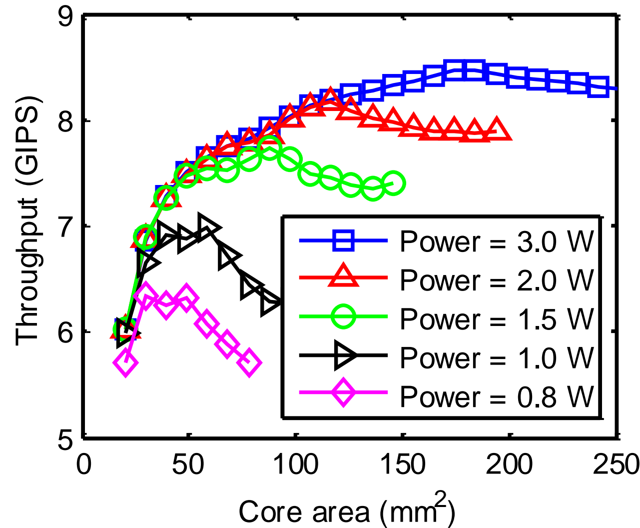


Figure 77: The optimal throughput of a single core using TFETs versus core area at various total power consumption budgets.

5.5 Optimization under Process Variation

In Section 4.2, a CMOS logic core is optimized at the nominal state without the process variation. Under the process variation, a throughput distribution can be obtained for each design point, according to the circuit-level leakage and delay distributions with a given device-level random and systematic variation σ/μ of 3% and 5%, respectively [48, 141]. To satisfy the power density constraint, two post-silicon tuning methods are applied, including frequency tuning and adaptive supply voltage tuning, to ensure that the power constraint is not violated. Once the throughput distributions are generated for each design point, one can choose the best design point that provides the maximal throughput under a given yield target. Here, the yield is defined as the percentage of the processors that can operate higher than the given frequency target.

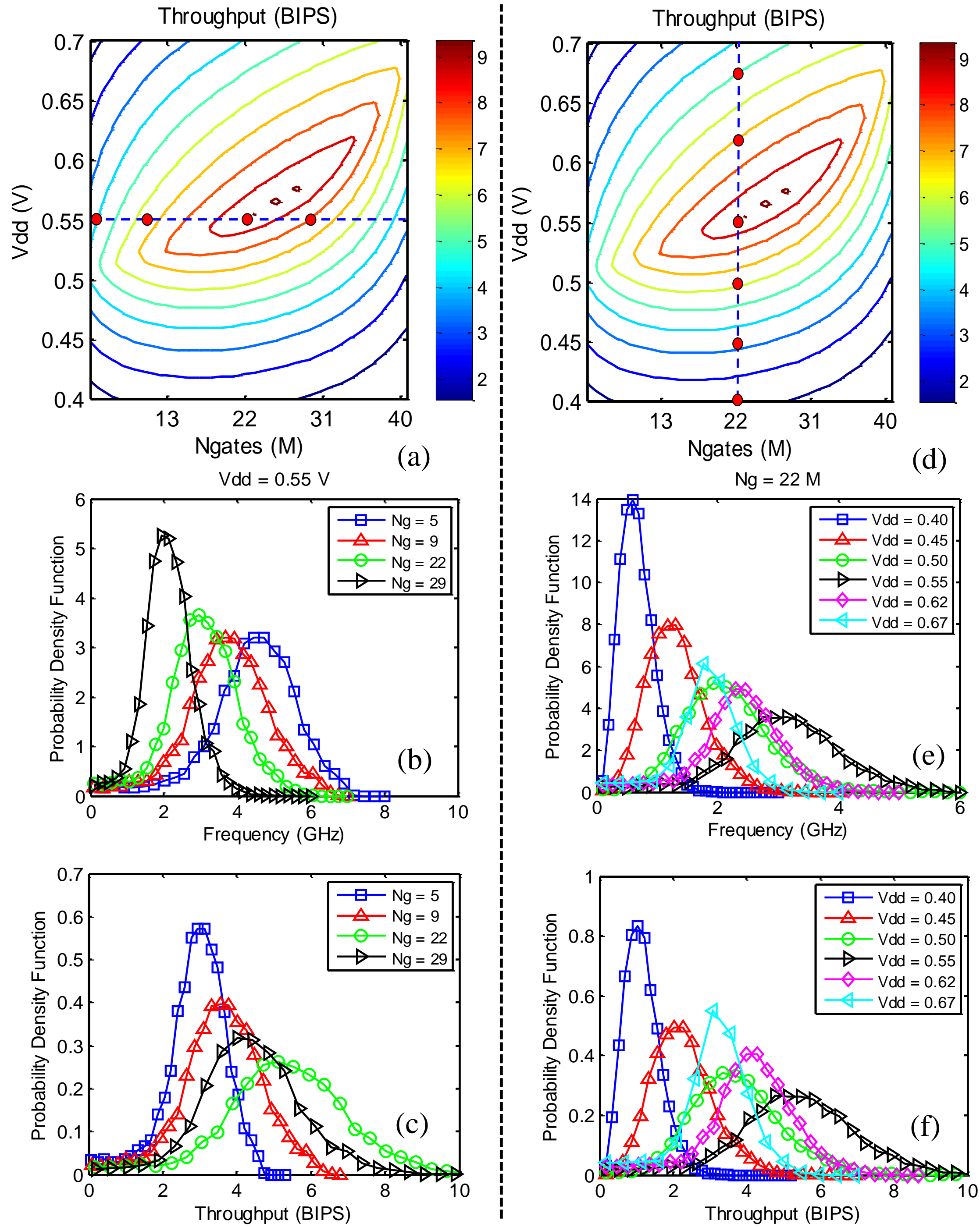


Figure 78: (a) and (d) show the optimal throughput versus the supply voltage and the number of logic gates at the nominal state, where the red points are the design points being investigated in the bottom four figures. (b) and (c) show the frequency and throughput distributions for various number of logic gates at vdd = 0.5V. (e) and (f) show the frequency and throughput distributions for various supply voltages at N_g = 22M.

5.5.1 Single-Core Optimization

In this subsection, a single core is optimized based on the 16nm planar CMOS technology with 5 mm^2 core area and 100W/cm^2 power density constraints. Figure 78 (a) and (b) show the optimal throughput versus the supply voltage and the number of logic gates at the nominal state. After the process variation is included, to evaluate the throughput distribution, one can take samples from the distributions of the critical path delay, the longest signal path delay, and the leakage power consumption, and put them into the optimized single core. Each point in Figure 78 (a) and (d) now becomes a throughput distribution instead of a fixed nominal value. The blue dash lines with red dots shown in Figure 78 (a) and (d) represent the frequency and throughput distributions at a given supply voltage and a fixed number of logic gates. For a given supply voltage of 0.55V, due to the increasingly complicated interconnect network, the frequency distribution, shown in Figure 78 (b), shifts to the left as the number of logic gates increases. However, from the throughput distribution shown in Figure 78 (c), an optimal curve is observed to achieve the maximum throughput at 22 million logic gates. This is because when the number of logic gates is small, the throughput is dominated by the CPI, which keeps decreasing as the number of logic gates increases. But when too many logic gates are put into the system, they introduce a significant penalty in maximum frequency due to the complicated interconnection network, causing the decrease of the overall throughput. For a given number of logic gates, the frequency distributions are also obtained for various supply voltages, shown in Figure 78 (e). An optimal supply voltage also exists to maximize the frequency distribution. Because when the supply voltage is too high, the leakage power increases dramatically, forcing the frequency to decrease to satisfy the power density constraint even though the maximum allowed clock frequency is high. On the other hand, if the supply voltage is too low, the high ON resistance significantly increase the RC delay, causing a low maximum frequency. Another

observation is that the deviation of the frequency distribution decreases as the supply voltage drops, which is counter-intuitive for the circuit-level results. That is because at the system-level optimization, the design points with large supply voltages are tremendously affected by the leakage deviation. To satisfy a given power density budget, the frequency needs to be adjusted according to the leakage power, which induces an additional deviation on top of the intrinsic frequency deviation. Since the number of logic gates is fixed, the throughput distributions, shown in Figure 78 (f), have the same shape as the frequency distribution. These distributions in Figure 78 are based on the power density cap of 100 W/cm^2. Only the frequency is adjusted to ensure the power density is below the constraint. If the yield is set as 80%, Figure 79 shows the new throughput versus the supply voltage and the number of logic gates. The top region is empty because the leakage due to variation is so high that more than 20% of the cores cannot operate. In addition, the optimal point shifts to the left bottom corner because the core designed at a smaller supply voltage with fewer logic transistors has a smaller deviation according to the throughput distributions shown in Figure 78 (c) and (f).

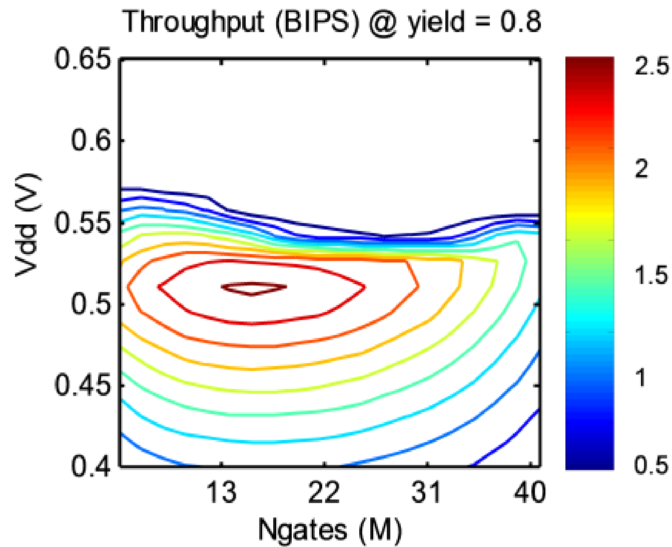


Figure 79: Throughput versus supply voltage and the number of logic gates for a 5 mm2 core with 100 W/cm2 power density constraints based on 80% yield implemented by frequency tuning.

If one adjusts not only the frequency but also the supply voltage at each design point, Figure 80 shows a significant improvement in throughput at 80% yield compared with Figure 79. If the leakage power of a core is too high, using supply voltage tuning not only reduces the operation frequency but also simultaneously reduces the subthreshold leakage.

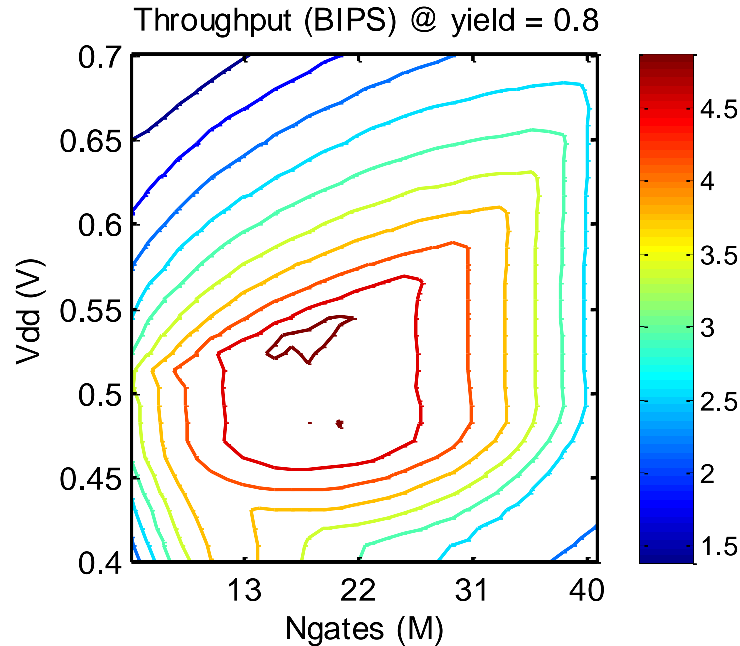


Figure 80: Throughput versus supply voltage and the number of logic gates for a 5 mm2 core with 100 W/cm2 power density constraints based on 80% yield implemented by adaptive supply voltage.

Figure 81 shows the leakage power versus the frequency at the optimal point in Figure 80. The red dots are the samples applied with adaptive supply voltage with a 100 W/cm^2 power density constraint and the blue dots are the raw data without applying any leakage control techniques. Therefore, some of the blue samples on the upper side of the figure violate the power constraint due to the large leakage current. To improve the overall yield, they are shifted to the left bottom direction. And the blue samples at the left bottom corner that consume less power are applied with a higher supply voltage, leading them to shift to the top right direction so that the overall performance increases. Figure 82 compares the cumulative throughput distributions for these two tuning methods,

indicating the advantages of using adaptive supply voltage technique. If the performance target is set as 4 BIPS, the yield can be improved from 68% to 87%; if the yield target is 80%, then the throughput can increase from 2.5 BIPS to 4.9 BIPS.

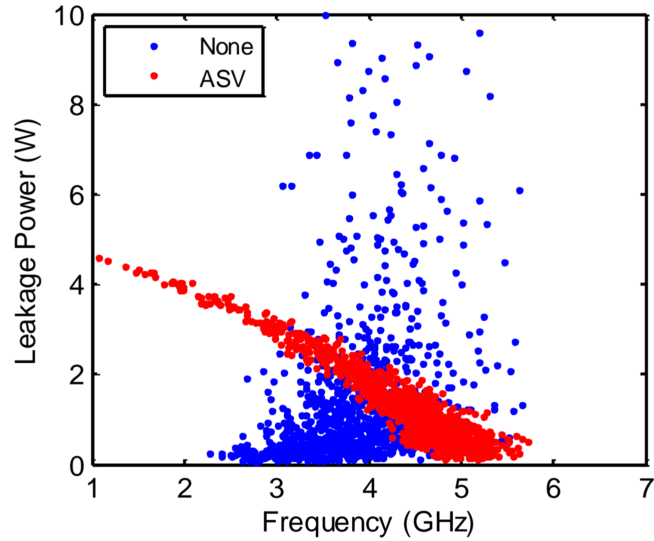


Figure 81: Leakage power versus the frequency of a core. The red dots are the samples applied with adaptive supply voltage with 100 W/cm^2 power density constraint and the blue dots are raw data.

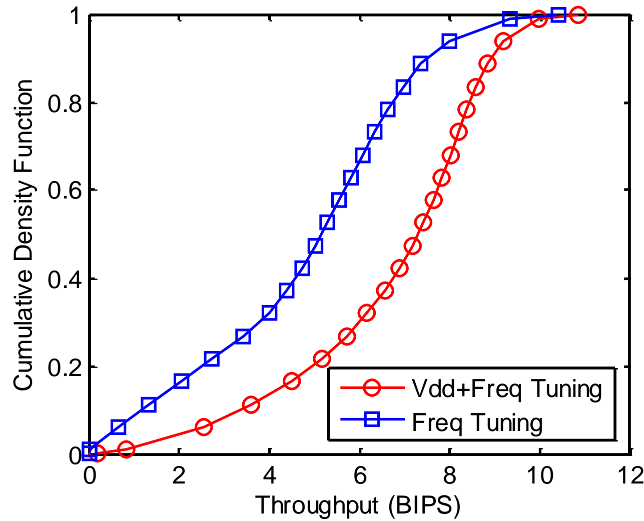
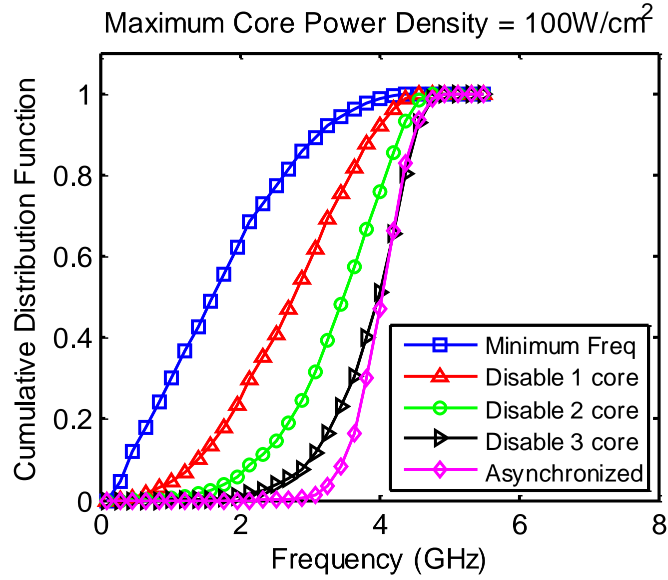


Figure 82: Comparison of cumulative throughput distributions for the core implemented with two post tuning methods.

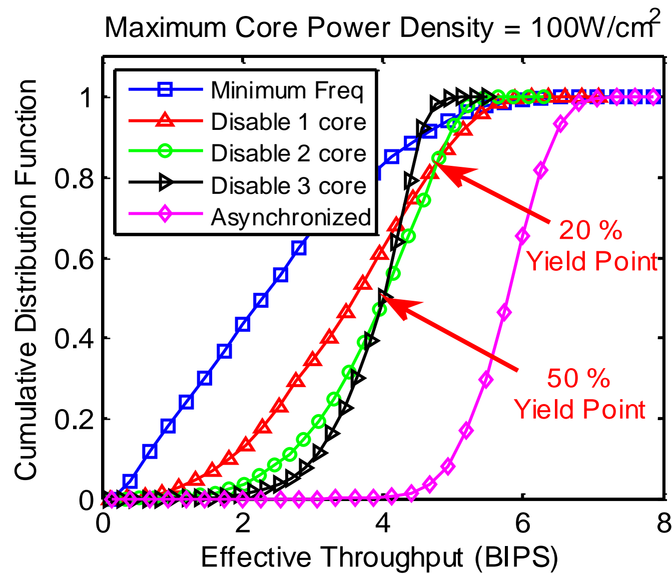
5.5.2 Multi-Core Optimization

To analyze the variation impact on a multi-core processor, a system with a total 50 mm^2 die size area constituting 10 cores is assumed. Multi-core processor samples are taken based on the throughput distribution obtained from the single-core design. Due to the systematic WID variation, some cores may operate faster and violate the power density constraint while other cores may operate slower and are unable to fully utilize the power budget. Therefore, from the comparison shown in the previous subsection, the better technique, adaptive supply voltage method, is applied to ensure that each core can work properly under the power density constraint of $100W/cm^2$.

For a synchronous multi-core processor, by disabling the slowest cores, Figure 83 shows the cumulative frequency and throughput distributions, illustrating that the frequency of the lowest core keeps increasing as more cores at worse process corners are disabled. At the same time, the deviation due to the variation decreases. However, because of the performance penalty due to the fewer active cores, the cumulative throughput distributions shift to the left. If one wants to achieve a higher yield, more disabled cores are required as the black curve indicates; if one wants to pursue a higher chip throughput by sacrificing the yield, fewer cores need to be disabled as the red curve reveals. For such a synchronous processor, the slowest cores at worse process corners consume more power than the cores at better process corners. If the system supports asynchronous operation and each core can operate at their maximum power budget, another 40% improvement in the average throughput can be achieved according to Figure 83 (b).



(a)



(b)

Figure 83: Cumulative distribution function for frequency and effective throughput per core of a multi-core processor with the assumption that the maximum power density for each core is 100 W/cm^2.

For an asynchronous processor with a 100 W/cm^2 average power density constraint, each core can operate at its maximum frequency, which offers a higher throughput

compared with the synchronous processor. Here, a power reallocating technique is proposed to even better utilize the fixed total power budget. If the processor has an excellent lateral thermal conductivity by using some advanced cooling methods [142, 143], this technique can be achieved by reducing the power budget for the cores at worse process corners and allocating more power to those at better process corners. Since the relation between power consumption and frequency can be obtained from the optimizer for each core, optimal clock frequencies can be obtained individually by using nonlinear programming [144]. The problem can be stated as follows:

$$\max \quad TP(\mathbf{f}) = \sum_{i=1}^{n} Perf(f_i)$$

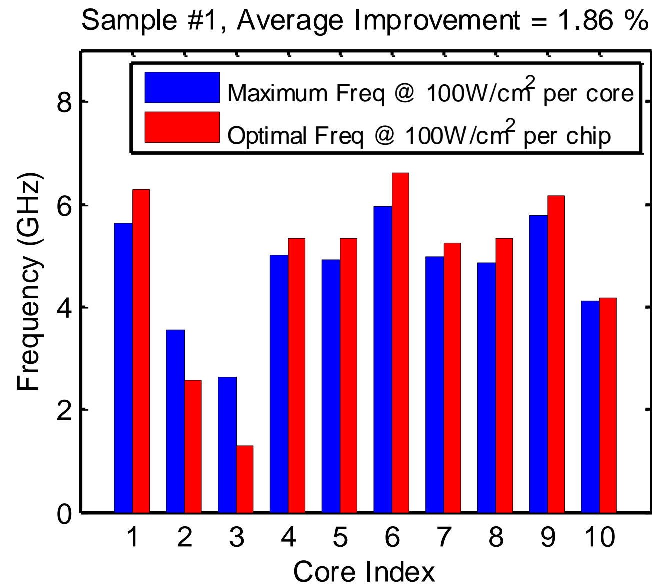
$$\text{subject to} \quad P_{tot}(\mathbf{f}) = \sum_{i=1}^{n} P_i(f_i) \leq P_0$$

$$f_i \geq 0, \quad i = 1, \ldots, n$$

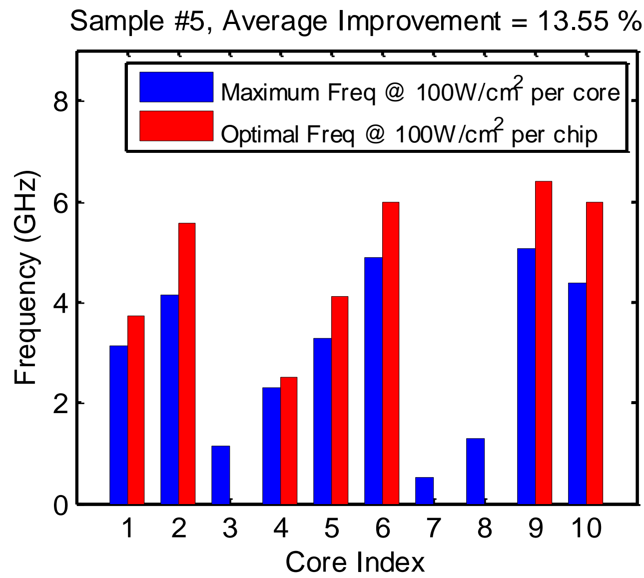
where TP is the overall chip throughput, which is the summation of the throughput of each core, *Perf*, assuming a fully parallel operation, P_{tot} is the total power consumption on chip, P_i is the function that relates the clock frequency and the power for the ith core, f is the clock frequency, and P_0 is the total power budget. After reallocating the power to each core, one can obtain the improved chip throughput for each multi-core processor sample. Finally, the overall chip throughput distribution can be generated and a comparison can be made among various variation control techniques.

From one of the sample results shown in Figure 84 (a), the frequency of the two slowest cores are reduced, and those power savings are reallocated to the other cores, especially to those at better process corners that can operate at a higher clock frequency. However, the average improvement is less than 2% due to the fact that this multi-core processor sample has a small deviation. If one inspects at another sample shown in Figure 84 (b), three cores that operate at much worse process corners are completely turned off

so that cores #2, #6, #9, and #10 gain a considerable frequency boosting, and the overall performance improvement reaches 13.6%.



(a)



(b)

Figure 84: Comparison of the frequency between two multi-core processors with and without applying the power reallocating technique for two samples (a) and (b).

To fully explore the potential benefits of these two power management techniques, 1000 processor samples are taken and the comparison of the frequency distributions is shown in Figure 85. The major frequency improvement after implementing power reallocating is at low performance corner and the mean throughput improvement is 5.4% compared with the original asynchronous system, where each core operates at its own maximum frequency. If the performance target is set as 4 GHz, the yield can be improved from 80.5% to 94.8%; if the yield target is 90%, then the average frequency can increase from 3.8GHz to 4.1GHz.

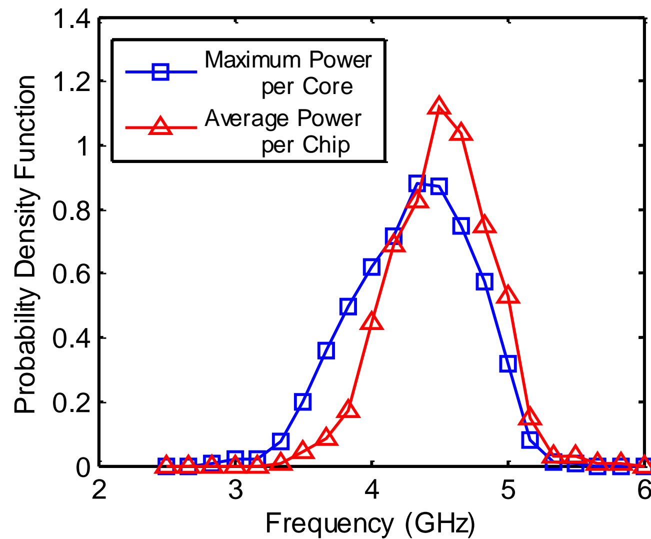


Figure 85: Comparison of the probability density function for the average frequency between two asynchronous processors.

As the number of cores increases, the overall performance improvement keeps increasing as the bar chart shows in Figure 86. That is because a system with more cores suffers more from the systematic variation, causing a larger performance/power asymmetry between the fast and slow cores. Therefore, it benefits more from using the power reallocation technique.

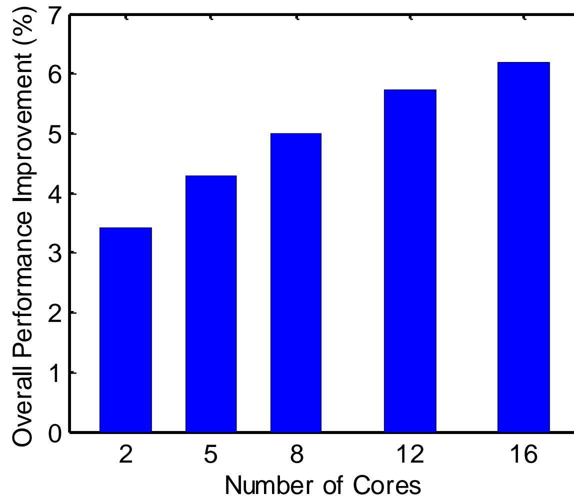


Figure 86: Overall performance improvement versus the number of cores in the multi-core processor.

5.6 Conclusions

In this chapter, benchmarking and optimization are performed for the conventional Si CMOS and two emerging device technologies based on the device-, circuit-, and system-level compact models and a hierarchical optimization engine developed in Chapters 2-4.

For the conventional Si CMOS devices, multiple device- and system-level design parameters are simultaneously optimized to maximize the chip throughput for a given device technology and an architecture family under certain power, thermal and die size budgets. In the single-core processor analysis, a high-performance 25 mm^2 FinFET processor can provide a 26% higher throughput than its planar high-performance CMOS counterpart, and a low-power TFET processor can offer more than 2X improvement in throughput compared with its low-power FinFET counterpart at the 16nm technology node. For various technology nodes, an accurate power-law relation is observed between the throughput and the die size area for a relatively small processor. However, the

optimal throughput saturates when the core area is beyond certain point, which indicates that a large core is not desirable.

For the GPNJ-based processors, the proposed design methodology is applied to efficiently perform device-, circuit-, and system-level co-optimization. For given power density and die size area budgets, various device-level parameters, including supply voltage, control voltage, gap distance, and oxide thickness, are optimized for a GPNJ core, where 2.1X throughput improvement is observed for a sharp-corner GPNJ core compared to its Si CMOS counterpart implemented at the 16nm technology node. This advantage is predominantly because of the smaller output resistance, which reduces both device and interconnect delay and saves the power for repeaters. For a curved-corner GPNJ core, the throughput improvement drops to 66% due to the larger input capacitance. For the scalability analysis, various design parameters are optimized, and the trends are shown down to the 7 nm technology node for both single- and multi-core processors.

For the TFET based processors, the results indicate that TFETs have excellent performance at the low power density range due to the low supply voltage. The limitations imposed by the interconnects and the large leakage current at high supply voltage restrict the driving current, leading to a lower performance of the TFETs at a high power density and a large core size compared with CMOS. The larger number of logic gates and higher clock frequency of the TFET processor eventually bring more than 2X throughput improvement compared with its low-power FinFET counterpart, indicating an outstanding performance for the ultralow-power applications.

For the device variation, multiple leakage and variation control techniques are investigated, and an optimization is performed under various power, area and yield targets. The results indicate that using adaptive supply voltage and frequency tuning techniques significantly improve the throughput and yield for a given power density constraint. For a multi-core processor, disabling the slowest cores helps to reduce the

deviation and increase the chip throughput. A system with asynchronous operations can provide the maximal throughput since each core runs at its own operational frequency. The proposed power reallocating technique can achieve further throughput improvement by allocating the power to those more power efficient cores. For a 10-core processor, the results show that the yield can be improved from 80.5% to 94.8% at a 4GHz frequency target; if the yield target is 90%, then the average frequency can increase from 3.8GHz to 4.1GHz.

CHAPTER 6 INTERCONNECT TECHNOLOGY OPTIMIZATION

6.1 Introduction

Power dissipation is one of the major limitations for high-performance processors. As the technology node scales down, the power consumption associated with the interconnects increases. At the 130nm technology node, a processor was reported to consume more than half of the total power on its interconnects [145]. To reduce the interconnect power, lowing the dielectric constant k of the ILD and reducing the interconnect aspect ratio are the top two most effective ways. However, for the first approach, the scaling of k for interconnect dielectrics has largely slowed down due to the severe challenges associated with manufacturability, cost, and reliability issues; for the second approach, it induces significant overhead of interconnect resistance due to the size effects as the semiconductor industry approaches sub-20 nm technology nodes [146]. These size effects include the surface and grain boundary scatterings, which are expected to dramatically increase the effective resistivity of copper interconnects [10]. Moreover, diffusion barriers, which are poor conductors, will take an ever-increasing fraction of the interconnect volume. To deal with these interconnect scaling problems, two novel interconnect structures and architectures are proposed, and the potential benefits are quantified in this chapter.

To overcome the scaling challenges of the conventional copper interconnects, carbon-based interconnects, such as graphene sheets and carbon nanotubes, are potential candidates because of their outstanding electrical properties, including a long electron

MFP, a large current conduction capacity, and a small capacitance per unit length. However, since graphene is a two-dimensional structure, increasing the interconnect pitch does not lower the resistance as fast as it does in copper interconnects. Hence, comparing graphene and copper interconnects strongly depends on the interconnect pitch. On the other hand, multilevel interconnects accommodate interconnects of various lengths routed in metal levels with different wiring pitches. As a result, system-level analyses are essential to better understand and evaluate the overall benefits of graphene interconnects.

Interconnect variation, which is contributed by lithography, alignment, etching, polishing, and orientation, also has a large impact on the chip performance [13, 147]. With the scaling of the technology, double patterning appears to be one of the cost-effective advanced patterning options [39]. One conventional double patterning technique is the litho-etch-litho-etch (LELE) process [40]. In this approach, two sequential steps of lithography and etch processes are performed, which make the LELE process suffer from the overlay variation, especially at a small technology node [40]. More recently, a more advanced double patterning technique, self-aligned double patterning (SADP) process, is developed to deal with the overlay variation by using a pattern-mask and a block-mask [41, 42]. As the technology scales down to sub-10nm technology nodes, the triple or even quadruple patterning schemes are required due to the single-exposure resolution limitation [43]. Recently, the self-aligned quadruple patterning (SAQP) process is proposed and demonstrated [43]. However, this process needs two steps of sidewall spacer processes, which makes it more sensitive to the CD spacer variation. Therefore, it is crucial to quantify the impact of the CD spacer variation, especially for the SAQP process, and compare it with other patterning options, assuming that the extreme ultra violet (EUV) may become available in the future.

To analyze the interconnect variation, previous work has focused on its impact on the interconnect resistance and capacitance [39, 148]. However, since the CD variation

for the line/core and spacer affects the resistance and capacitance of the interconnect in opposite directions, the intrinsic interconnect delay or a simple path delay cannot truly evaluate the impact of the variation. Both the driver resistance and capacitance as well as the length and pitch of the interconnects affect the variation impact significantly. For instance, a path with a large driver resistance and a small input capacitance, the delay will be dominated by the interconnect capacitance variation; if the driver resistance is small, then the resistance variation of the interconnect will have a larger impact. The benefits of various types of fabrication processes can be fully evaluated only though the system-level analysis of a processor in terms of the overall clock frequency. Previous research has focused on the device- and circuit-level interconnect variation analyses [39], [149]. In this chapter, for the first time, the impact of the interconnect variation on the system-level performance is quantified. Five sources of interconnect variations are analyzed, including CD line/core, CD spacer, etch, CMP, and overlay variations. The impact of these individual variations on the overall clock frequency for a single-core processor using three fabrication techniques are investigated and compared with each other at three technology generations down to the 7nm node.

The rest of this chapter is organized as follow: in Section 2, insightful trends are shown for both devices and interconnects as the technology scales down. Two implications are concluded in Section 3 to address the interconnect challenges at sub-20nm technology nodes. Section 4 proposes one potential solution to rebalance the resistance and capacitance of the interconnect by using the interconnect width that is larger than the half pitch. An observable improvement exists compared with the conventional interconnect structure. In Section 5, a novel Al-Cu hybrid interconnect architecture is proposed to alleviate the ever-increasing size effects at small technology nodes. In Section 6, a novel carbon-based multi-layer graphene interconnect is

benchmarked and optimized at the overall system-level chip throughput, and promising results are shown compared with the conventional copper interconnects.

6.2 Device and Interconnect Scaling Trends

To sustain the growing transistor performance and density, novel device structures have evolved from the conventional planar to multi-gate devices, such as FinFETs, Pi-gate, and Omega-FET devices, and even to the ultimate GAA devices [3, 6]. With each newer device structure, the superior electrostatic operation significantly improves the sub-threshold slope, but the input capacitance also jumps due to the larger intrinsic gate capacitance and parasitics as well as manufactural challenges at small nodes [98]. On the other hand, for the interconnection network, as the metal interconnects scale, the capacitance per unit length almost remains the same, leading to a smaller capacitance per gate pitch. The resistance per unit length, however, increases dramatically because of 1) the smaller cross-sectional area and 2) the severe size effects at sub-20nm nodes, which increases the overall resistance per gate pitch. It has been demonstrated that the surface and grain boundary scatterings largely increase the effective resistivity of copper interconnects [10]. Moreover, diffusion barriers, which are very poor conductors, will take an ever-increasing fraction of the interconnect volume. As a result, the delay associated with the interconnect resistance increases significantly. This chapter, for the first time, reports the shift in the importance of interconnect capacitance and resistance for FinFETs and GAA FETs, and studies the implications of this shift for the design of local interconnect technologies. The trend reported here is in agreement with the analysis presented in [150] that indicated an increase in importance of interconnect resistance compared to interconnect capacitance at the 14nm FinFET technology.

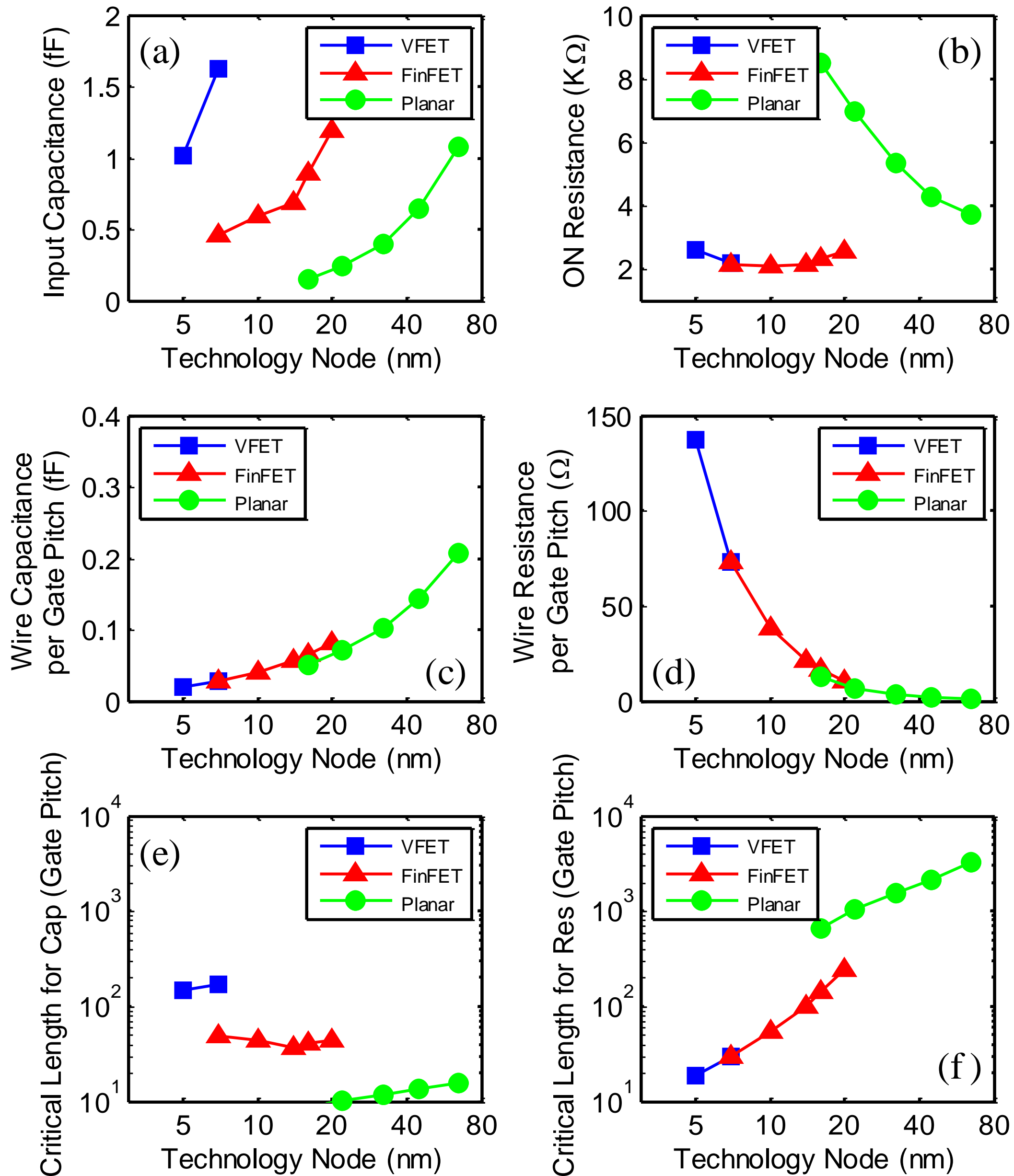


Figure 87: Various device and interconnect metrics for three device structures using copper interconnects at various technology nodes.

The device models for FinFET and planar CMOS transistors follow ASU Predictive Technology Model [67] as indicated in Chapter 2, which provides calibrated SPICE models based on TCAD simulations. The input capacitance and the ON resistance of the inverter are simulated by using HSPICE. The design rule and layout of minimum sized

FinFET devices are adopted from 9 track cell design at the 10nm technology node in [98]. For other nodes, the area is assumed to be proportionally sized. Beyond the 7nm technology node, GAA 3D transistors, such as the vertical field-effect transistor (VFET) devices, are assumed to be available [6]. The input capacitance, ON current, and footprint area can be directly taken from [6] based on the similar ON resistance as a FinFET device. The average fan-out is assumed to be 3, and the M1 pitch is assumed to be four times the technology node. The interconnect resistance model follows previous work [105] with a grain boundary reflectivity R of 0.5 and side wall specularity p of 0.5. The capacitance is obtained using the compact model from [93], which has been validated by a field solver RAPHAEL [94]. Figure 87 shows various device and interconnect metrics for three types of devices [151]. In Figure 87 (a), the input capacitance has a sudden jump with each newer device structure. Since the interconnect capacitance per unit length does not scale, the critical length of the interconnect to match the input device capacitance keeps increasing, shown in Figure 87 (e). However, for the interconnect resistance, the critical length keeps decreasing, and becomes shorter than that for matching the capacitance when the technology node scales below 10nm, shown in Figure 87 (f). This is caused by the increasingly severe size effects of the copper interconnect at small dimensions, such as the grain boundary scattering and edge reflection. Note that the parasitic capacitances between FETs and M1 are ignored in the aforementioned device-level models. Therefore, in reality, all of the input capacitance values would shift up, which makes the interconnect capacitance even less important.

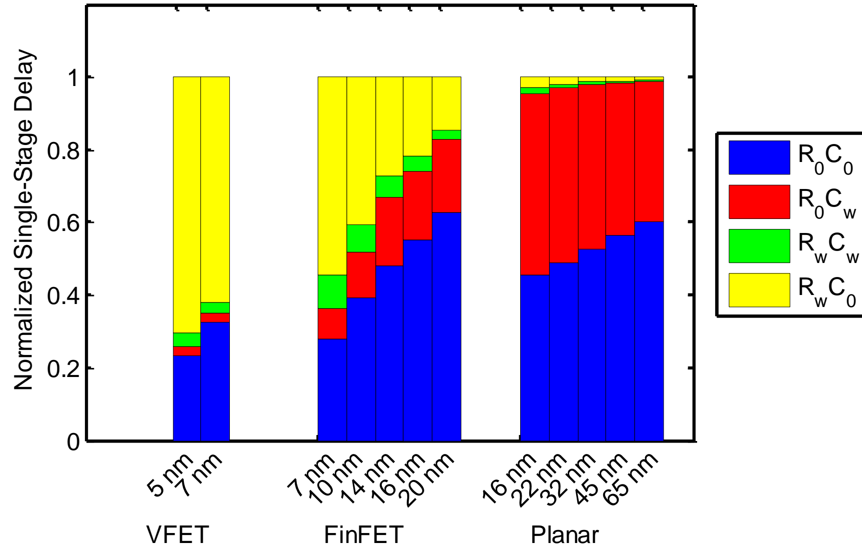


Figure 88: Normalized single-stage inverter delay for three device structures using copper interconnects at various technology nodes.

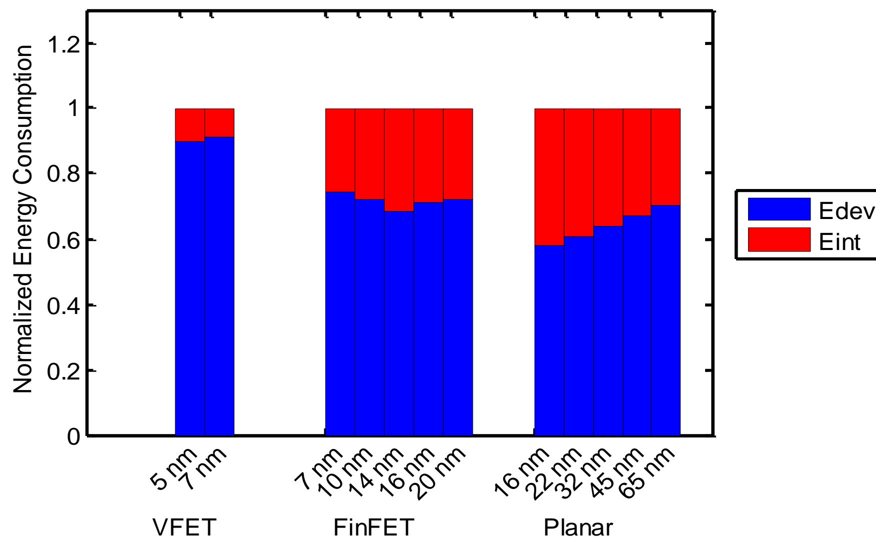


Figure 89: Normalized energy consumption for three device structures using copper interconnects at various technology nodes.

Figure 88 shows the delay breakdown of a single-stage inverter with a fan-out of 3, an average interconnect length of 10 gate pitches, and an aspect ratio of 2. The simulations are performed for three device structures covering 10 technology generations

using twice minimum-sized devices. The upper two stacks (yellow and green) and the middle two stacks (green and red) represent the delay associated with the interconnect resistance and capacitance, respectively. One clear trend is that the delay due to the interconnect resistance increases significantly as the technology scales, while the interconnect capacitance contributes to an ever decreasing portion. Figure 89 shows the normalized energy consumption breakdown for the local interconnection network and the switching energy of devices. It is assumed that 1) the circuit is made of 10 million gates; 2) the local interconnect consumes four metal levels with a wiring efficiency of 0.3; and 3) the placement efficiency is 0.5. Since the gate capacitance of the devices increases for each new device structure as shown in Figure 87, the energy associated with the local interconnects keeps decreasing as the technology scales.

6.3 Interconnect Implications

The trends in Figure 88 and Figure 89 have major implications for design and development of local interconnect technologies. First, lowering the dielectric constant and the interconnect capacitance at future technology nodes may not be as important as it has been in the past. Dielectric materials with slightly larger dielectric constants but significantly better mechanical and reliability properties may therefore be better options.

It is also greatly helpful if the interconnect resistance can be improved even if its capacitance increases. A conventional method to achieve this is to increase the interconnect aspect ratio. This helps to reduce the interconnect resistance at the cost of larger line-to-line capacitance. However, making large aspect ratio interconnects at small dimensions induces further difficulties in the fabrication processes, including etch, clean, and metal deposition [152], which cause non-ideal sidewall angles and induce even smaller Cu grain size at the bottom of the trenches. Moreover, the reliability degrades due

to the electromigration from the defective metallization and the TDDB from smaller top line-to-line spacings.

6.4 A Modified Interconnect Structure

Figure 90 shows a potential solution to balance the interconnect resistance and capacitance by increasing the interconnect width while keeping the interconnect pitch unchanged. This structure provides a smaller interconnect resistance without compromising the wiring efficiency. Since the input gate capacitance dominates the overall capacitance as the technology scales, the overhead of the larger line-to-line capacitance in delay and energy is small.

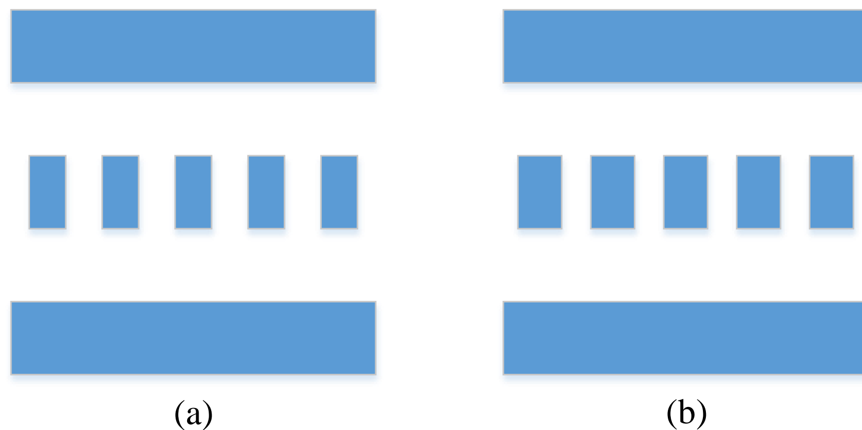


Figure 90: The cross-sectional view of the interconnect structure with a constant pitch and a width of (a) half pitch and (b) 0.6×pitch.

Based on the device and interconnect models described in Chapters 2 and 3, the relative delay versus the interconnect width is shown in Figure 91 for both VFET and FinFET circuits at the 7 nm technology node. An optimal relative interconnect width exists to achieve the minimum delay for both VFET and FinFET circuits, because as the width increases, the interconnect resistance decreases and shrink the yellow stacks. Meanwhile, the interconnect capacitance increases due to the smaller spacing and larger

line-to-line capacitance, leading to larger red stacks. Therefore, if the interconnect width is too large, the capacitance becomes dominant and overshadows the benefits of the small resistance. At the optimal width, the delay of the VFET circuit improves about 17%, larger than that of the FinFET, because the VFET circuit is more dominated by the device capacitance rather than the interconnect capacitance. Note that at the optimal relative width of 1.6 in Figure 91 (a), the interconnect may encounter major reliability issues such as the aforementioned TDDB, especially at the top line-to-line spacings; nevertheless, at a suboptimal design point, such as a relative width of 1.4, a noticeable improvement still exists.

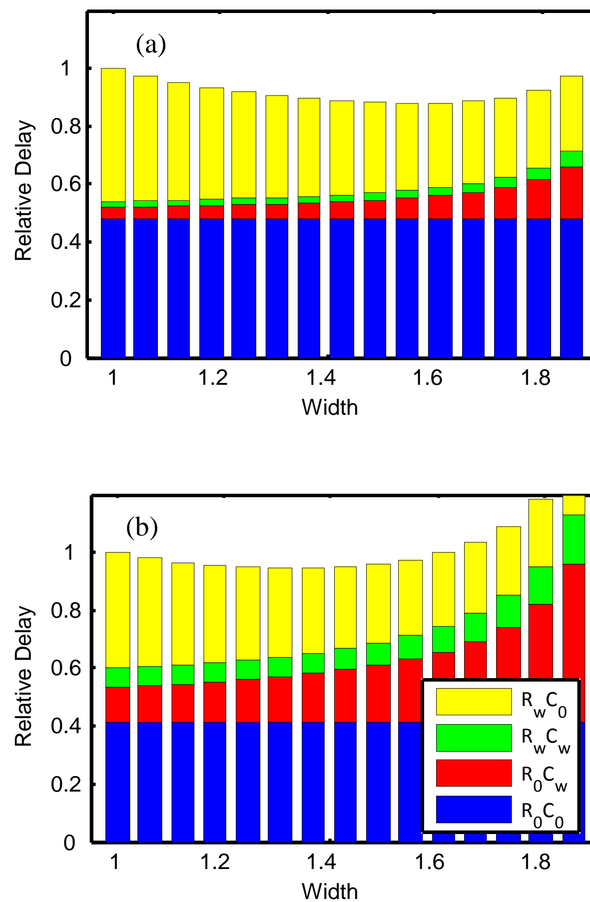


Figure 91: Single-stage delay breakdown versus the relative width of the interconnect for the circuits at the 7nm technology node using (a) VFET and (b) FinFET devices. Here, the length of the interconnect is 10 gate pitches.

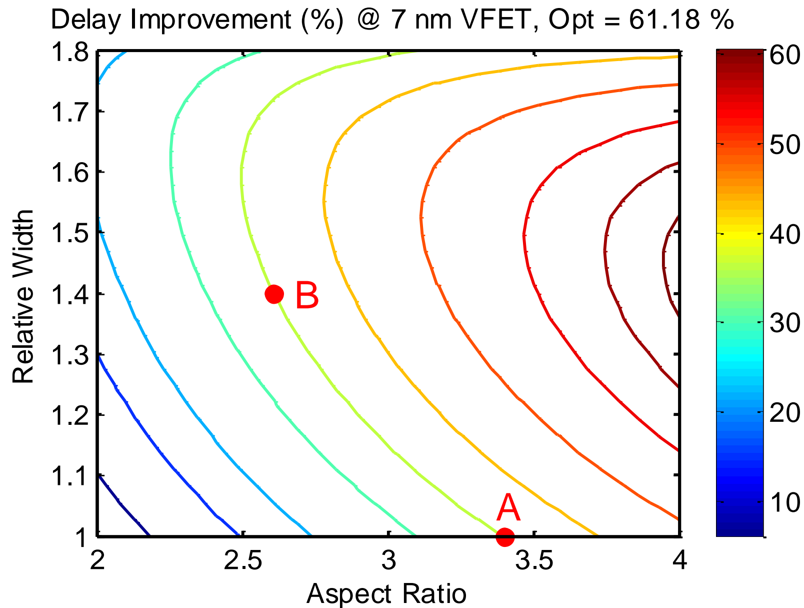
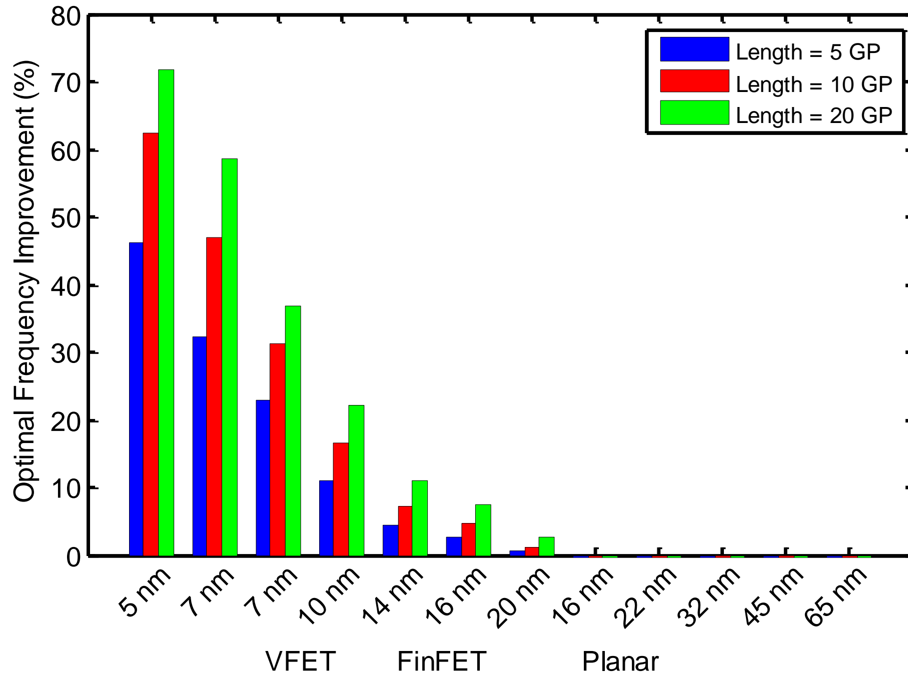


Figure 92: The contour figure of the delay improvement at various relative widths and aspect ratios of the interconnects.

The previous results are based on a fixed aspect ratio of 2. For various scenarios, Figure 92 shows the contour of the delay improvement of the circuits with various combinations of the relative width and aspect ratio. The improvement is compared with the default configuration at the relative width of 1 and aspect ratio of 2. For a given performance, the aspect ratio requirement can be reduced by increasing the relative width of the interconnect. For instance, an aspect ratio of 3.4 is required (point A) to achieve about 40% delay improvement, but with a relative width of 1.4, the aspect ratio only needs to be 2.6 (24% reduction at point B). Since increasing the interconnect capacitance affects both delay and energy, a more comprehensive figure of merit is EDP. At a fixed aspect ratio of 3, the improvement in EDP have been plotted for 10 technology generations. At the optimal width, the EDP improvements are about 5% to 15% less than the improvement in delay. The optimal width and EDP improvements both increase as technology advances.



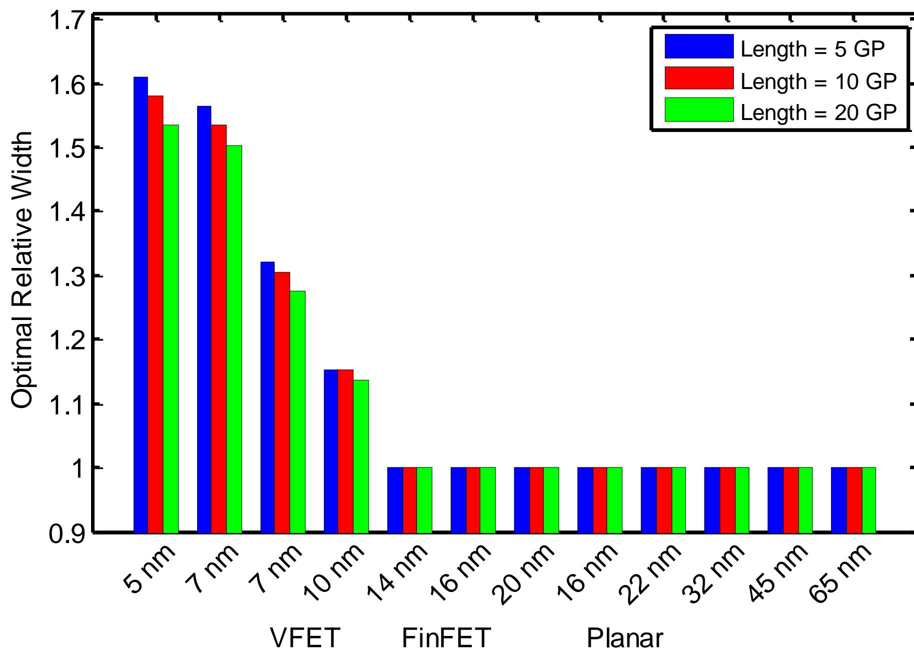
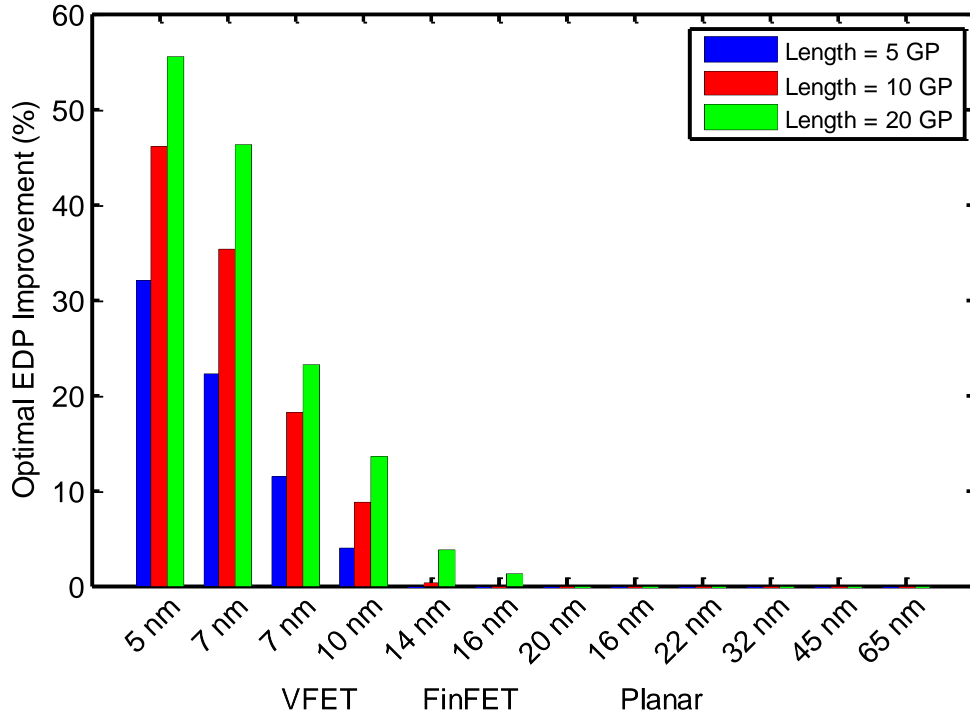


Figure 93: Optimal (a) clock frequency improvement and (b) relative width for three device structures at various technology nodes, assuming the aspect ratio is 3.



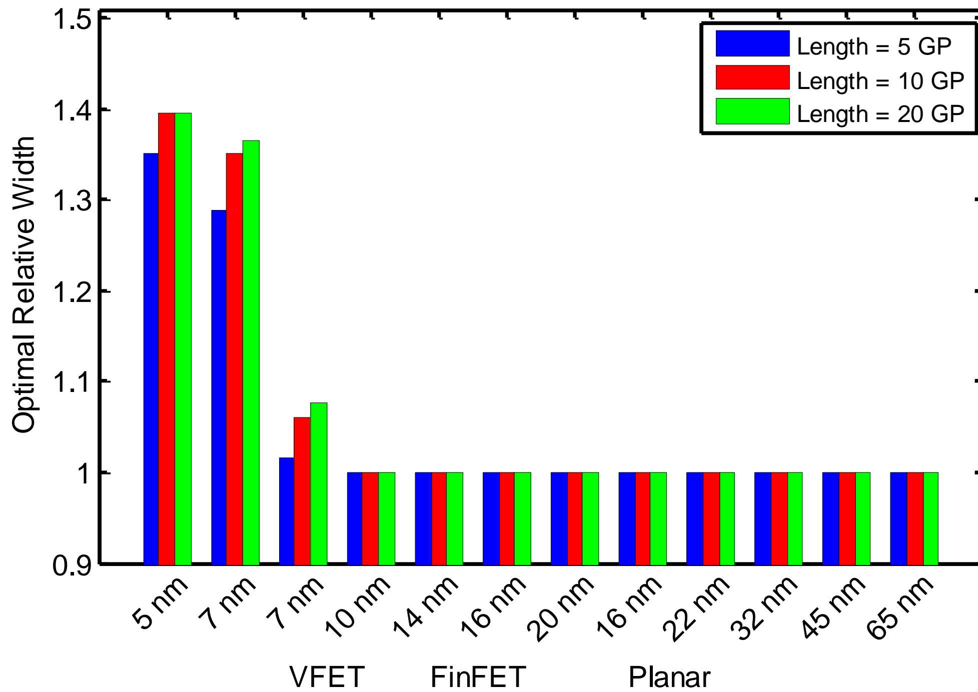


Figure 94: Optimal (a) EDP improvement and (b) relative width for three device structures at various technology nodes, assuming the aspect ratio is 3.

6.5 Al-Cu Hybrid Interconnect Architecture

While the search for novel interconnect materials such as carbon nanotubes and graphene is ongoing, there are no clear alternative materials at this point. Due to the long transition time from research to production, there is a great risk that no novel material will be ready to replace Cu/low k interconnects at 11nm or even smaller technology nodes. More than a decade ago, the semiconductor industry switched from aluminum to copper for its superior conductivity and its significantly better resistance to electromigration. However, it has long been known that Cu loses its conductivity advantage at nanoscale dimensions because the bulk electron MFP in Cu (40 nm) is larger than that in Al (16nm), leading to more pronounced size effects in Cu interconnects [153]. Furthermore, Al interconnects do not need diffusion barriers, which take a large fraction of the volume of an interconnect at nanoscale dimensions. Despite these facts, it is widely believed that it is improbable to switch back to Al because of the reliability problems caused by almost 5 times lower current conduction capacity of Al as compared to Cu [154]. However, electromigration is a major challenge mainly for power and ground interconnects where unidirectional currents exist, while signal interconnects that conduct bi-directional currents are immune to electromigration [155]. The mean-time-to-failure for Al interconnects conducting an AC current is more than 4 orders of magnitude higher compared to Al interconnects conducting a DC current equal to the root mean square value of the AC current [156]. This fact opens the possibility of a hybrid interconnect technology, where metals with high current conduction capacities (e.g. Cu or tungsten (W)) are used for local power distribution networks, and a low resistivity metal is used for signal interconnects. For short local signal interconnects that are minimum-sized, Al can be a promising candidate. For longer interconnect lengths when larger cross-sectional dimensions are used, again Cu can be used as it offers a better resistivity at larger dimensions.

6.5.1 Proposed Fabrication Process Flow

A simplified sequence of the fabrication steps for the proposed hybrid interconnect technology is shown in Figure 95. A thick layer of Al is deposited and annealed so that the grain size can be enlarged. This will result in an average grain size that is equal to the film thickness [157]. The Al film is then polished down to the desired thickness, which is followed by pattering minimum-sized Al interconnects with a subtractive process. Then, the dielectric material is deposited, and the trenches are made for the Cu power interconnects. Finally, liner/diffusion/seed layers are formed inside the trenches, and Cu is electroplated into trenches. Although this process requires two lithography/etch steps, the cost of the second step will be lower because of the relaxed pitch used for Cu interconnects. In principle, this process allows the signal and power interconnects to have different thicknesses. For power interconnects, important metrics are resistance and current density; thereby, larger thicknesses are desired. For signal interconnects, a larger thickness increases the interconnect capacitance, which adversely affects power dissipation and even signal delay for short interconnects because the resistance is dominated by driver resistance.

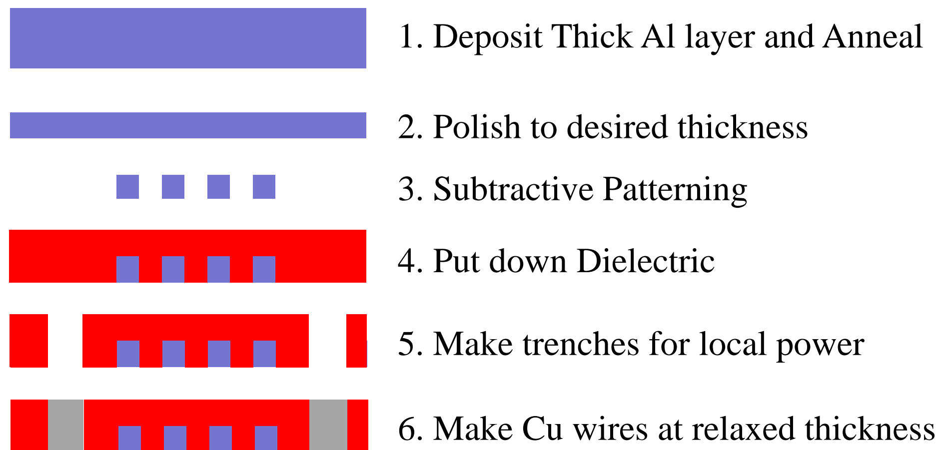


Figure 95: Flow chart of subtractive Al-Cu hybrid interconnect process.

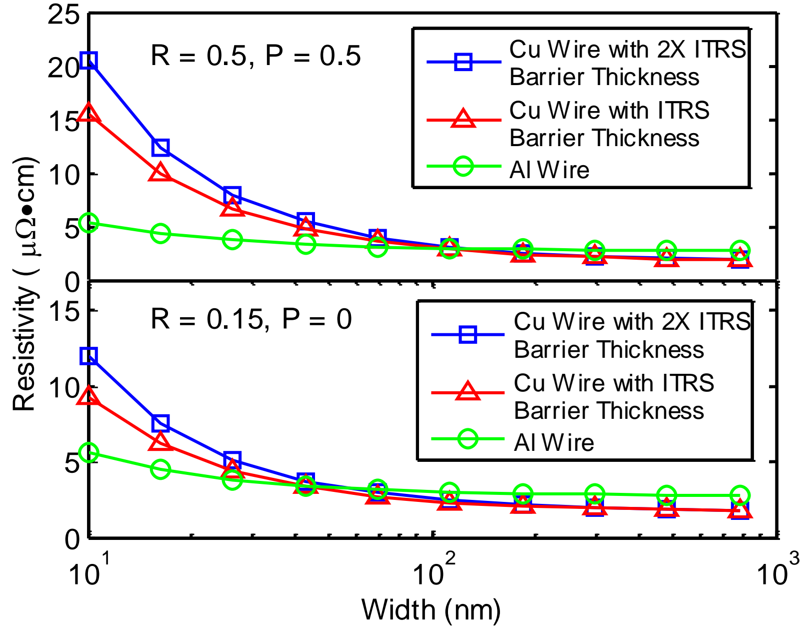


Figure 96: Resistivity versus the width of Cu and Al interconnects with two grain boundary reflectivity co-efficient *R* and surface specularity parameter *p*. Here, the aspect ratio is 2.

6.5.2 Simulation Results

Based on the interconnect resistivity model shown in Chapter 3, Figure 96 shows the resistivity versus the width of the interconnects. For Al interconnects, no barriers are needed and the grain size is assumed to be twice the interconnect pitch, which corresponds to the initial Al film thickness selected in Step 1 in Figure 95. The ITRS projections for diffusion barrier thickness [132] are known to be quite optimistic and challenging to achieve. Hence, for Cu interconnects, the ITRS projected barrier thicknesses and twice of those values are considered as references. Here, two sets of reflectivity co-efficient *R* and specularity parameter *p* are used based on the experimental data [10, 158]. The upper figure in Figure 96 shows that interconnects that are dominated by the grain boundary scattering suffer more from the size effects. Therefore, below 90nm interconnect widths, Al interconnects offer lower resistivities, where up to a 4×

smaller value can be observed at 10nm width. For those interconnects that are dominated by the surface scattering, the critical width decreases to 40nm, indicating a more moderate but still significant advantage of using Al interconnects at sub-20nm widths.

Based on the system-level design methodology illustrated in Chapter 4, the potential performance improvements for various interconnect configurations are quantified. The design rules, including fin pitches and M1 pitches for various technology nodes, are obtained from [132, 140]. The maximum number of metal levels is set based on the ITRS projection [132]. A single logic core with 20 million logic gates is analyzed under a 100 W/cm^2 power density constraint. The core area at the 16 nm technology node is assumed to be 5 mm^2, and for other technology nodes, the core area is proportionally sized based on the feature size squared.

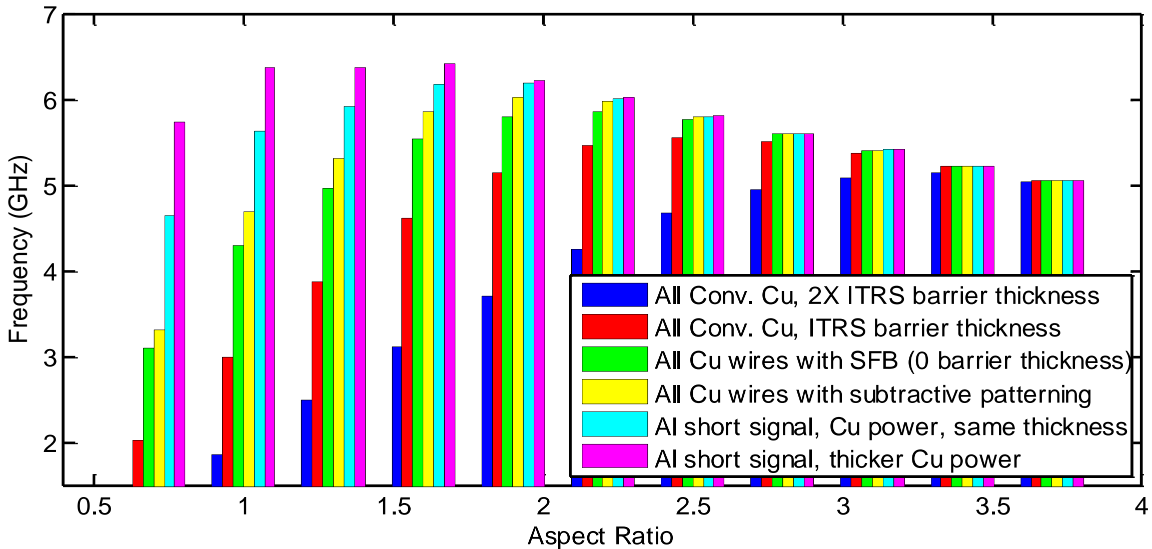


Figure 97: Chip clock frequency versus aspect ratio for six interconnect configurations at the 16nm technology node. Here, $R = p = 0.5$.

Six different types of interconnect architectures are investigated and compared. The first and second configurations are the conventional Cu interconnects with liner/barrier thicknesses of 5.2nm (2× ITRS projection) and 2.6nm (ITRS projection), respectively

[132]; the third is Cu interconnects without a barrier, which is the idealized SFB technology; the fourth is the Cu interconnect technology patterned using a subtractive process to achieve grain sizes twice as wide as the interconnect pitch; the fifth is the Al-Cu hybrid interconnect technology, where all the Cu signal interconnects whose width is narrower than the critical width (shown in Figure 96) are replaced by Al interconnects; the last one is based on the fifth option where the thickness of the power interconnects is twice the thickness of the local signal interconnects. From Figure 97, optimal aspect ratios are observed to maximize the clock frequency. This is because if the aspect ratio is too low, the frequency suffers significantly from the large resistance of the interconnect, particularly for the conventional Cu interconnects with thick diffusion barriers; if the aspect ratio is too high, the delay is dominated by the large interconnect-to-interconnect parasitic capacitance, which also increases the dynamic power dissipation of the interconnect network. Another observation from Figure 97 is that the optimal aspect ratio required to achieve the maximum chip frequency decreases when a better interconnect technology is used, which improves the manufacturability and the circuit-level performance such as crosstalk noise and delay variation. By using the hybrid interconnect technology, a 25% improvement in chip clock frequency is observed at the optimal aspect ratio for the same power budget. One reason for the improvement is that the aspect ratio of the copper interconnects in the local power distribution network is enlarged, which helps to reduce the power via density, increasing the overall wiring efficiency. Another reason is that the power dissipation of interconnects is saved by reducing the aspect ratio of the interconnects thanks to the low resistivity of the Al interconnects. For a fixed power density budget, this contributes to the performance increase. The simulation results also confirm that the AC current density in local signal interconnects is in the order of 10^5 A/cm^2, which is 10 times smaller than the DC current density limit set by the ITRS [132]. Because of the 10^4 increase in time-to-failure due to the self-healing effect for bi-

directional currents [156], there should not be electromigration issues for signal interconnects.

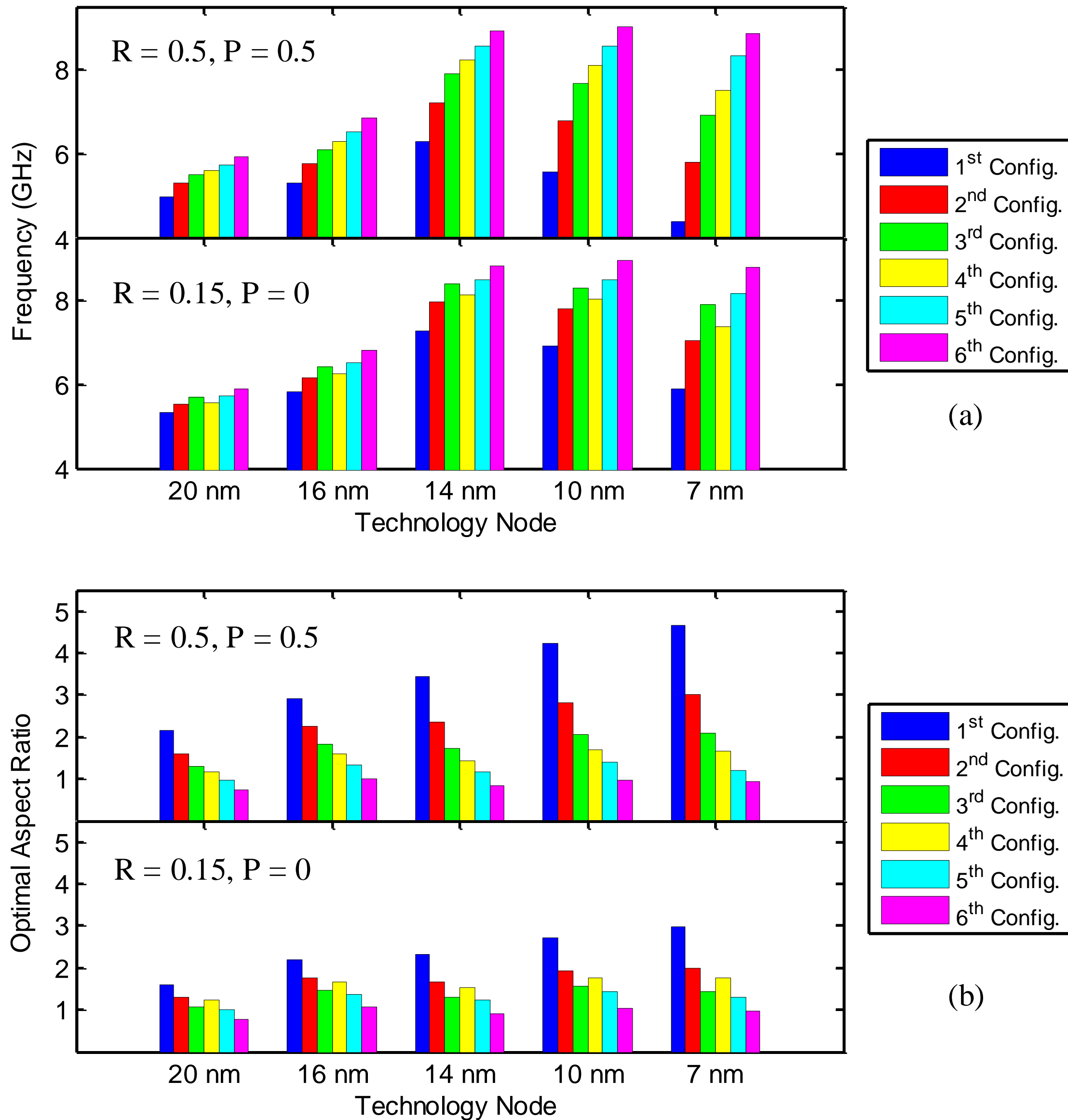


Figure 98: (a) Optimal clock frequency and (b) aspect ratio versus technology node for six different interconnect configurations that are defined in Figure 97.

For various technology nodes, Figure 98 (a) shows the optimal clock frequency at the optimal aspect ratio for six types of interconnect architectures. The optimal clock frequency begins to decrease below the 14 nm technology node, particularly for the

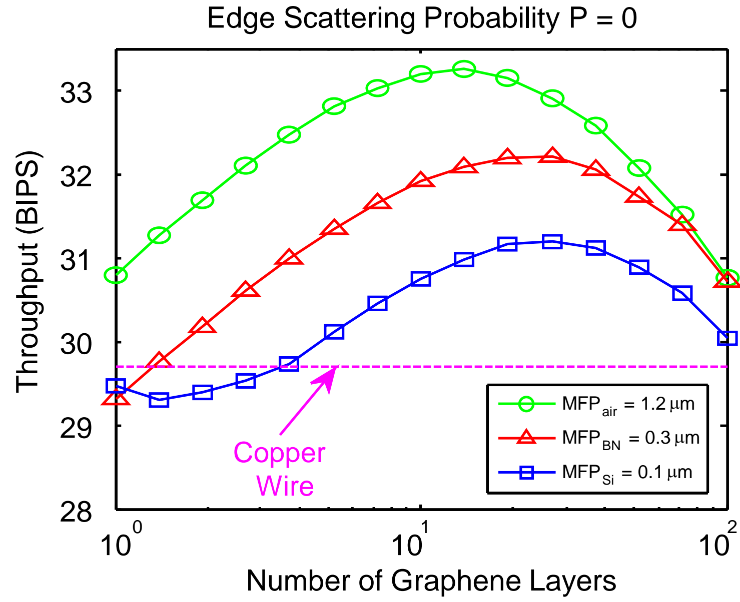
processors using Cu interconnects with thick diffusion barriers, in which the size effects are prominent. The proposed Al-Cu interconnect technology significantly suppresses the frequency drop. For the upper chart, in which grain boundary is dominant, up to 2× improvement in chip clock frequency is predicted by using the novel interconnect architecture at the end of the roadmap; for the lower chart, in which surface scattering is more dominant, there is a 50% improvement in clock frequency. Figure 98 (b) shows the optimal aspect ratio for local signal interconnects to reach the optimal clock frequency, corresponding to each bar in Figure 98 (a). The optimal aspect ratio in the proposed scheme is up to 70% smaller than that in a conventional Cu interconnect technology at the 7 nm technology node.

6.6 Graphene Interconnect

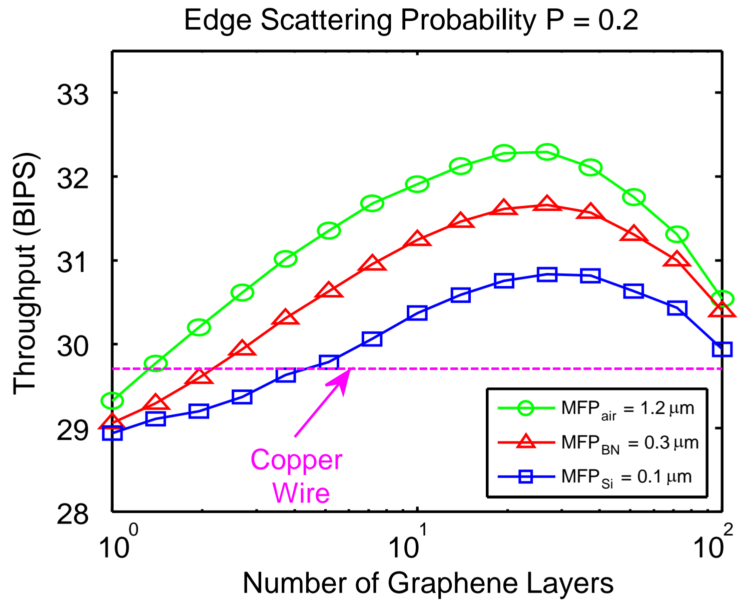
To further reduce the interconnect dissipation while maintaining the resistance under control, the multi-layer graphene interconnect is one of the promising candidates due to its thin structure and large current carrying capability. Based on the circuit-level models and simulation results shown in Chapter 3, graphene interconnects can provide smaller delay for short lengths. Therefore, *graphene is only used for short interconnects in the lower interconnect levels, where no repeaters are used*. To benchmark and optimize a logic core using multi-layer graphene interconnects, the hierarchical optimization engine developed in Chapter 4 needs to be modified to accommodate the power and delay saving brought by the graphene interconnects. After each evaluation of the power dissipation inside the multi-level interconnection network design module shown in Figure 64, the power consumption saved by the graphene interconnects without repeater insertion needs to be subtracted from the estimated power dissipation. The iteration continues until the total power dissipation converges. Then, the maximum clock frequency allowed from the critical path is updated based on the interconnect capacitance saving from graphene. This

outside loop keeps running until the maximum clock frequency converges as well. The rest of the optimization follows Figure 64.

Based on the modified hierarchical design engine, a 5 mm^2 GPNJ single-core processor with 100 W/cm^2 power density constraint is assumed for the performance comparison between multi-layer graphene interconnects and copper interconnects. Figure 99 shows the throughput versus the number of graphene layers at various MFP and edge roughness conditions, where optimal numbers of layers can be found to achieve the maximum throughput. At the optimal point, up to 15 % of the throughput improvement can be observed compared with the GPNJ core using copper interconnects. The improvement primarily comes from the smaller capacitance per unit length of the graphene sheet, which helps to reduce the critical path delay and the dynamic power dissipation of the interconnects. The reason for the drop of the throughput at a large number of layers is that although the resistance per unit length is inversely proportional to the number of layers, when there are too many graphene layers, the increase of the interconnect capacitance overshadows the benefits of the graphene. In addition, the optimal number of graphene layers increases with the increase of the MFP and edge scattering probability. For various core areas, the throughput improvement increases as the core size decreases, shown in Figure 100 (a), where up to 20% of the improvement can be observed for a small core. This is because a smaller core leads to shorter interconnects, which can better utilize the advantage of the graphene interconnects. In addition, since the resistance per unit length increases as the edge scattering probability increases, more optimal numbers of layers are required to reach the maximum throughput for a worse edge scattering probability as shown in Figure 100 (b).

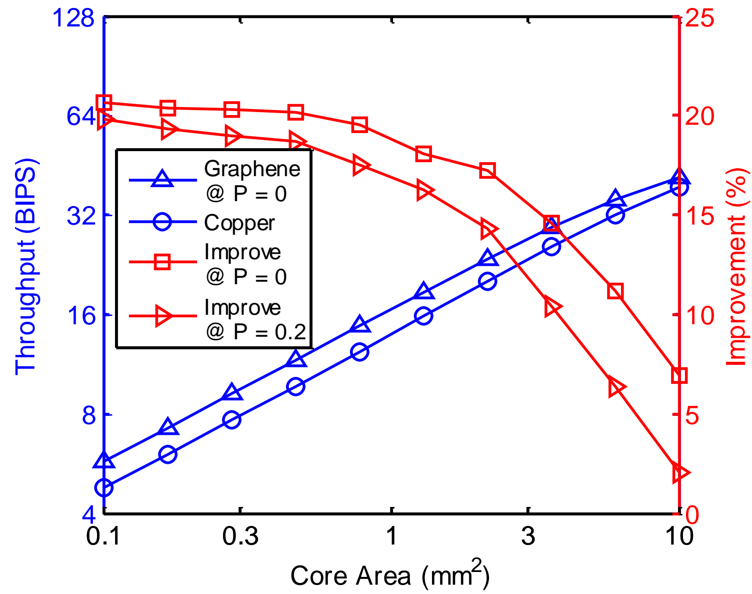


(a)

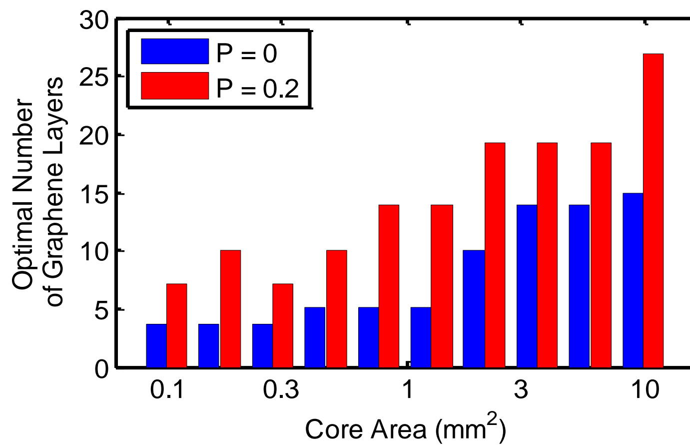


(b)

Figure 99: Throughput versus the number of graphene layers at various MFP using different materials as the substrates. (a) Smooth edge (b) Edge scattering probability P = 0.2.



(a)



(b)

Figure 100: (a) Optimal throughput versus core area for GPNJ processor using graphene and copper interconnects. The red curves show the improvement of using multi-layer graphene interconnects with perfect edge and edge scattering probability p = 0.2. (b) Optimal number of graphene layers versus core area for a single core using multi-layer graphene interconnects.

6.7 Optimization under Process Variation

In this section, the impact of various sources of interconnect variations are quantified at the overall system-level throughput. Based on the system-level variation-aware design methodology presented in Chapter 4, one thousand samples are generated for a single-core processor made of 25 million logic gates. The die area is assumed to be 5mm^2 at the 10nm technology node. For other technology nodes, the die areas are proportionally sized. Device models for FinFETs from 14 nm to 7 nm nodes are obtained from [67]. Since this work mainly focuses on the interconnect variation, the device-level variation is not included in this chapter. The device-level design rules such as the fin pitches and the footprint area follows [140], and the interconnect parameters such as the M1 pitch, spacings, aspect ratio are shown in Table 3. The maximum number of metal levels is set based on the ITRS projection [132].

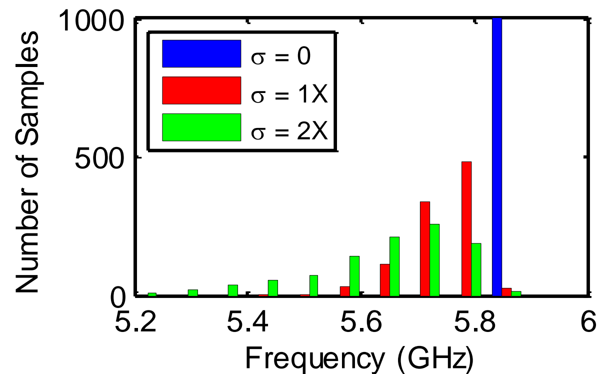


Figure 101: The histograms of the optimal clock frequency under each individual source of interconnect variation, including CD core/spacer, etch, CMP, and overlay variations, with various σ values using three fabrication processes such LELE, SADP, and SAQP.

The histograms of the corresponding optimal frequencies under the CD line/core variation at the 10nm node using the LELE process are plotted in Figure 101. The blue columns represent the performance of the processor at the nominal design; the red columns are the processors using default interconnect variation values shown in Figure 101; and the green columns are the ones using twice the values shown in Table 3 to take

into account the situation when the interconnect variability increases during the massive industrial production. In general, the spread of the clock frequency under various sources of the interconnect variation increases as the σ value increases, and the majority of the samples shift to the left. This is because the interconnect variability is assumed to be the same for all metal levels. Since the geometries of interconnects at higher metal levels are much larger than those at the local levels, the delay distributions for the upper levels are much narrower than those at the local levels as shown in Figure 102, which limits the maximum clock frequency that a core can operate and shifts the frequency distribution to the left.

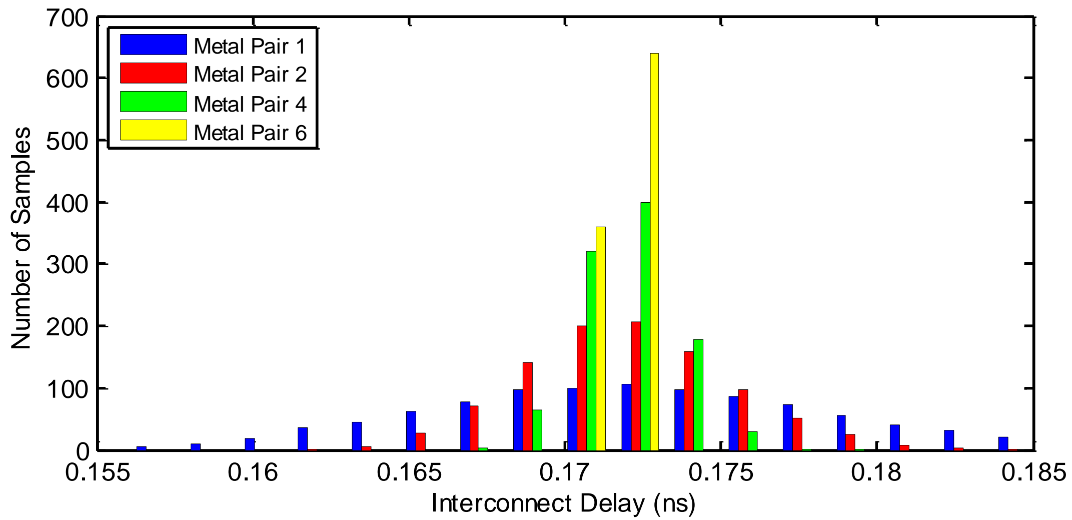


Figure 102: The histogram of the delay for the longest interconnects on various metal levels.

Table 12: The Standard Deviation and Mean Optimal Clock Frequency under Each Individual Source of Interconnect Process Variation Using Three Fabrication Processes

Variation Sources	**LELE**	**SADP**	**SAQP**
CD Line/Core (σ/μ)	0.061/5.739	0.076/5.743	0.073/5.745
CD Spacer (σ/μ)	0/5.826	0.049/5.763	0.092/5.714
Etch (σ/μ)	0.052/5.777	0.051/5.774	0.053/5.774
CMP (σ/μ)	0.097/5.733	0.097/5.727	0.096/5.730
Overlay (σ/μ)	0.026/5.799	0/5.826	0/5.826

Table 12 summarizes the standard deviation and the mean clock frequency for five sources of interconnect variations, including CD line/core, CD spacer, etch, CMP, and overlay variations, using three interconnect fabrication processes, including LELE, SADP, and SAQP. The variation assumptions follow Table 3. For the CMP variation, it has the largest 3σ values in Table 3, which causes the largest frequency deviations for all three types of interconnect fabrication processes. For the CD core/line variation, the comparison among LELE, SADP, and SAQP processes shows that the one using LELE process has the smallest frequency deviation. The reason is that in the LELE process, the CD line variation affects both the width and spacing of the interconnect, which changes the resistance and capacitance of the interconnect in opposite directions. Therefore, the impact on the overall interconnect delay is smaller. For the CD spacer variation, the first observation is that it causes the largest spread in the SAQP process. This is because two steps of spacer deposition are required to perform the quadruple patterning, which affects both the spacing between any two nearby interconnects as well as the width of every other interconnect, especially for the one in the middle of two cores. As can be seen from Figure 102, the majority of the delay variation exists at the local metal levels, where the length of the interconnects is relatively short and the number of the interconnects is large. Therefore, in most of the samples, the delay on the local interconnect is limited by the slowest paths that are associated with the worst case scenario during the SAQP process, leading to a large spread under the CD spacer variation. For the comparison between LELE and SADP processes, the overlay variation has a smaller impact for the LELE process than the CD spacer variation does for the SADP process. The reason is that the interconnect capacitance is less sensitive to the overlay variation when the interconnect is closer to the nominal case, whereas it increases more rapidly as the deviation becomes larger. This can be confirmed from the non-linear curve shown in Figure 22 (b). If the σ becomes twice of the default value shown in Table 3, the overlay variation has about the

same impact as the CD spacer variation. Since there is no spacer variation in the LELE process and no overlay variation in SADP and SAQP processes, the standard deviations are zero.

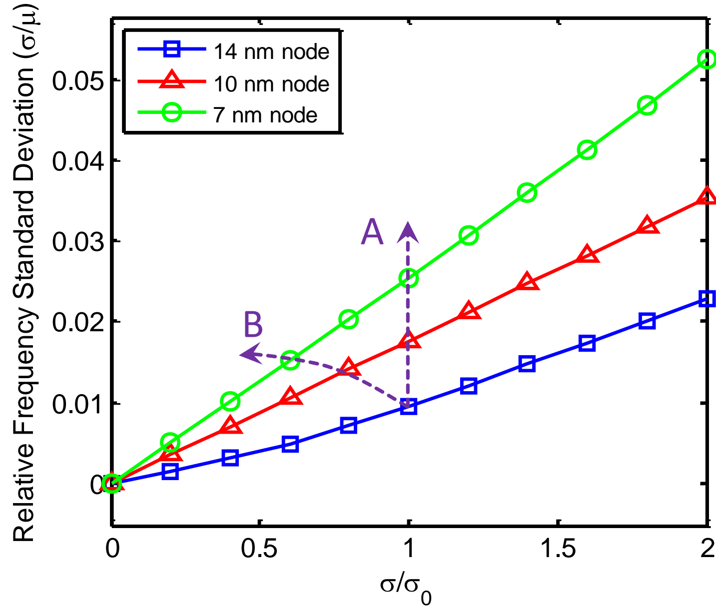


Fig. 1: The frequency standard deviation relative to mean frequency under the CMP variation at various σ values relative to the values assumed in Table 3.

For the sensitivity analysis, Fig. 1 shows the relative frequency standard deviation versus various variability of the CMP variation relative to the default value in Table 3. The frequency deviation keeps increasing as the CMP variability increases, especially at a small technology node. As the technology scales, if the CMP variability remains the same as assumed in Table 3, the corresponding frequency deviation will follow the line 'A'. However, if the process advances, the frequency deviation degradation may slow down and follow line 'B'. The fabrication process assumed for Fig. 1 is the LELE double patterning, and for SADP and SAQP processes under other sources of variation, such as the CD core/line and spacer variations, the same behaviors are observed.

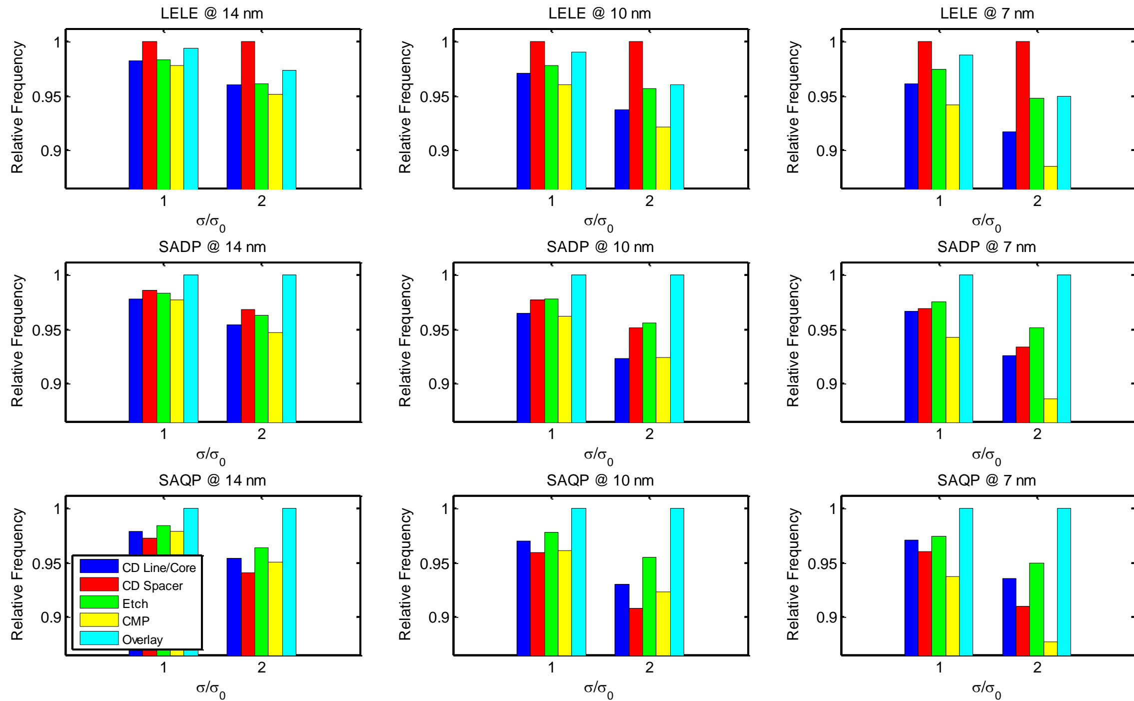


Figure 103: The relative frequency at 90% yield versus σ/σ_0 under five sources of interconnect variations at three technology nodes using LELE double patterning, SADP, and SAQP interconnect fabrication processes. Here, σ_0 represent the default interconnect variability values shown in Table 3.

To analyze the impact of the interconnect variations at various technology nodes, the clock frequency at 90% yield point relative to the nominal frequency are shown in Figure 103. Two σ values relative to the default values for five sources of interconnect variations are investigated at three technology nodes using the LELE, SADP, and SAQP fabrication processes. As the technology scales, in general, a larger impact on the clock frequency is observed under various sources of interconnect variations, especially for the CMP variation. It is because the variability for the CMP is larger, and it does not scale as the technology goes down, shown in Table 3. For the etch variation, its impact on the relative clock frequency only increases a little because the 3σ value scales 28.6% and 25% at the 10nm and 7nm technology nodes, respectively, which are quite aggressive. For SADP and SAQP processes, the major difference is in the CD spacer variation, which is one of

the major sources of the variation in the SAQP process that is comparable to the CMP variation.

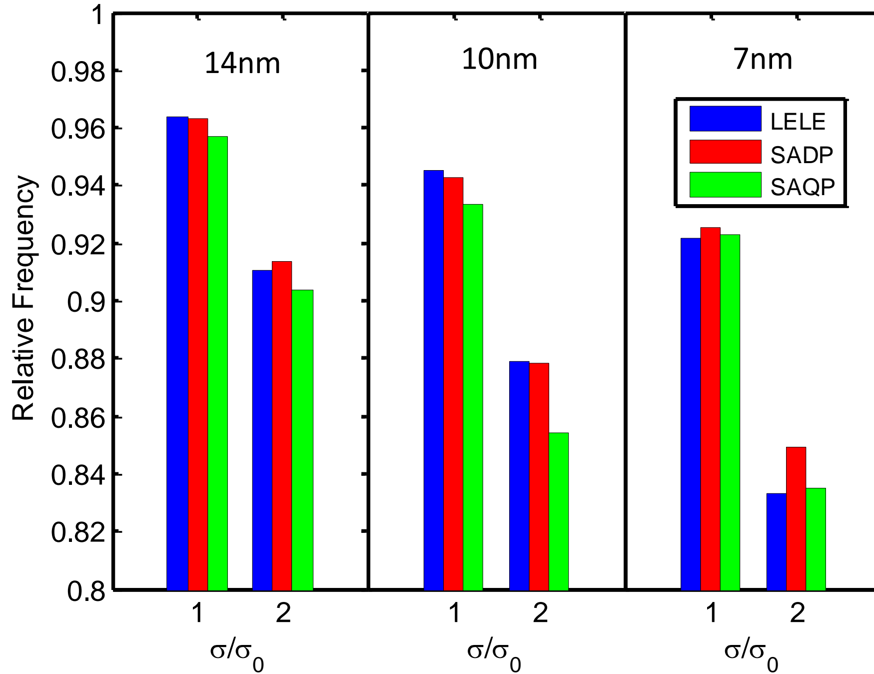


Figure 104: The relative clock frequency versus various σ values with the combination of five types of interconnect variations at the 14 nm, 10 nm, and 7 nm technology nodes using LELE, SADP, and SAQP fabrication processes. σ_0 represent the default interconnect variability values shown in Table 3.

In Figure 104, five sources of interconnect variations are included, and the relative clock frequency versus the σ values are plotted for three technology nodes. The leftmost set of columns represent the nominal cases, where no interconnect variations are included. Even though the dielectric permittivity decreases as the technology node scales down, the impact on the clock frequency due to the interconnect variation increases, especially for the LELE double patterning process. This is because the geometries such as the width and thickness of the interconnect scale 25% for each technology generation, but the overlay variation only scales 16% and 20% at the 10nm and 7nm technology nodes, respectively. Moreover, the variability of CD core/spacer and CMP remains the same at

all three nodes, leading to a significant impact on the 7 nm technology node. Since the overlay variation only increases the line-to-line capacitance without reducing the resistance, it shifts the entire interconnect delay distribution to the larger delay side. Therefore, it contributes to a larger frequency drop to the LELE process. Another observation is that at a large technology node, where the σ value is relatively small, the LELE process is the least sensitive to the interconnect variation; however, as the technology scales and with the increase of the variability values, the LELE process suffers more from the variation. At the 7nm technology node, even if the LELE process is available, it becomes worse than the SADP process due to the larger impact of the overlay variation, indicating the importance of reducing the overlay variation at small nodes. Up to 16% of the frequency drop is observed at the 7 nm technology node with twice the values assumed in Table 3 for the SAQP process. Compared with SADP process, SAQP process is more sensitive to the variation because of the additional spacer deposition step.

6.8 Conclusions

A paradigm shift in local interconnect technology design is presented based on the observation that the interconnect performance becomes more resistance dominated rather than capacitance dominated. Two implications are addressed, and one potential solution to rebalance the interconnect resistance and capacitance by increasing the interconnect width beyond half-pitch has been evaluated. At the optimal interconnect width with an aspect ratio of 3, the EDP of a VFET circuit improves up to 55% at the 5 nm technology node. The optimal interconnect width depends on the technology node and device structure.

To suppress the impact of size effects on sub-20nm wide interconnects, a novel aluminum-copper hybrid interconnect architecture is proposed and its potential

performance has been quantified based on the device-, circuit-, and system-level compact models and a hierarchical optimization engine developed in Chapters 2-4. Al interconnects offer lower resistivities at nanoscale dimensions because they do not need diffusion barriers, and size effects are less prominent in them due to their smaller bulk MFP. However, their current conduction capacity is substantially lower than that of Cu interconnects. To get around this limitation, this letter proposes a hybrid interconnect technology to replace only short narrow local signal interconnects by Al interconnects. This scheme takes advantage of the fact that signal interconnects conduct bi-directional currents and are therefore virtually immune to electromigration. Six interconnect architecture options are analyzed and their optimal aspect ratio and chip frequency are predicted for five technology generations. The optimization and benchmarking results indicate that the potential improvement in chip clock frequency can be between 50 to 100% for the 7nm technology node.

A comprehensive optimization/benchmarking is also performed for the multi-layer graphene interconnects. The results show that a single core using graphene interconnects can potentially have a higher throughput within the same power density and die size area because of the power saving offered by the low capacitance per unit length of graphene interconnects. Optimal numbers of graphene layers have also been observed to reach the maximum throughput, and the number of layers increases with the increase of the MFP and edge roughness. A smaller core can better utilize graphene interconnects because of the shorter interconnects.

For the interconnect variation, various sources of interconnect variations are compared, such as the CD line/core, CD spacer, etch, CMP, and overlay variations. The default 3σ values for these independent variation values are extracted from different fabrication processes, including LELE double patterning, SADP, and SAQP. The results indicate that the CMP variation has the largest impact on the overall clock frequency of

the processor. In addition, the impact of the interconnect variation increases as the technology node goes down, especially for the LELE process with a larger σ value because of the overlay variation. A larger frequency drop is observed for the SAQP process compared with the SADP process due to an additional spacer sidewall variation. Up to 8% and 16% of the frequency drops are observed based on extracted 3σ values and twice of those as the worst case scenario, respectively.

CHAPTER 7 SYSTEM-LEVEL BENCHMARKING AND OPTIMIZATION

7.1 Introduction

In this chapter, higher system-level benchmarking and optimization are performed based on the hierarchical optimization engine developed in Chapter 4. Two novel system configurations, 3D and heterogeneous integrations, are benchmarked and optimized.

The 3D ICs have intrigued many research in the past decade due to its superiorities of extending Moore's Law by increasing the device density, overcoming the barriers in the interconnect scaling, and providing further performance improvement with less power consumption [44]. However, challenges still exist, such as the high power density caused by the 3D stacking and the complex interconnect design methodology. To quantify the system-level benefits of a 3D processor, one has to design an entire processor with complicated interconnection network, which can be extremely time-consuming. In this chapter, the proposed hierarchical optimization engine is used to efficiently quantify the system-level throughput (in instructions per second) of MIV-based fine-grained 3D processors. The models are also used to determine the optimal number of transistors, size of the cache, number of stacked dice, supply voltage and interconnect network design. In addition, the impact of via diameter and capacitance on the overall chip throughput has been quantified, and the comparison between the conventional 2D and TSV-based 3D ICs suggests the advantage of fine-grained MIV-based 3D integration.

Another system-level technique analyzed in this chapter is the heterogeneous multi-core processor. To achieve a higher throughput under the limited power and thermal constraint, multi-core processors were developed in the past decade. Instead of using a

complex single core processor with a complicated interconnection network, a multi-core processor reduces the design complexity of each individual core and takes advantage of the parallel computing to achieve a better power efficiency. Recently, the heterogeneous integration is also proposed to further improve the energy and power efficiency by using asymmetric cores or even different technologies for different cores. Previous work analyzed the CMOS-TFET heterogeneous integration by using the cycle-accurate simulation method [46], demonstrating that the energy efficient TFET cores can save both leakage and dynamic energy dissipation with little performance degradations. This estimation is accurate, but it is time-consuming and only investigates one specific system configuration. To truly understand the potential benefit and advantage of the heterogeneous TFET technology, device- and system-level co-optimization is required to find the best design points, including the supply voltage, number of cores and logic transistors, logic-to-cache ratio, and big-to-small core ratio, to maximize the chip performance under certain area/power budgets.

Since one of the major limitations of both high-performance and low-power processors is the power dissipation, where the leakage power has become a major fraction at advanced technology nodes, the power-gating has emerged as a key technique to suppress the leakage power [159]. By adding a sleeping transistor between the virtual ground to the actual ground, the leakage current of the circuit can be suppressed significantly when the sleep transistor is turned off. However, power gating exacerbates the power delivery integrity issues due to the rush current during the wake-up period. One can use a larger sleep transistor to reduce the DC drop, but it also increases the area overhead and affects the power delivery network, which may increase the settling time. In addition, the voltage at the virtual power supply at the steady state during the idle period is determined by the ratio of the size between logic and sleep transistors. This voltage will influence the energy overhead associated with charging the parasitic and

decap capacitance. This trade-off between performance and energy will also change how the processor is designed, including the optimal supply voltage and the complexity of the logic core. Moreover, these changes may vary for various package configurations with different decap insertions. All of the aforementioned trade-offs can be analyzed only through a system-level chip and package co-optimization, which maximizes the overall chip throughput under certain power/area budgets. Many previous publications have presented PDN and power-gating analyses [160-163]; however, most of the prior work focus only on the power distribution network to minimize the maximum noise on the power supply or make a trade-off between the energy and performance of a processor chip using power-gating technique, which neglects the interaction between the chip performance and the package design [160, 161]. In this chapter, based on the proposed hierarchical optimization engine, multiple design parameters are co-optimized to maximize the overall chip throughput.

The rest of this chapter is organized as follow: in Section 2, the 3D integration is optimized and the overall chip throughput is quantified and compared with a conventional 2D processor under various via diameters and capacitance values. In addition, the MIV- and TSV-based multi-core processors are optimized and compared based on the same power constraint. Section 3 investigates the advantage of a heterogeneous multi-core processor by using different device technologies for different cores. The analyses are performed for both synchronous and asynchronous processors with various parallelisms of the program. In Section 4, the impact due to the process variation of both logic devices and interconnects on the overall chip throughput are quantified. Section 5 performs chip and package co-optimization to quantify the impact of the power delivery noise for a multi-core processor implemented with power-gating techniques. The position of the decap is optimized for various levels of parallelism in the program.

7.2 3D Technology Optimization

In this section, the potential benefits of using 3D integration are quantified based on the interconnection network models and the hierarchical optimization engine developed in Chapter 4. The impact due to the size of the vias is quantified, and the performance of a multi-core processor using TSV- and MIV-based technologies are compared.

7.2.1 Single logic core optimization

First, a logic core with a 5 mm^2 footprint area and 80 °C thermal constraint is optimized, as shown in Figure 105 [164]. The thermal model is taken from HotSpot [133], which is an open source thermal simulator used for estimating the temperature of each block and layer of the chip according to their corresponding power consumption. It can be seen in Figure 105 (a) and (b) that there exist optimal numbers of gates and supply voltages that maximize the computational throughput because processors with few transistors suffer from large CPI values as the empirical CPI model indicates; the ones with too many transistors suffer from low clock frequencies constrained by complicated interconnection networks and large switching energy of the logic gates. Likewise, processors with large supply voltages suffer from the large power dissipation; and the ones with very low V_{dd} suffer from the high ON resistance of the devices, which limit the clock frequency. Compared with the optimal point in the 2D logic core, a larger number of logic gates are required to reach the optimal throughput due to the larger area budget in the 3D case.

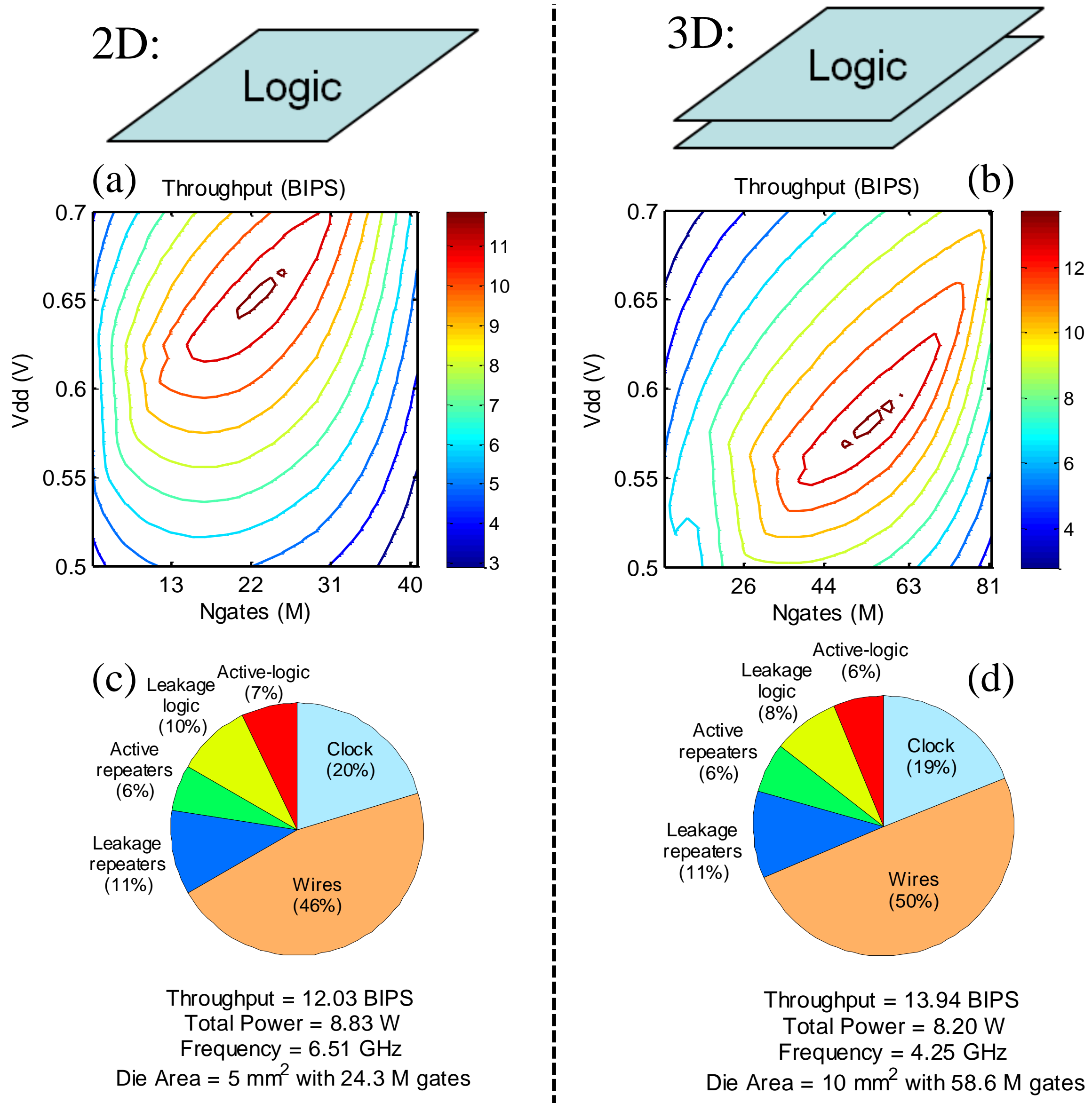


Figure 105: Optimization results for a single core implemented by the conventional 2D integration and 3D integration with MIVs. (a) and (b) show the throughput versus the supply voltage and the number of logic gates. (c) and (d) show the power and delay components for each part.

Since the maximum temperature is fixed at 80 °C, the power density per tier decreases significantly for the 3D chip due to the poorer thermal conductivity, which causes a smaller total power consumption, leading to a drop in the optimal supply voltage compared with its 2D counterpart. This can be also confirmed by Figure 105 (c) and (d),

where one can observe about 8% less power consumed by the 3D chip. Furthermore, the throughput of the 3D core is 16% higher and that is because although the total logic area is doubled, the average and the longest interconnects do not increase proportionally due to the implementation of MIV-based 3D integration. As a result, based on the same thermal constraint, the overall performance of the more complex 3D core can gain an improvement in CPI with no major frequency penalty. In Figure 106, an optimal number of logic dice can be observed to achieve the maximum throughput because as the number of logic dice increases, the allowed power consumption decreases due to the degradation of the thermal conductivity, thus reducing the overall throughput. Another reason is that the number of interlayer vias increases as the number of logic layers increases, which reduces the wiring efficiency significantly if there are too many logic levels.

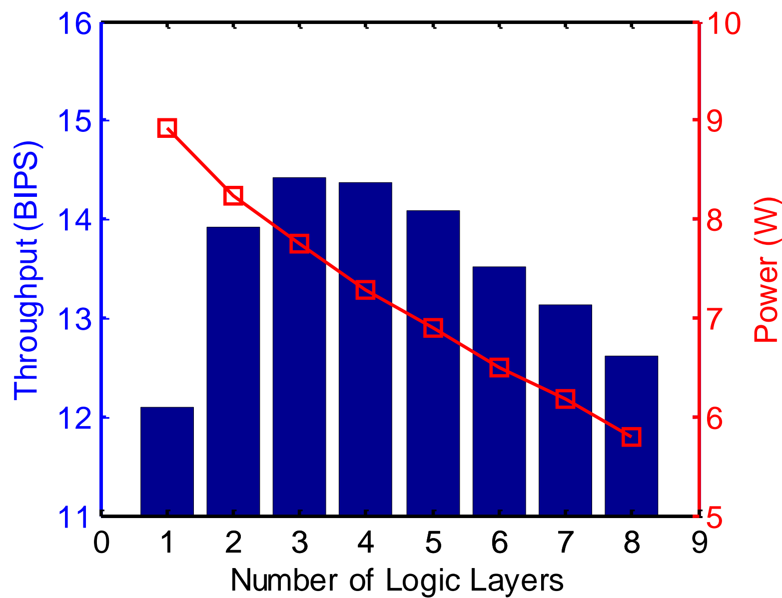


Figure 106: The optimal throughput and power consumption versus the number of logic layers for the system implemented by MIV-based 3D ICs, given D_{via} = 100 nm, C_{via} = 0.4fF.

7.2.2 Impact of the Via Diameter and Capacitance

The simulation results shown above are based on one specific dimension of the inter-tier via reported in the previous research [165]. Therefore, it is insightful to investigate how the via diameter and capacitance will affect the overall performance of the logic core. For simplicity, the via capacitance is assumed to follow the cylinder oxide capacitance, which can be expressed as

$$C_{ox} = \frac{2\pi\varepsilon_{ox} l_{via}}{\ln\left(1 + \frac{t_{ox}}{R_{metal}}\right)} \tag{35}$$

where ε_{ox} and t_{ox} are the permittivity and the thickness of the oxide, respectively, l_{via} and R_{metal} represent the length and the radius of the via, respectively.

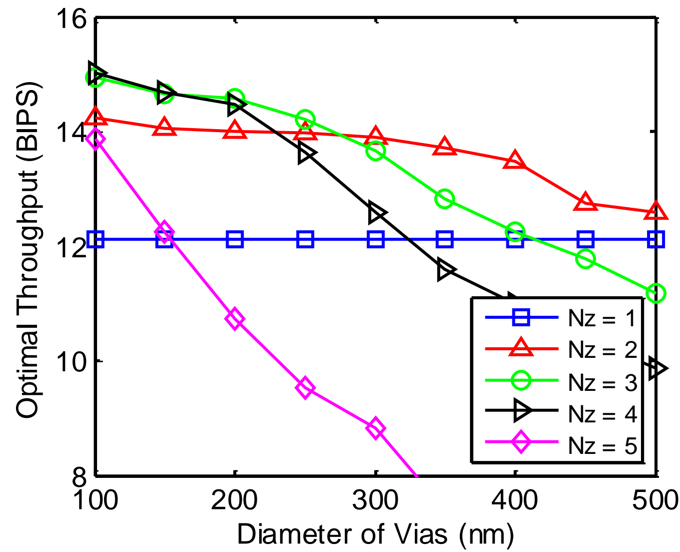


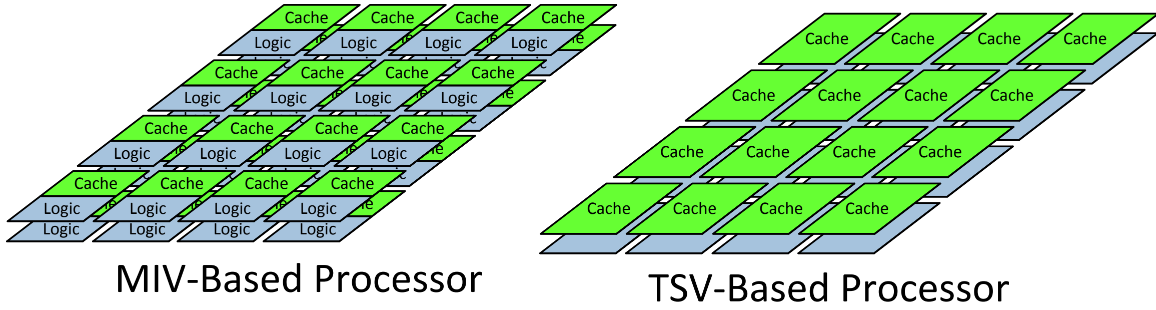
Figure 107: Optimal throughput of a single core versus the diameter of the vias for various numbers of logic layers.

In Figure 107, both the optimal throughput and number of tiers decrease as the via diameter increases, which is mainly caused by two reasons. First, a bigger via increases the via blockage factor, leading to a poorer routing efficiency. Hence, more repeaters are required at higher metal levels to make sure that all the interconnects can be placed

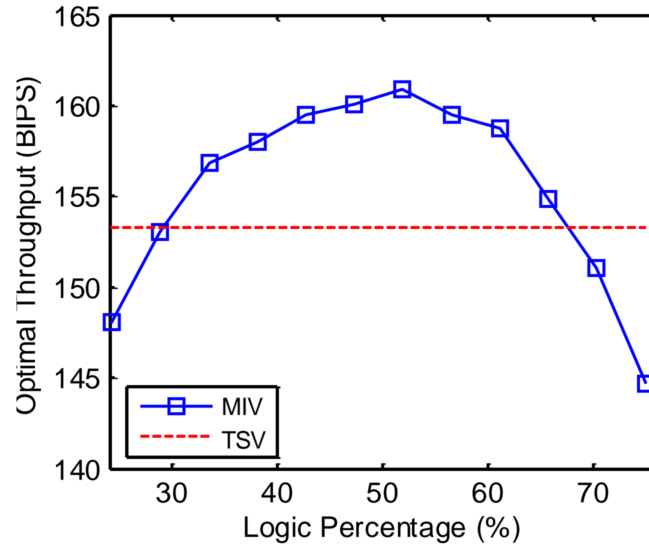
within the limited number of metal levels. Those additional repeaters consume more power and reduce the clock frequency and the overall throughput for the same thermal constraint. Second, an increase in the diameter increases the via capacitance as well, which enlarges the effective distance of the two nearby tiers, prompting the interconnect distribution to adjust a larger number of short interconnects on the same tier without utilizing the inter-tier vias. In short, the benefits brought by 3D integration become less remarkable if the via diameter increases.

7.2.3 Comparison between the TSV- and MIV-based systems

In this subsection, a multi-core system, including both logic and cache cores, is analyzed and a comparison is made between the TSV- and MIV-based processors as shown in Figure 108. Since the quality of TSV-based 3D-IC severely suffers if the TSV dimensions are large, in this chapter, TSV-based processors are limited to memory-on-logic or logic core-on-core designs with a relatively small number of core-to-core interconnects as shown in Figure 108 (a). In Figure 108 (b), an optimal logic percentage exists to reach the maximum throughput because either the computation or the memory access time become dominant if the logic percentage is too low or too high. The performance metrics at the optimal design points for both systems are shown in TABLE 1, indicating the total power consumption for the MIV-based system is 18% lower than that in the TSV-based system relying on the fact that the power density budget is not fully utilized in the cache region. However, the optimal throughput of the MIV-based system is 5% higher due to the interconnect length saved by the fine-grained within-core 3D technology, which can be observed in Figure 108 (b) in which both the average and the longest interconnect lengths are reduced by 30%.



(a)



(b)

Figure 108: (a) Basic configuration of MIV- and TSV- based systems. (b) Optimal throughput versus the logic percentage for both MIV- and TSV-based two-layer 16-core processors, assuming the program is fully parallelized.

Table 13: Comparison between Optimized MIV- and TSV-Based Two-Layer 16-Core Processors.

Parameters	MIV	TSV
Throughput (BIPS)	160.87	153.31
Frequency (GHz)	2.72	3
Ngates per Core (M)	96.79	74.7
Logic + Cache Power (W)	85.74	103.98
Average Interconnect Length (µm)	3.42	4.8
Longest Interconnect Length (mm)	5.7	8

In Figure 109, optimal numbers of stacked dice exist for both MIV- and TSV-based processors. In addition, for the MIV-based 3D system, more than 25% improvement in the throughput can be observed at the optimal number of dice because more power can be utilized in the cache region. Moreover, MIV-based 3D systems have the ability of tuning the logic-to-cache ratio, allowing a better trade-off between computation and memory access penalties. However, if too many stacked dice are used, there is a diminishing return in the overall throughput because of the poor thermal conductivity.

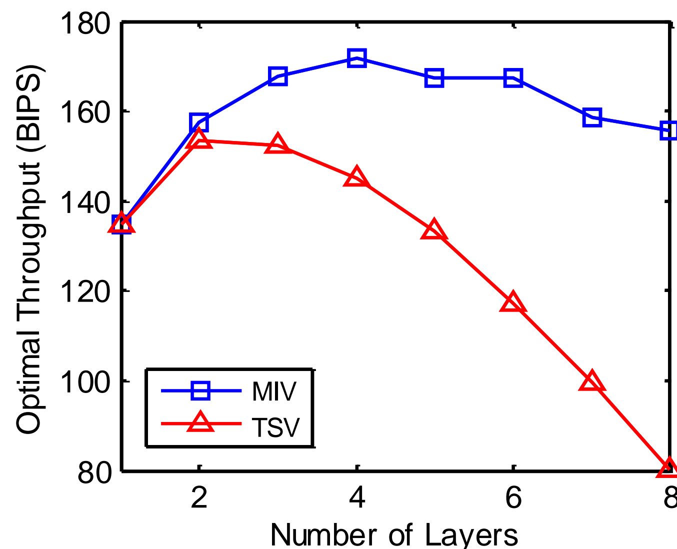


Figure 109: Comparison of the optimal throughput between MIV- and TSV-based processors for various number of logic layers.

7.3 Heterogeneous Multi-core Integration

Heterogeneous multi-core processors are proposed to further improve the power efficiency. In this section, both symmetric and asymmetric multi-core processors are investigated by using conventional planar CMOS, advanced FinFET technology, and emerging TFET technology for various given power budgets at various levels of parallelisms. Previous work analyzed the CMOS-TFET heterogeneous integration by

using cycle-accurate simulation method [46], demonstrating that the energy efficient TFET cores can save leakage and dynamic energy with little performance degradation. This estimation is accurate, but it is time-consuming and only investigates one specific system configuration. To truly understand the potential benefit and advantage of the TFET technology, device- and system-level co-optimization is required to find the best design points, including supply voltage, number of cores and logic transistors, logic-to-cache ratio, and big-to-small core ratio, to maximize the chip performance under certain area/power budgets.

7.3.1 Symmetric Multi-Core Analysis

In Figure 110 (a), an optimal number of cores can be observed to achieve the maximum chip throughput (blue curve). In Figure 110 (b), the stacked bar chart of the average time per instruction is plotted, which is the reciprocal of the chip throughput. Each bar has four components, including the time due to the serial and parallel execution of the logic core, DRAM latency, and the link latency. The reason for the existence of the optimal number of cores is that as the number of cores increases, the computational capability keeps increasing thanks to the highly parallelized program and a larger number of power-efficient small cores. It can be seen from the red segment in Figure 110 (b) that the execution time of the logic cores during the parallel part of the program keeps decreasing. Meanwhile, the miss rate also increases due to the smaller die size budget for the cache inside each core, causing a significant increase in the memory channel utilization rate during the parallel execution (black curve in Figure 110 (a)). The overall chip throughput suffers significantly from the cache miss penalty, which increases the execution time due to the DRAM and the link latency, as can be seen from the green and yellow segments in Figure 110 (b). This explains why the area percentage for logic keeps decreasing (red curve in Figure 110 (a)) and the system prefer to assign more area for the

cache. Moreover, during the serial part of the program, the execution time increases because the core becomes simpler, shown as the blue segment in Figure 110 (b). Therefore, the chip throughput starts decreasing when the number of cores is beyond a certain point.

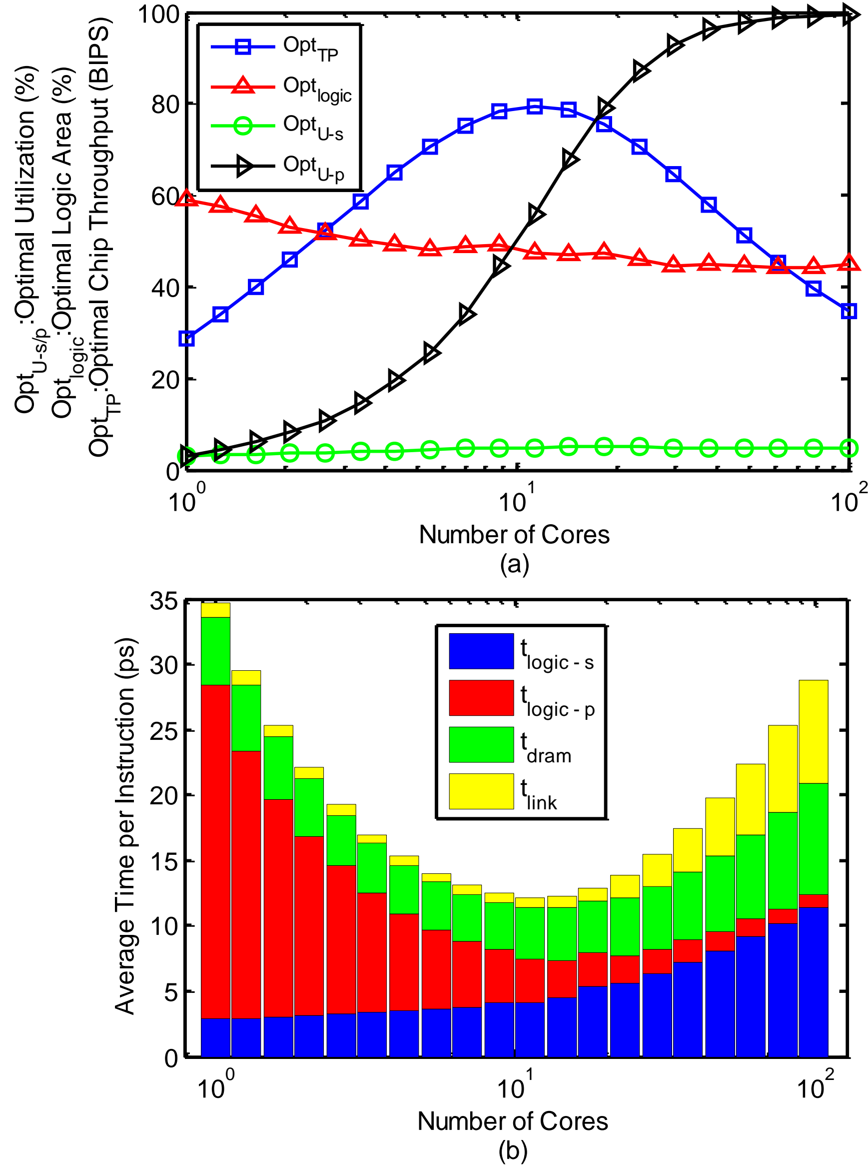


Figure 110: Various design parameters and performance metrics versus number of cores for a symmetric multi-core processor. (a) four bullets in the legend correspond to the optimal chip throughput, logic area percentage, link utilization during serial and parallel part of the program, respectively. Here, the parallelism is assumed to be 0.9; the die are is 200 mm^2; and the power budget is 200W.

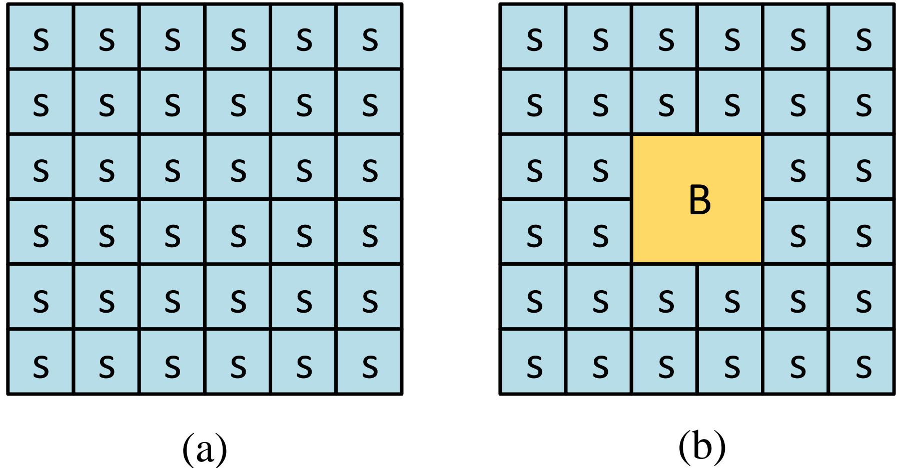


Figure 111: Diagram for the symmetric versus asymmetric multi-core processor.

7.3.2 Asymmetric Multi-core Analysis

Asymmetric multi-core is one configuration that can provide fast serial and parallel execution at the same time due to the big core and power-efficient small cores existing on the same die, as shown in Figure 111. In this subsection, one more parameter, big-to-small core ratio, is optimized based on the system-level design methodology. Here, it is assumed that during the serial execution, only the big core is active; during the parallel execution, only the small cores are active.

From Figure 112, an optimal big-to-small core ratio and optimal number of cores exist to reach the maximum chip throughput for a 200 mm^2 asymmetric multi-core processor. This is because if the ratio is too small, the big core is not powerful enough and the serial execution becomes the bottleneck; if the ratio is too big, the big core becomes too large and not power/area efficient as shown in Figure 67 (a). The speedup of the serial part is overshadowed by the performance degradation of the parallel execution due to the simpler small cores. For various parallelisms of the program, optimal design parameters and performance metrics are shown in Table 14, where one can observe that as the parallelism increases, the optimal number of cores keeps increasing, and the ratio

of the big-to-small core keeps decreasing. This is because for a processor designed for highly parallelized programs, the parallel computing power is more important. Therefore, a processor with a larger number of more powerful small cores is more desirable than one with a larger big core. This causes the decrease of serial performance and the increase of parallel performance, and the overall chip throughput increases as the parallelism increases. Another observation is that the logic percentage keeps decreasing for both big and small cores. The reason is that as the total chip throughput increases thanks to the highly parallelized program, the cache miss per second increases. Thus, a larger cache is required to suppress the increasing miss penalty associated with the DRAM and link latency.

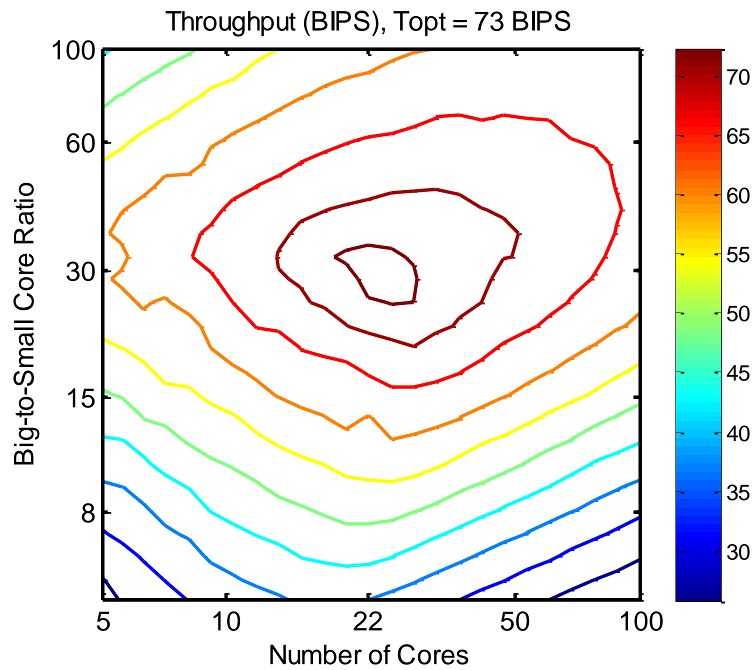


Figure 112: Contour figure of the chip throughput versus big-to-small core ratio and number of cores in an asymmetric multi-core processor. Here, the parallelism of the program is 0.8, and the total power budget is 150W.

Table 14: Design Parameters and Performance Metrics for an Asymmetric Multi-Core Processor at Various Parallelisms of the Program.

Parallelism		0.40	0.60	0.80	0.95
Number of Cores		23	25	29	37
Big-to-Small Core Ratio		75.2	47.0	22.7	12.4
Big Core Parameters	Core Area (mm^2)	154.3	132.1	90.4	51.5
	Logic Percentage	57.25	54.56	50.25	48.48
	Frequency (GHz)	3.89	4.19	3.82	4.76
	V_{dd} (V)	0.50	0.48	0.50	0.47
	N_{gate} (Million)	129.28	107.24	109.62	64.07
	Channel Width (F)	16.24	15.93	8.97	8.30
Small Core Parameters	Core Area (mm^2)	2.07	2.82	3.98	4.16
	Logic Percentage	37.81	32.30	26.83	16.53
	Frequency (GHz)	6.70	6.70	6.60	6.73
	V_{dd} (V)	0.46	0.46	0.46	0.46
	N_{gate} (Million)	3.49	4.06	4.77	3.11
	Channel Width (F)	3.78	3.74	3.71	3.61
Throughput (BIPS)	Serial	27.22	26.25	23.89	20.78
	Parallel	95.39	118.42	150.99	169.34
	Overall	34.64	42.97	73.14	124.73

7.3.3 Heterogeneous Multi-core Analysis

Heterogeneous processors have attracted much attention because of their better power efficiency. Basically, the high-performance big core can execute fast during the serial part of the program, while small cores can execute the parallel part more power efficiently.

Based on various power consumption budgets, four configurations of the multi-core processors are investigated. Three acronyms CMOSHP, CMOSLP, and TFET represent three technologies: CMOS high-performance devices, CMOS low-power devices and TFETs, respectively. The third configuration is a heterogeneous multi-core processor, which uses CMOS high-performance devices for the big core and TFETs for the small cores. With a high power consumption budget, processors using CMOS high-performance FinFETs can achieve maximum throughput due to their high driving current.

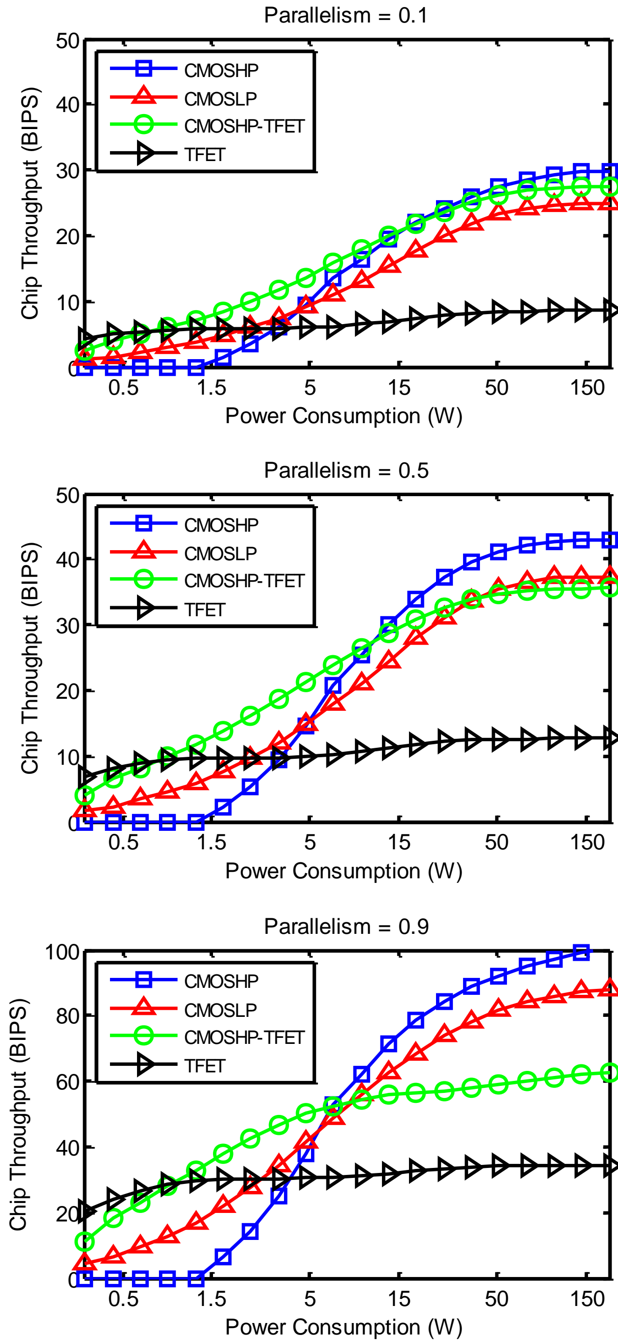


Figure 113: Chip throughput versus power consumption of heterogeneous multi-core processors with four configurations running a program with various parallelisms. In the legend, three acronyms CMOSHP, CMOSLP, and TFET represent three technologies: CMOS high-performance devices, CMOS low-power devices and TFETs, respectively. CMOSHP-TFET represents a heterogeneous multi-core processor with a big core implemented with CMOS high-performance devices and small cores implemented with TFETs.

Both TFET and heterogeneous CMOSHP-TFET processors perform worse than their CMOS high-performance and low-power FinFET counterparts because the low ON current of TFETs greatly suppress the parallel execution speed. As the power consumption budget goes down to below 40W, the heterogeneous CMOSHP-TFET processor surpasses the CMOS low-power processor, because the CMOS high-performance core can perform much better than the CMOS low-performance core during the serial execution even though the parallel execution speed of the TFET cores is still slower than its CMOS counterpart. When the power consumption budget is below 10W, the heterogeneous processor can provide the best performance due to the fact that the CMOS high-performance processor suffers significantly from the high leakage power, causing a substantial frequency drop. At a 5W power consumption budget, a CMOSHP-TFET processor can provide 45% throughput improvement compared with a CMOS high-performance processor. With an ultra-low power consumption budget, the TFET processor can provide maximum throughput due to its ultra-low supply voltage. For various parallelisms, the throughput increases as the parallelism increases for all technologies due to the advantage of parallel computing. Another observation is that the maximum power consumption budget for heterogeneous CMOSHP-TFET processors to surpass CMOS high-performance processors decreases as the parallelisms increases. This is because the parallel performance of the heterogeneous processors saturate due to the limited ON current of the small TFET cores, which makes them less competitive under a high power consumption budget, especially when they are running a program with high parallelism.

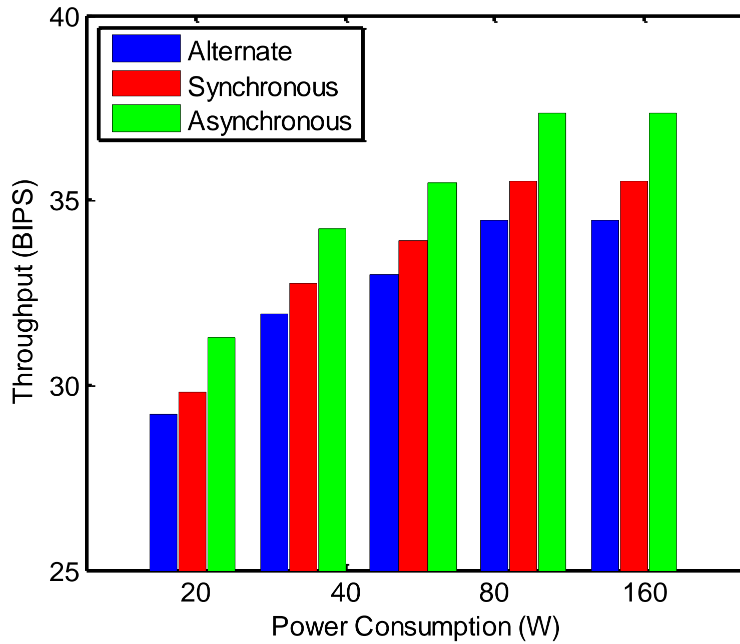


Figure 114: Optimal chip throughput versus power consumption for various multi-core scenarios, including switching big or small cores alternately and simultaneously.

The simulation above is based on the assumption that only small cores are running during the parallel execution. If both big and small cores are operating simultaneously, the throughput model shown in (20) needs to be extended as

$$TP_b = \frac{1}{\frac{CPI_b}{f_b} + M_b\left(\frac{t_{dram}}{N_{pr,b}} + t_{link}\right)} \tag{36}$$

$$TP_s = \frac{n-1}{\frac{CPI_s}{f_s} + M_s\left(\frac{t_{dram}}{N_{pr,s}} + t_{link}\right)} \tag{37}$$

$$t_{link} = \frac{D}{BW}\cdot\left[1 + \frac{U}{2(1-U)}\right] \tag{38}$$

$$U = (TP_s \cdot M_s + TP_b \cdot M_b)\frac{D}{BW} \tag{39}$$

where TP, CPI, f, N_{pr}, and$M$ are the throughput, computational CPI, clock frequency, average parallel memory requests per cycle, and cache miss rate, respectively. The subscripts 'b' and 's' represent the big and small cores, respectively. t_{link} is the link latency that depends on the bandwidth BW, data package size D, and link utilization U. By solving the equations above, we can get three solutions based on the cubic equations, but only one of the solutions of U is between 0 and 1, which is the meaningful solution for this problem.

The results for the asymmetric heterogeneous CMOS-TFET hybrid multi-core processors are shown in Figure 114, where blue bars represent the processor using only small cores during the parallel execution. With all cores active during the execution, the processors with synchronous operation (red bars) can have 3% throughput gain; and if the processors support asynchronous operation (green bars), the throughput improvement can increase to 8% compared with the processors running only small cores during the parallel execution.

Another important metric for a processor especially for battery operated systems is the total energy for a given task or energy per instruction. For the energy consumption of a multi-core processor, it is estimated as

$$Energy = [P_b + (n-1)k_s P_s] \cdot t_b + [(n-1)P_s + k_b P_b] \cdot t_s \quad (40)$$

where P_b and P_s are the power consumption for big and small cores, respectively, k_b and k_s are the leakage percentage of big and small cores, respectively, t_b and t_s are the serial and parallel execution time, respectively, and n is the total number of cores.

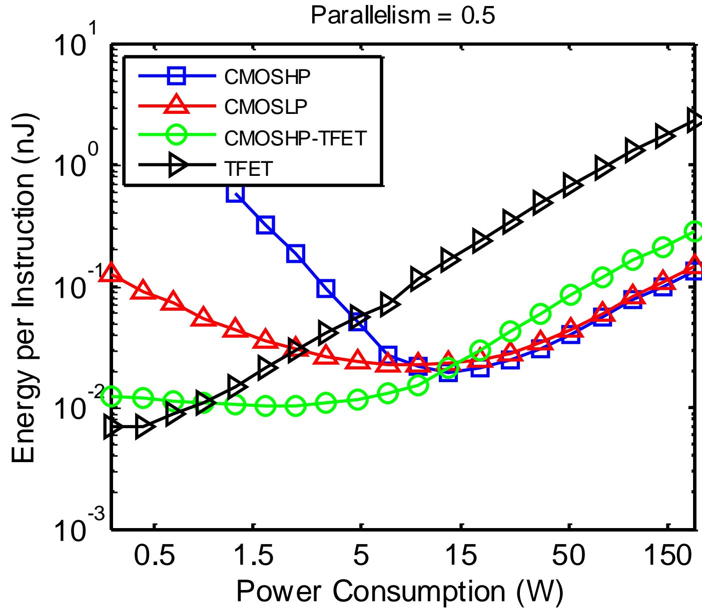


Figure 115: Energy consumption per instruction versus power consumption of heterogeneous multi-core processors with four configurations running a program with parallelism of 0.5.

The energy consumption per instruction is shown in Figure 115, where one can observe that TFET processors can provide the minimum energy due to their supreme performance at ultralow-power budget and small leakage power. However, as the power consumption increases, the performance saturates due to the limited ON current, and the long execution time causes TFET processors to have the largest energy consumption when the power consumption is beyond 5W. For the CMOSHP-TFET heterogeneous processor, it consumes the least energy consumption under low power consumption budget between 1W to 15W. This is because it has a faster serial execution compared with the TFET and CMOS low-power processors, and it consumes less leakage power compared with CMOS high-performance processors. However, when the power consumption is beyond 15W, both CMOS high-performance and low-power processors surpass CMOSHP-TFET processors due to the low performance of the small TFET cores. With a large power consumption budget, although CMOS low-power processors perform

worse than CMOS high-performance processors according to Figure 113, the energy consumption of two types of CMOS processors are close because of the smaller leakage of CMOS low-power processors. At a 5W power consumption budget, a CMOSHP-TFET processor can provide 50% less energy consumption compared with a CMOS low-power processor. Another observation is that optimal power consumption exists for four types of processors to achieve the minimum energy consumption. The reason is that if the power consumption budget is too small, the energy is limited by the slow execution time; when the power is too large, the processor becomes power inefficient, which increases the total energy consumption.

7.4 Chip/Package Co-Optimization

The original compact PDN model is developed and validated in [114], which provides a fast way to estimate the power supply noise based on certain parameters including resistance, capacitance, and inductance on both chip and package. In this section, this model is modified for analyzing the impact of the supply noise on the overall chip throughput of a processor implemented with the power-gating technique. Various packaging configurations are investigated and compared with each other. Optimal parameters are obtained for each configuration, including supply voltage, width of the logic gates and the sleep transistors. A smaller EDP is observed for the processors using the power-gating technique at the optimal design point. The optimal position of the decap is demonstrated to be application dependent, indicating that using center decap is more preferable if the power gating activity is relatively low.

7.4.1 Package and Circuit-Level Simulation

Figure 116 shows three packaging configurations that are investigated in this chapter. They are composed of a chip on a package through C4 connection with BGA to connect

to PCB, where the decap position differs, including no decap, side decap, and center decap, respectively. The basic parameters, including resistance, inductance, and capacitance for chip, package, C4s, and BGAs, are obtained based on the typical values reported in [114] and the ITRS projections [132].

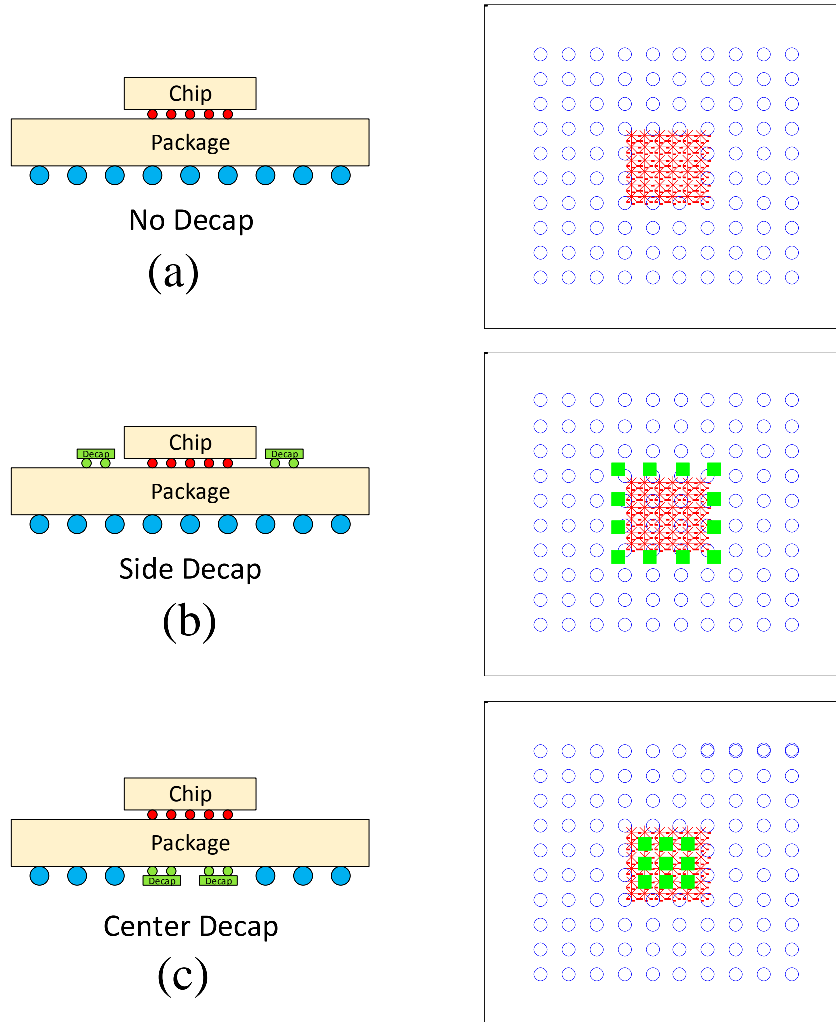


Figure 116: Three packaging configurations. (a) without off-chip decap (b) with side decap (c) with center decap.

The current density is assumed to be 100A/cm^2, and the decap is 250nF each. When a step signal is applies at t = 0, which represents the core transition between idle and active states, Figure 117 shows the envelops of the transient responses of the worse noise

on the power supply network on chip. One can see that without decap insertion, the settling time is longer compared with the one using side decap. Using center decap reduces the settling time, however, the DC drop becomes larger. This is because there is no direct current path from chip through package to PCB. The current from the chip has to horizontally traverse to the BGAs that locate peripherally on the package to reach PCB. Therefore, the large resistance causes a larger DC drop, which will degrade the chip performance at the steady state.

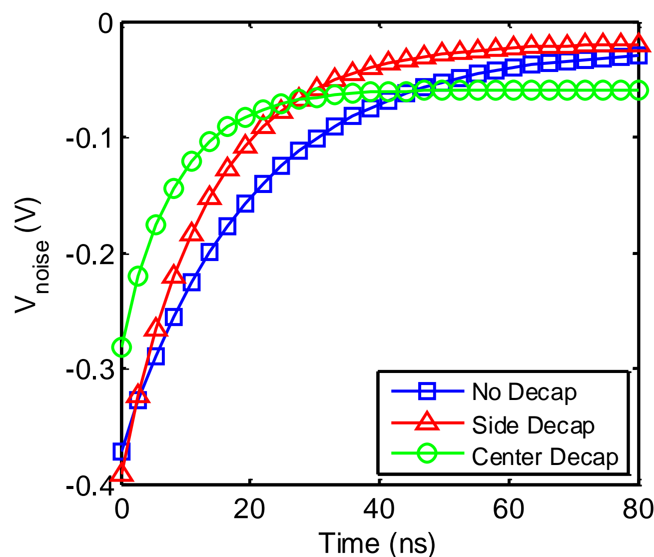


Figure 117: The envelop of the transient response of the power supply noise for three package configurations without inserting sleep transistors.

If the power-gating technique is implemented into the chip, the original PDN model in [114] needs to be modified. First, one more sleep transistor needs to be added between C4s and the PDN on chip. Second, the initial condition of the PDN needs to be obtained based on the IV characteristics and the width of the logic gates and sleep transistors. After these two modifications are made, the transient response of the virtual ground is shown in Figure 118. When t = 0, which represents the idle state of the logic core, the voltage of the virtual ground is close to the supply voltage because the leakage resistance

of the logic gates is much smaller than that of the gating transistor, which raises the voltage. When the logic core is turned on at t = 10 ns, the virtual ground is pulled down to a value that is close to 0V, but a certain DC offset still exists because of the DC drop on the sleep transistor. This DC drop will cause a performance penalty if the total power budget remains the same. Another performance penalty of performing power-gating comes from the settling time due to the supply noise. In this chapter, the settling time is estimated as the time beyond which the voltage difference between the virtual ground and the actual ground remains less than 1% of the supply voltage. The processor can only start executing instructions after the settling time. The third penalty of using power-gating is the energy overhead during the beginning of the idle state. The internal parasitic capacitance and the decap capacitance on the virtual ground will be charged due to the leakage of the logic gates. All the aforementioned performance and energy overheads will be included in the analyses and benchmarking.

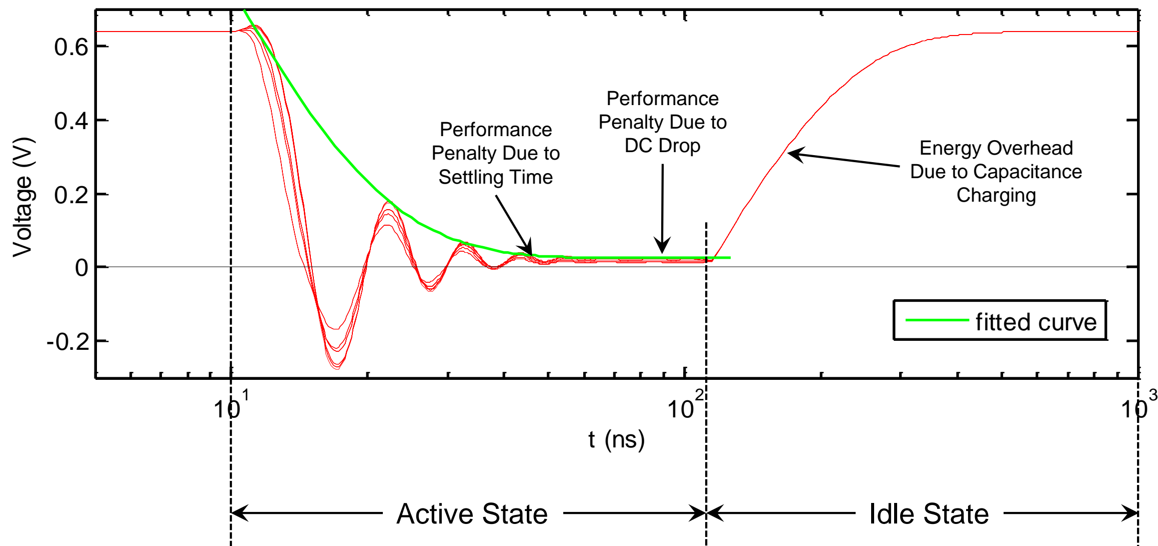


Figure 118: Transient response of the virtual ground during active and idle states.

The relations among DC drop, transistor area, and the width of the transistor are depicted in Figure 119. The area of the sleep transistor is estimated as $W \times L$, where W is

the width of the sleep transistor and L is taken as 8 times the minimum feature size F, which is based on the design rules in [50]. One trade-off here is that one can either design a logic core with a large sleep transistor to reduce the performance penalty associated with the DC drop or with a small sleep transistor to reduce the area overhead.

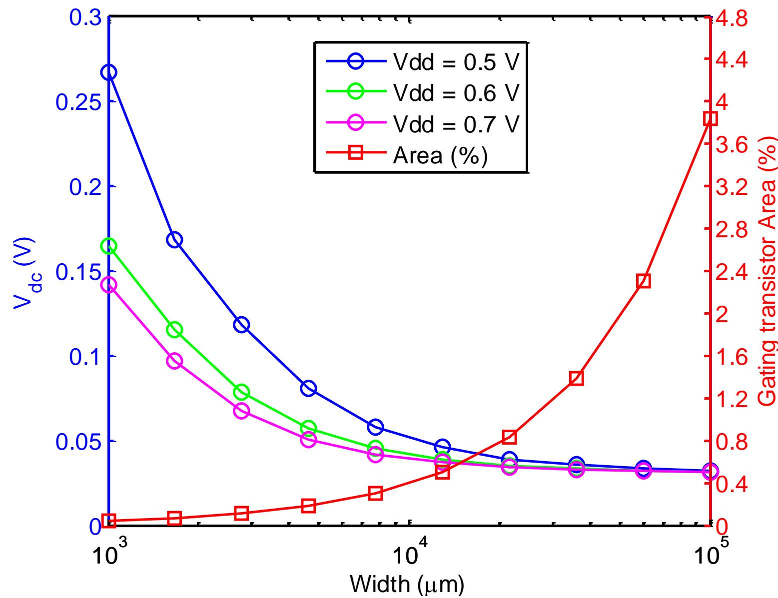


Figure 119: DC drop and sleep transistor area as a percentage of the chip area versus width of the transistor under various supply voltages.

7.4.2 System-Level Optimization Results

Based on the system-level design methodology and the modified power supply noise model described in Chapter 4, an optimization is performed for a 40 mm^2 8-core processor with a 100 W/cm^2 power density constraint. Figure 120 shows the relative throughput, energy, and EDP versus the width of the sleep transistor. The solid and dash lines represent processors implemented with power-gating and clock-gating techniques, respectively. These performance metrics are relative to those of a processor with no power gating. Various curves are simulated under different average numbers of sequential instructions. These instructions are assumed to be executed by a single logic

core continuously before changing to the parallel execution. During the sequential execution, only one core is at the active state, while the rest are at the idle state. A larger number of instructions means that the power-gating is performed less frequently. Also, in this chapter, the number of sequential instructions is assumed to be the same as the parallel instructions. For the power-gating processors, the idle cores are assumed to consume negligible power compared with the active cores; for the clock-gating processors, the idle cores are assumed to only have leakage power consumption, which is obtained from the optimal design during the active state.

From Figure 120 (a), an optimal throughput can be observed for various average numbers of sequential instructions. The reason is that the DC drop keeps decreasing as the width of the sleep transistor increases, which lowers the performance overhead. However, when the width becomes too large, the large area overhead of the sleep transistor significantly reduces the logic core area based on the fixed total die size budget, which decreases the overall throughput.

The optimal throughput decreases as the processor needs to switch between active and idle states more frequently, which is because of the delay due to the constant settling time causes a larger impact on the performance. Another observation is that when the processor needs to frequently perform power-gating, the optimal width of the sleep transistor decreases. Although a smaller sleep transistor provides a smaller current for charging those capacitances on the virtual power network, the settling time is dominated by the AC noise in the PDN. Therefore, a larger resistance of the sleep transistor reduces the settling time due to the larger damping factor. If the switching frequency is too high, the delay overhead due to the settling time becomes dominant, making a sleep transistor with a small width and a large resistance more preferable.

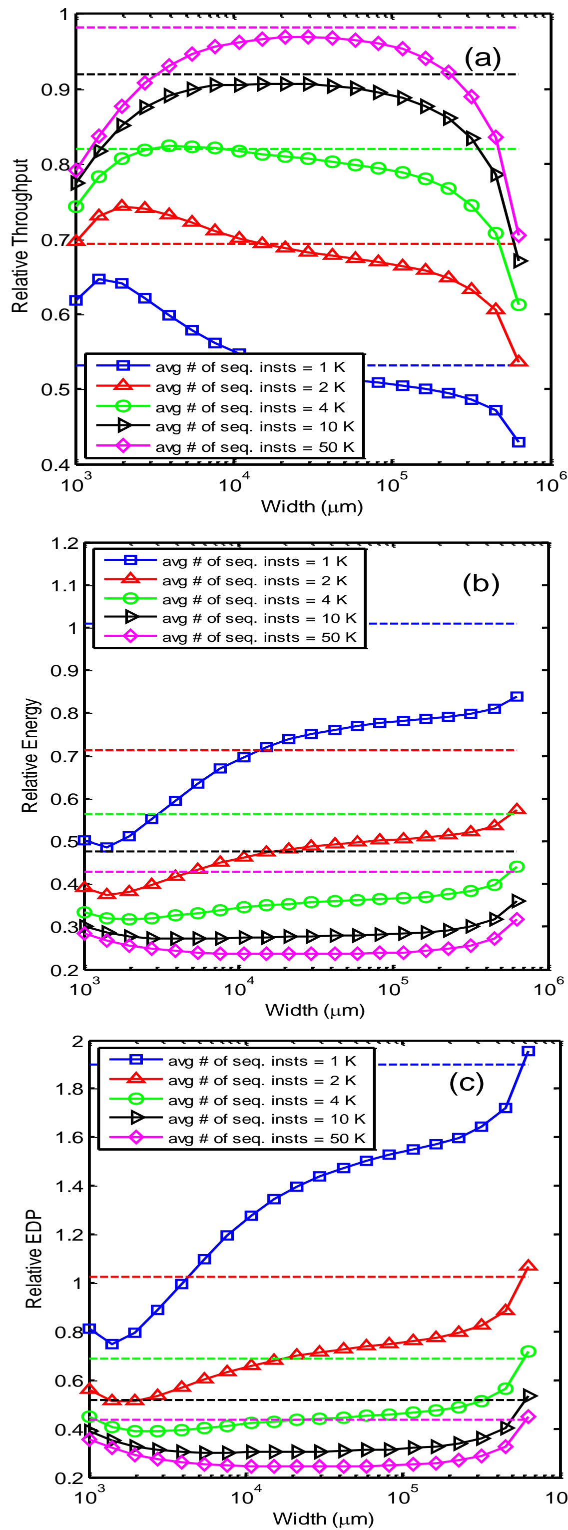


Figure 120: Relative throughput, energy, and EDP versus width of sleep transistors. The solid and dash lines represent processors implemented with power-gating and clock-gating techniques, respectively.

Compared with using clock-gating technique, using power-gating loses performance at a low switching frequency due to the DC drop caused by the sleep transistors. When the switching frequency increases, the optimal throughput of a processor using power-gating surpasses the one using clock-gating because the resistance of the sleep transistor reduces the settling time. The benefit of using the power-gating technique can be observed in the energy consumption, shown in Figure 120 (b). Up to 50% of the energy consumption can be saved by using power-gating compared with the processors using the clock-gating technique. An additional 25% of the energy saving can be observed compared with the processors without any gating techniques. The overall EDP saving for the power-gating processors varies from 25% to 75% compared with processors without gating techniques, which is based on how frequently the processors need to switch between active and idle states, according to Figure 120 (c).

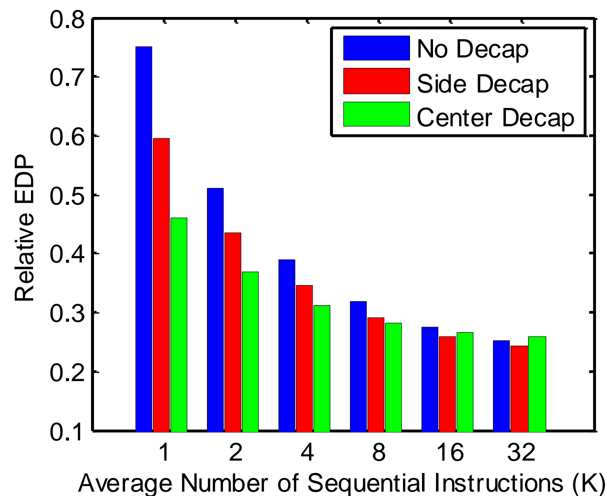


Figure 121: Optimal relative EDP versus average number of sequential instructions for three package configurations.

Figure 121 shows the optimal relative EDP versus the average number of sequential instructions for three package configurations. The blue column without using decap represents the optimal EDP shown in Figure 120 (c). One can observe that the processors

using side decap can always provide lower EDP than those without decap. Using center decap can provide the lowest EDP if the processor does not need to frequently switch between active and idle states. If the switching frequency becomes too high, the benefit of fast settling time is overshadowed by the large DC drop due to the lack of direct current path from the chip to the PCB board. Therefore, the optimal position of the decap insertion is influenced by the application. Using center decap is more preferable if the application does not need to frequently switch between sequential and parallel parts.

7.5 Conclusions

In this chapter, benchmarking and optimization are performed for various system-level innovations and technologies based on the device-, circuit-, and system-level compact models and the hierarchical optimization engine developed in Chapters 2-4.

For the 3D integration, analyses and comparisons are performed for systems implemented with TSV- and MIV-based 3D integration technologies. In addition, the impact of via diameter and capacitance on the overall system throughput has been quantified. It is demonstrated that for the same die area and thermal constraint, an MIV-based processor offers over 25% improvement in computational throughput as compared with its 2D counterpart.

For the multi-core processor, optimal numbers of cores and logic transistors, logic-to-cache ratios, big-to-small core ratio, and device-level design parameters are obtained to maximize the overall chip throughput. For a heterogeneous CMOS-TFET multi-core processor, about 45% throughput improvement and 50% energy reduction are observed compared with a FinFET processor at a 5W power budget.

Optimization and benchmarking are performed for a processor implemented with the power-gating technique under various package configurations. Optimal widths of the sleep transistors are obtained based on how frequently the processors need to switch

between active and idle states. Up to 75% of the EDP saving is observed for the processors using the power-gating technique with a low switching frequency. In addition, the optimal position of the decap insertion is demonstrated to be application dependent.

CHAPTER 8 CONCLUSIONS AND FUTURE WORK

This chapter concludes the dissertation. Main contributions and implications from the current research are listed. Future work that can be built upon this study is proposed.

8.1 Conclusions of dissertation

As the Si CMOS technology and copper interconnects approach their physical limits, many novel devices, interconnects, and systems are proposed to refine or even replace the conventional technologies. Since large amounts of capital investment, research, and engineering efforts are needed for each innovation before massive production, a fast and efficient hierarchical design methodology is proposed in this dissertation to alleviate the challenge. It allows us to benchmark and optimize any novel technology at the overall chip throughput instead of the just delay or energy dissipation of an individual device or interconnect or a simple logic circuit. Multiple device-, interconnect-, and system-level design parameters, such as the supply voltage, channel width and length, oxide thickness, interconnect width and aspect ratio, complexity and number of the logic cores, logic-to-cache ratio, and etc., are optimized simultaneously under a variety of constraints, including the power consumption, maximum operating temperature, die size area, yield, and power delivery noise. The design methodology and optimization approach are validated by comparing the simulation results with the measurements in commercial processors.

At the device-level benchmarking and optimization, multiple device- and system-level design parameters are simultaneously co-optimized to maximize the overall chip throughput using conventional Si CMOS devices under certain power, thermal and die size budgets. A high-performance 25 mm^2 FinFET single-core processor can provide 26%

more throughput than its planar high-performance CMOS counterpart at the 16nm technology node. For various technology nodes, an accurate power-law relation exists between the throughput and the die size area for a processor with relatively small area. When the core area is beyond a certain point, the optimal throughput saturates, indicating that a large core is not desirable. For the novel GPNJ-based processors, several device-level parameters including the supply voltage, control voltage, gap distance, and oxide thickness are optimized, where 2.1X throughput improvement is observed for a sharp-corner GPNJ core compared to its Si CMOS counterpart implemented at the 16nm technology node. This advantage is predominantly because of the smaller output resistance, which reduces both device and interconnect delay and saves the power for repeaters. The throughput improvement of a curved-corner GPNJ core drops to 66% due to the larger input capacitance. The scalability analyses are also performed, and insightful trends of various design parameters and performance metrics are shown down to the 7 nm technology node for both single- and multi-core processors. TFETs have excellent performance at the low power density range due to the low supply voltage. The limitations of the interconnection and large leakage current at a high supply voltage restrict the driving current, leading a lower performance of the TFETs at a high power density and a large core size compared with CMOS. For the process variation analyses, using adaptive supply voltage and frequency tuning can significantly improve the throughput and yield for a given power density constraint. For a multi-core processor, disabling the slowest cores helps to reduce the deviation and increase the chip throughput. The system with asynchronous operation can provide the maximal throughput. The proposed power reallocating technique can achieve further throughput improvement by allocating the power to those more power efficient cores. For a 10-core processor, the results show that the yield can be improved from 80.5% to 94.8% at a 4GHz frequency

target; if the yield target is 90%, then the average frequency can increase from 3.8GHz to 4.1GHz.

For the interconnect benchmarking and optimization, a paradigm shift in local interconnect technology design is presented, indicating that the interconnect performance becomes more resistance dominated rather than capacitance dominated. Two implications are addressed, and one potential solution to rebalance the interconnect resistance and capacitance by increasing the interconnect width beyond half-pitch has been evaluated. At the optimal interconnect width with an aspect ratio of 3, the EDP of a VFET circuit improves up to 55% at the 5 nm technology node. The optimal interconnect relative width depends on the technology node and device structure. In addition, a novel aluminum-copper hybrid interconnect architecture is proposed to suppress the impact of size effects on sub-20nm wide interconnects. Using Al for the short signal interconnect with a subtractive process can potentially achieve a smaller resistance per unit length at a small dimension without suffering from the electromigration because of the self-healing effect for interconnects that conduct alternating currents. Compared with the conventional Cu interconnect technology, the proposed scheme is projected to offer between 50 to 100% improvement in the clock frequency of a logic core implemented at the 7 nm technology node. A comprehensive optimization and benchmarking are also performed for the multi-layer graphene interconnects. The results show that a single core using graphene interconnects can potentially have a higher throughput within the same power density and die size area because of the power saving offered by the low capacitance per unit length of graphene interconnects. Optimal numbers of graphene layers have also been observed to reach the maximum throughput, and the number of layers increases with the increase of the MFP and edge roughness. A smaller core can better utilize graphene interconnects because of the shorter interconnects. For the interconnect variation analyses, various sources of variations are compared, such as the

CD line/core, CD spacer, etch, CMP, and overlay variations. The CMP variation has the largest impact on the overall clock frequency of the processor based on the default 3σ values extracted from different fabrication processes, including LELE double patterning, SADP, and SAQP. In addition, the impact of the interconnect variation increases as the technology node goes down, especially for the LELE process with a larger σ value because of the overlay variation. A larger frequency drop is observed for the SAQP process compared with the SADP process due to an additional spacer sidewall variation. Up to 8% and 16% of the frequency drops are observed based on extracted 3σ values and twice of those as the worst case scenario, respectively.

For the system-level benchmarking and optimization, MIV- and TSV-based 3D single- and multi-core systems are investigated and compared. The impact of the via diameter and capacitance has been quantified in terms of the system-level throughput. Due to the savings in both average and total length of the interconnects, an optimized MIV-based 3D multi-core processor provides more than 25% improvement in chip computational throughput compared with its 2D counterpart with the same footprint area and thermal budgets. For the heterogeneous multi-core processor analyses, optimal numbers of cores and logic transistors, logic-to-cache ratios, big-to-small core ratio, and device-level design parameters are obtained to maximize the overall chip throughput. A heterogeneous CMOS-TFET multi-core processor shows about 45% throughput improvement and 50% energy reduction compared with a FinFET processor at a 5W power budget. For the power delivery integrity analysis, chip and package co-optimization and benchmarking are performed for a processor implemented with power-gating technique. Optimal widths of the sleep transistors are obtained to maximize the overall chip throughput for various package configurations and switching frequency between active and idle states, indicating that using power-gating can provide energy

saving with small performance penalty. Optimal sleep transistor widths and the optimal position of the inserted decap varies, depending on the specific application.

In conclusion, a fast hierarchical optimization is developed and validated to efficiently quantify, optimize, and predict the performance of a VLSI system implemented with various novel device options, interconnect architectures, and system innovations.

8.2 Future Work

In this section, several potential extensions of this dissertation are projected on the device-, interconnect-, and system-level performance modeling and optimization.

8.2.1 Device-Level Modeling and Analyses

All emerging devices and interconnects studied in this dissertation use the voltage as the computational state variables. One possible extension of this work is to incorporate novel devices using other state variables, such as electric dipole, magnetic dipole, or even orbital state based devices. The representative devices include SWD, ASL, STO, spin torque based domain wall (STT/DW), and NML [17].

Note that these novel devices may have fundamentally different operation principles, and parts of them use majority gates to perform logic operation. Therefore, some of the assumptions and models in the current hierarchical optimization engine need to be updated. For example, the Rent's constant is assumed based on CMOS logic herein. With novel spin-based logic gates, this constant can be noticeably different because the number of devices and the connection among devices required to achieve the same logic function may vary.

8.2.2 Circuit- and Interconnect-Level Modeling and Analyses

In Chapter 3, an ARM core is investigated by performing synthesis, placement, and routing, and then the extracted parasitics from copper interconnects are updated with graphene interconnects. To better quantify the improvement of the multi-layer graphene interconnects, the placement and routing can be performed directly for the graphene interconnects. The improvement is expected to be even larger than the value reported in this dissertation since the placement and routing tool will optimize the circuit based on the graphene properties such as the contact resistance, sheet resistance, and capacitance.

Crosstalk models can be included into the design methodology because the signal integrity is crucial for today's VLSI systems. As the technology scales, the noise margin reduces due to the supply voltage scaling and higher operation frequency. In many cases, the dominant signal distortions and logic failures come from the interconnects. It would be insightful to analyze the impact due to the crosstalk of the interconnects on the overall throughput.

8.2.3 System-Level Modeling and Analyses

For the empirical CPI model, a better model can be developed to take into account the impact of the logic depth. It is also worthy of investigation the interaction between the device and interconnect technologies and the logic depth. Logic depth could be one more parameter that needs to be optimized to achieve the overall chip throughput. The optimal logic depth could be dependent on various constraints such as the power dissipation.

For a multi-core processor, one improvement that can be made is to develop a more elaborate model to take into account the core-to-core communication overhead during the evaluation of the overall chip throughput for executing a program with a certain parallelism, where the memory organization and topology can be investigated as well.

Reliability is another topic that can be investigated because as the technology scales down, lifetime reliability becomes a challenge. Many failure mechanism can be incorporated, such as the hot carrier injection, electromigration, negative bias temperature instability, stress migration, and time dependent dielectric breakdown [166-168]. Trade-offs can be made to achieve the maximum overall chip throughput.

The system-level performance of a processor using reconfigurable logic gates, such as the GPNJ devices, can be analyzed. One approach is to obtain the average equivalent number of logic gates for each specific type of device structure to achieve certain logic operations.

Some non-Boolean neuromorphic computing based architectures have been proposed, such as the artificial neuron network. Another possible extension of this work is to include the performance modeling and benchmarking for those systems.

8.2.4 Hierarchical Optimization Engine Development

A more user-friendly graphic user interface (GUI) can be developed to enable people who are working on emerging device or interconnect technology to better utilize the hierarchical optimization engine. It is crucial for professionals to understand the potential benefits of various novel technologies and to avoid development with little benefit.

LIST OF PUBLICATIONS

Journal Papers:

[1] **C. Pan**, P. Raghavan, D. Yakimets, P. Debacker, F. Catthoor, N. Collaert, Z. Tokei, D. Verkest, A. Thean, and A. Naeemi, "Technology/System Co-Design and Benchmarking for Lateral and Vertical GAA Nanowire FETs at 5nm Technology Node," to be published in *IEEE Transactions on Electron Devices (TED)*.

[2] **C. Pan**, R. Baert, I. Ciofi, Z. Tokei, and A. Naeemi, "System-level Variation Analysis for Interconnection Networks at Sub-10nm Technology Nodes," *IEEE Transactions on Electron Devices (TED),* July, 2015.

[3] **C. Pan**, P. Raghavan, A. Ceyhan, F. Catthoor, Z. Tokei, and A. Naeemi, "Technology/Circuit/System Co-Optimization and Benchmarking for Multilayer Graphene Interconnects at Sub-10nm Technology Node," *IEEE Transactions on Electron Devices (TED),* May, 2015.

[4] **C. Pan** and A. Naeemi, "A Fast System-Level Design Methodology for Heterogeneous Multi-core Processors Using Emerging Technologies," to be published in *IEEE Journal on Emerging and Selected Topics in Circuits and Systems (JETCAS),* March, 2015.

[5] **C. Pan** and A. Naeemi, "A Paradigm Shift in Local Interconnect Technology Design in the Era of Nanoscale Multi-Gate and Gate-All-Around Devices," *IEEE Electron Device Letters (EDL)*, March, 2015.

[6] **C. Pan** and A. Naeemi, "A Proposal for a Novel Aluminum-Copper Hybrid Interconnect Technology for the End of Roadmap," *IEEE Electron Device Letters (EDL)*, Feb, 2014.

Conference Papers:

[7] **C. Pan**, P. Raghavan, F. Catthoor, Z. Tokei, and A. Naeemi, "Technology/Circuit Co-Optimization and Benchmarking for Multilayer Graphene Interconnects at Sub-10nm Technology Node," *IEEE International Symposium on Quality Electronic Design (ISQED)*, March, 2015.

[8] **C. Pan** and A. Naeemi, "System-Level Chip/Package Co-Design for Multi-Core Processors Implemented with Power-Gating Technique," *IEEE Conference on Electrical Performance of Electronic Packaging and Systems (EPEPS)*, Oct, 2014.

[9] A. Naeemi, A. Ceyhan, V. Kumar, **C. Pan**, R. M. Iraei, and S. Rakheja, "BEOL Scaling Limits and Next Generation Technology Prospects," *IEEE/ACM Design Automation Conference (DAC)*, June, 2014.

[10] **C. Pan** and A. Naeemi, "System-level Variation Analysis for Interconnection Networks," *IEEE International Interconnect Technology Conference (IITC)*, May, 2014.

[11] **C. Pan**, S. Mukhopadhyay and A. Naeemi, "An Analytical Approach to System-level Variation Analysis and Optimization for Multi-core Processors," *IEEE International Symposium on Quality Electronic Design (ISQED)*, March, 2014.

[12] **C. Pan** and A. Naeemi, "System-level Analysis for 3D Interconnection Networks," *IEEE International Interconnect Technology Conference (IITC)*, June, 2013.

[13] **C. Pan,** A. Ceyhan, and A. Naeemi, "System-level Optimization and Benchmarking for InAs Nanowire Based Gate-All-Around Tunneling FETs," *IEEE International Symposium on Quality Electronic Design (ISQED)*, March, 2013.

[14] **C. Pan** and A. Naeemi, "System-Level Performance Optimization and Benchmarking for On-Chip Graphene Interconnects," *IEEE Conference on Electrical Performance of Electronic Packaging and Systems (EPEPS)*, Oct, 2012.

[15] **C. Pan** and A. Naeemi, "System-Level Optimization and Benchmarking of Graphene pn Junction Logic System Based on Empirical CPI Model", *IEEE International Conference on IC Design and Technology (ICICDT)*, May, 2012.

[16] **C. Pan** and A. Naeemi, "Device- and System-Level Performance Modeling for Graphene P-N Junction Logic," *IEEE International Symposium on Quality Electronic Design (ISQED)*, March, 2012.

REFERENCES

[1] J. M. Rabaey, A. P. Chandrakasan, and B. Nikolic, *Digital integrated circuits* vol. 2: Prentice hall Englewood Cliffs, 2002.

[2] K. J. Kuhn, "Considerations for ultimate CMOS scaling," *IEEE Trans. Electron Devices,* vol. 59, pp. 1813-1828, 2012.

[3] K. J. Kuhn, "CMOS scaling for the 22nm node and beyond: Device physics and technology," in *VLSI Technology, Systems and Applications (VLSI-TSA), 2011 International Symposium on*, 2011, pp. 1-2.

[4] W. K. Henson, N. Yang, S. Kubicek, E. M. Vogel, J. J. Wortman, K. De Meyer*, et al.*, "Analysis of leakage currents and impact on off-state power consumption for CMOS technology in the 100-nm regime," *Electron Devices, IEEE Transactions on,* vol. 47, pp. 1393-1400, 2000.

[5] H. Li, C.-Y. Cher, K. Roy, and T. Vijaykumar, "Combined circuit and architectural level variable supply-voltage scaling for low power," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on,* vol. 13, pp. 564-576, 2005.

[6] D. Yakimets, T. H. Bao, M. G. Bardon, M. Dehan, N. Collaert, A. Mercha*, et al.*, "Lateral versus vertical gate-all-around FETs for beyond 7nm technologies," in *Device Research Conference (DRC), 2014 72nd Annual*, 2014, pp. 133-134.

[7] J. Park and C. Hu, "Air spacer MOSFET technology for 20nm node and beyond," in *Solid-State and Integrated-Circuit Technology, 2008. ICSICT 2008. 9th International Conference on*, 2008, pp. 53-56.

[8] C. Yin, P. C. Chan, and M. Chan, "An air spacer technology for improving short-channel immunity of MOSFETs with raised source/drain and high-κ gate dielectric," *Electron Device Letters, IEEE,* vol. 26, pp. 323-325, 2005.

[9] J. W. McPherson, "Reliability challenges for 45nm and beyond," in *Proceedings of the 43rd annual Design Automation Conference*, 2006, pp. 176-181.

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

[11] M. Maenhoudt, J. Versluijs, H. Struyf, J. Van Olmen, and M. Van Hove, "Double patterning scheme for sub-0.25 k1 single damascene structures at NA= 0.75, λ= 193nm," in *Microlithography 2005*, 2005, pp. 1508-1518.

[12] K. Jeong, A. B. Kahng, and R. O. Topaloglu, "Is overlay error more important than interconnect variations in double patterning?," in *Proceedings of the 11th international workshop on System level interconnect prediction*, 2009, pp. 3-10.

[13] L. Scheffer, "An overview of on-chip interconnect variation," in *Proceedings of the 2006 international workshop on System-level interconnect prediction*, 2006, pp. 27-28.

[14] J. A. Davis, R. Venkatesan, A. Kaloyeros, M. Beylansky, S. J. Souri, K. Banerjee, *et al.*, "Interconnect limits on gigascale integration (GSI) in the 21st century," *Proceedings of the IEEE,* vol. 89, pp. 305-324, 2001.

[15] N. Z. Haron and S. Hamdioui, "Why is CMOS scaling coming to an END?," in *Design and Test Workshop, 2008. IDT 2008. 3rd International*, 2008, pp. 98-103.

[16] R. D. Isaac, "The future of CMOS technology," *IBM Journal of Research and Development,* vol. 44, pp. 369-378, 2000.

[17] D. E. Nikonov and I. A. Young, "Overview of Beyond-CMOS Devices and a Uniform Methodology for Their Benchmarking," *Proceedings of the IEEE,* vol. 101, pp. 2498-2533, 2013.

[18] C. Pan and A. Naeemi, "Device- and system-level performance modeling for graphene P-N junction logic," in *Quality Electronic Design (ISQED), 2012 13th International Symposium on*, 2012, pp. 262-269.

[19] C. Pan and A. Naeemi, "System-level optimization and benchmarking of graphene PN junction logic system based on empirical CPI model," in *IC Design & Technology (ICICDT), 2012 IEEE International Conference on*, 2012, pp. 1-5.

[20] J. Appenzeller, "Carbon nanotubes for high-performance electronics—progress and prospect," *Proceedings of the IEEE,* vol. 96, pp. 201-211, 2008.

[21] H. Hashempour and F. Lombardi, "Device model for ballistic CNFETs using the first conducting band," *IEEE Design & Test,* vol. 25, pp. 178-186, 2008.

[22] S. K. Banerjee, L. F. Register, E. Tutuc, D. Reddy, and A. H. MacDonald, "Bilayer pseudospin field-effect transistor (BiSFET): a proposed new logic device," *Electron Device Letters, IEEE,* vol. 30, pp. 158-160, 2009.

[23] B. Dellabetta and M. Gilbert, "Performance characteristics of strongly correlated bilayer graphene for Post-CMOS logic devices," in *Silicon Nanoelectronics Workshop (SNW), 2010*, 2010, pp. 1-2.

[24] S. Sugahara and M. Tanaka, "A spin metal–oxide–semiconductor field-effect transistor using half-metallic-ferromagnet contacts for the source and drain," *Applied Physics Letters,* vol. 84, pp. 2307-2309, 2004.

[25] A. Khitun, M. Bao, J.-Y. Lee, K. Wang, D. Lee, S. Wang*, et al.*, "Inductively coupled circuits with spin wave bus for information processing," *Journal of Nanoelectronics and Optoelectronics,* vol. 3, pp. 24-34, 2008.

[26] A. Khitun and K. L. Wang, "Nano scale computational architectures with Spin Wave Bus," *Superlattices and Microstructures,* vol. 38, pp. 184-200, 2005.

[27] R. Cowburn and M. Welland, "Room temperature magnetic quantum cellular automata," *Science,* vol. 287, pp. 1466-1468, 2000.

[28] B. Behin-Aein, D. Datta, S. Salahuddin, and S. Datta, "Proposal for an all-spin logic device with built-in memory," *Nature nanotechnology,* vol. 5, pp. 266-270, 2010.

[29] B. Behin-Aein, A. Sarkar, S. Srinivasan, and S. Datta, "Switching energy-delay of all spin logic devices," *Applied Physics Letters,* vol. 98, p. 123510, 2011.

[30] F. Macià, A. D. Kent, and F. C. Hoppensteadt, "Spin-wave interference patterns created by spin-torque nano-oscillators for memory and computation," *Nanotechnology,* vol. 22, p. 095301, 2011.

[31] J. Noguchi, T. Oshima, T. Matsumoto, S. Uno, and K. Sato, "Multilevel Interconnect With Air-Gap Structure for Next-Generation Interconnections," *Electron Devices, IEEE Transactions on,* vol. 56, pp. 2675-2682, 2009.

[32] D. Kondo, H. Nakano, B. Zhou, K. Hayashi, M. Takahashi, S. Sato*, et al.*, "Sub-10-nm-wide intercalated multi-layer graphene interconnects with low resistivity," in *Interconnect Technology Conference/Advanced Metallization Conference (IITC/AMC), 2014 IEEE International*, 2014, pp. 189-192.

[33] C. Pan and A. Naeemi, "System-level performance optimization and benchmarking for on-chip graphene interconnects," in *Electrical Performance of Electronic Packaging and Systems (EPEPS), 2012 IEEE 21st Conference on*, 2012, pp. 33-36.

[34] A. Naeemi and J. D. Meindl, "Compact physical models for multiwall carbon-nanotube interconnects," *Electron Device Letters, IEEE,* vol. 27, pp. 338-340, 2006.

[35] A. Naeemi and J. D. Meindl, "Physical modeling of temperature coefficient of resistance for single-and multi-wall carbon nanotube interconnects," *Electron Device Letters, IEEE,* vol. 28, pp. 135-138, 2007.

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

[37] I. O'Connor and F. Gaffiot, "On-chip optical interconnect for low-power," in *Ultra Low-Power Electronics and Design*, ed: Springer, 2004, pp. 21-39.

[38] M. Haurylau, G. Chen, H. Chen, J. Zhang, N. A. Nelson, D. H. Albonesi*, et al.*, "On-chip optical interconnect roadmap: challenges and critical directions," *Selected Topics in Quantum Electronics, IEEE Journal of,* vol. 12, pp. 1699-1705, 2006.

[39] M. Stucchi, Z. Tokei, S. Demuynck, and Y.-K. Siew, "Impact of advanced patterning options, 193nm and EUV, on local interconnect performance," in *Interconnect Technology Conference (IITC), 2012 IEEE International*, 2012, pp. 1-3.

[40] S. Demuynck, C. Huffman, M. Claes, S. Suhard, J. Versluijs, H. Volders*, et al.*, "Integration and Dielectric Reliability of 30 nm Half Pitch Structures in Aurora® LK HM," *Japanese Journal of Applied Physics,* vol. 49, p. 04DB05, 2010.

[41] Y. Ma, J. Sweis, C. Bencher, H. Dai, Y. Chen, J. P. Cain, *et al.*, "Decomposition strategies for self-aligned double patterning," in *SPIE Advanced Lithography*, 2010.

[42] C. Shyng-Tsong, K. Tae-Soo, N. Seo-woo, N. Lafferty, K. Chiew-Seng, N. Saulnier, *et al.*, "48nm Pitch cu dual-damascene interconnects using self aligned double patterning scheme," in *Interconnect Technology Conference (IITC), 2013 IEEE International*, 2013, pp. 1-3.

[43] P. Xu, Y. Chen, Y. Chen, L. Miao, S. Sun, S.-W. Kim, *et al.*, "Sidewall spacer quadruple patterning for 15nm half-pitch," in *SPIE Advanced Lithography*, 2011, pp. 79731Q-79731Q-12.

[44] J. U. Knickerbocker, P. Andry, B. Dang, R. Horton, C. Patel, R. Polastre, *et al.*, "3D silicon integration," in *Electronic Components and Technology Conference, 2008. ECTC 2008. 58th*, 2008, pp. 538-543.

[45] J. Y.-C. Sun, "Semiconductor innovation into the next decade," in *Solid-State Circuits Conference (A-SSCC), 2014 IEEE Asian*, 2014, pp. 117-120.

[46] K. Swaminathan, E. Kultursay, V. Saripalli, V. Narayanan, M. Kandemir, and S. Datta, "Improving energy efficiency of multi-threaded applications using heterogeneous CMOS-TFET multicores," in *Proceedings of the 17th IEEE/ACM international symposium on Low-power electronics and design*, 2011, pp. 247-252.

[47] E. Humenay, D. Tarjan, and K. Skadron, "Impact of process variations on multicore performance symmetry," in *Proceedings of the conference on Design, automation and test in Europe*, 2007, pp. 1653-1658.

[48] L. Jungseob, P. P. Ajgaonkar, and K. Nam Sung, "Analyzing throughput of GPGPUs exploiting within-die core-to-core frequency variation," in *Performance Analysis of Systems and Software (ISPASS), 2011 IEEE International Symposium on*, 2011, pp. 237-246.

[49] C. Pan and A. Naeemi, "An Analytical Approach to System-level Variation Analysis and Optimization for Multi-core Processor," in *Quality Electronic Design (ISQED), 2014 15th International Symposium on*, 2014.

[50] D. E. Nikonov and I. A. Young, "Uniform methodology for benchmarking beyond-CMOS logic devices," in *Electron Devices Meeting (IEDM), 2012 IEEE International*, 2012, pp. 25.4. 1-25.4. 4.

[51] H. Kam, T.-J. King-Liu, E. Alon, and M. Horowitz, "Circuit-level requirements for MOSFET-replacement devices," in *Electron Devices Meeting, 2008. IEDM 2008. IEEE International*, 2008, pp. 1-1.

[52] V. Saripalli, S. Datta, V. Narayanan, and J. P. Kulkarni, "Variation-tolerant ultra low-power heterojunction tunnel FET SRAM design," in *Proceedings of the 2011 IEEE/ACM International Symposium on Nanoscale Architectures*, 2011, pp. 45-52.

[53] L. Wei, S. Oh, and H.-S. Wong, "Performance benchmarks for Si, III–V, TFET, and carbon nanotube FET-re-thinking the technology assessment methodology for complementary logic applications," in *Electron Devices Meeting (IEDM), 2010 IEEE International*, 2010, pp. 16.2. 1-16.2. 4.

[54] L. Wei, D. J. Frank, L. Chang, and H.-S. Wong, "A non-iterative compact model for carbon nanotube FETs incorporating source exhaustion effects," in *Electron Devices Meeting (IEDM), 2009 IEEE International*, 2009, pp. 1-4.

[55] P. Solomon, D. Frank, and S. Koswatta, "Compact model and performance estimation for tunneling nanowire FET," in *Device Research Conference (DRC), 2011 69th Annual*, 2011, pp. 197-198.

[56] S. Borkar, "Thousand core chips: a technology perspective," in *Proceedings of the 44th annual Design Automation Conference*, 2007, pp. 746-749.

[57] M. D. Hill and M. R. Marty, "Amdahl's law in the multicore era," *Computer,* vol. 41, pp. 33-38, 2008.

[58] D. H. Woo and H.-H. Lee, "Extending Amdahl's law for energy-efficient computing in the many-core era," *Computer,* vol. 41, pp. 24-31, 2008.

[59] G. M. Amdahl, "Validity of the single processor approach to achieving large scale computing capabilities," in *Proceedings of the April 18-20, 1967, spring joint computer conference*, 1967, pp. 483-485.

[60] F. Schwierz, "Graphene transistors," *Nature nanotechnology,* vol. 5, pp. 487-496, 2010.

[61] M. Katsnelson, K. Novoselov, and A. Geim, "Chiral tunnelling and the Klein paradox in graphene," *Nature Physics,* vol. 2, pp. 620-625, 2006.

[62] S. Tanachutiwat, J. U. Lee, W. Wang, and C. Y. Sung, "Reconfigurable multi-function logic based on graphene pn junctions," in *Design Automation Conference (DAC),* , 2010, pp. 883-888.

[63] R. N. Sajjad, S. Sutar, J. Lee, and A. W. Ghosh, "Manifestation of chiral tunneling at a tilted graphene pn junction," *Physical Review B,* vol. 86, p. 155412, 2012.

[64] R. N. Sajjad and A. W. Ghosh, "Manipulating chiral transmission by gate geometry: switching in graphene with transmission gaps," *ACS Nano, 2013, 7 (11), pp 9808-9813,* 2013.

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

[66] J. Appenzeller, J. Knoch, M. T. Bjork, H. Riel, H. Schmid, and W. Riess, "Toward nanowire electronics," *Electron Devices, IEEE Transactions on,* vol. 55, pp. 2827-2845, 2008.

[67] Predictive Technology Model (PTM), available online at http://ptm.asu.edu, 2012.

[68] D. Yakimets, G. Eneman, P. Schuddinck, T. Huynh Bao, M. G. Bardon, P. Raghavan*, et al.*, "Vertical GAAFETs for the Ultimate CMOS Scaling," *Electron Devices, IEEE Transactions on,* vol. 62, pp. 1433-1439, 2015.

[69] T. H. Bao, D. Yakimets, J. Ryckaert, I. Ciofi, R. Baert, A. Veloso*, et al.*, "Circuit and process co-design with vertical gate-all-around nanowire FET technology to extend CMOS scaling for 5nm and beyond technologies," in *Solid State Device Research Conference (ESSDERC), 2014 44th European*, 2014, pp. 102-105.

[70] T. Bryllert, L. E. Wernersson, L. E. Froberg, and L. Samuelson, "Vertical high-mobility wrap-gated InAs nanowire transistor," *Electron Device Letters, IEEE,* vol. 27, pp. 323-325, 2006.

[71] H. Liu, D. Mohata, A. Nidhi, V. Saripalli, V. Narayanan, and S. Datta, "Exploration of vertical MOSFET and tunnel FET device architecture for Sub 10nm node applications," in *Device Research Conference (DRC), 2012 70th Annual*, 2012, pp. 233-234.

[72] C. Thelander, C. Rehnstedt, L. E. Froberg, E. Lind, T. Martensson, P. Caroff*, et al.*, "Development of a vertical wrap-gated InAs FET," *Electron Devices, IEEE Transactions on,* vol. 55, pp. 3030-3036, 2008.

[73] V. V. Cheianov and V. I. Fal'ko, "Selective transmission of Dirac electrons and ballistic magnetoresistance of np junctions in graphene," *Physical review b,* vol. 74, p. 041403, 2006.

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

[75] A. Das, B. Chakraborty, S. Piscanec, S. Pisana, A. Sood, and A. Ferrari, "Phonon renormalization in doped bilayer graphene," *Physical Review B,* vol. 79, p. 155417, 2009.

[76] T. Low and J. Appenzeller, "Electronic transport properties of a tilted graphene pn junction," *Physical Review B,* vol. 80, p. 155406, 2009.

[77] S. Datta, *Electronic transport in mesoscopic systems*: Cambridge university press, 1997.

[78] F. Xia, V. Perebeinos, Y.-m. Lin, Y. Wu, and P. Avouris, "The origins and limits of metal-graphene junction resistance," *Nature nanotechnology,* vol. 6, pp. 179-184, 2011.

[79] S. Sutar, E. Comfort, J. Liu, T. Taniguchi, K. Watanabe, and J. Lee, "Angle-Dependent Carrier Transmission in Graphene p–n Junctions," *Nano letters,* vol. 12, pp. 4460-4464, 2012.

[80] C. Pan, A. Ceyhan, and A. Naeemi, "System-level optimization and benchmarking for InAs nanowire based gate-all-around tunneling FETs," in *Quality Electronic Design (ISQED), 2013 14th International Symposium on*, 2013, pp. 196-202.

[81] S. M. Sze and K. K. Ng, *Physics of semiconductor devices*: John Wiley & Sons, 2006.

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

[83] N. 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," *Electron Devices, IEEE Transactions on,* vol. 56, pp. 2193-2201, 2009.

[84] U. E. Avci, R. Rios, K. J. Kuhn, and I. A. Young, "Comparison of power and performance for the TFET and MOSFET and considerations for P-TFET," in *Nanotechnology (IEEE-NANO), 2011 11th IEEE Conference on*, 2011, pp. 869-872.

[85] S. Xiong, T.-J. King, and J. Bokor, "Study of the extrinsic parasitics in nano-scale transistors," *Semiconductor science and technology,* vol. 20, p. 652, 2005.

[86] BSIM-CMG 107.0.0 Technical Manual. [Online]. Available: http://www-device.eecs.berkeley.edu/bsim/?page=BSIMCMG, 2013.

[87] Sentaurus Device Monte Carlo, Sentaurus Band Structure User Guide, I-2013.12, Synopsys, Inc., Mountain View, CA, USA, 2013.

[88] J. Ryckaert, P. Raghavan, P. Schuddinck, H. B. Trong, A. Mallik, S. S. Sakhare*, et al.*, "DTCO at N7 and beyond: patterning and electrical compromises and opportunities," 2015, pp. 94270C-94270C-8.

[89] W. Guo, M. Choi, A. Rouhi, V. Moroz, G. Eneman, J. Mitard*, et al.*, "Impact of 3D integration on 7nm high mobility channel devices operating in the ballistic regime," in *Electron Devices Meeting (IEDM), 2014 IEEE International*, 2014, pp. 7.1.1-7.1.4.

[90] D. Yakimets, D. Jang, P. Raghavan, G. Eneman, H. Mertens, P. Schuddinck*, et al.*, "Lateral NWFET Optimization for Beyond 7nm Nodes," *to be published in IC Design & Technology (ICICDT), 2015 IEEE International Conference on,* June, 2015.

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

[92] C. Pan, R. Baert, I. Ciofi, Z. Tokei, and A. Naeemi, "System-Level Variation Analysis for Interconnection Networks at Sub-10-nm Technology Nodes Using

Multiple Patterning Techniques," *Electron Devices, IEEE Transactions on,* vol. 62, pp. 2071-2077, 2015.

[93] J.-H. Chern, J. Huang, L. Arledge, P.-C. Li, and P. Yang, "Multilevel metal capacitance models for CAD design synthesis systems," *Electron Device Letters, IEEE,* vol. 13, pp. 32-34, 1992.

[94] RAPHAEL: Interconnect Analysis Program, TMA Inc, 1996.

[95] V. Kumar, S. Rakheja, and A. Naeemi, "Performance and Energy-per-Bit Modeling of Multilayer Graphene Nanoribbon Conductors," 2012.

[96] C. Pan, P. Raghavan, F. Catthoor, Z. Tokei, and A. Naeemi, "Technology/circuit co-optimization and benchmarking for graphene interconnects at Sub-10nm technology node," in *Quality Electronic Design (ISQED), 2015 16th International Symposium on*, 2015, pp. 599-603.

[97] C. Pan, P. Raghavan, A. Ceyhan, F. Catthoor, Z. Tokei, and A. Naeemi, "Technology/Circuit/System Co-Optimization and Benchmarking for Multilayer Graphene Interconnects at Sub-10-nm Technology Node," *Electron Devices, IEEE Transactions on,* vol. PP, pp. 1-1, 2015.

[98] P. R. J. Ryckaert, R. Baert, M. G. Bardon, M. Dusa*, *et al.*, "Design Technology Co-optimization for N10," *Proceedings of the Custom Integrated Circuits Conference (CICC),* 2014.

[99] S. Knowles, "A family of adders," in *Computer Arithmetic, 1999. Proceedings. 14th IEEE Symposium on*, 1999, pp. 30-34.

[100] K. I. Bolotin, K. Sikes, Z. Jiang, M. Klima, G. Fudenberg, J. Hone*, et al.*, "Ultrahigh electron mobility in suspended graphene," *Solid State Communications,* vol. 146, pp. 351-355, 2008.

[101] C. Dean, A. Young, I. Meric, C. Lee, L. Wang, S. Sorgenfrei*, et al.*, "Boron nitride substrates for high-quality graphene electronics," *Nature nanotechnology,* vol. 5, pp. 722-726, 2010.

[102] K. M. S. Sakhare, J. Ryckaert, P. Raghavan, *et al.*, "Simplistic simulation based device VT targeting technique to determine technology high density LELE gate

patterned FinFET SRAM in sub-10nm era," *IEEE Transactions on Electron Devices, 2014.*

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

[104] H. Liu, Y. Liu, and D. Zhu, "Chemical doping of graphene," *Journal of Materials Chemistry,* vol. 21, pp. 3335-3345, 2011.

[105] C. Pan and A. Naeemi, "A Proposal for a Novel Hybrid Interconnect Technology for the End of Roadmap," *Electron Device Letters, IEEE,* vol. 35, pp. 250-252, 2014.

[106] Virtuoso Liberate Reference Manual Version 13.1, Cadence, Inc., San Jose, CA, USA, 2014.

[107] T. Austin, E. Larson, and D. Ernst, "SimpleScalar: An infrastructure for computer system modeling," *Computer,* vol. 35, pp. 59-67, 2002.

[108] S. Li, J. H. Ahn, R. D. Strong, J. B. Brockman, D. M. Tullsen, and N. P. Jouppi, "McPAT: an integrated power, area, and timing modeling framework for multicore and manycore architectures," in *Microarchitecture, 2009. MICRO-42. 42nd Annual IEEE/ACM International Symposium on*, 2009, pp. 469-480.

[109] J. Eble, "A generic system simulator with novel on-chip cache and throughput models for GSI," *Ph.D. Dissertation, Dept. Elect. Eng., Georgia Institute of Technology, Atlanta, GA,* 1998.

[110] J. A. Davis, V. K. De, and J. D. Meindl, "A stochastic wire-length distribution for gigascale integration (GSI). I. Derivation and validation," *Electron Devices, IEEE Transactions on,* vol. 45, pp. 580-589, 1998.

[111] T. Ghani, K. Mistry, P. Packan, S. Thompson, M. Stettler, S. Tyagi*, et al.*, "Scaling challenges and device design requirements for high performance sub-50 nm gate length planar CMOS transistors," in *VLSI Technology, 2000. Digest of Technical Papers. 2000 Symposium on*, 2000, pp. 174-175.

[112] S. Borkar, "Design perspectives on 22nm CMOS and beyond," in *Design Automation Conference (DAC), 2009 46th ACM/IEEE*, 2009, pp. 93-94.

[113] S. K. Springer, S. Lee, N. Lu, E. J. Nowak, J.-O. Plouchart, J. S. Watts, *et al.*, "Modeling of variation in submicrometer CMOS ULSI technologies," *Electron Devices, IEEE Transactions on,* vol. 53, pp. 2168-2178, 2006.

[114] H. Gang, "Compact Physical Models for Power Supply Noise and Chip/Package Co-design in Gigascale Integration (GSI) and Three-Dimensional (3-D) Integration Systems," *Ph.D. Dissertation, Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA,* 2008.

[115]D. Blaauw, K. Chopra, A. Srivastava, and L. Scheffer, "Statistical Timing Analysis: From Basic Principles to State of the Art," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on,* vol. 27, pp. 589-607, 2008.

[116] K. Roy, S. Mukhopadhyay, and H. Mahmoodi-Meimand, "Leakage current mechanisms and leakage reduction techniques in deep-submicrometer CMOS circuits," *Proceedings of the IEEE,* vol. 91, pp. 305-327, 2003.

[117] A. Srivastava, D. Sylvester, and D. Blaauw, "Statistical optimization of leakage power considering process variations using dual-Vth and sizing," in *Proceedings of the 41st annual Design Automation Conference*, 2004, pp. 773-778.

[118] K. A. Bowman, A. R. Alameldeen, S. T. Srinivasan, and C. B. Wilkerson, "Impact of Die-to-Die and Within-Die Parameter Variations on the Clock Frequency and Throughput of Multi-Core Processors," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on,* vol. 17, pp. 1679-1690, 2009.

[119] N. M. S. Thoziyoor, J. H. Ahn and N. P. Jouppi, "CACTI 5.1," *Technical Report HPL-2008-20, HP Laboratories,* 2008.

[120] S. Wasson, "Intel's 'Sandy Bridge' Core Processors", Tech Report, available online at http://techreport.com/review/20188/intel-sandy-bridge-core-processors, 2011.

[121] Intel Corp., vendor data, available online at http://ark.intel.com, 2012.

[122] G. Varghese, J. Sanjeev, T. Chao, K. Smits, D. Satish, S. Siers, *et al.*, "Penryn: 45-nm next generation Intel Core 2 processor," in *Solid-State Circuits Conference, 2007. ASSCC '07. IEEE Asian*, 2007, pp. 14-17.

[123] A. L. Shimpi, "Intel's Sandy Bridge Architecture Exposed," *Anandtech, Sept,* 2010.

[124] N. A. Kurd, S. Bhamidipati, C. Mozak, J. L. Miller, T. M. Wilson, M. Nemani, *et al.*, "Westmere: A family of 32nm IA processors," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2010 IEEE International*, 2010, pp. 96-97.

[125] CPU World, available online at http://www.cpu-world.com/forum/viewtopic.php?t=19888, 2012.

[126] Intel Corp., "Intel Tick Tock Model", available online at http://www.intel.com/content/www/us/en/silicon-innovations/intel-tick-tock-model-general.html, 2012.

[127] J. Lee and N. S. Kim, "Optimizing throughput of power-and thermal-constrained multicore processors using DVFS and per-core power-gating," in *Design Automation Conference (DAC), 2009 46th ACM/IEEE*, 2009, pp. 47-50.

[128] D. C. Sekar, A. Naeemi, R. Sarvari, J. A. Davis, and J. D. Meindl, "Intsim: a CAD tool for optimization of multilevel interconnect networks," in *Computer-Aided Design (ICCAD), IEEE/ACM International Conference on*, 2007, pp. 560-567.

[129] P. Batude, M. Vinet, A. Pouydebasque, C. Le Royer, B. Previtali, C. Tabone, *et al.*, "3D monolithic integration," in *Circuits and Systems (ISCAS), 2011 IEEE International Symposium on*, 2011, pp. 2233-2236.

[130] A. Rahman and R. Reif, "System-level performance evaluation of three-dimensional integrated circuits," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on,* vol. 8, pp. 671-678, 2000.

[131] C. Pan, S. Mukhopadhyay, and A. Naeemi, "System-level chip/package co-design for multi-core processors implemented with power-gating technique," in *Electrical Performance of Electronic Packaging and Systems (EPEPS), 2014 IEEE 23rd Conference on*, 2014, pp. 11-14.

[132] International Technology Roadmap for Semiconductors (ITRS), available online at http://www.itrs.net/, 2012.

[133] H. Wei, S. Ghosh, S. Velusamy, K. Sankaranarayanan, K. Skadron, and M. R. Stan, "HotSpot: a compact thermal modeling methodology for early-stage VLSI design," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on,* vol. 14, pp. 501-513, 2006.

[134] C. Pan and A. Naeemi, "A Fast System-Level Design Methodology for Heterogeneous Multi-Core Processors Using Emerging Technologies," *Emerging and Selected Topics in Circuits and Systems, IEEE Journal on,* vol. 5, pp. 75-87, 2015.

[135] Standard Performance Evaluation Corporation (SPEC), available online at http://www.spec.org, 2012.

[136] R. Singhal, "Inside Intel next generation Nehalem microarchitecture," in *Hot Chips*, 2008.

[137] S. Kottapalli and J. Baxter, "Nehalem-ex cpu architecture," in *Hot chips*, 2009.

[138] K. A. Bowman, S. G. Duvall, and J. D. Meindl, "Impact of die-to-die and within-die parameter fluctuations on the maximum clock frequency distribution for gigascale integration," *Solid-State Circuits, IEEE Journal of,* vol. 37, pp. 183-190, 2002.

[139] J. W. Tschanz, J. T. Kao, S. G. Narendra, R. Nair, D. A. Antoniadis, A. P. Chandrakasan*, et al.*, "Adaptive body bias for reducing impacts of die-to-die and within-die parameter variations on microprocessor frequency and leakage," *Solid-State Circuits, IEEE Journal of,* vol. 37, pp. 1396-1402, 2002.

[140] S. Sinha, G. Yeric, V. Chandra, B. Cline, and C. Yu, "Exploring sub-20nm FinFET design with Predictive Technology Models," in *Design Automation Conference (DAC), 49th ACM/EDAC/IEEE*, 2012, pp. 283-288.

[141] S. R. Sarangi, B. Greskamp, R. Teodorescu, J. Nakano, A. Tiwari, and J. Torrellas, "VARIUS: A model of process variation and resulting timing errors for microarchitects," *Semiconductor Manufacturing, IEEE Transactions on,* vol. 21, pp. 3-13, 2008.

[142] C. E. Green, A. G. Fedorov, and Y. K. Joshi, "Dynamic thermal management of high heat flux devices using embedded solid-liquid phase change materials and solid state coolers," in *Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm), 2012 13th IEEE Intersociety Conference on*, 2012, pp. 853-862.

[143] V. Sahu, Y. K. Joshi, and A. G. Fedorov, "Experimental investigation of hotspot removal using superlattice cooler," in *Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm), 2010 12th IEEE Intersociety Conference on*, 2010, pp. 1-8.

[144] R. H. Byrd, J. C. Gilbert, and J. Nocedal, "A trust region method based on interior point techniques for nonlinear programming," *Mathematical Programming,* vol. 89, pp. 149-185, 2000.

[145] N. Magen, A. Kolodny, U. Weiser, and N. Shamir, "Interconnect-power dissipation in a microprocessor," in *Proceedings of the 2004 international workshop on System level interconnect prediction*, 2004, pp. 7-13.

[146] A. Naeemi, A. Ceyhan, V. Kumar, C. Pan, R. M. Iraei, and S. Rakheja, "BEOL scaling limits and next generation technology prospects," in *Proceedings of the 51st Annual Design Automation Conference*, 2014, pp. 1-6.

[147] C. Pan and A. Naeemi, "System-level variation analysis for interconnection networks," in *Interconnect Technology Conference/Advanced Metallization Conference (IITC/AMC), 2014 IEEE International*, 2014, pp. 303-306.

[148] M. van Veenhuizen, G. Allen, M. Harmes, T. Indukuri, C. Jezewski, B. Krist*, et al.*, "Demonstration of an electrically functional 34 nm metal pitch interconnect in ultralow-k ILD using spacer-based pitch quartering," in *Interconnect Technology Conference (IITC), 2012 IEEE International*, 2012, pp. 1-3.

[149] V. Mehrotra, S. L. Sam, D. Boning, A. Chandrakasan, R. Vallishayee, and S. Nassif, "A methodology for modeling the effects of systematic within-die interconnect and device variation on circuit performance," in *Proceedings of the 37th Annual Design Automation Conference*, 2000, pp. 172-175.

[150] N. Inoue, "Challenges in low-k integration of advanced Cu BEOL beyond 14 nm node," in *Electron Devices Meeting (IEDM), 2013 IEEE International*, 2013, pp. 29.1.1-29.1.4.

[151] C. Pan and A. Naeemi, "A Paradigm Shift in Local Interconnect Technology Design in the Era of Nanoscale Multigate and Gate-All-Around Devices," *Electron Device Letters, IEEE,* vol. 36, pp. 274-276, 2015.

[152] R. H. Havemann and J. A. Hutchby, "High-performance interconnects: An integration overview," *Proceedings of the IEEE,* vol. 89, pp. 586-601, 2001.

[153] P. Kapur, J. P. McVittie, and K. C. Saraswat, "Technology and reliability constrained future copper interconnects. I. Resistance modeling," *Electron Devices, IEEE Transactions on,* vol. 49, pp. 590-597, 2002.

[154] J. Lienig, "Interconnect and current density stress: an introduction to electromigration-aware design," in *Proceedings of the 2005 international workshop on System level interconnect prediction*, 2005, pp. 81-88.

[155] M. H. Lin and A. S. Oates, "AC and Pulsed-DC Stress Electromigration Failure Mechanisms in Cu Interconnects," in *Interconnect Technology Conference (IITC), 2013 IEEE International on*, 2013, pp. 28-30.

[156] B. Liew, N. Cheung, and C. Hu, "Electromigration interconnect lifetime under AC and pulse DC stress," in *Reliability Physics Symposium, 1989. 27th Annual Proceedings., International*, 1989, pp. 215-219.

[157] K. Schuegraf, M. C. Abraham, A. Brand, M. Naik, and R. Thakur, "Semiconductor Logic Technology Innovation to Achieve Sub-10 nm Manufacturing," *Electron Devices Society, IEEE Journal of the,* vol. 1, pp. 66-75, 2013.

[158] S. Maîtrejean, R. Gers, T. Mourier, A. Toffoli, and G. Passemard, "Experimental measurements of electron scattering parameters in Cu narrow lines," *Microelectronic engineering,* vol. 83, pp. 2396-2401, 2006.

[159] M. Powell, S.-H. Yang, B. Falsafi, K. Roy, and T. Vijaykumar, "Gated-V dd: a circuit technique to reduce leakage in deep-submicron cache memories," in *Proceedings of the 2000 international symposium on Low power electronics and design*, 2000, pp. 90-95.

[160] H. Jiang, M. Marek-Sadowska, and S. R. Nassif, "Benefits and costs of power-gating technique," in *Computer Design: VLSI in Computers and Processors, 2005. ICCD 2005. Proceedings. 2005 IEEE International Conference on*, 2005, pp. 559-566.

[161] W. Cui, P. Parmar, J. M. Morgan, and U. Sheth, "Modeling the network processor and package for power delivery analysis," in *Electromagnetic Compatibility, 2005. EMC 2005. 2005 International Symposium on*, 2005, pp. 690-694.

[162] Z. Hu, A. Buyuktosunoglu, V. Srinivasan, V. Zyuban, H. Jacobson, and P. Bose, "Microarchitectural techniques for power gating of execution units," in *Proceedings of the 2004 international symposium on Low power electronics and design*, 2004, pp. 32-37.

[163] P. A. Daly and J. Morrison, "Understanding the potential benefits of distributed generation on power delivery systems," in *Rural Electric Power Conference, 2001*, 2001, pp. A2/1-A213.

[164] C. Pan and A. Naeemi, "System-level analysis for 3D interconnection networks," in *Interconnect Technology Conference (IITC), 2013 IEEE International*, 2013, pp. 1-3.

[165] Y.-J. Lee, P. Morrow, and S. K. Lim, "Ultra high density logic designs using transistor-level monolithic 3D integration," in *Computer-Aided Design (ICCAD), 2012 IEEE/ACM International Conference on*, 2012, pp. 539-546.

[166] J. Srinivasan, S. V. Adve, P. Bose, and J. A. Rivers, "Lifetime reliability: Toward an architectural solution," *Micro, IEEE,* vol. 25, pp. 70-80, 2005.

[167] W. Wang, S. Yang, S. Bhardwaj, S. Vrudhula, F. Liu, and Y. Cao, "The impact of NBTI effect on combinational circuit: modeling, simulation, and analysis," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on,* vol. 18, pp. 173-183, 2010.

[168] M. White, *Microelectronics reliability: physics-of-failure based modeling and lifetime evaluation*: Pasadena, CA: Jet Propulsion Laboratory, National Aeronautics and Space Administration, 2008.

VITA

Chenyun Pan was born in Shanghai, China in 1988. He received a Bachelor of Science in Microelectronics from Shanghai Jiaotong University in 2010. In Spring 2011, he had the privilege to join Professor Azad Naeemi's Nanoelectronic Research Lab in the School of Electrical and Computer Engineering, the Georgia Institute of Technology. He received a Master of Science in Spring 2013. In Summer 2014 and Spring 2015, he worked in IMEC, Belgium, as a visiting scholar, and conducted research on circuit- and system-level benchmarking and optimization for multilayer graphene interconnects and vertical and lateral gate-all-around FETs. During his Ph.D. years, he published sixteen international conference and peer-reviewed journal papers, and won two best paper awards in the IEEE International Symposium on Quality Electronic Design (ISQED 2012) and in the IEEE International Conference on IC Design and Technology (ICICDT 2012), respectively. His research focuses on the device and interconnect modeling, and the system-level benchmarking and performance optimization for various emerging beyond-CMOS technologies.