Temperature Adaptive and Variation Tolerant CMOS Circuits

by

Ranjith Kumar

A dissertation submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy

Supervised by
Professor Volkan Kursun

Department of Electrical and Computer Engineering University of Wisconsin-Madison, Madison, Wisconsin

May 2008
Dedication

To Thatha and Mami... for providing the aspiration to do my higher studies
and to Appa, Amma, and Archana... for providing the needed love, affection, and support
Curriculum Vitae

The author was born in Chennai, India on July 30th 1981. He attended P. S. G. College of Technology, Coimbatore, India from 1998 to 2002 and graduated with a Bachelor of Engineering degree in Electronics and Communication Engineering in 2002. He received a Master of Science degree in Electrical and Computer Engineering from the University of Wisconsin, Madison, Wisconsin in December 2004. Since 2005 he has been working toward a Ph.D. degree in Electrical and Computer Engineering at the University of Wisconsin.

He worked as a software engineer at Cognizant Technology Solutions, Chennai, India from May 2002 to July 2003. During the summer of 2006, he was with Qualcomm, San Diego, California, providing a comprehensive analysis of the supply voltages required to implement static voltage scaling in mobile station modems. He evaluated the maximum substrate/well tap distance in a row-based standard cell layout during his internship with AMD, Austin, Texas, in the summer of 2007. His current research interests include process parameter, supply voltage, and temperature variation tolerant low-power CMOS circuit design.
Acknowledgements

Foremost, I extend my sincere gratitude to my academic advisor Professor Volkan Kursun for his enthusiasm and encouragement during all my years as a doctoral student. His professionalism and commitment makes him a true mentor. I admire his attention to the details while his brilliant research ideas earned my deepest respect. I will always remember his constant motivation to become a successful researcher.

I would like to express my appreciation to the University of Wisconsin and the Department of Electrical and Computer Engineering for their unique environment that supports high quality research. I am grateful Professors Leon Shohet, Michael Schulte, Robert Blick, and Willis J. Tompkins for serving in my proposal and defense committees. I would also like to express my gratitude to several ECE faculty members, especially, Professors Michael Morrow, Paul Milenkovic and Ian A. Hiskens for their valuable support in my teaching assignments. I am also grateful to the administrative staffs in the Department of Electrical and Computer Engineering for their assistance during my tenure as a teaching assistant.

I appreciate the support from all my colleagues and friends. Specifically, Arul Sundaramoorthy, Devesh Ranjan, Kamal Srinivasan, Magesh Thiyagarajan, Naval Shanware, Niveditha Sundaram, Sankara Hari, Shyam Bharat, Sherif Tawfik, Sivakumar Shanmugam, Sunil Sunkara, and Zhiyu Liu for their constant motivation throughout my Ph.D. work. I thank my mentor Matt Severson of Qualcomm for his support during the internship in summer 2006. I also thank Michael Kent and Aaron Rogers of AMD for their encouragement during my internships in summer 2007.

All this was made possible by the love and encouragement from my family. I am deeply indebted to my grand parents and parents for teaching me the value in accumulating the wealth of knowledge. I am also fortunate to have the collective guidance of my immediate relatives, Mohan, Sridhar, Kanchana, Meena, and Senthil. Each one of them along with their family had an irreplaceable influence in my life, for which I am grateful. I want to acknowledge my parents-in-law, Dr. Ezhilarasu and Shanthi, for their moral support. Thanks to my brothers, sisters, sister-in-law, brother-in-law, and all well-wishers for their constant encouragement throughout my career. Finally,
I am exceedingly thankful for the immense love and friendship from my soul mate, Archana.
Abstract

The imbalanced utilization and the diversity of circuitry cause on-chip temperature gradients. Different sections of high-performance integrated circuits typically operate at different temperatures. Furthermore, environmental temperature fluctuations can cause significant variations in the die temperature. Temperature fluctuations alter the speed characteristics of CMOS circuits. Several techniques to reduce the sensitivity of circuit speed to the fluctuations of die temperature are proposed in this dissertation. The techniques simultaneously target temperature variation resilience and enhanced energy efficiency in CMOS circuits. A generic power measurement methodology to accurately evaluate the energy savings provided with the proposed techniques is also presented.

The drain current produced by a MOSFET operating at the prescribed nominal supply voltage is reduced at elevated temperatures primarily due to the degradation of carrier mobility. New voltage optimization techniques are described in this dissertation to provide temperature variation insensitive constant transistor current with the standard CMOS technologies. The optimum supply and threshold voltages that achieve temperature variation insensitive constant circuit speed are identified for different technologies. The energy savings provided with the two proposed temperature variation tolerant voltage optimization techniques are compared.

Low power design methodologies are typically aimed to reduce the energy consumption or the energy-delay product in CMOS integrated circuits. The propagation delays of circuits optimized for minimum energy consumption and minimum energy-delay product are highly sensitive to temperature fluctuations. Alternatively, the voltage optimization techniques explored in this dissertation simultaneously achieve enhanced energy efficiency and temperature variation tolerance. The speed and energy tradeoffs in circuits operating at the supply voltages that provide temperature variation insensitive performance, minimum energy consumption, and minimum energy-delay product are presented.

Integrated circuits with ultra-low-voltage power supplies exhibit reversed temperature dependence. The speed of logic circuits and memory arrays optimized for
minimum energy consumption is enhanced at elevated temperatures. New temperature-adaptive dynamic supply and threshold voltage tuning techniques are proposed in this dissertation to enhance the high temperature energy efficiency of ultra-low-voltage CMOS circuits. The objective is achieved by either dynamically scaling the power supply voltage or dynamically increasing the device threshold voltages through reverse body-bias at elevated temperatures. The energy savings provided with the two temperature-adaptive dynamic voltage tuning techniques are presented.
# Contents

Dedication ................................................................. ii
Curriculum Vitae .......................................................... iii
Acknowledgements ......................................................... iv
Abstract ........................................................................... vi
List of Tables ..................................................................... xii
List of Figures .................................................................... xv

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1</td>
<td>Scaling Trends of Integrated Circuits</td>
<td>7</td>
</tr>
<tr>
<td>1.2</td>
<td>Outline of the Dissertation</td>
<td>16</td>
</tr>
<tr>
<td>2</td>
<td>Sources of Power Consumption in Digital Circuits</td>
<td>20</td>
</tr>
<tr>
<td>2.1</td>
<td>Dynamic Switching Power</td>
<td>20</td>
</tr>
<tr>
<td>2.1.1</td>
<td>Power Consumption Due to Glitches</td>
<td>24</td>
</tr>
<tr>
<td>2.2</td>
<td>Leakage Currents in Nano-CMOS Technologies</td>
<td>26</td>
</tr>
<tr>
<td>2.2.1</td>
<td>Subthreshold Leakage Current</td>
<td>26</td>
</tr>
<tr>
<td>2.2.1.1</td>
<td>Short Channel Effects</td>
<td>27</td>
</tr>
<tr>
<td>2.2.1.2</td>
<td>Drain Induced Barrier Lowering</td>
<td>28</td>
</tr>
<tr>
<td>2.2.1.3</td>
<td>Characterization of the Subthreshold Leakage Current</td>
<td>30</td>
</tr>
<tr>
<td>2.2.2</td>
<td>Junction Leakage Current</td>
<td>32</td>
</tr>
<tr>
<td>2.2.3</td>
<td>Gate Tunneling Leakage Current</td>
<td>33</td>
</tr>
<tr>
<td>2.3</td>
<td>Short Circuit Power</td>
<td>36</td>
</tr>
<tr>
<td>2.4</td>
<td>Static DC Power</td>
<td>38</td>
</tr>
</tbody>
</table>
Die Temperature Variations

3.1. Impact of Temperature Variations on Carrier Concentration

3.2. Device Parameter Variations with Temperature Fluctuations

3.3. Modeling of Parameter Variations

3.4. Temperature Effects in 180nm and 65nm CMOS Technologies

3.5. Chapter Summary

Power Measurement Techniques with CAD Tools

4.1. Generic Power Measurement Methodology

4.2. Test Circuits and Experimental Set-up

4.3. Comparison of the Power Measurements

4.3.1. Active Mode Power Consumption

4.3.2. Stand-by Mode Power Consumption

4.3.3. Power Consumption of Body-Biased Circuits

4.4. Sources of Error with the Built-in Power Estimation Commands

4.5. Chapter Summary

Temperature Variation Insensitive CMOS Circuits

5.1. Supply Voltage Optimization for Temperature Variation Insensitive CMOS Circuits

5.2. Threshold Voltage Optimization Technique

5.3. Speed and Energy Efficiency of the Voltage Optimization Techniques

5.4. Chapter Summary
6 Temperature-Aware Low Power Design

6.1 Low Power Design with Supply Voltage Scaling

6.2 Energy Efficient Temperature Variation Resilient CMOS Circuits

6.3 Chapter Summary

7 High Temperature Energy Reduction in Low-Voltage Circuits

7.1 Supply Voltage Optimization for Minimum Energy Consumption

7.2 Techniques for High Temperature Energy Reduction

7.2.1 Temperature-Adaptive Dynamic Supply Voltage Scaling

7.2.2 Temperature-Adaptive Body Bias

7.3 Effectiveness of the Temperature-Adaptive Voltage Tuning Schemes

7.3.1 Characteristics of the Temperature-Adaptive Schemes

7.3.2 Impact of Process-Parameter and Supply Voltage Variations

7.4 Chapter Summary

8 Temperature-Adaptive Low-Voltage Memory Banks

8.1 Sizing of a Subthreshold SRAM Bit Cell

8.2 Supply Voltage for Minimum Energy Consumption

8.3 High Temperature Energy Reduction in Memory Banks

8.3.1 Temperature-Adaptive Supply Voltage Scaling

8.3.2 Temperature-Adaptive Body Bias

8.4 Effectiveness of the TA-DVS Technique

8.4.1 Influence of the TA-DVS Technique on the Noise Margins

8.4.2 Impact of Process-Parameter and Supply Voltage Variations

8.5 Chapter Summary
Conclusions…………………………………………………………………………………….. 163

Future Research…………………………………………………………………………… 172

10.1. Temperature-Adaptive Voltage Scaling Power Supplies………………… 173
10.2. Thermal Variation Aware Interconnect Design in Nano-CMOS
Technologies………………………………………………………………………………… 176
10.3. Temperature-Adaptive Dynamic Memory…………………………………… 181
10.4. Thermal Aware FinFET Device Optimizations…………………………… 183
10.5. 3-D Stacked Integrated Circuits………………………………………………. 186

Bibliography…………………………………………………………………………………. 191

Appendix A Publications…………………………………………………………………… 207
# List of Tables

1.1 Scaling trends of high performance microprocessors from Intel [81] ............ 10

1.2 The ITRS projection of the die area of microprocessors and memory chips in future technologies .......................................................... 14

2.1 Semiconductor device scaling trends .................................................. 34

3.1 Model coefficients that effect the MOSFET drain current when the temperature fluctuates .......................................................................................................................... 55

4.1 Comparison of the power measurements for an NMOS device with zero
body-bias at various temperatures .......................................................... 74

4.2 Comparison of the power measurements for a PMOS device with zero
body-bias at various temperatures .......................................................... 75

4.3 Comparison of the power measurements for a PMOS device at different body-bias voltages and temperature .......................................................... 76

4.4 Comparison of the power measurements for an NMOS device at different body-bias voltages and temperature .......................................................... 77

4.5 Comparison of the power measurements for 32-bit MUX at different body-bias voltages and temperature .......................................................... 78

4.6 Comparison of the power measurements for NAND4 at different body-bias voltages and temperature .......................................................... 78

4.7 Currents and the power consumption of the NMOS device in Fig. 4.9a 
measured with HSPICE for different body-bias voltages ................................ 80

4.8 Currents and the power consumption of the NMOS device in Fig. 4.9b 
measured with HSPICE for different body-bias voltages ................................ 80

4.9 Currents and the power consumption of the NMOS device in Fig. 4.9a 
measured with CADENCE-SPECTRE for different body-bias voltages ...... 82
5.1 Gate-overdrive variations at different supply voltages for devices in a 65nm CMOS technology

5.2 Normalized energy at the supply voltages providing temperature variation insensitive speed in 180nm and 65nm CMOS technologies

5.3 Normalized energy with the threshold voltages providing temperature variation insensitive speed in 180nm and 65nm CMOS technologies

6.1 Gate-overdrive and carrier mobility variations at different supply voltages for devices in a 180nm CMOS technology

6.2 Delay and energy at the nominal supply voltage in 180nm and 65nm CMOS technologies

6.3 Normalized Delay and energy at the supply voltages providing temperature variation insensitive circuit performance in 180nm and 65nm CMOS technologies

6.4 Normalized Delay and energy at the supply voltages providing minimum energy delay product at 25°C in 180nm and 65nm CMOS technologies

6.5 Normalized Delay and energy at the supply voltages providing minimum energy delay product at 125°C in 180nm and 65nm CMOS technologies

6.6 Normalized Delay and energy at the supply voltages providing minimum energy at 25°C in 180nm and 65nm CMOS technologies

6.7 Normalized Delay and energy at the supply voltages providing minimum energy at 125°C in 180nm and 65nm CMOS technologies

7.1 Supply voltages that achieve minimum energy in a constant-V_{DD} and constant-f_{i}, Brent-Kung Adder

7.2 Supply voltages for achieving minimum energy consumption in standard constant-V_{DD} and constant-f_{i}, circuits

7.3 Maximum-f_{i} and energy consumption for a Brent-Kung adder at different supply voltages (Temperature = 125°C)
7.4 Propagation delay comparison of constant-\( V_{\text{DD}} \) circuits at \( V_{\text{DD-25}} \) and circuits with TA-DVS capability

7.5 Propagation delay comparison of constant-\( V_{\text{DD}} \) circuits at \( V_{\text{DD-125}} \) and circuits with TA-DVS capability

7.6 Normalized energy savings with the temperature-adaptive voltage scaling scheme

7.7 Propagation delay comparison of zero-body-biased circuits operating at \( V_{\text{DD-25}} \) and circuits with TA-BB capability

7.8 Propagation delay comparison of zero-body-biased circuits operating at \( V_{\text{DD-125}} \) and circuits with TA-BB capability

7.9 Normalized energy reduction with the temperature-adaptive voltage scaling scheme

7.10 Post-layout current measured at the different terminals of the PMOS device for various body-bias voltages

7.11 Percent energy reduction with the temperature-adaptive voltage tuning schemes

8.1 Supply voltages that achieve minimum energy in a constant-\( V_{\text{DD}} \) and constant-\( f \), 64-bit x 64-bit SRAM array

8.2 Read delay, write delay, and energy consumption of the SRAM array at various temperatures and power supply voltages

8.3 Read delay, write delay, and energy consumption of the SRAM array at various temperatures and body-bias voltages

8.4 Post-layout current measured at the different terminals of the PMOS device for various body-bias voltages

8.5 Hold and read static margins of the SRAM bit-cells operating at different power supply voltages and temperatures
List of Figures

1.1 The first transistor [2] ................................................................. 2
1.2 First integrated circuits. (a) Jack Kilby’s hybrid IC with discrete wires (1957). (b) Robert Noyce’s monolithic IC with vapor deposited metal connections (1959).............................. 3
1.3 Microphotographs comparing the evolution of integrated circuits. Sizes of dies are not to scale. (a) Intel 4004, 1971 [81]. (b) Intel Pentium IV microprocessor, 2002 [81] ................................................................. 4
1.4 Microphotographs of recent multi-core processors. Die sizes are not to scale. (a) Intel Xeon 5100 series dual-core processor, 2006 [81]. (b) AMD’s quad-core Opteron processor, 2007 [13].............................. 5
1.5 A typical heat removal system for a high-performance microprocessor…… 6
1.6 General form of Moore’s law [1], [81] .............................................. 8
1.7 The reduction in unit transistor price with time [18]......................... 8
1.8 Number of transistors shipped every year [18].................................. 9
1.9 Scaling of the minimum feature size of transistors [18].................... 12
1.10 Scaling of the MOSFET effective (electrical) dielectric thickness [18].... 12
1.11 Supply voltage scaling [18].............................................................. 13
1.12 Power consumption of lead Intel microprocessors [1], [117].............. 14
1.13 Power density of lead Intel microprocessors..................................... 15
1.14 Leakage power consumed by an NMOSFET in a 180nm CMOS technology at different temperatures. Width = 600nm ........................................... 16
2.1 A CMOS gate driving an output capacitor. (a) Equivalent representation for a low-to-high output node transition. (b) Equivalent representation for a high-to-low output node transition........................................ 21
2.2 Block diagram of a 16-bit ripple carry adder…………………………………….. 25

2.3 Waveforms indicating the glitches at the outputs of a 16-bit ripple carry adder. The numbers indicate the sum output voltage at the corresponding bit position. Sum bits 0 and 1 experience partial glitching. The glitching voltage rises all the way to $V_{\text{HIGH}}$ at bit positions SUM[2] to SUM[15]…… 25

2.4 Variation of the threshold voltage with channel length for devices in a 65nm CMOS technology. $N_{\text{width}} = 600\text{nm}$. $P_{\text{width}} = 1200\text{nm}$. Temperature = 25°C. The threshold voltage is measured as the gate-to-source voltage required for producing $1\times10^{-4}\text{A}$ drain current……………………………… 28

2.5 Variation of the threshold voltage with the drain voltage for devices in a 65nm CMOS technology. $N_{\text{width}} = 600\text{nm}$. $P_{\text{width}} = 1200\text{nm}$. Temperature = 25°C. The threshold voltage is measured as the gate-to-source voltage required for producing $1\times10^{-4}\text{A}$ drain current……………………………… 29

2.6 Drain current versus gate-to-source voltage curve for an NMOSFET in a 65nm CMOS technology. $N_{\text{width}} = 600\text{nm}$. Drain-to-source voltage = 1V. Temperature = 25°C………………………………………………... 31

2.7 Parasitic diodes in a CMOS circuit………………………………………………… 33

2.8 Different components of gate dielectric tunneling current in a MOSFET….. 34

2.9 The primary mechanism of gate dielectric direct tunneling current in PMOS and NMOS transistors. (a) PMOS transistor in inversion. (b) NMOS transistor in inversion…………………………………………………………………… 36

2.10 Short-circuit currents during input signal transients. $V_m$ is the threshold voltage of the NMOS device. $V_{\text{tp}}$ is the threshold voltage of the PMOS device…………………………………………………………………… 37

2.11 Impact of the output capacitance on the short-circuit current. (a) CMOS inverter with a large capacitive load and a sharp input signal. (b) CMOS inverter with a small capacitive load and a slowly rising input signal……… 38
2.12 Static CMOS inverter driven by a low swing signal. $I_{\text{STATIC}}$ is the static DC current produced by the inverter. $V_{\text{tn}}$ is the threshold voltage of the NMOS device. $V_{\text{tp}}$ is the threshold voltage of the PMOS device.

3.1 Die temperature variation of an Intel Itanium processor in a 180nm CMOS technology [81]. The heavily utilized core area of the die is at 120°C while the caches are at 70°C, thereby indicating a 50°C temperature gradient within the same die.

3.2 The silicon bonding model.

3.3 The energy band diagram of a semiconductor [68].

3.4 Band diagrams indicating the energy levels corresponding to the impurity atoms. (a) n-type extrinsic semiconductor [$E_d$]. (b) p-type extrinsic semiconductor [$E_a$].

3.5 Typical temperature dependence of the majority-carrier concentration in a doped semiconductor. A phosphorous-doped Si sample ($N_D = 10^{15} / \text{cm}^3$).

3.6 Source of the majority charge carriers in a donor-doped semiconductor material at different temperatures (T). (a) At $T = 0K$. (b) At low temperatures ($T < 150K$). (c) At very high temperatures ($T > 450K$). $E_C$: Conduction-band energy. $E_D$: Energy level of donor atoms. $E_V$: Valence-band energy [68].

3.7 Threshold voltage variations with temperature for devices in 180nm and 65nm CMOS technologies.

3.8 Gate overdrive variation with temperature for devices in 180nm and 65nm CMOS technologies.

3.9 Mobility variation with temperature for devices in 180nm and 65nm CMOS technologies.

3.10 Variation of MOSFET drain current with supply voltage and temperature in the TSMC 180nm CMOS technology. $|V_{\text{DS}}| = |V_{\text{GS}}| = V_{\text{DD}}$ and $|V_{t0}(T_0)| = 0.46V$. 


3.11 Variation of MOSFET drain current with supply voltage and temperature in
the predictive 65nm CMOS technology. $|V_{DS}| = |V_{GS}| = V_{DD}$ and $|V_{t0}(T_0)| =
0.22V$……………………………………………………………………... 58

3.12 Percent delay variation with temperature for circuits operating at the
nominal supply voltage ($V_{DD} = 1.8V$) in the TSMC 180nm CMOS
technology…………………………………………………………………... 59

3.13 Percent delay variation with temperature for circuits operating at the
nominal supply voltage ($V_{DD} = 1V$) in the predictive 65nm CMOS
technology………………………………………………………………..…. 59

3.14 Percent energy variation with temperature for circuits operating at the
nominal supply voltage ($V_{DD} = 1.8V$) in the TSMC 180nm CMOS
technology…………………………………………………………………... 60

3.15 Percent energy variation with temperature for circuits operating at the
nominal supply voltage ($V_{DD} = 1V$) in the PTM 65nm CMOS technology… 61

4.1 MOSFETs biased with different power supplies. (a) An n-channel
MOSFET. (b) A p-channel MOSFET………………………………………. 66

4.2 An integrated circuit with multiple I/O, power supply, ground, and body
contact terminals…………………………………………………………….. 67

4.3 3-stage zero-body-biased inverter chain…………………………………….. 69

4.4 Circuit set-up for power measurement using the proposed methodology….. 70

4.5 Comparison of the average power consumption measured with the HSPICE
built-in command AVG POWER and the actual power consumption when
the input is oscillating at 1GHz. $V_{DD} = 1.0V$ (super-threshold operation)….. 72

4.6 Comparison of the average power consumption measured with the HSPICE
built-in command AVG POWER and the actual power consumption of the
subthreshold logic circuits. Input signal frequency is 1MHz. $V_{DD} = 0.2V$…. 73
4.7 Comparison of the stand-by mode power consumption measured with the HSPICE built-in command AVG POWER and the actual power consumption when the inputs are biased at zero. $V_{DD} = 1.0V$………………… 74

4.8 Comparison of the stand-by mode power consumption measured with the HSPICE built-in command AVG POWER and the actual power consumption when the inputs are biased at $V_{DD} = 1.0V$……………..…..………. 75

4.9 NMOS devices in a 65nm CMOS technology. Temperature = 125°C. Width = 300nm. Length = 65nm. (a) Drain, gate, and source terminals biased at 0V, 1V, and 0V respectively. (b) Drain, gate, and source terminals biased at 1V, 1V, and 0V respectively…………………………… 79

4.10 Comparison of subthreshold and gate-oxide leakage currents produced by an NMOS transistor for various supply voltages at three different temperatures. $V_{Gate} = V_{Source} = V_{Bulk} = 0V$, $V_{Drain} = V_{DD}$…………………………… 83

4.11 Percent error in the power measured using the HSPICE IN-CAD and the ACTUAL for devices in a 65nm CMOS technology with $\lvert V_{GS} \rvert = \lvert V_{SB} \rvert = 0$ and $\lvert V_{DS} \rvert = 1V$……………………………………………………………………… 84

4.12 The percent error in the power measurement with HSPICE IN-CAD for an NMOS device with different body-bias voltages………………………… 85

5.1 Variation of the MOSFET drain current with supply voltage and temperature in the TSMC 180nm CMOS technology. $\lvert V_{DS} \rvert = \lvert V_{GS} \rvert = V_{DD}$ and $\lvert V_{t0(T0)} \rvert = 0.46V$. $V_{DD,insensitive}$ for NMOSFET is 0.71V. $V_{DD,insensitive}$ for PMOSFET is 1.13V……………………………………………………………………… 91

5.2 Variation of the MOSFET drain current with supply voltage and temperature in the predictive 65nm CMOS technology. $\lvert V_{DS} \rvert = \lvert V_{GS} \rvert = V_{DD}$ and $\lvert V_{t0(T0)} \rvert = 0.22V$. $V_{DD,insensitive}$ for NMOSFET is 0.33V. $V_{DD,insensitive}$ for PMOSFET is 0.29V……………………………………………………………………… 91

5.3 Supply voltages ($V_{DD,insensitive}$) that achieve temperature variation insensitive speed characteristics in the TSMC 180nm CMOS technology…. 92
5.4 Supply voltages ($V_{DD,\text{insensitive}}$) that achieve temperature variation insensitive speed characteristics in the predictive 65nm CMOS technology.

5.5 Variation of the MOSFET drain current with threshold voltage and temperature in the TSMC 180nm CMOS technology. $|V_{DS}| = |V_{GS}| = V_{DD} = 1.8V$. $V_{t0,\text{insensitive}}$ for NMOSFET is 1.58V. $V_{t0,\text{insensitive}}$ for PMOSFET is 1.28V.

5.6 Variation of the MOSFET drain current with threshold voltage and temperature in the predictive 65nm CMOS technology. $|V_{DS}| = |V_{GS}| = V_{DD} = 1.0V$. $V_{t0,\text{insensitive}}$ for NMOSFET is 0.93V. $V_{t0,\text{insensitive}}$ for PMOSFET is 0.98V.

5.7 Threshold voltages ($V_{t0,\text{insensitive}}$) that achieve temperature variation insensitive speed characteristics in the TSMC 180nm CMOS technology.

5.8 Threshold voltages ($V_{t0,\text{insensitive}}$) that achieve temperature variation insensitive speed characteristics in the predictive 65nm CMOS technology.

6.1 Normalized energy per cycle, delay, and energy-delay product as a function of the supply voltage at the room temperature (Temperature = 25°C) for an inverter in the TSMC 180nm CMOS technology.

6.2 The energy-delay product at two different temperatures and the percent delay variation as a function of the supply voltage. The temperature is increased from 25°C to 125°C for an 8-bit array multiplier in the TSMC 180nm CMOS technology.

6.3 Normalized switching, leakage, and total energy as a function of the supply voltage at 25°C for a 16-bit Brent-Kung adder in the TSMC 180nm CMOS technology.

6.4 Normalized switching, leakage, and total energy as a function of the supply voltage at 125°C for a 16-bit Brent-Kung adder in the TSMC 180nm CMOS technology.
6.5 Normalized switching, leakage, and total energy as a function of the supply voltage at 25°C for a 16-bit Brent-Kung adder in the predictive 65nm CMOS technology…………………………………………………………... 107

6.6 Normalized switching, leakage, and total energy as a function of the supply voltage at 125°C for a 16-bit Brent-Kung adder in the predictive 65nm CMOS technology………………………..…………………………………. 107

7.1 Flow-chart for identifying the supply voltage that achieves minimum energy consumption at a specific temperature ($T_{\text{specific}}$) for a standard constant-$V_{\text{DD}}$ and constant-$f_s$ IC……………………………………………………………………………….. 118

7.2 The input and output waveforms of an inverting circuit……………………... 120

7.3 Supply voltages that minimize the energy consumption at different temperatures in the 16-bit Brent-Kung adder. $V_{\text{DD-25}}$, $V_{\text{DD-50}}$, $V_{\text{DD-75}}$, $V_{\text{DD-100}}$, and $V_{\text{DD-125}}$ are the supply voltages providing minimum energy consumption at 25°C, 50°C, 75°C, 100°C, and 125°C, respectively……….. 122

7.4 Output signal ($\text{SUM}[15]$) of a 16-bit Brent-Kung adder operating at $V_{\text{DD-25}}$ (0.26V) and various temperatures…………………………………………………………... 126

7.5 Output signal swing of $\text{SUM}[15]$ for a 16-bit Brent-Kung adder at various scaled supply voltages………………………………………………………………………….. 127

7.6 Temperature-adaptive body-bias technique. $V_{DD}$: standard constant-supply voltage providing minimum energy ($V_{DD-25}$ or $V_{DD-125}$)…………………………. 132

7.7 A PMOS device in the TSMC 180nm CMOS technology. The gate and source terminals are biased at 0.27V. Temperature = 125°C. The device is reverse-body-biased by applying a voltage higher than 0.27V to the body terminal…………………………………………………………………………… 135
<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.8</td>
<td>Delay versus energy consumption plots for the 16-bit ripple carry adders operating at $V_{DD-25}$ and at the optimized high temperature supply voltage with the TA-DVS in the presence of process parameter and supply voltage variations. $N_{CH}$, $L_{GATE}$, $T_{OX}$, and $V_{DD}$ are assumed to have independent normal Gaussian statistical distributions with a three-sigma variation of 10%</td>
</tr>
<tr>
<td>8.1</td>
<td>On-die memory capacity in Intel Xeon microprocessors [81]</td>
</tr>
<tr>
<td>8.2</td>
<td>Six transistor (6T) SRAM bit-cell</td>
</tr>
<tr>
<td>8.3</td>
<td>Voltage transfer characteristics (VTC) of the cross-coupled inverters to measure the static noise margin. The length of the side of the largest embedded square in the butterfly curve is the SNM</td>
</tr>
<tr>
<td>8.4</td>
<td>Read static noise margins of the 6T-SRAM cells for different supply voltages and cell ratios. $W_{A1} = W_{A2} = W_{P1} = W_{P2} = 360nm$. $W_{N1} = W_{N2} = (Memory\ Cell\ Ratio) \times 360nm$. TSMC 180nm CMOS technology</td>
</tr>
<tr>
<td>8.5</td>
<td>Layout of the 6T-SRAM bit-cell</td>
</tr>
<tr>
<td>8.6</td>
<td>The simulation setup of a 64-bit x 64-bit SRAM array. The read and the write propagation delays are measured with respect to the shaded SRAM cell</td>
</tr>
<tr>
<td>8.7</td>
<td>The write and read drivers of the SRAM array. (a) The write driver. (b) The read driver</td>
</tr>
<tr>
<td>8.8</td>
<td>Signals indicating the read and the write operations in an SRAM bit-cell</td>
</tr>
<tr>
<td>8.9</td>
<td>CLK and DATA_OUT of the 64th column in a 64-bit x 64-bit SRAM array operating at $V_{DD-25}$ (0.28V) and various temperatures</td>
</tr>
<tr>
<td>8.10</td>
<td>A PMOS device in the TSMC 180nm CMOS technology. The gate and source terminals are biased at 0.28V. Temperature = 125°C. The device is reverse-body-biased by applying a voltage higher than 0.28V to the body terminal</td>
</tr>
</tbody>
</table>
8.11 Delay versus energy consumption plots for the 64-bit x 64-bit SRAM arrays operating at $V_{DD-25}$ and at the optimized high temperature supply voltage with the TA-DVS in the presence of process parameter and supply voltage variations. $N_{CH}$, $L_{GATE}$, $T_{OX}$, and $V_{DD}$ are assumed to have independent normal Gaussian statistical distributions with a three-sigma variation of 10%.

10.1 Temperature-adaptive dynamic supply voltage scaling technique

10.2 Simplified buck converter schematic. M1 and M2 are the power MOSFETs. $L_f$ is the filter inductance. $C_f$ is the filter capacitance. $R_{Load}$ is the load resistance. $V_{Out}$ is the output voltage [143].

10.3 Percent energy loss due to the output voltage ripple [122].

10.4 A complex integrated circuit with global interconnects connecting the different functional blocks.

10.5 A complex integrated circuit with repeaters.

10.6 Interconnects. (a) Without repeaters. (b) With repeaters.

10.7 Typical integrated circuit with on-die temperature gradients.

10.8 The supply and threshold voltages in different CMOS technology generations.

10.9 DRAM refresh cycles at different temperatures.

10.10 Comparator circuitry for temperature-adaptive DRAM refresh cycle [134]. Trigger is activated to refresh the DRAM cells.

10.11 FinFET architectures. (a) Tied-gate FinFET. (b) Independent-gate FinFET.

10.12 Cross-sectional top-view of a FinFET.

10.13 2-input static NAND gates. (a) Single-gate planar MOSFET implementation. (b) Independent gate FinFET implementation.
10.14 Typical gate and interconnect delays at different technology nodes. An optimally repeated line is assumed and the delays of the repeaters are included in the interconnect delay [132].

10.15 Schematic diagrams of Systems-on-Chip. (a) Top view of a 2-D planar IC. (b) Cross-sectional view of a 3-D SoC.

10.16 Heat flow path in typical ICs. (a) Schematic view of heat flow in a 2-D planar IC. (b) Schematic of an ‘n’ layer 3-D IC with heat sink at the bottom.

10.17 Temperature profile of a 3-D chip with increasing number of silicon layers. The temperature and the power density increase as more silicon layers are packed in the 3-D IC.
Chapter 1
Introduction

Studies on semiconductor materials date back to 1833 when Michael Faraday investigated the temperature dependence of the electrical conductivity of silver sulphide [3]. With these experiments, Michael Faraday witnessed an enhancement in the electrical conductivity of silver sulphide with increasing temperature, a signature property of semiconductors [2]-[4]. This first noted experimental observation of semiconducting behavior was followed by the discovery of the photoconductivity of selenium by Willoughby Smith in 1873. In 1874, Ferdinand Braun discovered that certain materials would allow current only in one direction (rectification). Braun observed that the resistance of lead sulphide displayed a dependence on the magnitude and the sign of the applied voltage that did not obey the Ohm’s law. Furthermore, in 1883, Charles Fritts produced the first large area dry rectifier. Thus by 1885, the four fundamental properties of semiconductors had been discovered: the negative temperature coefficient of resistance, the rectification, the photoconductivity, and the photo-electromotive force. However, no theory could explain these distinctive properties of the semiconductor materials for another fifty years until the early 1930s [2]-[4].

In 1931, Alan Wilson developed a quantum mechanical model of the semiconductors [4]. According to quantum mechanics, the energy states in a crystal exist as bands separated by energy gaps. This understanding was able to explain the difference between insulators, metals, and semiconductors based on the idea of filled and empty energy bands. In addition, the theory also showed how the controlled introduction of small amounts of impurity atoms into a semiconductor could strongly modulate the electrical conductivity. The exponential increase of electrical conductivity with increasing temperature and the existence of bipolar conduction (electrons and holes) in semiconductors were explained using the band theory [4].

In 1935, metallurgist Russell Ohl from the Bell Laboratories initiated an effort to understand the formation and the electrical characteristics of silicon crystals [2]. By 1940, Ohl discovered that depending on how a single crystal of silicon is prepared,
semiconductors capable of allowing current only in a specific direction could be produced. Semiconductors rectifying current in the positive direction (allowing current with positive polarity) were called p-type. Alternatively, the semiconductors rectifying current in the negative direction (allowing current with negative polarity) were named n-type. Ohl displayed that a sample melt prepared with an n-type semiconductor at one end and a p-type semiconductor at the other end exhibits excellent rectifier characteristics with phenomenal photo-electromotive force [4]. Ohl’s experiments were instrumental in the eventual development of the p-n junction diode. The p-n junction diodes continue to be the most important building blocks for present-day semiconductor devices.

In January 1946, a team of scientists from the Bell laboratories (Walter Brattain, John Bardeen, John Pearson, Bert Moore, Bill Shockley, Stanley Morgan, and Robert Gibney) initiated an extensive research project to explore the semiconducting properties of silicon and germanium [2]. In December 1947, Bardeen and Brattain managed to produce the first working transistor, as shown in Fig. 1.1. The transistor consists of a block of germanium (base) with two very closely spaced gold contacts held by a spring, as shown in Fig. 1.1. In the experiments, Bardeen and Brattain found that a small voltage applied on one of the contacts modulated the current flow between the other two terminals, amplifying the input signal by orders of magnitude [2].

![Fig. 1.1. The first transistor [2].](image-url)
The invention of the transistor marked the beginning of the revolutionary developments in the semiconductor industry. As the transistor manufacturing process matured, more complex systems with increasing number of discrete transistors were produced. Interconnecting the increased number of discrete transistors however posed higher design and manufacturing complexity. In 1957, Jack Kilby of Texas Instruments developed the first integrated circuit (IC). Kilby used discrete wires for interconnecting a transistor, a resistor, and a capacitor [2]. This type of hybrid integration with discrete (off-chip) metal lines is shown in Fig. 1.2a. Later in 1959, Robert Noyce managed to implement the first monolithic IC with vapor deposited metal connections on-chip. Noyce interconnected the integrated components using a metal vaporization process which allowed the planar integration of the wires. An integrated circuit with this planar fabrication technique is shown in Fig. 1.2b. Today, nearly fifty years later, the planar processing continues to be the primary technology used to fabricate the integrated circuits [2]-[4].

Fig. 1.2. First integrated circuits. (a) Jack Kilby’s hybrid IC with discrete wires (1957). (b) Robert Noyce’s monolithic IC with vapor deposited metal connections (1959).

The semiconductor industry has grown tremendously over the past five decades [1], [13], [81], [117]. In order to meet the customer demand for integrated systems with higher performance supporting a wider range of applications, the semiconductor industry releases a new, faster, and more compact process technology generation every two to three years. The performance and the complexity of today’s integrated systems have
increased manifolds as compared to the very first integrated systems of the early 1960s. Microphotographs of the Intel’s first microprocessor (Intel 4004, 1971) [81] and a recent single core microprocessor (Pentium IV, 2002) [81] are shown in Fig. 1.3.

(a)                                                               (b)

Fig. 1.3. Microphotographs comparing the evolution of integrated circuits. Sizes of dies are not to scale. (a) Intel 4004, 1971 [81]. (b) Intel Pentium IV microprocessor, 2002 [81].

The IC industry has been consistently scaling the design rules, increasing the chip areas, and manufacturing larger wafers for over forty-eight years [5]. Consequently, the semiconductor industry has enjoyed phenomenal enhancement in circuit speed and functionality combined with a steady decline in the cost per function [5]. In high-end processors, increasing the clock frequency has traditionally been the primary objective. In addition to the performance enhancements due to technology scaling, the performance of integrated circuits have also been enhanced by utilizing the growing number of transistors to develop novel circuit techniques and microarchitectures [8]. In recent years, the performance of integrated systems is further enhanced by increasing the number of computational units in a package [14]. The microphotographs of an Intel dual-core processor (Xeon 5100 series, 2006 [81]) and an AMD quad-core processor (Opteron, 2007 [13]) are shown in Fig. 1.4. Dual core processors provide two processing units
within a single package. Alternatively, quad-core processors provide four computing units in one package, thereby enhancing the performance.

![Microphotographs of recent multi-core processors.](image)

**Fig. 1.4.** Microphotographs of recent multi-core processors. Die sizes are not to scale. (a) Intel Xeon 5100 series dual-core processor, 2006 [81]. (b) AMD’s quad-core Opteron processor, 2007 [13].

One primary side effect of enhanced circuit performance and functionality is typically an increase in the power consumption. Using power hungry circuit techniques and microarchitectures (with continuously increasing levels of speculative execution often translated into an inefficient use of energy) have increased the power consumption of the high performance circuits many folds over the years [1], [9], [117]. The power consumed by a circuit is dissipated as heat through the substrate. The heat removal is traditionally achieved with inexpensive packages, passive heat sinks, and air fan flows. A representation of a typical heat removal system in a high-performance microprocessor is shown in Fig. 1.5. With power consumption rising well above 100W, however, more expensive packaging and cooling technologies will soon be required for advanced integrated circuits [1], [117].

A second family of ICs are utilized in systems that target miniaturization and portability. Traditionally, portable applications are aimed at reducing the power
consumption. The enhanced computational throughput goals used to play a secondary role in the portable systems of the 1990s due to the strict requirements for an extended battery lifetime. However, particularly since the late 1990s, customer demand has been growing for higher performance and a wider variety of applications in mobile systems as well. Today, a typical customer expects and prefers to invest in mobile devices with computational capabilities comparable to the desktop systems. While the performance requirements in portable applications increase at a fast pace in response to this market demand, battery technologies evolve at a much slower rate [17].

![Diagram of a high-performance microprocessor heat removal system](image)

Circuits employed in high performance systems typically cannot be used in portable applications due to the power hungry characteristics of high speed ICs. Alternatively, circuits designed for battery operated devices typically cannot be utilized in high performance systems because of the low throughput characteristics of the portable ICs. Today, the semiconductor industry experiences a shift in the requirements of both the high performance and the portability ends of the market. Power dissipation is no longer a secondary issue in high performance systems. Similarly, enhancing throughput is as important as reducing power in many portable applications. Energy efficient semiconductor devices, circuit techniques, and micro-architectures are necessary to maintain the pace of growth in the semiconductor industry [1], [117].
The scaling of devices and wires with each new technology generation paves the way for the growth of the semiconductor industry. The scaling trends of IC technologies are reviewed in Section 1.1. An outline of the dissertation is presented in Section 1.2.

1.1. Scaling Trends of Integrated Circuits

ICs were expensive during the 1960s, limiting the use of ICs only for specific military applications with severe restrictions on weight and size and strict requirement for reliability. In 1965, only six years after the invention of the IC, Gordon Moore noticed that the cost of integrated circuits was steadily decreasing as the technology evolved and the fabrication techniques matured [19], [20]. Moore saw that the shrinking transistor sizes, the increasing manufacturing yield, and the larger wafer and die sizes would make ICs increasingly cheaper, more powerful, and abundant. In 1965, witnessing the power of integrated circuits, Moore declared that “the future of integrated electronics is the future of electronics itself”.

The general form of Moore’s law [19], [20], [81] is depicted in Fig. 1.6. As more components are added onto an IC at a particular process technology generation (or a technology node), the relative manufacturing cost per component decreases (assuming that the same amount of semiconductors and the same package are used to incorporate more components). However, the complexity (at the process, circuits, and layout levels) increases, thereby degrading the yield with integrating more components onto an IC. There is, therefore, an optimum number of components per IC that minimizes the fabrication cost per component. Meanwhile, the unit price of a transistor decreases with the scaling of technology. The optimum number of transistors per IC that minimizes the manufacturing cost per component increases from one technology generation to the next as shown in Fig. 1.6.

Following Moore’s remarkable prediction in 1965, the total number of transistors that can be integrated onto a piece of semiconductor material has increased by more than a million times. The increase in the number of transistors on a single die has reduced the manufacturing cost of the individual transistors. The reduction in the average price of a transistor with time is shown in Fig. 1.7 [18]. The reduction in the average price of
transistors has allowed an increase in the number of transistors shipped per year parallel to the expansion of the semiconductor industry as shown in Fig. 1.8 (data from Intel) [18]. The transistors and the integrated electronics have found a way into virtually all the products that characterize the modern society, ranging from automobiles to the greeting cards.

![Graph showing Moore's law](image)

Fig. 1.6. General form of Moore’s law [1], [81].

![Graph showing reduction in transistor price](image)

Fig. 1.7. The reduction in unit transistor price with time [18].
High performance microprocessors represent one end of the electronics industry where enhanced performance has traditionally been the key parameter driving the market success. The impact of technology scaling on the high-end microprocessors from Intel is examined in this section [81]. The scaling trends of some of the important technological parameters among the different microprocessor generations from Intel are listed in Table 1.1. Intel’s first microprocessor (Intel 4004) was shipped in 1971. The number of transistors integrated on a microprocessor has increased from 2300 (in 1971) to 820,000,000 (in 2007), as listed in Table 1.1. The primary facilitator for this remarkable increase in the number of integrated transistors is the reduction in the feature size of transistors from 10µm in 1971 to 45nm in 2007, as listed in Table 1.1.
TABLE 1.1.
SCALING TRENDS OF HIGH-PERFORMANCE MICROPROCESSORS FROM INTEL [81]

<table>
<thead>
<tr>
<th>Processor Name</th>
<th>Year</th>
<th>Clock Speed</th>
<th>No. of Transistors (in millions)</th>
<th>Feature Length</th>
<th>Process Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>4004</td>
<td>1971</td>
<td>108 KHz</td>
<td>0.0023</td>
<td>10µm</td>
<td>PMOS</td>
</tr>
<tr>
<td>8080</td>
<td>1974</td>
<td>2 MHz</td>
<td>0.0045</td>
<td>6µm</td>
<td>NMOS</td>
</tr>
<tr>
<td>8086</td>
<td>1978</td>
<td>5 MHz</td>
<td>0.029</td>
<td>3µm</td>
<td>NMOS</td>
</tr>
<tr>
<td>286</td>
<td>1982</td>
<td>6 MHz</td>
<td>0.134</td>
<td>1.5µm</td>
<td>CMOS</td>
</tr>
<tr>
<td>386</td>
<td>1985</td>
<td>16 MHz</td>
<td>0.275</td>
<td>1.5µm</td>
<td>CMOS</td>
</tr>
<tr>
<td>486</td>
<td>1989</td>
<td>25 MHz</td>
<td>1.2</td>
<td>1µm</td>
<td>CMOS</td>
</tr>
<tr>
<td>Pentium</td>
<td>1993</td>
<td>66 MHz</td>
<td>3.1</td>
<td>800nm</td>
<td>Bi-CMOS</td>
</tr>
<tr>
<td>Pentium Pro</td>
<td>1995</td>
<td>200 MHz</td>
<td>5.5</td>
<td>600nm</td>
<td>CMOS</td>
</tr>
<tr>
<td>Pentium II</td>
<td>1997</td>
<td>300 MHz</td>
<td>7.5</td>
<td>250nm</td>
<td>CMOS</td>
</tr>
<tr>
<td>Pentium III</td>
<td>1999</td>
<td>500 MHz</td>
<td>9.5</td>
<td>180nm</td>
<td>CMOS</td>
</tr>
<tr>
<td>Pentium IV</td>
<td>2000</td>
<td>1.5 GHz</td>
<td>42</td>
<td>180nm</td>
<td>CMOS</td>
</tr>
<tr>
<td>Pentium M</td>
<td>2002</td>
<td>1.7 GHz</td>
<td>55</td>
<td>90nm</td>
<td>CMOS</td>
</tr>
<tr>
<td>Itanium II</td>
<td>2002</td>
<td>1 GHz</td>
<td>220</td>
<td>130nm</td>
<td>CMOS</td>
</tr>
<tr>
<td>Pentium D</td>
<td>2005</td>
<td>3.2 GHz</td>
<td>291</td>
<td>65nm</td>
<td>CMOS</td>
</tr>
<tr>
<td>Core 2 Duo</td>
<td>2006</td>
<td>2.93 GHz</td>
<td>291</td>
<td>65nm</td>
<td>CMOS</td>
</tr>
<tr>
<td>Dual Core</td>
<td>2006</td>
<td>1.66 GHz</td>
<td>1720</td>
<td>90nm</td>
<td>CMOS</td>
</tr>
<tr>
<td>Itanium II</td>
<td>2006</td>
<td>2.66 GHz</td>
<td>582</td>
<td>65nm</td>
<td>CMOS</td>
</tr>
<tr>
<td>Quad-Core Xeon</td>
<td>2007</td>
<td>&gt; 3 GHz</td>
<td>820</td>
<td>45nm</td>
<td>CMOS</td>
</tr>
</tbody>
</table>

The three approaches to technology scaling are constant-field scaling, constant-voltage scaling, and general scaling [26]. In constant-field scaling, all the device dimensions including the channel length, the device width, and the gate-oxide thickness are reduced by a factor of $1/S$ [44]. The supply voltage ($V_{DD}$) and the threshold voltage ($V_t$) are also reduced by $1/S$. The substrate doping is however increased by $S$. Since the distances between the device terminals and the voltages applied to the devices are scaled by the same factor, the electric fields within the devices remain constant [44]. Constant-field scaling leads to higher device integration density, enhanced circuit performance, and reduced power consumption [26]. In order to keep the new devices compatible with the existing systems that employ previous IC technology generations, however, the voltage...
levels cannot be scaled arbitrarily. Unlike constant-field scaling, in constant-voltage scaling, only the lateral dimensions of the devices are scaled leaving the voltage levels unchanged. Constant-voltage scaling is therefore a geometrical scaling process. This approach provides quadratic improvement in the gate delay while ensuring compatibility with the existing I/O standards [44]. Constant-voltage scaling, however, increases the electric fields within the devices thereby causing the velocity of the charge carriers to saturate (velocity saturation) [26], [44]. Furthermore, the higher voltage levels with the constant-voltage scaling approach increases the power consumption of circuits [26]. Devices in the nanometer regime are therefore scaled using a more generalized scaling approach that is a compromise between the device reliability, the I/O voltage compatibility, the power consumption, and the performance goals. With the general scaling, device dimensions are scaled by a factor of $S$ while the voltages are scaled by a different factor of $U$. The general scaling approach typically enhances the circuit speed and the energy efficiency while also ensuring the reliability of the scaled devices [26], [44].

The historical scaling trends of the device channel length, the gate-oxide thickness, and the power supply voltage are shown in Figs. 1.9, 1.10, and 1.11, respectively [18]. A transistor operates faster and becomes cheaper when the device dimensions are scaled. Furthermore, the advancements in the fabrication technologies have enabled the manufacturing of ICs with larger die areas [21], [44]. The higher number of transistors per IC provides enhanced circuit performance and functionality at the cost of increased power consumption [1], [7], [15], [21], [117].
Fig. 1.9. Scaling of the minimum feature size of transistors [18].

Fig. 1.10. Scaling of the MOSFET effective (electrical) dielectric thickness [18].
The power consumption trends of lead Intel microprocessors are shown in Fig. 1.12 [1], [117]. The maximum power consumption of the Intel microprocessors has been increasing over the past forty years, as shown in Fig. 1.12. The circuitry of the first Intel microprocessor is PMOS. Starting with Intel 8080 in 1974, NMOS became the preferred circuit technology due to the speed and area advantages of NMOS as compared to PMOS. NMOS circuits however suffer from intrinsic static DC power consumption and low noise margins [1], [117]. The use of NMOS technology in the IC industry was therefore terminated by the end of 1970s. The CMOS technology was adapted by the IC industry in the early 1980s due to the intrinsically lower power dissipation and better scaling characteristics of CMOS as compared to NMOS. CMOS has become the preferred technology in the lead Intel microprocessors since 1982 [1], [117]. The transition from NMOS to CMOS temporarily reduced the power consumption of Intel microprocessors, as shown in Fig. 1.12. However, the need to enhance performance has maintained the higher power consumption trends with each new technology generation even after the transition to the CMOS technology, as shown in Fig 1.12.
The ITRS projected die areas of memory banks and microprocessors are listed in Table 1.2 [35]. As listed in Table 1.2, the die size of memory banks and microprocessors increase with each new technology generation. However, the rate of growth of chip area is smaller as compared to the rate at which the number of transistors and the power consumption per chip increase. The significant increase in the power consumption causes a surge in the power density (power consumed per unit area) of ICs in each new technology generation as shown in Fig. 1.13.

**TABLE 1.2**

**THE ITRS PROJECTION OF THE DIE AREA OF MICROPROCESSORS AND MEMORY CHIPS IN FUTURE TECHNOLOGIES**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>DRAM G-bits/Chip</td>
<td>2.2</td>
<td>4.3</td>
<td>8</td>
<td>24</td>
<td>64</td>
<td>192</td>
</tr>
<tr>
<td>DRAM Chip Area (mm²)</td>
<td>400</td>
<td>480</td>
<td>530</td>
<td>630</td>
<td>710</td>
<td>860</td>
</tr>
<tr>
<td>MPU Chip Area (mm²)</td>
<td>340</td>
<td>370</td>
<td>400</td>
<td>470</td>
<td>540</td>
<td>620</td>
</tr>
<tr>
<td>Wafer Size (mm)</td>
<td>300</td>
<td>300</td>
<td>300</td>
<td>450</td>
<td>450</td>
<td>450</td>
</tr>
</tbody>
</table>
The increase in the overall power consumption and the power density of microprocessors raises the temperature of microprocessors well above the ambient temperature. The heat generated by the circuits is injected into the substrate and dissipated through the surface of the microprocessor. The surface temperature of today’s microprocessors can exceed 120°C [1], [28], [117]. Variations in the die temperature alter the circuit performance and the power consumption [78].

Leakage power consumption increases at elevated temperatures as illustrated in Fig. 1.14 [1], [117]. The increase in the leakage power with temperature increases the total power consumption of a circuit. The increase in the total power consumption, in turn, further elevates the die temperature. This positive feedback between the die temperature, the leakage current, and the total power consumption accelerates the degradation of the device and circuit reliability due to excessive heating and can even cause thermal runaway in extreme environments. Thermal management and temperature induced variations in the circuit performance and power consumption pose significant challenges in the design of current integrated systems. Techniques to eliminate the performance sensitivity of circuits to temperature fluctuations are outlined in this dissertation. New temperature-aware power reduction techniques are presented.
1.2. Outline of the Dissertation

Several techniques to reduce the sensitivity of circuit speed to the fluctuations of the die temperature are described in this dissertation. The goals of the proposed techniques are to simultaneously achieve temperature variation resilience and enhanced energy efficiency in CMOS circuits. New temperature-adaptive design methodologies for enhanced energy efficiency are presented. The speed and energy tradeoffs with the different techniques are provided.

An analysis of the power and heat dissipation related problems faced by the semiconductor industry starts with the identification of the sources of power consumption. The primary sources of power consumption in CMOS integrated circuits are described in Chapter 2. The impact of technology scaling on the different sources of power consumption is also briefly explained.

To reduce the sensitivity of the circuit speed to the die temperature variations, circuit designers must be aware of the physical mechanisms that alter the MOSFET drain current when the temperature fluctuates. The fluctuations in the drain current with temperature are due to the variations in the semiconductor properties. Fluctuations in the
die temperature alter the number of charge carriers in a semiconductor device [68]. The physical mechanisms that cause a variation in the carrier concentration of semiconductors with the temperature are reviewed in Chapter 3. The MOSFET device parameters that are modulated when the die temperature fluctuates are identified. The temperature effects on the device and circuit parameters are explored using BSIM [42], [72], [73].

Computer-aided design (CAD) tools are used for the pre-fabrication characterization of integrated circuits [55], [60]. Design objectives such as speed, area, reliability, and power consumption can be verified with the aid of the CAD tools. Reducing the power consumption is a primary objective in the design of digital integrated circuits. A significant portion of the research work presented in this dissertation is aimed at achieving enhanced energy efficiency in CMOS circuits. Power consumption of CMOS circuits can be lowered by employing several techniques described in [1], [10]-[12], and [117]. Accurate power estimation with the circuit simulators is critical to be able to correctly identify the most effective techniques that satisfy the design objectives.

A generic methodology to accurately measure the power and energy consumption with the circuit simulators is described in Chapter 4. The actual power consumption measured using the proposed method is compared with the power measurements using the built-in functions of the two most-popular commercial circuit simulators: HSPICE [55], [56] and CADENCE-SPECTRE [60], [61]. Results indicate that the power measurements with the built-in functions can introduce errors exceeding 8540. The assumptions and the simplifications that cause significant errors in power estimation with the built-in functions of the commonly used CAD tools are identified.

Weakening the sensitivity of the circuit speed to the die temperature fluctuations is desirable for reducing the uncertainty in the propagation delay characteristics of CMOS circuits. A design methodology for suppressing the drain current and the propagation delay variations due to temperature fluctuations is described in Chapter 5 [28]. There exists a bias voltage at which the temperature fluctuation induced gate-overdrive variations counterbalance the carrier mobility variations experienced by the transistors when the temperature fluctuates [26], [75]-[78], [89], [90]. MOSFETs biased at this voltage produce temperature variation insensitive constant drain current [26], [75]-[78], [89], [90]. The optimum bias voltages that achieve temperature variation insensitive drain
current are identified for devices in 180nm and 65nm CMOS technologies. An alternative design methodology based on threshold voltage optimization for providing temperature variation insensitive speed is also evaluated [91]. The energy efficiency and the propagation delay characteristics of the two techniques are compared.

Supply voltage scaling is an effective technique to lower all the primary components of power consumption in CMOS circuits [1], [117]. As the supply voltage is reduced, the energy per cycle decreases while the propagation delay increases [102], [103]. The energy-delay product, therefore, has a minimum, as described in [102] and [103]. Furthermore, subthreshold operation minimizes the total energy consumption of a CMOS circuit [66]. In Chapter 6, the supply voltages that achieve minimum energy-delay product and minimum energy consumption are identified for circuits in 180nm and 65nm CMOS technologies. Results indicate that these supply voltages are lower than the prescribed nominal supply voltage. The speed and energy tradeoffs in circuits operating at the supply voltages that provide temperature variation insensitivity, minimum energy consumption, and minimum energy-delay product are compared.

Integrated circuits with ultra-low-voltage power supplies are highly sensitive to process and temperature variations [120]. As the supply voltage is scaled to minimize the energy consumption, the supply voltage to threshold voltage ratio is reduced. The temperature fluctuation induced threshold voltage variations therefore determine the MOSFET drain current variations when the temperature fluctuates in circuits with extremely low power supply voltages. Contrary to the standard-higher-voltage circuits designed for high-speed, low-voltage circuits optimized for minimum energy operate faster when the die temperature increases [101], [121].

New temperature-adaptive dynamic supply and threshold voltage tuning techniques are proposed in Chapter 7 for reducing the active-mode energy consumption of ultra-low-voltage CMOS circuits at elevated temperatures. The high temperature energy efficiency is enhanced while maintaining a constant-clock-frequency by dynamically scaling the supply voltage or by dynamically increasing the device threshold voltages of a subthreshold logic circuit. The active mode energy savings provided by the two temperature-adaptive voltage tuning techniques are presented.
The performance of digital circuits can be enhanced by increasing the integrated on-chip memory capacity. Larger embedded memories improve the effective memory bandwidth by reducing the average memory access time [128]. Larger on-chip memories, however, increase the total power consumption of integrated circuits [107]. The large switching capacitance in the bit-lines and the word-lines of an SRAM array contribute to the high dynamic power consumption during a memory access. Furthermore, the leakage currents of large memory banks can dominate the total power consumption in ultra-low-voltage circuits [107]. To achieve higher reliability and longer battery life-time in portable applications, the power consumed by the memory banks should be reduced.

Similar to the circuits employed in the logic core, the read and write propagation delay of subthreshold memory banks are reduced at elevated temperatures. The effectiveness of the new temperature-adaptive voltage scaling schemes for improving the high-temperature energy efficiency of ultra-low-voltage memory banks is evaluated in Chapter 8. The influence of process parameter and environmental variations on the energy savings provided by the temperature-adaptive schemes is also investigated.

The research results presented in this dissertation are summarized in Chapter 9. Finally, several future research ideas for enhancing the reliability and the energy efficiency of nano-CMOS integrated circuits operating in environments subject to significant temperature fluctuations are described in Chapter 10.
Chapter 2
Sources of Power Consumption in Digital Circuits

In recent years, the focus of digital circuit design has shifted increasingly towards power reduction and further away from the traditional goal of higher clock frequency. The desirability of extended battery-life time in portable devices and the challenges of cooling in non-portable systems have generated significant interest in developing power reduction techniques at architecture, microarchitecture, circuit, and fabrication technology levels. There are four sources of the power consumption in CMOS circuits. The total power consumption of an integrated circuit is

\[ P_{\text{total}} = P_{\text{dynamic}} + P_{\text{leakage}} + P_{\text{short-circuit}} + P_{\text{DC}}, \]  

where \( P_{\text{dynamic}} \) is the dynamic switching power dissipated while charging or discharging the parasitic capacitances when the node voltages transition. \( P_{\text{leakage}} \) is the power dissipated by various leakage mechanisms. \( P_{\text{short-circuit}} \) is the transitory power dissipated during an input signal transition when both the pull-up and the pull-down networks of a CMOS gate are simultaneously active. \( P_{\text{DC}} \) is the static DC power consumed when a CMOS circuit is driven by low voltage swing input signals.

The four sources of power consumption in a CMOS integrated circuit are studied in this chapter. The dynamic switching power is discussed in Section 2.1. The different sources of leakage power consumption are identified in Section 2.2. The short circuit and the static DC power consumption are analyzed in Sections 2.3 and 2.4, respectively.

2.1. Dynamic Switching Power

The dominant component of the power consumption of a typical CMOS circuit is the switching power [1], [6], [10]-[12], [15], [22]-[25], [117]. The dynamic switching power is consumed during the charging and discharging of parasitic capacitances when the node voltages transition. The switching power is determined by the supply voltage, the switching frequency, the voltage swing, and the equivalent capacitance of the
switching node [1], [12], [15], [117]. Analytical expressions for the switching power consumption are derived next for low-to-high and high-to-low output node voltage transitions. In this section, the input waveform is assumed to have zero rise and fall times. The pull-up and pull-down networks are therefore never active simultaneously. CMOS equivalent circuits for the low-to-high and high-to-low output voltage transitions are shown in Figs. 2.1a and 2.1b, respectively. $C_L$ is the capacitive load at the output node. The drain-to-body junction capacitance, the interconnect capacitance, and the input capacitance of the load are lumped into one capacitive load, $C_L$. At time $t$, the current drawn from the power supply, the current entering the load capacitance, and the voltage of the output node are $I_{VDD}(t)$, $I_C(t)$, and $V_{OUT}(t)$, respectively.

![Fig. 2.1. A CMOS gate driving an output capacitor. (a) Equivalent representation for a low-to-high output node transition. (b) Equivalent representation for a high-to-low output node transition.](image)

The voltage of the output node is initially assumed to be at $V_{LOW}$ ($V_{OUT(0)} = V_{LOW}$). During the low-to-high transition of the output node ($V_{LOW} \rightarrow V_{HIGH}$), the current drawn from the power supply, $I_{VDD}$, passes through the pull-up transistors and charges the load capacitance. Since the pull-down network is switched off, the current drawn from
the power supply is entirely used for charging the load capacitor, as shown in Fig. 2.1a \([I_{VDD}(t) = I_C(t)]\). The current flowing into the load capacitor is

\[
I_{VDD}(t) = I_C(t) = C_L \frac{dV_{OUT}(t)}{dt}.
\] (2.2)

The instantaneous power consumed by a circuit element is the product of the voltage across the element and the current flowing into the element [30]. The energy consumed by the element is derived by integrating the instantaneous power over the time period of interest [1], [26], [117].

The instantaneous power delivered by the power supply \((P_{VDD})\) and the energy drawn from the power supply for an output voltage transition of \(V_{LOW} \rightarrow V_{HIGH}\) are

\[
P_{VDD}(t) = V_{DD}I_{VDD}(t),
\] (2.3)

\[
E_{V_{DD}} = \int_{t_1}^{t_2} P_{VDD}(t)dt = V_{DD} \int_{t_1}^{t_2} I_{VDD}(t)dt = C_L V_{DD} \int_{V_{LOW}}^{V_{HIGH}} dV_{OUT}(t),
\]

\[
= C_L V_{DD} (V_{HIGH} - V_{LOW}),
\] (2.4)

\[
V_{Swing} = V_{HIGH} - V_{LOW},
\] (2.5)

\[
E_{V_{DD}} = C_L V_{DD} V_{Swing},
\] (2.6)

where \(E_{V_{DD}}\) is the energy drawn from the power supply for charging the output node from \(V_{LOW}\) to \(V_{HIGH}\) while \(t_1\) and \(t_2\) are the times when the voltage of the load capacitor is at \(V_{LOW}\) and \(V_{HIGH}\), respectively. The energy stored in the capacitor from \(t_1\) to \(t_2\) is

\[
E_{CAP} = \int_{t_1}^{t_2} P_{CAP}(t)dt = \int_{t_1}^{t_2} I_C(t)V_{OUT}(t)dt = C_L \int_{V_{LOW}}^{V_{HIGH}} V_{OUT}(t)dV_{OUT}(t),
\]

\[
= \frac{1}{2} C_L (V_{HIGH}^2 - V_{LOW}^2),
\] (2.7)
where $E_{\text{CAP}}$ and $P_{\text{CAP}}$ are the energy stored in the capacitor and the instantaneous power consumption of the capacitor, respectively. The total energy dissipated in the pull-up network during the low-to-high output voltage transition is

$$E_{\text{charge}} = E_{V_{\text{DD}}} - E_{\text{CAP}} = C_L V_{DD} V_{\text{Swing}} - \frac{1}{2} C_L (V_{\text{HIGH}}^2 - V_{\text{LOW}}^2). \quad (2.8)$$

During the high-to-low voltage transition of the output node, the pull-up network is switched off and the pull-down network is activated, as shown in Fig. 2.1b. The capacitor current, $I_C(t)$, flows in the reverse direction to discharge the voltage stored in the output load. The energy dissipated in the pull-down network ($E_{\text{discharge}}$) during the output high-to-low transition is

$$E_{\text{discharge}} = \int_{t_1}^{t_2} P_{\text{discharge}}(t) \, dt = -\int_{t_1}^{t_2} I_C(t) V_{\text{OUT}}(t) \, dt = -C_L \int_{V_{\text{HIGH}}}^{V_{\text{LOW}}} V_{\text{OUT}}(t) \, dV_{\text{OUT}}(t),$$

$$= -\frac{1}{2} C_L (V_{\text{LOW}}^2 - V_{\text{HIGH}}^2) = \frac{1}{2} C_L (V_{\text{HIGH}}^2 - V_{\text{LOW}}^2), \quad (2.9)$$

where $P_{\text{discharge}}$ is the instantaneous power dissipated in the pull-down network while $t_1$ and $t_2$ are the times when the load capacitor voltage reaches $V_{\text{HIGH}}$ and $V_{\text{LOW}}$, respectively. The total energy consumption for one complete switching cycle (considering one low-to-high and one high-to-low output voltage transition) is

$$E_{\text{total}} = E_{\text{charge}} + E_{\text{discharge}},$$

$$= C_L V_{DD} V_{\text{Swing}} - \frac{1}{2} C_L (V_{\text{HIGH}}^2 - V_{\text{LOW}}^2) + \frac{1}{2} C_L (V_{\text{HIGH}}^2 - V_{\text{LOW}}^2),$$

$$= C_L V_{DD} V_{\text{Swing}}. \quad (2.10)$$

Power is the energy stored or dissipated per unit time [30]. Assuming that the output node transitions periodically between $V_{\text{LOW}}$ and $V_{\text{HIGH}}$ with a period of $T_s$, the
power consumption of a CMOS circuit driving an output load capacitance $C_L$ is [1], [10]-[12], [15], [30], [117]

$$P_{\text{dynamic}} = \frac{C_L V_{DD} V_{\text{Swing}}}{T_s} = C_L V_{DD} V_{\text{Swing}} f_s.$$  \hspace{1cm} (2.11)

In synchronous systems, all the internal nodes of a CMOS circuit do not switch every clock cycle (excluding the clock buffers) [1], [117]. If $\alpha_0 \rightarrow 1$ is the average number of times that a node with capacitance $C_L$ makes a power consuming transition ($V_{\text{LOW}} \rightarrow V_{\text{HIGH}}$) in each clock cycle, then the resulting average dynamic power consumption is

$$P_{\text{dynamic}} = \alpha_0 C_L V_{DD} V_{\text{Swing}} f_s.$$ \hspace{1cm} (2.12)

The activity factor of a circuit, $\alpha_0$, can be evaluated taking into account the logic function, signal statistics, logic style (static circuits versus dynamic circuits), and the circuit topology [12]. In a typical full voltage swing CMOS circuit where the output voltage transitions from $0V \rightarrow V_{DD} \ (V_{\text{Swing}} = V_{DD})$, the average dynamic power consumption per switching period is

$$P_{\text{dynamic}} = \alpha_0 C_L V_{DD}^2 f_s.$$ \hspace{1cm} (2.13)

### 2.1.1. Power Consumption Due to Glitches

When measuring the dynamic power consumption, considering only the final transition at the output nodes of a circuit may not be adequate. The timing behavior of the circuit has to be considered for accurately measuring the total switching power. Timing skew between the signals can cause spurious transitions, called glitches, resulting in extra power consumption. In this section, the additional power consumed due to the presence of glitches in an integrated circuit is illustrated with an example.

The block diagram of a 16-bit ripple carry adder is shown in Fig. 2.2. The actual waveforms illustrating the glitching behavior in the 16-bit ripple carry adder are shown in
Fig. 2.3. With the input excitation scenario illustrated in Fig. 2.2, all the bits of one of the input vectors and the carry-in ($C_{in}$) from the least significant bit position transition from low-to-high while all the bits of the second input vector are held at $V_{LOW}$ (0V). The sum corresponding to these input vectors is zero with the carry output asserted. The propagation delay of the carry signal from the least significant bit position to the most significant bit position however causes a logic high voltage ($V_{HIGH}$) to temporarily appear at most of the outputs, as shown in Fig. 2.3. These spurious transitions cause additional switching power consumption. Note that most of the sum outputs experience full-voltage swing glitching in a ripple carry adder, as shown in Fig. 2.3.

Fig. 2.2. Block diagram of a 16-bit ripple carry adder.

Fig. 2.3. Waveforms indicating the glitches at the outputs of a 16-bit ripple carry adder. The numbers indicate the sum output voltage at the corresponding bit position. Sum bits 0 and 1 experience partial glitching. The glitching voltage rises all the way up to $V_{HIGH}$ at bit positions SUM[2] to SUM[15].
Glitches are a function of the input patterns, internal state assignments in the logic circuit, skew among different input signal delay paths, and the logic depth [12]. Glitches can be eliminated through careful logic design with balanced delay paths. Glitches are observed only in static CMOS circuits with skewed inputs. The dynamic CMOS circuit family is intrinsically immune to glitches and the additional switching power consumption associated with the spurious output events [12].

2.2. Leakage Currents in Nano-CMOS Technologies

A MOSFET is a resistive-capacitive switch. The MOS device is switched off when the gate-to-source voltage is less than the threshold voltage of the device. However, due to the non-ideal off-state characteristics of a transistor, current is drawn from the power supply even when a transistor operates in the cut-off region (gate-to-source voltage is less than the threshold voltage of the device). Subthreshold leakage and junction leakage are the dominant leakage mechanisms in long channel devices [1], [6], [10], [12], [45], [46], [117]. In nano-CMOS technologies, however, other leakage mechanisms such as gate-oxide tunneling also play a significant role in the total leakage current produced by a device [6], [45]-[49].

The primary leakage mechanisms in the nano-CMOS devices are reviewed in this section. The sources of subthreshold leakage current are discussed in Section 2.2.1. The mechanisms for junction leakage current are described in Section 2.2.2. The gate oxide tunneling current is characterized in Section 2.2.3.

2.2.1. Subthreshold Leakage Current

A MOSFET with a gate-to-source voltage less than the magnitude of the threshold voltage operates in the weak inversion (subthreshold) region [50]. The current produced in the weak inversion region (subthreshold current) is primarily due to the diffusion of the charge carriers [6], [10], [25], [46], [50]. The subthreshold current is influenced by the threshold voltage, the channel length, the channel width, the depletion width beneath the channel area, the channel/surface doping profiles, the drain/source junction depths, the gate-oxide thickness, the supply voltage, and the temperature [45]. The threshold voltage
has an exponential dependence on the subthreshold leakage current [44]. The various mechanisms that alter the threshold voltage, thereby effecting the subthreshold leakage current produced by a MOSFET, are described in this section.

Device dimensions scale with each new technology generation [19]. Scaling the channel length degrades the capability of the gate terminal to control the charge distribution in the channel area [1], [117]. The device threshold voltage, therefore, degrades with each new technology generation [45], [46]. The modulation of threshold voltage with the channel length is called short-channel effects. The short-channel effects are described in Section 2.2.1.1. While the gate loses control of the channel region, the influence of the drain voltage to alter the charge distribution of the channel is enhanced with technology scaling. The impact of the drain voltage on the threshold voltage of short-channel devices is discussed in Section 2.2.1.2. The subthreshold leakage current is characterized with analytical expressions in Section 2.2.1.3.

2.2.1.1. Short Channel Effects

In older technology generations with long channel devices, the source and drain depletion regions occupy only a small fraction of the channel area. The effects of the source and drain depletion region extensions on the threshold voltage are, therefore, negligible in a long channel device [6], [46]. As the channel length is reduced to enhance the device performance and the number of components per chip, the effective channel length becomes comparable to the total depth of the source and drain depletion regions. More charge is contributed to the depletion region beneath the gate by the source-to-substrate/well and the drain-to-substrate/well depletion layers in a short channel device as compared to a long channel device [46], [50], [51]. A lesser gate voltage is therefore sufficient to form the channel inversion layer beneath the gate area in a short channel device. The modification of the threshold voltage and the resulting fluctuations in the device electrical characteristics with the channel length are called short channel effects [1], [117].

The variation of the threshold voltage with the channel length for the devices in a 65nm predictive CMOS technology (PTM) [52] is shown in Fig. 2.4. Due to the short channel effects, the threshold voltages of the MOSFETs vary by up to 4.6x when the
channel length is scaled, as shown in Fig. 2.4. The presence of die-to-die and intra-die process variations enhances the impact of short-channel effects in nano-CMOS devices.

Fig. 2.4. Variation of the threshold voltage with channel length for devices in a 65nm CMOS technology. \( N_{\text{width}} = 600\text{nm} \). \( P_{\text{width}} = 1200\text{nm} \). Temperature = 25°C. The threshold voltage is measured as the gate-to-source voltage required for producing \( 1\times10^{-4}\text{A} \) drain current.

2.2.1.2. Drain Induced Barrier Lowering (DIBL)

In long channel MOSFETs, the channel depletion region is solely due to the applied gate voltage. Approximately all of the depletion charge beneath the gate originates from the field effect of the gate terminal. However, in short channel devices, the applied drain voltage significantly influences the channel formation. Raising the absolute value of the drain-to-bulk reverse bias voltage (\( |V_{DB}| \)) increases the drain-junction depletion width. The increase in the depletion width reduces the effective channel length and lowers the energy barrier height [1], [117]. Consequently, the threshold voltage decreases with increasing drain-to-body voltage in a MOSFET. This effect is called the drain induced barrier lowering (DIBL). For a sufficiently high reverse bias voltage across the drain junction, the source and drain regions can even be shorted,
thereby eliminating the normal transistor operation. The sharp increase in the drain current when the source and the drain depletion regions are shorted in a MOSFET is called punch-through [26].

The variation of the threshold voltage with the drain-to-body voltage for the devices in a 65nm CMOS technology [52] is shown in Fig. 2.5. Due to DIBL, the threshold voltages of the transistors vary by up to 24.5% with the drain voltage, as shown in Fig. 2.5.

![Fig. 2.5. Variation of the threshold voltage with the drain voltage for devices in a 65nm CMOS technology. N\text{\_width} = 600\text{nm}. P\text{\_width} = 1200\text{nm}. Temperature = 25^\circ\text{C}. The threshold voltage is measured as the gate-to-source voltage required for producing 1x10^{-4}\text{A} drain current.](image)

The variations in the die temperature and the substrate/well bias voltage also alter the threshold voltage of a MOSFET. The impact of temperature fluctuations on the device threshold voltages and the CMOS circuit performance are described in Chapter 3.
2.2.1.3. Characterization of Subthreshold Leakage Current

The subthreshold leakage current produced by a MOSFET is [1], [46], [117]

\[
I_{\text{subthreshold}} = \frac{\mu WC_{OX}}{L} V_T^2 e^{\frac{|V_{GS}| - |V_T|}{\eta V_T}} \left( 1 - e^{-\frac{|V_{DS}|}{V_T}} \right),
\]

(2.14)

\[
V_T = \frac{kT}{q},
\]

(2.15)

where \(I_{\text{subthreshold}}\), \(\mu\), \(W\), \(C_{OX}\), \(L\), \(V_T\), \(V_{GS}\), \(V_T\), \(\eta\), \(V_{DS}\), \(k\), \(T\), and \(q\) are the subthreshold current, carrier mobility, transistor width, oxide capacitance per unit area, channel length, thermal voltage, gate-to-source voltage, device threshold voltage, subthreshold swing coefficient, drain-to-source voltage, Boltzmann constant \((1.38 \times 10^{-23} \text{ J/K})\), absolute temperature in Kelvin, and the unit charge \((1.6 \times 10^{-19} \text{C})\), respectively. As given by (2.14), the subthreshold current increases exponentially with the reduction of the threshold voltage [1], [12], [26], [42], [51], [117]. The subthreshold swing coefficient \(\eta\) for a bulk MOSFET is [1], [117]

\[
\eta \equiv 1 + \frac{\varepsilon_{Si} t_{OX}}{\varepsilon_{OX} t_{Si}},
\]

(2.16)

where \(\varepsilon_{Si}\), \(t_{OX}\), \(\varepsilon_{OX}\), and \(t_{Si}\) are the permittivity of silicon, physical thickness of the gate oxide, permittivity of gate-oxide, and the thickness of the depletion layer of the substrate, respectively.

Subthreshold conduction is the most significant source of device leakage in the current CMOS technologies [46]. Subthreshold slope \((S)\) is a widely used parameter to characterize the weak inversion current produced by a MOSFET [1], [6], [44], [117]. The subthreshold slope is the variation in the gate-to-source voltage \((V_{GS})\) that is required to alter the subthreshold leakage current by one decade. The weak inversion current of an n-channel device in a 65nm CMOS technology is shown in Fig. 2.6. The subthreshold slope
of the device can be evaluated by choosing two points in the subthreshold region of the $I_D-V_{GS}$ curve such that the leakage current changes by a factor of 10. The subthreshold slope of this n-channel device is 80mV/decade, as shown in Fig. 2.6. Using equation 2.14,

$$\frac{I_{\text{subthreshold}} @ V_{GS1}}{I_{\text{subthreshold}} @ V_{GS2}} = e^{\frac{V_{GS1} - V_{GS2}}{nV_T}} = 10$$

(2.17)

$$S = |V_{GS1} - V_{GS2}| = \eta V_T \ln 10,$$

(2.18)

where $V_{GS1}$ and $V_{GS2}$ are the two gate-to-source voltages between which the weak inversion current varies by one decade. The subthreshold slopes of MOSFETs in typical CMOS technologies are in the range of 80mV/decade to 120mV/decade [1], [117]. The subthreshold slope for an ideal MOSFET (with zero depletion capacitance) is 60mV/decade at the room temperature.

![Graph showing drain current versus gate-to-source voltage curve for an NMOSFET in a 65nm CMOS technology. Nwidth = 600nm. Drain-to-source voltage = 1V. Temperature = 25°C.](image)

Fig. 2.6. Drain current versus gate-to-source voltage curve for an NMOSFET in a 65nm CMOS technology. Nwidth = 600nm. Drain-to-source voltage = 1V. Temperature = 25°C.
2.2.2. Junction Leakage Current

The causes for the junction leakage current are reviewed in this section. A MOSFET is composed of a number of diodes that directly influence the behavior of the device. The different p-n junction diodes in a CMOS circuit are shown in Fig. 2.7. The substrate and the well of a circuit fabricated in an n-well CMOS process are typically tied to $V_{GND}$ and $V_{DD}$, respectively, to ensure that the body diodes are never forward-biased, as shown in Fig. 2.7. The reverse-bias current produced by these parasitic diodes contribute to the leakage current. The junction leakage current is

$$\begin{align*}
I_{\text{Junction}} &= I_S \left( \frac{V_D}{V_T} - 1 \right),
\end{align*}$$

(2.19)

where $I_S$ depends on the doping levels and the area perimeter of the diffusion region. $V_D$ is the diode voltage (e.g. the source-to-body voltage or the drain-to-body voltage of a MOSFET). Junction leakage current is a strong function of the source-to-body and drain-to-body bias voltages. Even when the junction is reverse biased by a source significantly higher than the thermal voltage, the junction leakage is typically low (in the order of 0.1-0.01fA/µm²) [44]. Excessive reverse-body bias with unusually high voltages however may cause significant junction leakage current due to enhanced band-to-band tunneling [53].

Junction leakage has historically been a primary limiter of data storage time on dynamic circuit nodes. However, in modern transistors with low threshold voltages and thin gate insulators, the weak inversion conduction and the gate tunneling leakage current typically far exceed the junction leakage current [1], [117].
2.2.3. Gate Tunneling Leakage Current

Technology scaling has been driven by Moore’s law [19], [20]. Scaling the gate insulator thickness is crucial for enhancing the performance of MOSFETs in each new technology generation [1], [31]-[34], [117]. The impact of thinner gate insulators on the power consumption of CMOS circuits is discussed in this section.

Reducing the oxide thickness increases the gate capacitance, thereby enhancing the drain current produced by a MOSFET. Furthermore, a thinner gate-oxide layer suppresses the short-channel effects [1], [117]. Future scaling trends of the gate oxide thickness \( t_{ox} \) and some other important device parameters extracted from the projections of the International Technology Roadmap for Semiconductors, are listed in Table 2.1 [33], [35].

According to quantum mechanics, the probability for tunneling of charge carriers increases exponentially with the scaling of oxide thickness [1], [33], [36]-[40], [117]. This results in tunneling leakage current flowing through the gate terminal of a MOSFET. The \( t_{ox} \) in current CMOS technologies range from 8Å to 12Å, as listed in Table 2.1. Such a thin \( t_{ox} \) leads to a significant tunneling current between the gate and the other device terminals [1], [117].

The tunneling current is composed of several components as illustrated in Fig. 2.8 [1], [36], [117]. \( I_{gb} \) is the gate-to-substrate leakage current produced by the electron tunneling from the valence band in both NMOS and PMOS devices [41]. \( I_{gs0} \) and \( I_{gd0} \) are the leakage currents through the gate-to-source and gate-to-drain overlap regions, respectively. \( I_{gc} \) is the gate-to-channel tunneling current during operation in the inversion
region. A portion of the $I_{gc}$ is collected by the source ($I_{gs}$) while the remaining portion is collected by the drain ($I_{gcd}$), as shown in Fig. 2.8.

### TABLE 2.1.
**SEMICONDUCTOR DEVICE SCALING TRENDS [33]**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum feature size (nm)</td>
<td>180</td>
<td>130</td>
<td>100</td>
<td>70</td>
<td>50</td>
<td>35</td>
</tr>
<tr>
<td>Gate length (nm)</td>
<td>100</td>
<td>70</td>
<td>50</td>
<td>35</td>
<td>24</td>
<td>18</td>
</tr>
<tr>
<td>DRAM (bits/chip)</td>
<td>1G</td>
<td>3G</td>
<td>8G</td>
<td>24G</td>
<td>64G</td>
<td>192G</td>
</tr>
<tr>
<td>Physical gate oxide thickness ($t_{ox}$) (nm)</td>
<td>1.9-2.5</td>
<td>1.5-1.9</td>
<td>1.0-1.5</td>
<td>0.8-1.2</td>
<td>0.6-0.8</td>
<td>0.5-0.6</td>
</tr>
<tr>
<td>Power supply (V)</td>
<td>1.5-1.8</td>
<td>1.2-1.5</td>
<td>0.9-1.2</td>
<td>0.6-0.9</td>
<td>0.5-0.6</td>
<td>0.5</td>
</tr>
<tr>
<td>Dielectric constant of DRAM capacitor</td>
<td>22</td>
<td>50</td>
<td>250</td>
<td>700</td>
<td>1500</td>
<td>1500</td>
</tr>
<tr>
<td>DRAM chip size (mm²)</td>
<td>400</td>
<td>460</td>
<td>530</td>
<td>630</td>
<td>710</td>
<td>860</td>
</tr>
</tbody>
</table>

![Fig. 2.8](image.png)

Fig. 2.8. Different components of gate dielectric tunneling current in a MOSFET.
Gate tunneling current in a MOSFET depends on the voltage across the gate dielectric, the thickness of the dielectric, the tunneling barrier height, the effective mass of the carriers, and the number of free carriers available for tunneling on the MOS electrodes [33], [41]. The different mechanisms of gate dielectric tunneling in MOSFETs are illustrated in Fig. 2.9. In a technology with SiO$_2$ as the gate dielectric material, the tunneling current in a PMOS device is primarily due to the hole tunneling from the valence band (HVB) in the silicon substrate, as shown in Fig 2.9a. The electron tunneling from the polysilicon conduction band is negligible since the electron concentration in the p$^+$ polysilicon gate is low in a PMOSFET [1], [36], [42], [117]. Alternatively, the electron tunneling from the conduction band (ECB) of the silicon substrate is the dominant tunneling mechanism in an NMOS device, as shown in Fig. 2.9b. The HVB of the polysilicon is negligible due to the smaller number of holes as compared to the electrons in an NMOS device. The energy barrier heights for the HVB and ECB are 4.5eV and 3.1eV, respectively, as shown in Fig. 2.9 [40]. The probability of hole tunneling is, therefore, much smaller than the probability of electron tunneling through the gate oxide. The $I_{\text{gate}}$ for a PMOS device is significantly lower as compared to an NMOS device with similar physical dimensions (width, length, and tox) and similar voltage difference across the gate insulator [1], [40], [43], [117].

Silicon dioxide has been the material of choice as the gate insulator in many CMOS technologies [1], [34], [40], [43], [117]. Silicon dioxide is relatively easy to grow on silicon with near ideal electrical characteristics [1], [117]. However, the dielectric thicknesses projected in Table 2.1 will soon be unrealizable if SiO$_2$ is maintained as the dielectric material [1], [117]. As discussed in [43], gate tunneling leakage current is expected to dominate the total leakage current produced by future CMOS circuits. New materials with higher dielectric constants (high-K) as compared to SiO$_2$ are required in order to keep the gate tunneling leakage current under control with reverse gate insulator scaling (thicker tox) in the future technology generations. Several new materials such as aluminum oxide (Al$_2$O$_3$), hafnium dioxide (HfO$_2$), tantalum pentoxide (Ta$_2$O$_5$) are being explored as a suitable replacement for SiO$_2$ [16].
2.3. Short-Circuit Power

Short-circuit power is consumed by a CMOS circuit during the time periods when the input signals transition. The assumption of zero rise and fall times for an input waveform is invalid in real designs [26]. The finite slopes of the input signals cause a direct current path between the power supply and the ground terminal for a short period of time when the input transitions. The pull-up and the pull-down networks of a CMOS gate are simultaneously active during this time period. The short-circuit power is observed when the input voltage is between $V_{tn}$ and $(V_{DD} - |V_{tp}|)$ in a CMOS inverter, as illustrated in Fig. 2.10 $[V_{tn} \leq V_{in} \leq (V_{DD} - |V_{tp}|)]$.

Approximating the short-circuit current spikes produced by a CMOS circuit as triangles, the energy consumed per switching period is [1], [117]

$$E_{\text{short-circuit}} = V_{DD} \frac{I_{\text{peak}}t_{sc}}{2} + V_{DD} \frac{I_{\text{peak}}t_{sc}}{2} = t_{sc}V_{DD}I_{\text{peak}},$$

(2.20)

where $E_{\text{short-circuit}}$, $t_{sc}$, $I_{\text{peak}}$, and $V_{DD}$ are the short-circuit energy consumption, the time during which the NMOS and the PMOS devices are simultaneously active, the peak
The average short-circuit power \( P_{\text{short-circuit}} \) consumed per switching period \( t_{\text{period}} \) is [26]

\[
P_{\text{short-circuit}} = \frac{t_{\text{sc}} V_{\text{DD}} I_{\text{peak}}}{t_{\text{period}}},
\]

Fig. 2.10. Short-circuit currents during input signal transients. \( V_{tn} \) is the threshold voltage of the NMOS device. \( V_{tp} \) is the threshold voltage of the PMOS device.

The peak current \( I_{\text{peak}} \) produced during the short circuit period is dependent on the rise and fall times of the input and the output waveforms and the output load capacitance. The impact of the load capacitance on the short-circuit current is explained next. CMOS inverters with different output loads are shown in Fig. 2.11. Consider a low-to-high input voltage transition. In the circuit with a large output capacitance, the input transition is complete before the output transition effectively begins, as shown in Fig. 2.11a. Since the source-to-drain voltage of the PMOS device is approximately zero during the input transition, the PMOS transistor shuts off before getting a chance to deliver any short-circuit current. The short circuit current in this case is negligible. Alternatively, in a circuit with a small output load, the output fall time can be substantially smaller as compared to the input rise time, as shown in Fig. 2.11b. The drain-to-source voltage of the active PMOS device is \( V_{\text{DD}} \) for most of the transition period, producing maximum short-circuit current. The short-circuit current can be
significant when the rise and fall times of the input signals are significantly larger than the output rise and fall times since the short-circuit current path exists for a longer period of time [1], [12], [26], [27], [117].

As technology scales, the ratio of the power-supply-voltage to the device threshold voltages is reduced [28]. The contribution of the short-circuit current to the total power consumed by a CMOS circuit is therefore reduced with technology scaling, as discussed in [29]. Furthermore, the short circuit power can be completely eliminated by operating an integrated circuit at supply voltages lower than the sum of the threshold voltages of PMOS and NMOS transistors ($V_{DD} < V_{th} + |V_{tp}|$) [1], [117].

![Fig. 2.11. Impact of the output capacitance on the short-circuit current. (a) CMOS inverter with a large capacitive load and a sharp input signal. (b) CMOS inverter with a small capacitive load and a slowly rising input signal.](image)

### 2.4. Static DC Power

Static DC power is consumed when there is a direct path between the power supply ($V_{DD}$) and the ground ($V_{GND}$) through the simultaneously activated pull-up and pull-down networks of a CMOS gate at steady state. Static power (excluding the leakage currents) is negligible in integrated circuits where all the internal nodes have a full-rail voltage swing ($V_{GND} \rightarrow V_{DD}$). However, in CMOS circuits with reduced voltage swings ($V_{OL} > V_{GND}$ and/or $V_{OH} < V_{DD}$), the static DC power can have a significant contribution to the overall power consumption of the circuit [1], [117].
Non-full rail voltages are encountered in CMOS circuits due to the usage of NMOS and/or PMOS pass-transistors. A CMOS inverter driven by a low swing signal is shown in Fig. 2.12. When the low swing signal drives a CMOS gate connected to $V_{DD}$ and $V_{GND}$, static DC current is produced by the receiver, as shown in Fig. 2.12. Furthermore, the low voltage signaling techniques used in CMOS circuits to reduce the dynamic switching power consumption can also produce static DC currents in the receiver gates [1], [44], [117], [118]. Some low voltage signaling techniques currently employed in integrated circuits are discussed in [12].

Fig. 2.12. Static CMOS inverter driven by a low swing signal. $I_{STATIC}$ is the static DC current produce by the inverter. $V_{tn}$ is the threshold voltage of the NMOS device. $V_{tp}$ is the threshold voltage of the PMOS device.
Chapter 3

Die Temperature Variations

Process and environment parameter variations pose greater challenges in the design of high performance integrated circuits in scaled CMOS technologies [79], [80]. Variations can be categorized into die-to-die variations and within-die variations. Die-to-die fluctuations affect every element in an integrated circuit similarly. Alternatively, within-die variations cause a non-uniformity of physical characteristics among the devices in an integrated circuit. The accuracy of estimating the variations relates to the manufacturing cost of an integrated circuit. An overestimation of variations increases the design effort, thereby delaying the time-to-market. Alternatively, an underestimation of variations compromises the performance and functionality, thereby degrading yield [79], [80].

Process related variations in the channel length, gate-oxide thickness, and doping concentration of MOSFETs cause fluctuations in the speed and power consumption characteristics of CMOS circuits [26]. Furthermore, variations in the operating environment, such as the variation of the supply voltage and the die temperature, dynamically influence the behavior of the MOSFETs. In this chapter, the different sources of temperature variations are identified. The impact of temperature fluctuations on the speed and power consumption characteristics of CMOS integrated circuits is investigated.

The heat generated due to the activity of devices and interconnect flows through the substrate and the package. The ambient temperature on an integrated circuit die is raised due to the circuit operation [44]. Advancements in heat sinks, air flow fans, and packaging technologies have raised the practical limit for heat removal from approximately 8W in 1985 to more than a 100W in the current integrated systems without a significant increase in the system cost [44].

Not all internal nodes of an integrated circuit switch every clock cycle [1], [10], [117]. The circuit activity factor directly influences the power consumption and the heat generation, as explained in Chapter 2. In high performance microprocessors, the
imbalanced utilization and the diversity of circuitry at different sections of an integrated circuit cause temperature variations from one die area to another [1], [117]. The simulated temperature distribution of an Intel Itanium processor is shown in Fig. 3.1 [81]. There is a temperature gradient across the chip from the heavily utilized hot integer core near the center to the cooler level-2 caches characterized with a lower activity factor around the periphery, as shown in Fig. 3.1. Heat generation due to localized higher activity circuitry produces hot-spots, as shown in Fig. 3.1. The non-uniform temperature profile leads to fluctuations in the delay and power consumption characteristics of the circuits across a chip.

Fig. 3.1. Die temperature variation of an Intel Itanium processor in a 180nm CMOS technology [81]. The heavily utilized core area of the die is at 120°C while the caches are at 70°C, thereby indicating a 50°C temperature gradient within the same die.

In ultra-low-voltage circuits, temperature gradients due to imbalanced switching activity within a die are typically small. The primary source of temperature fluctuations in low-voltage circuits are the variations in the ambient temperature. For example, electronic systems mounted on automobile engines operate at a temperature range from -
40°C to 150°C [82]. Similarly, the ambient temperatures for integrated circuits employed in robotic explorations vary from -180°C to 486°C [64]. Changes in the ambient temperature tend to affect all of the devices in an IC in a similar way.

Fluctuations in the die temperature alter the number of charge carriers in a semiconductor device [68]. The physical mechanisms that cause a variation in the carrier concentration of semiconductors with temperature are reviewed in this chapter. The MOSFET device parameters that are altered when the die temperature fluctuates are identified. The temperature effects on the device parameters are characterized using BSIM [42], [72], [73]. The temperature fluctuation induced variations in the device and circuit characteristics are examined for 180nm and 65nm CMOS technologies [52], [71].

The chapter is organized as follows. The impact of temperature variations on the carrier concentration of doped semiconductors is described in Section 3.1. The device parameters that are influenced by temperature fluctuations are identified in Section 3.2. The industry-standard models to capture the temperature related variations in nano-CMOS devices are presented in Section 3.3. The effects of temperature variations on the device and circuit characteristics are examined in Section 3.4 for two different CMOS technologies (TSMC 180nm and PTM 65nm). A summary of the effects of temperature fluctuations on CMOS integrated circuits is provided in Section 3.5.

3.1. Impact of Temperature Variations on Carrier Concentration

Carriers are the entities that transport charge inside a material, thereby producing electrical currents. In a semiconductor, the charge carriers are electrons and holes. The impact of temperature variations on the charge carrier concentration is studied in this section. The semiconductor bonding model leading to the derivation of the analytical expressions for carrier concentration is briefly discussed.

An isolated silicon (Si) atom (a Si atom not interacting with other atoms) has four valence electrons. Each valence electron is shared (bonded) with four neighboring silicon atoms to form a crystalline state. The resulting covalent bonding in crystalline silicon is represented in Fig. 3.2.
Isolated silicon atoms are brought into close proximity during the formation of the crystalline silicon structure, as illustrated in Fig. 3.2. As the silicon atoms get closer, the energy states of the valence electrons are modified, as discussed in [68]. The proximity of the electrons results in a progressive spread in the allowed energies, thereby giving rise to closely spaced sets of allowed states called the energy bands [68]-[70]. At the inter-atomic distance corresponding to the Si lattice spacing, the distribution of the allowed energies consists of two bands of allowed states separated by an energy gap. A representation of the energy band diagram is shown in Fig. 3.3. The upper band of allowed states is the conduction band ($E_C$). Alternatively, the lower band of allowed states is the valence band ($E_V$), as shown in Fig. 3.3. The energy gap separating the upper and the lower energy bands is called the forbidden gap or the band gap. The band gap of silicon at 300K (room temperature) is 1.12eV [69].
When a semiconductor is at 0K, the valence band is completely filled with electrons while the conduction band is devoid of electrons. The lack of free charge carriers in the semiconductor at 0K eliminates the possibility of current flow. A minimum amount of energy, equivalent to the forbidden energy gap of the semiconductor, is required to excite the electrons from the valence band into the conduction band. As the ambient temperature is increased, thermal energy is supplied to the carriers. The thermal energy can provide an electron with the necessary energy to transition into the conduction band. The excitation of an electron with sufficient energy generates a hole in the valence band along with an electron in the conduction band (electron-hole pair). In a semiconductor with electron-hole pairs

\[ n = p = n_i, \]  

(3.1)

where \( n \) is the number of electrons/cm\(^3\), \( p \) is the number of holes/cm\(^3\), and \( n_i \) is the intrinsic semiconductor concentration. A semiconductor that has equal quantities of oppositely charged carriers (electrons and holes) is classified as an intrinsic semiconductor \[68\]. The current conducting capability of intrinsic semiconductors is inherently low \[69\], \[70\]. The current flow in a semiconductor however can be enhanced by increasing the concentration of either the electrons or the holes through a process called doping \[68\]-\[70\].

Doping, in semiconductor terminology, is the addition of controlled amounts of specific impurity atoms to increase either the electron or the hole concentration. Doped semiconductors are referred to as extrinsic semiconductors. To increase the electron concentration, silicon is doped with donor elements from column V of the periodic table (donors such as phosphorous, arsenic, and antimony). Similarly, to increase the hole concentration, acceptor elements from column III of the periodic table (acceptors such as boron, gallium, indium, and aluminum) are doped into the silicon atoms \[68\]. The added foreign donor (acceptor) atoms have one more (less) valence electron for perfect bonding with the silicon. Doping with donor elements thus creates an electron rich semiconductor. Alternatively, doping with an acceptor element creates a semiconductor abundant with
holes [70]. Semiconductors with very high doping concentrations are said to be degenerate.

The impact of semiconductor doping on the energy band diagram of a semiconductor is illustrated next. The band diagrams of doped semiconductors are shown in Fig. 3.4. In an extrinsic n-type semiconductor, the “extra” electron of a donor atom corresponds to an energy level $E_d$ slightly below the conduction band, as shown in Fig. 3.4a. Similarly, in a p-type extrinsic semiconductor, the vacancy in an acceptor atom corresponds to an energy level of $E_a$ slightly above the valence band, as shown in Fig. 3.4b. The relative closeness of $E_d$ ($E_a$) to $E_C$ ($E_V$) reduces the energy required for the excitation of electrons [68]-[70].

![Diagram of Energy Bands](image)

**Fig. 3.4.** Band diagrams indicating the energy levels corresponding to the impurity atoms. (a) n-type extrinsic semiconductor [$E_d$]. (b) p-type extrinsic semiconductor [$E_a$].

The reduction in the band gap energy shifts the Fermi energy level ($E_F$) of the semiconductor. Fermi Energy, $E_F$, refers to the energy of the highest occupied quantum state in a semiconductor at absolute zero temperature. The position of the Fermi energy in
the band diagram determines the relative magnitude of the carrier concentrations in the semiconductor. When $E_F$ is positioned in the upper half of the band gap (or higher), the electron concentration exceeds the hole concentration [69]. Alternatively, the predominance of holes results when the $E_F$ lies below the middle of the band gap [69]. According to Maxwell-Boltzmann statistics, the electron and hole concentrations at equilibrium are [68]-[70]

$$n = n_i e^{\frac{(E_F - E_i)}{kT}},$$  \hspace{2cm} (3.2)

$$p = n_i e^{\frac{(E_i - E_F)}{kT}},$$  \hspace{2cm} (3.3)

where $n$, $p$, $n_i$, $E_i$, $E_F$, $k$, and $T$ are the concentration of the electrons, the concentration of the holes, the intrinsic carrier concentration, the intrinsic energy level, the Fermi energy, the Boltzmann constant ($k = 8.617 \times 10^{-5}$ eV/K), and the absolute temperature in Kelvin. For intrinsic semiconductors, $E_i = E_F$. For an n-type extrinsic semiconductor, $E_i < E_F$. Alternatively, in a p-type extrinsic semiconductor, $E_i > E_F$.

As the temperature is increased, the number of electrons that can transition through the forbidden energy gap into the upper energy levels is also increased due to the enhanced thermal energy. Each electron excited from the conduction band leaves behind a hole in the valence band. The intrinsic carrier concentration ($n_i$), therefore, increases with temperature, as explained in [68]. Furthermore, the Fermi level is also temperature dependent [68]. An increase in temperature shifts the Fermi energy ($E_F$) further up (further down) in a donor (acceptor) doped semiconductor thereby increasing $|E_F - E_i|$ at elevated temperatures [68]-[70]. The increase in the intrinsic carrier concentration ($n_i$) and $|E_F - E_i|$ with temperature tends to increase the concentration of the charge carriers (electrons and holes), as given by equations 3.2 and 3.3.

The variation of carrier concentration with temperature is illustrated in Fig. 3.5 with a phosphorous doped Si sample ($N_D = 10^{15}/\text{cm}^3$) [68]. As shown in Fig. 3.5, the concentration of charge carriers ($n$) is fixed at approximately $N_D$ over a broad temperature range extending from approximately 150K to 450K for the given Si sample. In this
temperature range (150K to 450K, referred as the “extrinsic temperature region”), the
temperature fluctuation induced variations in the intrinsic carrier concentration, Fermi
energy, and the absolute temperature counterbalance each other for this semiconductor
sample. The concentration of the charge carriers is therefore less sensitive to the
temperature fluctuations, as shown in Fig. 3.5. Below 100K, however, $n$ drops
significantly below $N_D$ and approaches zero as $T \to 0\text{K}$ (the “freeze-out zone”).
Alternatively, in the “intrinsic temperature region” at the opposite end of the temperature
spectrum, $n$ rises above $N_D$, due to the significant increase in $n_i$ (intrinsic concentration)
well beyond $N_D$ at very high temperatures.

![Fig. 3.5. Typical temperature dependence of the majority-carrier concentration in a doped semiconductor. A phosphorous-doped Si sample ($N_D = 10^{15}/\text{cm}^3$).](image)

The temperature dependence of carrier concentration is explained next with a
donor-doped material. The source of the majority-carriers in a donor-doped material at
various temperatures is qualitatively explained using Fig. 3.6. The majority carriers in a
donor-doped material are the electrons donated by the donor atoms and the valence elec-
trons excited across the band-gap into the conduction band. At $T \to 0\text{K}$, the thermal en-
ergy available in the system is insufficient to release the weakly bound fifth electrons on the donor sites or to excite the electrons across the band-gap, as shown in Fig. 3.6a [68]. The \( n/N_D \) is therefore zero at \( T = 0K \), as shown in Fig. 3.5 [68]. Slightly increasing the temperature above 0K frees some of the electrons that are weakly bound to the donor sites, as illustrated in Fig. 3.6b. Band-to-band excitation, however, remains low. The release of the weakly bound donor electron into the conduction band increases the Fermi level \( (E_F) \) thereby enhancing the concentration of the charge carriers (electrons), as given by equation 3.2. The number of observed electrons in the freeze-out temperature region of Fig. 3.5 is approximately equal to the number of ionized donors [68].

Continuing to increase the temperature eventually frees almost all of the weakly bound electrons on the donor sites and starts to excite electrons across the band-gap. This corresponds to the “extrinsic temperature region” where \( n \) approaches \( N_D \). In progressing through the extrinsic temperature region, an increasing number of electrons are excited across the band gap. The number of electrons supplied in this fashion however remains well below \( N_D \). The temperature fluctuation induced variations in the intrinsic carrier concentration, Fermi energy, and the absolute temperature effectively counterbalance each other in this temperature range. The concentration of the charge carriers (electrons) therefore remains approximately equal to \( N_D \) in the extrinsic temperature region, as shown in Fig. 3.5.

With further increases in the temperature, the carrier concentration in the conduction band (due to the high number of electrons excited across the band gap, as illustrated in Fig. 3.6c) exceeds \( N_D \) and pushes the \( n/N_D \) curve into the “intrinsic temperature region”, as shown in Fig. 3.5. The concentration of charge carriers asymptotically approaches the enhanced intrinsic carrier concentration at very high temperatures [68]. The hole concentration in an acceptor-doped semiconductor also exhibits a similar behavior when the temperature is increased [68].
Fig. 3.6. Source of the majority charge carriers in a donor-doped semiconductor material at different temperatures (T). (a) At T = 0K. (b) At low temperatures (T < 150K). (c) At very high temperatures (T > 450K). E_C: Conduction-band energy. E_D: Energy level of donor atoms. E_V: Valence-band energy [68].

3.2. Device Parameter Variations with Temperature Fluctuations

As described in Section 3.1, the concentration of charge carriers varies with the temperature. The variation of carrier concentrations alters the behavior of semiconductor devices with temperature. The device parameters that fluctuate due to temperature variations are identified in this section.

The four primary MOSFET device parameters that are affected by temperature fluctuations are the threshold voltage, carrier mobility, saturation velocity, and the parasitic drain/source resistances [74]. Threshold voltage is the gate-to-source voltage required to form a strong inversion region in the channel area underneath the gate. The threshold voltage of a MOSFET is [26]
\[ V_t = \Phi_{GC} - 2\Phi_F - \frac{Q_{BO}}{C_{OX}} - \frac{Q_{OX}}{C_{OX}}, \]  

(3.4)

where \( V_t \), \( \Phi_{GC} \), \( \Phi_F \), \( Q_{BO} \), \( C_{OX} \), and \( Q_{OX} \) are the device threshold voltage, voltage representing the difference in the work function between the gate material and the channel material, Fermi potential, depletion region charge, gate-oxide capacitance per unit area, and the fixed (immobile) parasitic charge in the gate oxide and in the silicon-oxide interface, respectively. As explained in [68], the Fermi energy is temperature dependent. The variation of the Fermi potential with temperature alters the device threshold voltage. The Fermi potential is [69]

\[ \Phi_F = \frac{E_F - E_i}{-q} = \frac{kT}{q} \ln \left( \frac{n_i}{n} \right) = \frac{kT}{q} \ln \left( \frac{p}{n_i} \right), \]  

(3.5)

where \( T \), \( E_F \), and \( E_i \) are the temperature, Fermi energy level, and intrinsic energy level, respectively. An increase in the temperature shifts the Fermi-level further up towards the conduction band in a donor-doped semiconductor material [68]. Alternatively, in an acceptor-doped material, the Fermi-level is shifted further down towards to the valence band when the temperature increases. The shift in the Fermi-level effectively reduces the band-gap energy in a semiconductor material, thereby reducing the threshold voltage of a MOSFET.

The drift velocity of charge carriers in the transistor channel is directly proportional to the applied electric field. The drift velocity \( (\nu) \) is [26]

\[ \nu = \mu_{eff} \xi, \]  

(3.6)

where \( \mu_{eff} \) and \( \xi \) are the effective carrier mobility and the electric field, respectively. The variation in the carrier concentration with temperature alters the scattering events experienced by the majority charge carriers [70]. Three different scattering mechanisms have been identified to account for the mobility behavior when the gate voltage is above
the threshold voltage of a MOSFET (strong inversion region) [83]. Phonon scattering occurs due to the various modes of lattice vibration including surface acoustic phonons and optical phonons [86]. Phonon scattering is important at room temperature but can be ignored at very low temperatures [86]. Coulomb scattering occurs due to the collision of the majority carriers with the charged impurities including the fixed oxide charge, interface-state charge, and the localized charge due to ionized impurities [85], [86]. The effects of coulomb scattering are important for lightly inverted surfaces. Higher surface-charge densities and higher substrate doping concentrations imply enhanced coulomb scattering [85]. The third scattering mechanism, surface-roughness scattering, is observed due to the roughness of the Si/SiO₂ interface [84], [87]. The scattering due to surface roughness is important under strong inversion conditions because the strength of the interaction is governed by the distance of the carriers from the surface. The surface roughness scattering events become more significant as the carriers get closer to the surface due to the stronger vertical fields within the scaled devices.

The relative importance of these scattering mechanisms depends on the operating temperature and the strength of the surface electric field. At low temperatures, the mobility is determined by the combined effects of Coulomb scattering (which dominates the behavior in the low-gate-field region) and surface roughness scattering (which dominates the device behavior in the high-vertical-field region) [83]. At high temperatures, the mobility is governed by coulomb scattering and phonon scattering in the low-field region. Alternatively, the surface roughness and the phonon scattering dominate the carrier mobility in the strong inversion region at elevated temperatures [83]. Due to the increased phonon scattering at elevated temperatures, the effective carrier mobility decreases with temperature, as described in [70] and [88].

The carrier mobility expression given by equation 3.6 is valid only for weak electric fields [44]. As the technologies scale into the nanometer regime, the drift velocity rolls off due to carrier scattering (surface roughness scattering). The drift velocity saturates at $v_{SAT}$. The saturation velocity is

$$v_{SAT} = \mu_{eff} E_c,$$  \hspace{1cm} (3.7)
where $v_{SAT}$, $\mu_{eff}$, and $E_c$ are the saturation velocity, effective carrier mobility, and the electric field at which the carrier drift velocity saturates, respectively. Reduction in the effective mobility with temperature lowers the saturation velocity of MOSFETs, as indicated by equation 3.7. Although both saturation velocity and mobility have a negative temperature dependence, saturation velocity displays a relatively weaker dependence since $E_c$ increases with the temperature [74]. Furthermore, as the MOSFET currents become higher while the supply voltages shrink, the drain and source series resistances become increasingly effective on the $I-V$ characteristics of devices in scaled CMOS technologies [74]. The drain and source resistances increase approximately linearly with the temperature [74].

3.3. Modeling of Parameter Variations

The influence of temperature variations on the carrier concentration and the various scattering mechanism within MOSFETs are described in Sections 3.1 and 3.2. In this section, the industry standard Berkeley Short-channel Insulated gate field effect transistor Model (BSIM) is presented [42], [72], [73]. In Section 3.4, BSIM is used to characterize the effects of temperature fluctuations on the critical device parameters.

As explained in Section 3.2, the threshold voltage, the carrier mobility, and the saturation velocity degrade when the die temperature increases. Alternatively, the source and drain resistances increase at elevated temperatures. The device parameter variations are modeled using BSIM3 and BSIM4 MOSFET equations. The threshold voltage, carrier mobility, saturation velocity, and source/drain resistances are [42], [72], [73]

$$NMOS: \quad V_i(T) = V_i(T_0) + \left( K_{T1} + \frac{K_{T1L}}{L_{eff}} + V_{bseff} K_{T2} \right) \left( \frac{T}{T_0} - 1 \right), \quad (3.8)$$

$$PMOS: \quad V_i(T) = V_i(T_0) - \left( K_{T1} + \frac{K_{T1L}}{L_{eff}} + V_{bseff} K_{T2} \right) \left( \frac{T}{T_0} - 1 \right), \quad (3.9)$$
\[
\mu_{\text{eff}}(T) = \left(U_0 \left(\frac{T}{T_0}\right)^{U_w}\right) \left[1 + \left(\frac{V_{\text{gsteff}} + 2V_t(T)}{T_{\text{OXE}}}\right)^2\right] U_b(T) \\
+ (U_c(T)V_{\text{bsteff}} + U_a(T) \left(\frac{V_{\text{gsteff}} + 2V_t(T)}{T_{\text{OXE}}}\right))^{-1},
\]

where \( V_t, K_{T1}, K_{T1L}, K_{T2}, L_{\text{eff}}, V_{\text{bsteff}}, \mu_{\text{eff}}, V_{\text{gsteff}}, U_0, U_{\text{te}}, T_{\text{OXE}}, U_a, U_b, U_c, V_{\text{SAT}}, A_T, R_{\text{dsw}}, P_{RT}, T_0, \) and \( T \) are the threshold voltage, temperature coefficient for threshold voltage, channel length dependence of the temperature coefficient for threshold voltage, body-bias coefficient of threshold voltage temperature effect, effective channel length, effective substrate bias voltage, carrier mobility, effective gate overdrive (\(|V_{GS} - V_t|\)), mobility at the reference temperature, mobility temperature exponent, electrical gate-oxide thickness, first order mobility degradation coefficient, second order mobility degradation coefficient, body effect of mobility degradation coefficient, saturation velocity, temperature coefficient of saturation velocity, drain-to-source resistance, temperature coefficient of drain/source resistance, reference temperature, and the operating temperature, respectively. \( K_{T1}, K_{T1L}, K_{T2}, A_T, \) and \( P_{RT} \) are temperature independent empirical parameters while \( U_a, U_b, \) and \( U_c \) are temperature dependent [42], [72], [73]. \( U_a, U_b, \) and \( U_c \) are

\[
U_a(T) = U_a(T_0) + U_{a1} \left(\frac{T}{T_0} - 1\right),
\]

\[
U_b(T) = U_b(T_0) + U_{b1} \left(\frac{T}{T_0} - 1\right).
\]
where \( U_{a1}, U_{b1}, \) and \( U_{c1} \) are the temperature coefficients of \( U_a, U_b, \) and \( U_c, \) respectively. The drain current of a MOSFET in the super-threshold region is [42], [72], [73]

\[
I_{ds} \propto \frac{I_{ds0}}{1 + \frac{R_{ds}}{V_{dseff}}},
\]

\[
I_{ds0} \propto V_{gseff} \mu_{eff} V_{dseff} \left( 1 - \frac{A_{bulk} V_{dseff}}{2(V_{gseff} + 2V_T)} \right) \left( 1 + \frac{V_{dseff}}{E_{SAT} L_{eff}} \right),
\]

where \( I_{ds}, I_{ds0}, V_{dseff}, R_{ds}, A_{bulk}, V_T, E_{SAT}, \) and \( L_{eff} \) are the drain current with short-channel effects, drain current of a long channel device, effective drain-to-source voltage, parasitic drain-to-source resistance, parameter to model the bulk charge effect, thermal voltage, electric field at which the carrier drift velocity saturates, and the effective channel length, respectively.

### 3.4. Temperature Effects in 180nm and 65nm CMOS Technology

Influence of temperature fluctuations on the device and circuit characteristics are evaluated in this section for the TSMC 180nm and Berkeley Predictive 65nm CMOS technologies [75]-[78]. The nominal supply voltages are 1.8V and 1.0V for the 180nm and 65nm CMOS technologies, respectively. The device threshold voltages excluding the short-channel effects \(|V_{t0}|\) are 0.46V and 0.22V for the 180nm and 65nm CMOS technologies, respectively [52], [71].

As given by (3.8), (3.9), (3.10), and (3.11) absolute values of threshold voltage, carrier mobility, and saturation velocity degrade as the temperature is increased [42], [72], [73]. Alternatively, the source and drain resistances increase at elevated temperatures, as given by (3.12) [42], [72], [73]. The model parameter coefficients that
determine the temperature fluctuation induced MOSFET drain current variations are listed in Table 3.1 [75]-[78]. The saturation velocity is typically a weak function of temperature [74]. Furthermore, the temperature coefficient of drain/source resistance \((P_{RT})\) is zero in these TSMC and PTM models, as listed Table 3.1. Therefore, the threshold voltage and carrier mobility fluctuations are the primary parameters that determine the MOSFET drain current variations with the temperature in the currently available CMOS technology models based on BSIM.

<table>
<thead>
<tr>
<th>Model Parameters</th>
<th>180nm CMOS Technology</th>
<th>65nm CMOS Technology</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PMOS</td>
<td>NMOS</td>
</tr>
<tr>
<td>(K_{T1})</td>
<td>-0.214</td>
<td>-0.196</td>
</tr>
<tr>
<td>(K_{T1L})</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(K_{T2})</td>
<td>-0.035</td>
<td>-0.039</td>
</tr>
<tr>
<td>(U_{te})</td>
<td>-0.599</td>
<td>-1.945</td>
</tr>
<tr>
<td>(U_{a1})</td>
<td>1.22E-09</td>
<td>1.22E-09</td>
</tr>
<tr>
<td>(U_{b1})</td>
<td>-1.44E-18</td>
<td>-3.08E-18</td>
</tr>
<tr>
<td>(U_{c1})</td>
<td>1.97E-10</td>
<td>-2.39E-10</td>
</tr>
<tr>
<td>(A_T)</td>
<td>10000</td>
<td>20000</td>
</tr>
<tr>
<td>(P_{RT})</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Threshold voltage degradation with temperature tends to enhance the drain current because of the increase in gate overdrive \((V_{GS} - V_t)\). Alternatively, degradation in carrier mobility tends to lower the drain current as given by (3.16) and (3.17). Effective variation of MOSFET drain current is, therefore, determined by the variation of the dominant device parameter when the temperature fluctuates. The temperature fluctuation induced threshold voltage variation for the devices in the 180nm and 65nm CMOS technologies is shown in Fig. 3.7. The gate overdrive and carrier mobility variations with temperature are shown in Figs. 3.8 and 3.9, respectively, for devices operating at the nominal supply voltage. Variation of the drain current \((I_{DS})\) of transistors in 180nm and
65nm CMOS technologies with the temperature are shown in Figs. 3.10 and 3.11, respectively [78].

At the nominal supply voltage, variations of the gate overdrive are smaller as compared to the carrier mobility variations when the temperature is increased from 25°C to 125°C, as shown in Figs. 3.8 and 3.9. The drain current of devices operating at the nominal supply voltage is, therefore, degraded as shown in Figs. 3.10 and 3.11.

![Fig. 3.7. Threshold voltage variations with temperature for devices in 180nm and 65nm CMOS technologies.](image)

![Fig. 3.8. Gate overdrive variation with temperature for devices in 180nm and 65nm CMOS technologies.](image)
Gate overdrive variations with temperature are similar in both technologies, as shown in Fig. 3.8. Alternatively, the variations in carrier mobility are higher for devices in the 65nm CMOS technology as compared to the variations observed in the 180nm CMOS technology, as shown in Fig. 3.9. Therefore, for a die temperature spectrum of
25°C to 125°C, degradation of the drain current in the 65nm CMOS technology is more significant than the drain current degradation observed with the TSMC 180nm CMOS technology [78].

![Graph showing MOSFET drain current variation with supply voltage and temperature.](image)

Fig. 3.11. Variation of MOSFET drain current with supply voltage and temperature in the predictive 65nm CMOS technology. \( |V_{DS}| = |V_{GS}| = V_{DD} \) and \( |V_{th}(T_0)| = 0.22V \).

Variation in the current produced by the devices alters the circuit speed. Test circuits are designed to have equal low-to-high and high-to-low propagation delays at the nominal supply voltage for temperature = 125°C. Propagation delay variations with temperature for circuits operating at the nominal supply voltage in 180nm and 65nm CMOS technologies are shown in Figs. 3.12 and 3.13, respectively. The circuit speed at the nominal supply voltage degrades primarily due to the reduction of MOSFET currents following the degradation of carrier mobilities when the temperature is increased. When operating at the nominal supply voltage, the speed of circuits degrades by up to 15.9% and 54.5% as the temperature is increased from 25°C to 125°C in 180nm and 65nm CMOS technologies as shown in Figs. 3.12 and 3.13, respectively [78].
Fig. 3.12. Percent delay variation with temperature for circuits operating at the nominal supply voltage ($V_{DD} = 1.8V$) in the TSMC 180nm CMOS technology.

Fig. 3.13. Percent delay variation with temperature for circuits operating at the nominal supply voltage ($V_{DD} = 1V$) in the predictive 65nm CMOS technology.

Power consumption is the product of the instantaneous voltage and current [30]. Variation in the drain current produced by MOSFETs alters the power dissipated by an integrated circuit when the temperature fluctuates [1], [117]. The energy consumed by the circuit is derived by integrating the power dissipated over the time period of interest [1],
Energy consumption variations with temperature for circuits operating at the nominal supply voltage in 180nm and 65nm CMOS technologies are shown in Figs. 3.14 and 3.15, respectively. Unlike the circuit speed, the energy consumption increases with temperature, as shown in Figs. 3.14 and 3.15. The increase in energy consumption is primarily due to the increase in the leakage currents. When operating at the nominal supply voltage, the energy consumption increases by up to 6.1% and 26% as the temperature is increases from 25°C to 125°C in 180nm and 65nm CMOS technologies, respectively.

The node voltages in circuits operating in the subthreshold region switch with the subthreshold currents. Subthreshold leakage current is characterized in Chapter 2. According to equation 2.14, subthreshold leakage current is a linear function of carrier mobility. Alternatively, the leakage current is exponentially increased with the reduction in the device threshold voltage. The temperature fluctuation induced threshold voltage variation is the dominant parameter influencing the subthreshold leakage current. In contrast to the degradation of drain current in super-threshold circuits, the subthreshold current increases with an increase in the die temperature. Developing novel temperature-
adaptive voltage scaling techniques to enhance the high temperature energy efficiency of ultra-low-voltage subthreshold logic circuits will be an important research topic, as discussed in Chapter 7.

![Energy Efficiency Graph](image)

Fig. 3.15. Percent energy variation with temperature for circuits operating at the nominal supply voltage ($V_{DD} = 1V$) in the PTM 65nm CMOS technology.

### 3.5. Chapter Summary

Imbalanced utilization and diversity of circuitry at different sections of an integrated circuit can cause significant variations in the die temperature from one die area to another. Furthermore, environmental temperature fluctuations can cause significant variations in the die temperature. An increase in the die temperature alters the concentration of the charge carriers in a semiconductor device. The physical mechanisms that cause a variation in the carrier concentration of semiconductors with temperature are reviewed in this chapter.

The variation in the carrier concentration with temperature alters the behavior of semiconductor devices. The four primary MOSFET device parameters that are affected by temperature fluctuations are the threshold voltage, carrier mobility, saturation velocity, and the parasitic drain/source resistances. The effect of temperature fluctuations
on critical device parameters is evaluated in this chapter using the industry standard device model BSIM.

The temperature fluctuation induced variations in the speed characteristics of the circuits in 180nm and 65nm CMOS technologies are examined. Variation of carrier mobility with temperature dominates the propagation delay variations in circuits operating at the nominal supply voltage for both 180nm and 65nm CMOS technologies. The MOSFET currents together with the circuit switching speed degrade following the degradation of carrier mobility as the temperature is increased in both technologies. When operating at the prescribed nominal supply voltage, the propagation delay increases by up to 19.6% and 54.5% as the temperature is increased from 25°C to 125°C in 180nm and 65nm CMOS technologies, respectively.

Variation in the drain current with temperature also alters the power consumed by a circuit when the temperature fluctuates. Fluctuation in the power consumption with temperature necessitates a comprehensive power measurement methodology to accurately characterize the total power consumption of circuits at different die temperatures. A generic methodology to accurately measure the power and energy consumption with the circuit simulators is described in Chapter 4. Furthermore, a design methodology for achieving temperature variation insensitive circuit performance is described in Chapter 5.
Chapter 4

Power Measurement Techniques with CAD Tools

Technology scaling has been the primary driving force behind the evolution of integrated circuits. As described in Chapter 1, the feature sizes of transistors and interconnects have continuously been scaled thereby increasing the integration density in each new process technology generation [1], [19], [117]. Furthermore, the reduction in the defect density due to the maturing fabrication technology has enabled manufacturing integrated circuits (IC) with larger dies. As a result of the reduced physical dimensions of the transistors and the increased die area, the total number of transistors in an integrated circuit increases with each new technology generation [20]. The higher number of transistors per IC provides enhanced circuit performance and functionality at the cost of increased power consumption [1], [9], [18], [35], [45], [117].

Computer-aided design (CAD) tools are used for the pre-fabrication characterization of integrated circuits. Design objectives such as speed, area, reliability, and power consumption can be verified with the aid of CAD tools. Reducing the power consumption is a primary objective in the design of digital integrated circuits. The different sources of power consumption and the impact of temperature variations on the power dissipation of a circuit are described in Chapters 2 and 3, respectively. Power consumption of CMOS circuits can be lowered by employing several techniques as described in [1], [10]-[12], [54], and [117]. During the design process, accurate power estimation with the circuit simulators is critical to be able to correctly identify the best techniques that satisfy the design objectives.

Commercial circuit simulators provide built-in functions to measure the power consumption. The research community typically assumes that the built-in functions provide the accurate power figures, overlooking the often inappropriate approximations and assumptions made by the existing built-in power estimation tools. For example, the built-in command for power measurement in Star-HSPICE calculates only the dissipated power within the semiconductor devices [55], [56]. The energy stored in the device parasitic capacitances is excluded from the device power calculations. The total power consumption in a circuit is the sum of the dissipated and the stored power [62]. The
exclusion of the power stored in the parasitic capacitances may introduce significant error in the computation of the total power consumption. Furthermore, the methodologies used by the power estimation tools are typically not well documented thereby forcing the circuit designers to blindly trust the results produced by the CAD tools [55]-[57].

An explicit power measurement method using a current-controlled current source or a voltage-controlled current source is described in [58]. This method calculates the power consumption based on the current drawn from the power supply. The method, however, excludes the currents drawn from the input and the output (I/O) terminals that contribute to the total power consumption. Currents drawn from the I/O terminals cannot be excluded from the power calculations in deeply scaled CMOS technologies due to the high gate-tunneling leakage currents [1], [43], [62], [117]. Furthermore, the currents drawn from the power supplies that provide the bias voltages for the substrate and the wells also contribute to the total power consumption. The contribution of the body contact currents can be significant particularly in the explicitly reverse or forward body-biased circuits [1], [117]. A comprehensive power measurement methodology is therefore highly desirable to accurately characterize the total power consumption of circuits considering all the circuit terminals for different modes of operation (active versus standby and zero-body-biased versus reverse/forward body-biased) with the CAD tools.

A generic methodology to accurately measure the power and energy consumption with the circuit simulators is described in this chapter. An equation to calculate the device power consumption based on the different current conduction paths in a MOSFET is presented. An expression for the total power consumption of a complex circuit is derived by explicitly considering all the circuit terminals including the inputs, the outputs, and the body-contacts. The actual power consumption measured using the proposed method is compared with the power measured with the built-in functions of the two most-popular commercial circuit simulators: HSPICE [55], [56] and CADENCE-SPECTRE [60], [61]. The percent errors introduced by the built-in power measurement functions are reported for different test circuits. The assumptions and simplifications that cause significant errors in power estimation with the built-in functions of the commonly used CAD tools are also identified.
The chapter is organized as follows. A generic methodology to measure the power consumption of CMOS circuits is described in Section 4.1. The test circuits and the experimental set-up for measuring the actual power consumption of CMOS circuits are presented in Section 4.2. The power measurements with the generic methodology and the power estimation using the built-in functions of circuit simulators are compared in Section 4.3 for different test circuits. The sources of errors in the built-in power estimation commands are identified in Section 4.4. A summary of the research results presented in this chapter is provided in Section 4.5.

4.1. Generic Power Measurement Methodology

In this section, an equation to calculate the total device power consumption based on the different current conduction paths in a MOSFET is presented. An expression for the total power consumption of a complex circuit considering all the circuit terminals is also derived.

The total instantaneous power consumed by a device is calculated by algebraically summing the product of the voltage and the net current flowing at the different device terminals [62]. The currents that contribute to the total device power consumption are supplied by the independent voltage and/or current sources connected to the different terminals of a device, as illustrated in Fig. 4.1. The total power consumed by a device is the summation of the power drawn from the independent sources connected to the different device terminals.

An expression for the instantaneous power consumption of a MOSFET is derived here based on the power drawn from the independent voltage sources at the different device terminals. Power is supplied to a load when the current flows out of the positive terminal of a voltage source. Alternatively, power is absorbed by a voltage source when the current flows into the positive terminal of the source. The total power consumed by a device is equal to the algebraic sum of the power supplied or absorbed by the voltage sources at the different device terminals. The instantaneous power consumed by the MOSFETs shown in Fig. 4.1 is [62]
\[ P_{\text{Device}} = V_{\text{Drain}} I_{\text{Drain}} + V_{\text{Gate}} I_{\text{Gate}} - V_{\text{Source}} I_{\text{Source}} - V_{\text{Body}} I_{\text{Body}}, \]  

(4.1)

where \( V_{\text{Body}} \), \( V_{\text{Drain}} \), \( V_{\text{Gate}} \), and \( V_{\text{Source}} \) are the power supply voltages attached to the different device terminals. \( I_{\text{Body}} \), \( I_{\text{Drain}} \), \( I_{\text{Gate}} \), and \( I_{\text{Source}} \) are the currents at the bulk, drain, gate, and source terminals, respectively, of the devices with the polarities as shown in Fig. 4.1. Equation 4.1 is valid for the current polarities indicated in Fig. 4.1. Provided that the net current through a terminal is in the opposite direction to the direction indicated in Fig. 4.1, the current polarity in equation 4.1 is reversed.

Fig. 4.1. MOSFETs biased with different power supplies. (a) An n-channel MOSFET. (b) A p-channel MOSFET.

Similar to the expression for the total power consumption of individual MOSFETs, an expression for measuring the total power consumed by a complex circuit is derived next considering the voltages and currents at all the terminals of the circuit. An integrated circuit with multiple I/O, power supply, ground, and bulk-contact terminals is shown in Fig. 4.2. The instantaneous voltages at the different terminals of the circuit are indicated in Fig. 4.2 along with the assumed direction of the instantaneous current in that terminal. The total power consumption of the circuit is measured by algebraically
summing the power drawn from or absorbed at all the circuit terminals. The instantaneous power consumed by the integrated circuit shown in Fig. 4.2 is [62]

\[
P_{\text{inst}} = V_1 I_1 + V_2 I_2 + V_3 I_3 + V_9 I_9 + V_{10} I_{10} - V_4 I_4 - V_5 I_5 - V_6 I_6 - V_7 I_7 - V_8 I_8. \tag{4.2}
\]

![Integrated Circuit](image)

Fig. 4.2. An integrated circuit with multiple I/O, power supply, ground, and body contact terminals.

Equation 4.2 can be expanded to derive a generic expression for the precise measurement of the power consumption of any circuit without any approximations. The instantaneous power consumed by a circuit with \( i + j \) terminals is

\[
P_{\text{inst}} = \sum_i V_i I_i - \sum_j V_j I_j, \tag{4.3}
\]

where \( V \) and \( I \) are the voltages and the currents at the different terminals of the circuit, respectively. The indices \( i \) and \( j \) cover all the terminals where the current enters and exits the circuit, respectively.
The energy consumption of a circuit is measured by integrating the instantaneous power consumed by the circuit over the time period of interest [62], [30]. The average power consumed by the circuit is the energy consumed per unit time. The energy and average power consumption of the circuit are

\[
\text{Energy} = \int_{t_1}^{t_2} P_{\text{inst}} \, dt, \\
\text{Average Power} = \frac{\text{Energy}}{t_2 - t_1},
\]

where \( t_1 \) and \( t_2 \) are the initial and the final points, respectively, of the time period within which the energy and the average power consumption of the circuit are measured [62].

The total power consumption of a circuit can be accurately measured by modeling the circuit as a black-box with independent current or voltage sources at the different terminals of the circuit, as illustrated in Fig. 4.2. At the circuit terminals without any explicit external source, an independent voltage sources producing zero output voltage (dummy voltage source) or an independent current sources producing zero output current (dummy current source) can be attached to measure the power at that terminal using a CAD tool.

### 4.2. Test Circuits and the Experimental Set-up

The design of the test circuits and the experimental set-up for measuring the actual power consumption using the generic methodology presented in Section 4.1 (ACTUAL) are described in this section.

Minimum size p-channel and n-channel MOSFETs (\( W = W_{\text{min}} \) and \( L = L_{\text{min}} \)), a 9-stage inverter chain, a 4-input static NAND gate (NAND4), and a 32-bit dynamic multiplexer (32-bit MUX) are the test circuits employed in this study. Test circuits are designed in a 65nm CMOS technology [52], [63]. A load capacitance equivalent to the gate capacitance of 4-minimum sized inverters [fan-out-of-four (\( C_L = 2.63 \text{fF} \)) in this 65nm CMOS technology] is connected to the circuit output terminals.
The ambient temperatures for integrated circuits employed in robotic explorations vary from -180°C to 486°C [64]. Similarly, ultra-low-power sensor-net modules are designed for functionality at a temperature range of -25°C to 125°C in security applications [65]. The die temperature spectrum is assumed to vary from -40°C to 125°C in this study.

Circuits are sized for equal low-to-high and high-to-low propagation delays at the worst-case die temperature (Temperature = 125°C). The nominal supply voltage is assumed to be 1.0V. The long-channel zero-body-bias threshold voltage ($|V_{t0}|$) of the devices in this 65m CMOS technology is 0.22V [52].

A 3-stage zero-body-biased inverter chain with a pulsed input is shown in Fig. 4.3. The average power consumed by the circuit for the time interval of interest can be measured by using the built-in power estimation commands provided by the CAD tool. Alternatively, with the ACTUAL, the circuit is treated as a black-box considering the currents and voltages at all the circuit terminals, as illustrated in Fig. 4.4. The name of the independent voltage sources along with the voltage supplied by the source (with respect to the positive terminal) are indicated in Fig. 4.4. To measure the output current with the proposed technique, a dummy voltage source ($V_{\text{dummy-1}}$) is connected to the output terminal (OUT), as shown in Fig. 4.4.

![Fig. 4.3. 3-stage zero-body-biased inverter chain.](image-url)
Fig. 4.4. Circuit set-up for power measurement using the proposed methodology.

The power consumed by the circuit is drawn from the external sources. The instantaneous power supplied or drawn by an independent source can be measured by multiplying the voltage and the net current flowing into the positive terminal of the source. In HSPICE, the command $I(name_{ins})$ provides the magnitude of the current flowing into the positive terminal of an independent source with instance name $name_{ins}$ [56]. Similarly, the command $V(node_{-1})$ gives the instantaneous voltage at the node labeled as $node_{-1}$. The instantaneous power consumption of the circuit in Fig. 4.4 is

$$P_{inst} = -I(V_{supply})V_{DD} - I(V_{p-bias})V_{DD} - I(V_{IN})V(IN) - I(V_{gnd})V_{gnd}$$

$$-I(V_{n-bias})V_{n-bias} - I(V_{dummy-1})V(OUT),$$

where $V_{supply}$, $V_{p-bias}$, $V_{IN}$, $V_{gnd}$, $V_{dummy-1}$, and $V_{n-bias}$ are the instance names of the independent voltage sources shown in Fig. 4.4. Alternatively, the power consumption of the independent sources can also be measured using the built-in commands provided that the power measurement of the CAD tool is accurate. Star-HSPICE calculates the power consumed by the independent current and voltage sources without any approximations [56]. In HSPICE, the instantaneous power consumed by the independent sources are
measured using the $P(\text{name}_\text{ins})$ built-in power estimation command where $\text{name}_\text{ins}$ is the instance name of the independent source. For the dummy voltage (current) sources, however, the power cannot be measured directly using the built-in commands due to the intentional zero output voltage (zero output current) assigned to the source. The instantaneous power consumption of the circuit in Fig. 4.4 based on the second technique that accurately employs the independent-source built-in power measurement command is

$$P_{\text{inst}} = -P(V_{\text{sup}}) - P(V_{p\text{-bias}}) - P(V_{\text{IN}}) - P(V_{\text{gnd}}) - P(V_{n\text{-bias}}) - I(V_{\text{dummy}}-1)V(\text{OUT}).$$  \hspace{1cm} (4.7)

### 4.3. Comparison of the Power Measurements

In this section, the actual power consumption measurements with the generic methodology (ACTUAL) are compared with the power figures obtained using the built-in functions of a typical commercial CAD tool (IN-CAD). HSPICE (Ver. 2005.03-SP1) and CADENCE-SPECTRE (Ver. 5.1.41) are the circuit simulators used in this study [55], [61]. These versions of HSPICE and CADENCE-SPECTRE evaluate the gate tunneling currents through BSIM4 device equations [59]. The built-in power estimation commands of HSPICE and CADENCE-SPECTRE are POWER and PWR, respectively [55], [61]. Both DC and TRANSIENT analysis are performed at different temperatures to identify the error introduced by the HSPICE built-in power estimation command (POWER). The power measurements of the standard zero-body-biased active circuits are presented in Section 4.3.1. The error in the stand-by mode power measurement with the built-in function of HSPICE is evaluated in Section 4.3.2. The power measurements for the body-biased circuits are presented in Section 4.3.3.

#### 4.3.1. Active Mode Power Consumption

The active mode power consumption measured with the built-in command of HSPICE (POWER) and the actual power consumption are compared in this section. The average power consumption measured using the IN-CAD and the ACTUAL for zero-body-biased circuits operating at the nominal supply voltage ($V_{DD} = 1.0V$) are compared
in Fig. 4.5. The inputs of the different circuits are excited with a 1GHz pulse. The power measurement with the HSPICE built-in command AVG POWER introduces a -3.7% (32-bit MUX at -40°C) to +4.1% (32-bit MUX at 125°C) error, as shown in Fig. 4.5.

Reducing energy consumption is a primary objective in the design of digital integrated circuits [1], [10]-[12], [66], [117]. Circuits optimized for minimum energy consumption operate typically in the subthreshold regime with ultra-low power-supply-voltages [66]. The active-mode power consumption of an ultra-low-voltage circuit is due to the weak inversion current. The active-mode power consumption of subthreshold logic circuits measured with the IN-CAD and the ACTUAL are compared in Fig. 4.6. The supply voltage applied for subthreshold operation is 0.2V \((V_{DD} < |V_{t0}|)\). The input terminals of the different static circuits are excited with a 1MHz pulse. The power measurement with the HSPICE built-in command AVG POWER produces a +1.2% (9-stage inverter at 125°C) to +44.9% (NAND4 at -40°C) error for the ultra-low-voltage subthreshold logic circuits, as shown in Fig. 4.6.

![Fig. 4.5. Comparison of the average power consumption measured with the HSPICE built-in command AVG POWER and the actual power consumption when the input is oscillating at 1GHz. \(V_{DD} = 1.0V\) (super-threshold operation).](image-url)
Fig. 4.6. Comparison of the average power consumption measured with the HSPICE built-in command AVG POWER and the actual power consumption of the subthreshold logic circuits. Input signal frequency is 1MHz. $V_{DD} = 0.2V$.

### 4.3.2. Stand-by Mode Power Consumption

The stand-by mode power consumption measured using the IN-CAD and the ACTUAL are compared in this section. The instantaneous power consumption measured using the HSPICE built-in command AVG POWER and the actual power consumption for an n-channel and a p-channel device are compared in Tables 4.1 and 4.2, respectively, for different gate voltages and temperature. The power measurements are performed with the NMOS (PMOS) source, drain, and bulk terminals biased at 0V, $V_{DD}$, and 0V ($V_{DD}$, 0V, and $V_{DD}$), respectively. As listed in Table 4.1, the power measurements for an NMOSFET with the HSPICE built-in command AVG POWER produces an error of up to 433% (at -40°C with the gate biased at 0V). Similarly, for the PMOSFET, the error with the built-in command is up to 8.35% (at -40°C with the $G_v$ at 1.0V), as listed in Table 4.2.

The power consumption in the stand-by mode (no switching activity) is due to the leakage currents (subthreshold, junction, and gate-tunneling leakage currents), as described in Chapter 2. The leakage power consumption of a circuit is strongly dependent on the input vectors [67]. For circuits operating at the nominal supply voltage ($V_{DD} = 1.0V$), the stand-by mode power consumption measured using the two power measurement techniques with the inputs biased at zero and $V_{DD}$ are shown in Figs. 4.7.
and 4.8, respectively. When the input is biased at zero, the power measured with the HSPICE built-in command AVG POWER produces up to 16x (32-bit MUX at -40°C) error, as shown in Fig. 4.7. Alternatively, when the input is biased at $V_{DD}$, the power is underestimated by up to 105x (32-bit MUX at -40°C) with the HSPICE built-in power function, as shown in Fig. 4.8.

TABLE 4.1
COMPARISON OF THE POWER MEASUREMENTS FOR AN NMOS DEVICE WITH ZERO-BODY-BIAS AT VARIOUS TEMPERATURES

<table>
<thead>
<tr>
<th>$G_s$ (V)</th>
<th>NMOSFET</th>
<th>At -40°C</th>
<th>At 25°C</th>
<th>At 125°C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IN-CAD</td>
<td>ACTUAL</td>
<td>% Error</td>
<td>IN-CAD</td>
</tr>
<tr>
<td>0.0</td>
<td>0.019</td>
<td>0.100</td>
<td>433</td>
<td>0.109</td>
</tr>
<tr>
<td>0.1</td>
<td>0.453</td>
<td>0.495</td>
<td>9.21</td>
<td>1.281</td>
</tr>
<tr>
<td>0.2</td>
<td>7.359</td>
<td>7.380</td>
<td>0.28</td>
<td>11.67</td>
</tr>
<tr>
<td>0.3</td>
<td>62.01</td>
<td>62.02</td>
<td>0.01</td>
<td>69.80</td>
</tr>
<tr>
<td>0.4</td>
<td>273.1</td>
<td>273.1</td>
<td>0.0</td>
<td>250.9</td>
</tr>
<tr>
<td>0.6</td>
<td>1110</td>
<td>1110</td>
<td>0.0</td>
<td>927.0</td>
</tr>
<tr>
<td>0.8</td>
<td>2003</td>
<td>2003</td>
<td>0.0</td>
<td>1663</td>
</tr>
<tr>
<td>1.0</td>
<td>2846</td>
<td>2846</td>
<td>0.0</td>
<td>2349</td>
</tr>
</tbody>
</table>

Fig. 4.7. Comparison of the stand-by mode power consumption measured with the HSPICE built-in command AVG POWER and the actual power consumption when the inputs are biased at zero. $V_{DD} = 1.0V$.  

![Graph showing power consumption comparison](image-url)
**TABLE 4.2**

**COMPARISON OF THE POWER MEASUREMENTS FOR A PMOS DEVICE WITH ZERO-BODY-BIAS AT VARIOUS TEMPERATURES**

<table>
<thead>
<tr>
<th>Gv (V)</th>
<th>PMOSFET</th>
<th>At -40°C</th>
<th></th>
<th>At 25°C</th>
<th></th>
<th>At 125°C</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IN-CAD</td>
<td>ACTUAL</td>
<td>% Error</td>
<td>IN-CAD</td>
<td>ACTUAL</td>
<td>% Error</td>
<td>IN-CAD</td>
</tr>
<tr>
<td>0.0</td>
<td>1735</td>
<td>1735</td>
<td>0.0</td>
<td>1183</td>
<td>1183</td>
<td>0.0</td>
<td>696.2</td>
</tr>
<tr>
<td>0.2</td>
<td>1247</td>
<td>1247</td>
<td>0.0</td>
<td>859.3</td>
<td>859.3</td>
<td>0.0</td>
<td>517.8</td>
</tr>
<tr>
<td>0.4</td>
<td>711.5</td>
<td>711.5</td>
<td>0.0</td>
<td>497.7</td>
<td>497.7</td>
<td>0.0</td>
<td>311.2</td>
</tr>
<tr>
<td>0.6</td>
<td>198.9</td>
<td>198.9</td>
<td>0.0</td>
<td>153.0</td>
<td>153.0</td>
<td>0.0</td>
<td>106.8</td>
</tr>
<tr>
<td>0.7</td>
<td>52.36</td>
<td>52.36</td>
<td>0.0</td>
<td>48.85</td>
<td>48.85</td>
<td>0.0</td>
<td>40.66</td>
</tr>
<tr>
<td>0.8</td>
<td>7.513</td>
<td>7.514</td>
<td>0.01</td>
<td>9.520</td>
<td>9.521</td>
<td>0.01</td>
<td>10.92</td>
</tr>
<tr>
<td>0.9</td>
<td>0.571</td>
<td>0.572</td>
<td>0.22</td>
<td>1.189</td>
<td>1.190</td>
<td>0.11</td>
<td>2.234</td>
</tr>
<tr>
<td>1.0</td>
<td>0.026</td>
<td>0.028</td>
<td>8.35</td>
<td>0.108</td>
<td>0.110</td>
<td>2.01</td>
<td>0.381</td>
</tr>
</tbody>
</table>

IN-CAD: Power measured using the HSPICE built-in command [AVG POWER] (x10⁻⁷ W). ACTUAL: Actual power consumption measured using the proposed generic methodology (x10⁻⁷ W). Gv: is the voltage applied to the gate terminal of the device. The NMOS (PMOS) source, drain, and bulk terminals are biased at 0V, 1.0V, and 0V (1.0V, 0V, and 1.0V), respectively.

**Fig. 4.8.** Comparison of the stand-by mode power consumption measured with the HSPICE built-in command AVG POWER and the actual power consumption when the inputs are biased at Vdd = 1.0V.
4.3.3. Power Consumption of Body-Biased Circuits

The currents drawn from the power supplies that provide the bias voltages of the substrate and the wells also contribute to the total power consumption. The contribution of the body contact currents can be significant particularly in the explicitly body-biased (reverse or forward body-bias) circuit [1], [117]. The power consumption measured using the IN-CAD and the ACTUAL are presented in this section for the body-biased circuits.

The instantaneous power consumption of MOSFETs measured using the HSPICE built-in command and the actual power consumption are compared in Tables 4.3 and 4.4 for different body-bias voltages and temperature. The power measurements are performed with the PMOS (NMOS) source, drain, and gate terminals biased at V_Dd, 0V, and V_Dd (0V, V_Dd, and 0V), respectively. A negative V_Bs in Table 4.3 (Table 4.4) indicates that the PMOS (NMOS) device is forward (reversed) body-biased. Alternatively, a positive V_Bs in Table 4.3 (Table 4.4) indicates that the PMOS (NMOS) device is reversed (forward) body-biased. For a PMOSFET, power measurements with the HSPICE built-in command AVG POWER introduce a +0.28% (V_Bs = -0.3V at 125°C) to +21.43% (V_Bs = 0.3V at -40°C) error, as listed in Table 4.3. Similarly, for an NMOSFET, the power is underestimated by up to 1819% (V_Bs = -0.3V at -40°C) with the HSPICE built-in command, as listed in Table 4.4.

**TABLE 4.3**
COMPARISON OF THE POWER MEASUREMENTS FOR A PMOS DEVICE AT DIFFERENT BODY-BIAS VOLTAGES AND TEMPERATURE

<table>
<thead>
<tr>
<th>V_Bs (V)</th>
<th>PMOSFET At -40°C</th>
<th>PMOSFET At 25°C</th>
<th>PMOSFET At 125°C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IN-CAD</td>
<td>ACTUAL</td>
<td>% Error</td>
</tr>
<tr>
<td>-0.3</td>
<td>6.33</td>
<td>6.55</td>
<td>3.44</td>
</tr>
<tr>
<td>-0.2</td>
<td>4.87</td>
<td>5.08</td>
<td>4.47</td>
</tr>
<tr>
<td>-0.1</td>
<td>3.62</td>
<td>3.84</td>
<td>5.99</td>
</tr>
<tr>
<td>0.0</td>
<td>2.60</td>
<td>2.82</td>
<td>8.35</td>
</tr>
<tr>
<td>0.1</td>
<td>1.85</td>
<td>2.07</td>
<td>11.73</td>
</tr>
<tr>
<td>0.2</td>
<td>1.36</td>
<td>1.57</td>
<td>16.08</td>
</tr>
<tr>
<td>0.3</td>
<td>1.01</td>
<td>1.23</td>
<td>21.43</td>
</tr>
</tbody>
</table>
TABLE 4.4
COMPARISON OF THE POWER MEASUREMENTS FOR AN NMOS DEVICE AT DIFFERENT BODY-BIAS VOLTAGES AND TEMPERATURE

<table>
<thead>
<tr>
<th>$V_{BS}$ (V)</th>
<th>NMOSFET</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>At -40°C</td>
<td>At 25°C</td>
<td>At 125°C</td>
<td>At -40°C</td>
<td>At 25°C</td>
<td>At 125°C</td>
<td>At -40°C</td>
<td>At 25°C</td>
<td>At 125°C</td>
<td>At -40°C</td>
<td>At 25°C</td>
</tr>
<tr>
<td></td>
<td>IN-CAD</td>
<td>ACTUAL</td>
<td>% Error</td>
<td>IN-CAD</td>
<td>ACTUAL</td>
<td>% Error</td>
<td>IN-CAD</td>
<td>ACTUAL</td>
<td>% Error</td>
<td>IN-CAD</td>
<td>ACTUAL</td>
</tr>
<tr>
<td>-0.3</td>
<td>0.45</td>
<td>8.55</td>
<td>1819</td>
<td>3.73</td>
<td>11.83</td>
<td>217.6</td>
<td>24.70</td>
<td>32.81</td>
<td>32.8</td>
<td>24.70</td>
<td>32.81</td>
</tr>
<tr>
<td>-0.2</td>
<td>0.70</td>
<td>8.81</td>
<td>1155</td>
<td>5.23</td>
<td>13.34</td>
<td>155.2</td>
<td>31.21</td>
<td>39.32</td>
<td>25.9</td>
<td>31.21</td>
<td>39.32</td>
</tr>
<tr>
<td>-0.1</td>
<td>1.13</td>
<td>9.24</td>
<td>716.8</td>
<td>7.47</td>
<td>15.58</td>
<td>108.6</td>
<td>39.95</td>
<td>48.05</td>
<td>20.3</td>
<td>39.95</td>
<td>48.05</td>
</tr>
<tr>
<td>0.0</td>
<td>1.87</td>
<td>9.98</td>
<td>433.0</td>
<td>10.88</td>
<td>18.99</td>
<td>74.54</td>
<td>51.85</td>
<td>59.96</td>
<td>15.6</td>
<td>51.85</td>
<td>59.96</td>
</tr>
<tr>
<td>0.1</td>
<td>3.05</td>
<td>11.16</td>
<td>265.4</td>
<td>15.68</td>
<td>23.79</td>
<td>51.72</td>
<td>66.78</td>
<td>74.88</td>
<td>12.1</td>
<td>66.78</td>
<td>74.88</td>
</tr>
<tr>
<td>0.2</td>
<td>4.75</td>
<td>12.86</td>
<td>170.7</td>
<td>21.76</td>
<td>29.87</td>
<td>37.27</td>
<td>84.10</td>
<td>92.20</td>
<td>9.64</td>
<td>84.10</td>
<td>92.20</td>
</tr>
<tr>
<td>0.3</td>
<td>7.08</td>
<td>15.19</td>
<td>114.4</td>
<td>29.24</td>
<td>37.35</td>
<td>27.74</td>
<td>115.9</td>
<td>124.0</td>
<td>6.99</td>
<td>115.9</td>
<td>124.0</td>
</tr>
</tbody>
</table>

IN-CAD: Power measured using the HSPICE built-in command [AVG POWER] ($\times10^{-9}$ W). ACTUAL: Actual power consumption measured using the proposed generic methodology ($\times10^{-9}$ W). The PMOS (NMOS) source, drain, and gate terminal are biased at 1.0V, 0V, and 1.0V (0V, 1.0V, and 0V), respectively. A negative $V_{BS}$ indicates that the PMOS (NMOS) device is forward-body-biased (reverse-body-biased).

The average power consumption measured using the HSPICE built-in command AVG POWER and the actual power consumption for 32-bit MUX and NAND4 are compared in Tables 4.5 and 4.6, respectively, for different body-bias voltages and temperature. The input terminals are fixed at the nominal supply voltage ($V_{DD} = 1.0V$). A negative $V_{Bias}$ in Tables 4.5 and 4.6 indicates that an intentional forward-body-bias is applied to all the devices in the circuit. Alternatively, a positive $V_{Bias}$ in Tables 4.5 and 4.6 indicates that an intentional reverse-body-bias is applied to all the devices in the circuit. The power measurements with the IN-CAD (HSPICE built-in command AVG POWER) for 32-bit MUX introduces up to 110x (at $V_{Bias} = 0.3V$, Temp = -40°C) error, as listed in Table 4.5. Similarly, the error introduced by IN-CAD in NAND4 is up to 25.5x (at $V_{Bias} = 0.3V$, Temp = -40°C), as listed in Table 4.6.
### TABLE 4.5
COMPARISON OF THE POWER MEASUREMENTS FOR 32-BIT MUX AT DIFFERENT BODY-BIAS VOLTAGES AND TEMPERATURE

<table>
<thead>
<tr>
<th>V_{Bias} (V)</th>
<th>32-BIT MUX</th>
<th>At -40°C</th>
<th>At 25°C</th>
<th>At 125°C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IN-CAD</td>
<td>ACTUAL</td>
<td>% Error</td>
<td>IN-CAD</td>
</tr>
<tr>
<td>-0.3</td>
<td>0.58</td>
<td>1.55</td>
<td>2.7x</td>
<td>0.07</td>
</tr>
<tr>
<td>-0.2</td>
<td>0.01</td>
<td>1.01</td>
<td>77.7x</td>
<td>0.04</td>
</tr>
<tr>
<td>-0.1</td>
<td>0.01</td>
<td>1.03</td>
<td>103x</td>
<td>0.04</td>
</tr>
<tr>
<td>0.0</td>
<td>0.01</td>
<td>1.05</td>
<td>105x</td>
<td>0.04</td>
</tr>
<tr>
<td>0.1</td>
<td>0.01</td>
<td>1.06</td>
<td>106x</td>
<td>0.04</td>
</tr>
<tr>
<td>0.2</td>
<td>0.01</td>
<td>1.08</td>
<td>108x</td>
<td>0.04</td>
</tr>
<tr>
<td>0.3</td>
<td>0.01</td>
<td>1.10</td>
<td>110x</td>
<td>0.04</td>
</tr>
</tbody>
</table>

IN-CAD: Power measured using the HSPICE built-in command [AVG POWER] (x10^{-6} W). ACTUAL: Actual power consumption measured using the proposed generic methodology (x10^{-6} W). The inputs terminals are fixed at 1.0V. A negative V_{Bias} indicates that an intentional forward-body-bias is applied to all the devices in the circuit.

### TABLE 4.6
COMPARISON OF THE POWER MEASUREMENTS FOR NAND4 AT DIFFERENT BODY-BIAS VOLTAGES AND TEMPERATURE

<table>
<thead>
<tr>
<th>V_{Bias} (V)</th>
<th>NAND4</th>
<th>At -40°C</th>
<th>At 25°C</th>
<th>At 125°C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IN-CAD</td>
<td>ACTUAL</td>
<td>% Error</td>
<td>IN-CAD</td>
</tr>
<tr>
<td>-0.3</td>
<td>0.07</td>
<td>0.20</td>
<td>2.8x</td>
<td>0.12</td>
</tr>
<tr>
<td>-0.2</td>
<td>0.03</td>
<td>0.16</td>
<td>5.7x</td>
<td>0.10</td>
</tr>
<tr>
<td>-0.1</td>
<td>0.02</td>
<td>0.16</td>
<td>7.4x</td>
<td>0.08</td>
</tr>
<tr>
<td>0.0</td>
<td>0.02</td>
<td>0.15</td>
<td>10.1x</td>
<td>0.06</td>
</tr>
<tr>
<td>0.1</td>
<td>0.01</td>
<td>0.15</td>
<td>14.0x</td>
<td>0.05</td>
</tr>
<tr>
<td>0.2</td>
<td>0.01</td>
<td>0.15</td>
<td>19.0x</td>
<td>0.04</td>
</tr>
<tr>
<td>0.3</td>
<td>0.01</td>
<td>0.15</td>
<td>25.5x</td>
<td>0.03</td>
</tr>
</tbody>
</table>
4.4. Sources of Error with the Built-in Power Estimation Commands

The results presented in Section 4.3 indicate that the power measurement using the HSPICE built-in command AVG POWER is significantly off as compared to the actual power consumption of a CMOS circuit. The approximations that cause significant errors in the power measured with the built-in functions of CAD tools are identified in this section.

N-channel MOSFETs with different bias conditions are shown in Fig. 4.9. The drain of the MOSFET in Fig. 4.9a is biased at 0V. Alternatively, the device in Fig. 4.9b has a drain voltage of 1.0V (nominal supply voltage). The gate and source voltages for these MOSFETs are fixed at 1.0V and 0V, respectively. The currents observed at the different terminals of the device in Fig. 4.9a at 125°C for various body voltages ($V_B$) are listed in Table 4.7 along with the total device power consumption measured with the built-in power command of HSPICE and the ACTUAL. Similarly, for the device in Fig. 4.9b, the currents observed at the different terminals and the power consumption measured with the built-in power command of HSPICE and the ACTUAL are listed in Table 4.8.

![Diagram of NMOS devices](image)

Fig. 4.9. NMOS devices in a 65nm CMOS technology. Temperature = 125°C. Width = 300nm. Length = 65nm. (a) Drain, gate, and source terminals biased at 0V, 1V, and 0V respectively. (b) Drain, gate, and source terminals biased at 1V, 1V, and 0V respectively.
### TABLE 4.7
CURRENTS AND THE POWER CONSUMPTION OF THE NMOS DEVICE IN Fig. 4.9A
MEASURED WITH HSPICE FOR VARIOUS BODY VOLTAGES

<table>
<thead>
<tr>
<th>$V_B$ (V)</th>
<th>$I_D$ (nA)</th>
<th>$I_G$ (nA)</th>
<th>$I_S$ (nA)</th>
<th>$I_B$ (nA)</th>
<th>$I_{SUM}$ (A)</th>
<th>Total Device Power Consumption</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>IN-CAD (nW)</td>
</tr>
<tr>
<td>-1.0</td>
<td>-15.73</td>
<td>31.54</td>
<td>15.73</td>
<td>-0.07229</td>
<td>7.71p</td>
<td>17.4p</td>
</tr>
<tr>
<td>-0.8</td>
<td>-15.09</td>
<td>30.23</td>
<td>15.09</td>
<td>-0.05620</td>
<td>-6.20p</td>
<td>13.6p</td>
</tr>
<tr>
<td>-0.6</td>
<td>-14.47</td>
<td>28.99</td>
<td>14.47</td>
<td>-0.04400</td>
<td>6.00p</td>
<td>9.98p</td>
</tr>
<tr>
<td>-0.4</td>
<td>-13.89</td>
<td>27.82</td>
<td>13.89</td>
<td>-0.03487</td>
<td>5.13p</td>
<td>6.5p</td>
</tr>
<tr>
<td>-0.2</td>
<td>-13.33</td>
<td>26.69</td>
<td>13.33</td>
<td>-0.02807</td>
<td>1.93p</td>
<td>3.16p</td>
</tr>
<tr>
<td>0.0</td>
<td>-12.80</td>
<td>25.62</td>
<td>12.80</td>
<td>-0.00775</td>
<td>12.24p</td>
<td>0.0</td>
</tr>
<tr>
<td>0.2</td>
<td>-14.84</td>
<td>24.44</td>
<td>14.84</td>
<td>5.232</td>
<td>-8.00p</td>
<td>1.05n</td>
</tr>
<tr>
<td>0.4</td>
<td>-904.50</td>
<td>23.09</td>
<td>904.50</td>
<td>1786.00</td>
<td>90.0p</td>
<td>0.714u</td>
</tr>
<tr>
<td>0.6</td>
<td>-281300</td>
<td>21.62</td>
<td>281300</td>
<td>562600</td>
<td>21.62n</td>
<td>0.337m</td>
</tr>
<tr>
<td>0.8</td>
<td>-8952000</td>
<td>20.30</td>
<td>8952000</td>
<td>17900000</td>
<td>-3.9797u</td>
<td>14.39m</td>
</tr>
<tr>
<td>1.0</td>
<td>-24550000</td>
<td>19.15</td>
<td>24550000</td>
<td>49110000</td>
<td>10.019u</td>
<td>49.11m</td>
</tr>
</tbody>
</table>

* undefined: divide by 0. $I_{SUM} = I_D + I_G + I_B - I_S$.

### TABLE 4.8
CURRENTS AND THE POWER CONSUMPTION OF THE NMOS DEVICE IN Fig. 4.9B
MEASURED WITH HSPICE FOR VARIOUS BODY VOLTAGES

<table>
<thead>
<tr>
<th>$V_B$ (V)</th>
<th>$I_D$ (nA)</th>
<th>$I_G$ (nA)</th>
<th>$I_S$ (nA)</th>
<th>$I_B$ (nA)</th>
<th>$I_{SUM}$ (nA)</th>
<th>Total Device Power Consumption</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>IN-CAD (nW)</td>
</tr>
<tr>
<td>-1.0</td>
<td>136900</td>
<td>18.97</td>
<td>136900</td>
<td>-0.081</td>
<td>18.89</td>
<td>136.9</td>
</tr>
<tr>
<td>-0.8</td>
<td>142100</td>
<td>18.10</td>
<td>142200</td>
<td>-0.063</td>
<td>-81.96</td>
<td>142.1</td>
</tr>
<tr>
<td>-0.6</td>
<td>148000</td>
<td>17.28</td>
<td>148000</td>
<td>-0.049</td>
<td>17.23</td>
<td>148.0</td>
</tr>
<tr>
<td>-0.4</td>
<td>154500</td>
<td>16.50</td>
<td>154600</td>
<td>-0.039</td>
<td>-83.54</td>
<td>154.5</td>
</tr>
<tr>
<td>-0.2</td>
<td>161900</td>
<td>15.76</td>
<td>161900</td>
<td>-0.031</td>
<td>15.73</td>
<td>161.9</td>
</tr>
<tr>
<td>0.0</td>
<td>170000</td>
<td>15.05</td>
<td>170100</td>
<td>-0.018</td>
<td>-84.97</td>
<td>170.1</td>
</tr>
<tr>
<td>0.2</td>
<td>177800</td>
<td>14.28</td>
<td>177800</td>
<td>2.604</td>
<td>16.88</td>
<td>177.8</td>
</tr>
<tr>
<td>0.4</td>
<td>184000</td>
<td>13.38</td>
<td>184900</td>
<td>893.00</td>
<td>6.38</td>
<td>184.4</td>
</tr>
<tr>
<td>0.6</td>
<td>189000</td>
<td>12.41</td>
<td>472400</td>
<td>283300</td>
<td>-87.59</td>
<td>359.1</td>
</tr>
<tr>
<td>0.8</td>
<td>192800</td>
<td>11.46</td>
<td>9837000</td>
<td>96440000</td>
<td>-188.54</td>
<td>7908.0</td>
</tr>
<tr>
<td>1.0</td>
<td>195800</td>
<td>10.54</td>
<td>26690000</td>
<td>26490000</td>
<td>-4189.46</td>
<td>26690.0</td>
</tr>
</tbody>
</table>

*I_{SUM} = I_D + I_G + I_B - I_S*.
For the NMOS device shown in Fig. 4.9a, the currents that contribute to the device power consumption are supplied from the voltage sources connected to the gate and the body terminals (non-zero voltage sources). The currents drawn from these voltage sources exit the device through the drain and the source terminals due to the non-zero gate-to-source, gate-to-drain, body-to-source, and body-to-drain voltages. The drain-to-source voltage and current are zero. Provided that a non-zero gate-to-body voltage is applied, current flows between the gate and the body terminals (gate-tunneling leakage current). When the body terminal is grounded, the power consumption measured with the HSPICE built-in command AVG POWER is zero, as listed in Table 4.7. The actual device however still consumes power due to the current drawn from the voltage source connected to the gate terminal ($I_G$). ACTUAL accurately measures this power consumption without any approximations ($I_G \cdot V_G$), as listed in Table 4.7. Alternatively, the power measured with the IN-CAD clearly totally neglects the current at the gate terminal. The power consumption is underestimated by up to 8450x with the built-in power function of HSPICE, as listed in Table 4.7.

For the NMOS device shown in Fig. 4.9b, $V_{DS} = 1V$. The drain-to-source current produced by this MOSFET is orders of magnitude larger as compared to the currents at the gate and the body terminals, as listed in Table 4.8. The relatively small gate current in this device reduces the error introduced in the power measurements with the IN-CAD, as listed in Table 4.8. Alternatively, for the NMOS device shown in Fig. 4.9a, power consumption is primarily due to the gate-tunneling leakage current. The exclusion of the gate-tunneling current in the power measurements with the IN-CAD introduces significant error in this device, as listed in Table 4.7.

CADENCE-SPECTRE is another widely used commercial circuit simulator. The accuracy of the power measurements with the built-in circuit power function PWR [61] of CADENCE-SPECTRE is evaluated next [60], [61]. The different currents and the power measurements (with both IN-CAD and ACTUAL) with CADENCE-SPECTRE for the device in Fig. 4.9a are listed in Table 4.9 at various body voltages ($V_B$) [Temp = 125°C]. The data listed in Tables 4.7 and 4.9 are measured for the same NMOS device with the same technology parameters using two different circuit simulators (HSPICE and CADENCE-SPECTRE). As listed in Tables 4.7 and 4.9, the device currents measured
with HSPICE and CADENCE-SPECTRE are equal. The power consumption measured using the built-in commands of the two simulators are, however, significantly different, as listed in Tables 4.7 and 4.9. In CADENCE-SPECTRE, the power consumption is underestimated by up to \(113715046604527x\) with the built-in power estimation function PWR, as listed in Table 4.9.

**TABLE 4.9**
CURRENTS AND THE POWER CONSUMPTION OF THE NMOS DEVICE IN FIG. 4.9A MEASURED WITH CADENCE-SPECTRE FOR VARIOUS BODY VOLTAGES

<table>
<thead>
<tr>
<th>(V_B) (V)</th>
<th>(I_D) (nA)</th>
<th>(I_G) (nA)</th>
<th>(I_S) (nA)</th>
<th>(I_B) (nA)</th>
<th>(I_{SUM}) (A)</th>
<th>Total Device Power Consumption</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.0</td>
<td>-15.73</td>
<td>31.54</td>
<td>15.73</td>
<td>-0.07229</td>
<td>7.71p</td>
<td>70.29p 31.61n</td>
</tr>
<tr>
<td>-0.8</td>
<td>-15.09</td>
<td>30.23</td>
<td>15.09</td>
<td>-0.05620</td>
<td>6.20p</td>
<td>43.68p 30.27n</td>
</tr>
<tr>
<td>-0.6</td>
<td>-14.47</td>
<td>28.99</td>
<td>14.47</td>
<td>-0.04400</td>
<td>6.00p</td>
<td>25.68p 29.02n</td>
</tr>
<tr>
<td>-0.4</td>
<td>-13.89</td>
<td>27.82</td>
<td>13.89</td>
<td>-0.03487</td>
<td>5.13p</td>
<td>13.63p 27.83n</td>
</tr>
<tr>
<td>-0.2</td>
<td>-13.33</td>
<td>26.69</td>
<td>13.33</td>
<td>-0.02807</td>
<td>1.93p</td>
<td>5.53p 26.70n</td>
</tr>
<tr>
<td>0.0</td>
<td>-12.80</td>
<td>25.62</td>
<td>12.80</td>
<td>-0.00775</td>
<td>12.25p</td>
<td>0.0002253a 25.62n</td>
</tr>
<tr>
<td>0.2</td>
<td>-14.84</td>
<td>24.44</td>
<td>14.84</td>
<td>5.232</td>
<td>-8.00p</td>
<td>1.046n 25.49n</td>
</tr>
<tr>
<td>0.4</td>
<td>-904.50</td>
<td>23.09</td>
<td>904.50</td>
<td>1786</td>
<td>90.0p</td>
<td>0.714u 0.737u</td>
</tr>
<tr>
<td>0.6</td>
<td>-281300</td>
<td>21.62</td>
<td>281300</td>
<td>562700</td>
<td>121.62n</td>
<td>0.337m 0.338m</td>
</tr>
<tr>
<td>0.8</td>
<td>-8952000</td>
<td>20.30</td>
<td>8952000</td>
<td>17900000</td>
<td>-3.9797u</td>
<td>13.72m 14.32m</td>
</tr>
<tr>
<td>1.0</td>
<td>-24550000</td>
<td>19.14</td>
<td>24550000</td>
<td>49110000</td>
<td>10.019u</td>
<td>44.59m 49.11m</td>
</tr>
</tbody>
</table>

*\(I_{SUM} = I_D + I_G + I_B - I_S\).

The results presented above indicate that the gate-tunneling currents are completely ignored by the built-in power measurement functions of both HSPICE and CADENCE-SPECTRE. The conformity of the results presented in Section 4.3 with the above observation is verified next. The data presented in Tables 4.1 and 4.2 indicate that the percent error in the power measurement with IN-CAD is significant for devices biased with zero or small gate-to-source voltages (\(|V_{GS}|\)). The currents that contribute to the power consumption at low gate-to-source voltages are the leakage currents. The relative significance of the gate tunneling leakage current in the total leakage current produced by the MOSFET contributes to the error observed in the power measurements.
with the IN-CAD. The percent error in the power measurement with IN-CAD, however, decreases with increasing gate-to-source voltages due to the reduced significance of the gate tunneling current in the total current produced by a MOSFET operating in the moderate to strong inversion regions.

The reason for the significant error with the IN-CAD at low temperatures is illustrated next with Fig. 4.10. The variation of the subthreshold leakage and gate-tunneling leakage currents with supply voltage and temperature is shown in Fig. 4.10 for an NMOS device. For an NMOS device operating at the nominal supply voltages ($V_{DD} = 1.0V$), the gate tunneling leakage current at -40°C is up to 4.3x higher as compared to the subthreshold leakage current, as shown in Fig. 4.10. Absolute value of the threshold voltage degrades as the temperature increases, as discussed in Chapter 3 [1], [28], [117]. Degradation of the threshold voltage and the enhancement of the thermal voltage exponentially increase the subthreshold leakage current at elevated temperatures, as shown in Fig. 4.10. Alternatively, the gate-tunneling leakage current is relatively less sensitive to temperature fluctuations [67]. At elevated temperatures, therefore, subthreshold leakage current dominates the total leakage current produced by a MOSFET, as shown in Fig. 4.10.

![Fig. 4.10. Comparison of subthreshold and gate-oxide leakage currents produced by an NMOS transistor for various supply voltages at three different temperatures. $V_{Gate} = V_{Source} = V_{Bulk} = 0V, V_{Drain} = V_{DD}$.](image-url)
The percent error observed in power consumption measured using the IN-CAD and the ACTUAL for a PMOS and an NMOS device with $|V_{GS}| = |V_{SB}| = 0$ and $|V_{DS}| = 1\text{V}$ is shown in Fig. 4.11 for different temperatures (data from Tables 4.1 and 4.2). The relative significance of the gate-tunneling current at lower temperatures introduces significantly higher error with the IN-CAD, as shown in Fig. 4.11. A similar trend is observed in the results shown in Figs. 4.6, 4.7, and 4.8 and the results listed in Tables 4.1, 4.2, 4.3, 4.4, 4.5, and 4.6. Furthermore, the gate tunneling current of a PMOS device is significantly smaller as compared to an NMOS device with similar physical dimensions (width, length, and $t_{ox}$) and similar voltage difference across the gate insulator in a Si-SiO$_2$ based CMOS technology, as explained in Chapter 2 [43], [67]. The error in the power measurement with the IN-CAD is therefore more significant for the NMOS devices, as listed in Tables 4.1, 4.2, 4.3, and 4.4 and as shown in Fig. 4.11.

![Graph showing percent error for different temperatures and device types](image)

Fig. 4.11. Percent error in the power measured using the HSPICE IN-CAD and the ACTUAL for devices in a 65nm CMOS technology with $|V_{GS}| = |V_{SB}| = 0$ and $|V_{DS}| = 1\text{V}$.

The percent error in the power measurements with the IN-CAD for an NMOS device with different body-bias voltages is shown in Fig. 4.12 (data is obtained from Table 4.4). $V_{BS}$ is the body-to-source voltage of the device. A positive (negative) $V_{BS}$ in Fig. 4.12 indicates that the NMOS device is forward (reverse) body-biased. Applying reverse body-bias increases the device threshold voltage thereby reducing the
subthreshold leakage current [1], [117]. The reduction in the subthreshold leakage current increases the relative significance of the gate tunneling leakage current thereby causing significant errors in the power measurements with the IN-CAD, as shown in Fig. 4.12. Alternatively, applying forward-body-bias lowers the device threshold voltage, thereby increasing the subthreshold leakage current [1], [117]. The increase in the subthreshold leakage current reduces the percent error observed in the power measurements with the IN-CAD, as shown in Fig. 4.12.

![Bar graph showing percent error at -40°C](image)

**Fig. 4.12.** The percent error in the power measurement with HSPICE IN-CAD for an NMOS device with different body-bias voltages.

The overall reliability of the data produced by the commercial circuit simulators is evaluated next with the Kirchhoff’s current law. In Fig. 4.9, current enters the device through the drain, the gate, and the bulk terminals and exits through the source terminal. According to the Kirchhoff’s current law, the sum of the currents entering and exiting the device must be equal. The algebraic sum of the currents at the different terminals of the devices shown in Fig. 4.9 are, however, not zero ($I_{SUM} = I_D + I_G + I_B - I_S$), as listed in Tables 4.7, 4.8, and 4.9. The violation of the Kirchhoff’s current law in these devices could be due to the inappropriate numerical resolution of the commercial circuit simulators and/or the inaccurate and inconsistent modeling of the different device currents due to different physical phenomenon (such as the gate-oxide leakage...
mechanisms versus the subthreshold conduction). A closer inspection of the data listed in Tables 4.7, 4.8, and 4.9 indicates that the power measurements with the IN-CAD are significantly off as compared to the actual power consumption (measured with the ACTUAL) when the magnitude of $I_{\text{SUM}}$ is comparable to the magnitude of the power measured with the IN-CAD. For example, as listed in Table 4.9, the magnitude of $I_{\text{SUM}}$ and the power measurements with the IN-CAD are comparable for the NMOS shown in Fig. 4.9a when the bulk terminal voltage ranges from -1.0V to 0.2V. The power measurements with the IN-CAD at these bulk voltages are underestimated by up to $113715046604527x$, as listed in Table 4.9. The results presented above indicate that the violation of the Kirchhoff’s current law can also contribute to the error introduced in the power measurements with the built-in power commands of the circuit simulators. Furthermore, both HSPICE and CADENCE-SPECTRE do not provide any warning messages when the Kirchhoff’s current law is violated. This study clearly uncovers an existing big question mark on the overall reliability of the data and the measurements produced by the commercial circuit simulators.

Accurate power estimation is critical in battery operated devices to predict the lifetime of the battery. Battery operated integrated circuits with very low temperature specifications are discussed in [64] and [65]. Prefabrication power estimation with the built-in command would significantly overestimate the battery life-time of these portable devices. Gate-oxide thickness is reduced with each new technology generation to enhance the control that the gate terminal exerts on the channel area, as explained in Chapter 2 [35]. Reducing the gate-oxide thickness increases the probability of carrier-tunneling through the thin insulator layer [67]. Gate-oxide leakage increases with technology scaling. The error introduced by the built-in power measurement functions will become more significant in the future technology generations. For deeply scaled nano-CMOS circuits with thin gate-oxides, power and energy measurement with the proposed generic methodology is therefore strongly recommended for an accurate pre-fabrication circuit characterization.
4.5. Chapter Summary

Computer-aided design tools are used for pre-fabrication characterization of integrated circuits. An important circuit parameter that requires accurate characterization is the power consumption due to the strict constraints on the acceptable power envelope of integrated systems. Circuit simulators typically provide built-in functions to measure the power consumption. However, the accuracy of the measured power is mostly overlooked since the approximations and the methodologies used by the existing built-in power estimation tools are not well documented.

A generic methodology to accurately measure the power and energy consumption of a circuit with the circuit simulators is described in this chapter. An equation to calculate the device power consumption based on the different current conduction paths in a MOSFET is presented. An expression for the total power consumption of a complex circuit is derived by explicitly including the voltages and currents at all the circuit terminals.

Results indicate that the power consumption is drastically underestimated with the built-in power estimation functions of two widely-used commercial circuit simulators CADENCE-SPECTRE and HSPICE. For the deeply scaled CMOS circuits in the nanometer regime, power and energy measurement with the proposed explicit methodology is strongly recommended for an accurate pre-fabrication circuit characterization.

The focus of this dissertation is the development of temperature aware power reduction techniques for nano-CMOS circuits. The actual power consumption and the estimated power reduction achievable with the techniques explored in this dissertation are measured using the generic methodology proposed in this chapter.
Chapter 5
Temperature Variation Insensitive CMOS Circuits

Integrated circuits are designed for functionality at the worst-case process and environmental parameter corners [1], [117]. Propagation delay of a circuit is primarily a function of the drain saturation current produced by active transistors [26], [44], [45], [69], [75]-[78]. Die temperature variations alter the current produced by MOSFETs operating at the prescribed nominal supply voltage [69], [78]. Variations in the drain current with temperature can cause significant fluctuations in the speed characteristics of CMOS circuits as discussed in Chapter 3 with various examples. The speed variation with temperature necessitates the verification of the timing constraints at different die temperatures [78].

Decreasing the sensitivity of the circuit speed to die temperature fluctuations is desirable for reducing the uncertainty in the propagation delay characteristics of CMOS circuits. Design methodologies for suppressing the drain current and propagation delay variations due to temperature fluctuations are described in this chapter. There exists a bias voltage at which the temperature fluctuation induced gate-overdrive variations counterbalance the carrier mobility variations experienced by the transistors when the temperature fluctuates [26], [75]-[78], [89], [90]. MOSFETs biased at this voltage produce temperature variation insensitive constant drain current [26], [75]-[78], [89], [90]. The bias voltages that achieve temperature variation insensitive drain current are identified for devices in 180nm and 65nm CMOS technologies. An alternative design methodology based on threshold voltage optimization for providing temperature variation insensitive speed is also evaluated [91]. The energy per switching cycle and the propagation delay at the supply and threshold voltages providing temperature variation insensitive circuit performance are compared.

The chapter is organized as follows. A design technique based on optimizing the supply voltage for achieving temperature variation insensitive circuit performance is described in Section 5.1. An alternative technique based on threshold voltage optimization is presented in Section 5.2. The energy efficiency and propagation delay
characteristics of the two techniques are compared in Section 5.3. A summary of the research results are provided in Section 5.4.

5.1. Supply Voltage Optimization for Temperature Variation Insensitive CMOS Circuits

Operating an integrated circuit at the prescribed nominal supply voltage is not preferable for reliable circuit operation under temperature fluctuations. As shown in Chapter 3, MOSFET drain current degrades by up to 10% and 41% for devices in 180nm and 65nm CMOS technologies, respectively, when the die temperature increases from 25°C to 125°C at the nominal supply voltage. The degradation of MOSFET current with temperature also reduces the circuit speed in both the 180nm and 65nm CMOS technologies [78]. In this section, a design methodology based on scaling the supply voltage for suppressing the drain current and propagation delay variations due to temperature fluctuations is described.

Mobility of the charge carriers is the dominant device parameter that determines the drain current variations when the die temperature fluctuates. In order to compensate for the variation of carrier mobility, the sensitivity of gate overdrive to temperature fluctuations should be enhanced. Scaling the supply voltage enhances the gate overdrive variations when the temperature fluctuates [28]. The temperature fluctuation induced gate overdrive variations for devices in a 65nm CMOS technology are presented in Table 5.1 at different supply voltages. As listed in Table 5.1, scaling the supply voltage from 1V to 0.3V enhances the gate-overdrive variations from 4.71% to 46.3%.

MOSFET drain current variation (I_DDS) with the temperature for transistors in 180nm and 65nm CMOS technologies is shown in Figs. 5.1 and 5.2, respectively [78]. Scaling the supply voltage reduces the sensitivity of the drain current to the die temperature variations in both the 180nm and 65nm CMOS technologies, as shown in Figs. 5.1 and 5.2 [75]-[78]. For a particular lower-supply-voltage (V_DD = 1.13V for PMOS and V_DD = 0.71V for NMOS in a 180nm CMOS technology - V_DD = 0.29V for PMOS and V_DD = 0.33V for NMOS in a 65nm CMOS technology), the temperature fluctuation induced gate overdrive variation completely counterbalances the carrier
mobility variation, thereby providing temperature variation insensitive constant MOSFET drain current, as shown in Figs. 5.1 and 5.2. $V_{DD,\text{insensitive}}$ is the bias voltage that achieves temperature variation insensitive constant drain current, as illustrated in Figs. 5.1 and 5.2.

**TABLE 5.1**

GATE OVERDRIVE VARIATIONS AT DIFFERENT SUPPLY VOLTAGES FOR DEVICES IN A 65NM CMOS TECHNOLOGY

<table>
<thead>
<tr>
<th>Supply Voltage (V)</th>
<th>Temperature ($^\circ$C)</th>
<th>Gate Overdrive (V)</th>
<th>PMOS</th>
<th>NMOS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>25</td>
<td>-0.78</td>
<td>0.78</td>
<td></td>
</tr>
<tr>
<td></td>
<td>125</td>
<td>-0.82</td>
<td>0.82</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Variation (%)</td>
<td>4.71</td>
<td>4.71</td>
<td></td>
</tr>
<tr>
<td>0.7</td>
<td>25</td>
<td>-0.48</td>
<td>0.48</td>
<td></td>
</tr>
<tr>
<td></td>
<td>125</td>
<td>-0.52</td>
<td>0.52</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Variation (%)</td>
<td>7.65</td>
<td>7.65</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>25</td>
<td>-0.28</td>
<td>0.28</td>
<td></td>
</tr>
<tr>
<td></td>
<td>125</td>
<td>-0.32</td>
<td>0.32</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Variation (%)</td>
<td>13.13</td>
<td>13.13</td>
<td></td>
</tr>
<tr>
<td>0.3</td>
<td>25</td>
<td>-0.08</td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td></td>
<td>125</td>
<td>-0.12</td>
<td>0.12</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Variation (%)</td>
<td>46.3</td>
<td>46.3</td>
<td></td>
</tr>
</tbody>
</table>

Similar to the MOSFET drain currents, the speed characteristics of CMOS circuits can also be made insensitive to temperature variations [28], [78]. The supply voltages that provide temperature variation insensitive speed characteristics for circuits in 180nm and 65nm CMOS technologies are shown in Figs. 5.3 and 5.4, respectively. Circuits display a temperature variation insensitive behavior when operated at a supply voltage 44% to 47% lower than the nominal supply voltage ($V_{DD,\text{nominal}} = 1.8V$) in a 180nm CMOS technology. Similarly, the supply voltages achieving temperature variation insensitivity are 67% to 68% lower than the nominal supply voltage ($V_{DD,\text{nominal}} = 1.0V$) for circuits in a 65nm CMOS technology. The supply voltages observed with the proposed supply voltage optimization technique are similar for a diverse set of circuits in
both technologies. The proposed technique of operating large scale designs at a single supply voltage for diminishing the performance sensitivity to temperature fluctuations is, therefore, feasible.

Fig. 5.1. Variation of the MOSFET drain current with supply voltage and temperature in the TSMC 180nm CMOS technology. $|V_{DS}| = |V_{GS}| = V_{DD}$ and $|V_{t0}(T_0)| = 0.46V$. $V_{DD,\text{insensitive}}$ for NMOSFET is 0.71V. $V_{DD,\text{insensitive}}$ for PMOSFET is 1.13V.

Fig. 5.2. Variation of the MOSFET drain current with supply voltage and temperature in the predictive 65nm CMOS technology. $|V_{DS}| = |V_{GS}| = V_{DD}$ and $|V_{t0}(T_0)| = 0.22V$. $V_{DD,\text{insensitive}}$ for NMOSFET is 0.33V. $V_{DD,\text{insensitive}}$ for PMOSFET is 0.29V.
5.2. Threshold Voltage Optimization Technique

The results presented in Section 5.1 indicate that temperature variation resilience in CMOS integrated circuits can be achieved by optimizing the supply voltage at which
the circuit is operated. A new design methodology based on optimizing the threshold voltages to suppress the propagation delay variations when the temperature fluctuates is described in this section.

For circuits operating at the nominal supply voltage, the gate overdrive sensitivity to temperature fluctuations can be enhanced by increasing the device threshold voltage [91]. The threshold voltages of MOSFETs can be altered during fabrication by varying the substrate doping density [92]-[94] and/or varying the gate-oxide thickness [95]. The variation of the drain current ($I_{DS}$) with the device threshold voltage ($|V_{t0}|$) for devices in 180nm and 65nm CMOS technologies is shown in Figs. 5.5 and 5.6, respectively. For a particular higher threshold voltage ($|V_{t0}| = 1.28V$ for PMOS and $|V_{t0}| = 1.58V$ for NMOS in a 180nm CMOS technology - $|V_{t0}| = 0.98V$ for PMOS and $|V_{t0}| = 0.93V$ for NMOS in a 65nm CMOS technology), the temperature fluctuation induced gate overdrive variation completely counterbalances the carrier mobility variation, thereby providing temperature variation insensitive constant MOSFET drain current, as shown in Figs. 5.5 and 5.6 [91]. Similar to the $V_{DD, insensitive}$ operation, MOSFETs with the $V_{t0, insensitive}$ are also insensitive to temperature fluctuations, as shown in Figs. 5.5 and 5.6.

![Fig. 5.5. Variation of the MOSFET drain current with threshold voltage and temperature in the TSMC 180nm CMOS technology. $|V_{DS}| = |V_{GS}| = V_{DD} = 1.8V$. $V_{t0, insensitive}$ for NMOSFET is 1.58V. $V_{t0, insensitive}$ for PMOSFET is 1.28V.](image-url)
Fig. 5.6. Variation of the MOSFET drain current with threshold voltage and temperature in the predictive 65nm CMOS technology. $|V_{DS}| = |V_{GS}| = V_{DD} = 1.0V$. $V_{t0,\text{insensitive}}$ for NMOSFET is 0.93V. $V_{t0,\text{insensitive}}$ for PMOSFET is 0.98V.

The device threshold voltages ($V_{t0,\text{insensitive}}$) that provide temperature variation insensitive speed are presented in Figs. 5.7 and 5.8. The NMOS and PMOS transistor threshold voltages are assumed to be equal ($V_{t0,\text{NMOS}} = |V_{t0,\text{PMOS}}|$) and scaled together in this study. As shown in Fig. 5.7, circuits operating at the nominal supply voltage in a 180nm CMOS technology are insensitive to temperature variations when the threshold voltage is 2.7x to 2.9x higher than the nominal device threshold voltage ($|V_{t0,\text{nominal}}| = 0.46V$). Similarly, the delay variations of circuits operating at the nominal supply voltage in a 65nm CMOS technology are suppressed when the device threshold voltage is 4.2x to 4.3x higher than the nominal device threshold voltage ($|V_{t0,\text{nominal}}| = 0.22V$), as shown in Fig. 5.8 [91].

5.3. Speed and Energy Efficiency of the Voltage Optimization Techniques

Sections 5.1 and 5.2 illustrate the supply and threshold voltage optimization techniques to achieve temperature variation insensitive CMOS speed characteristics. The
tradeoffs of the proposed voltage optimization techniques are presented in this section. The speed and energy consumption with the supply and threshold voltage optimization techniques are compared.

Fig. 5.7. Threshold voltages ($V_{t0,\text{insensitive}}$) that achieve temperature variation insensitive speed characteristics in the TSMC 180nm CMOS technology.

Fig. 5.8. Threshold voltages ($V_{t0,\text{insensitive}}$) that achieve temperature variation insensitive speed characteristics in the predictive 65nm CMOS technology.
The drain currents at the supply and threshold voltages achieving temperature variation insensitivity are marked in Figs. 5.1, 5.2, 5.5, and 5.6. In a 180nm CMOS technology, the PMOSFET and NMOSFET drain currents at the $V_{DD, insensitive}$ are $0.064\sqrt{A}$ and $0.061\sqrt{A}$, respectively, as shown in Fig. 5.1. Alternatively, at the $V_{t0, insensitive}$, the PMOSFET and NMOSFET drain currents are $0.060\sqrt{A}$ and $0.058\sqrt{A}$, respectively, as shown in Fig. 5.5. Similarly, even in the 65nm CMOS technology, the temperature variation insensitive drain current produced by a MOSFET with the threshold voltage optimization technique is smaller as compared to the drain current with the supply voltage optimization technique, as shown in Figs. 5.2 and 5.6. The speed penalty with the threshold voltage optimization technique will therefore be higher as compared to the supply voltage optimization technique.

The propagation delay of circuits with the supply and threshold voltage optimization techniques are shown in Figs. 5.3, 5.4, 5.7, and 5.8. As shown in Figs. 5.3 and 5.7, the speed of circuits with the supply voltage optimization technique is up to 3x higher, as compared to the speed at the threshold voltages achieving temperature variation insensitive circuit performance in a 180nm CMOS technology. Similarly, as shown in Figs. 5.4 and 5.8, the propagation delay is up to 25.6x higher when circuits in a 65nm CMOS technology are operated at $V_{t0, insensitive}$ as compared to the circuit operation at the $V_{DD, insensitive}$.

The energy consumption with the supply and threshold voltage optimization techniques are listed in Tables 5.2 and 5.3, respectively [78]. The energy per switching cycle at different supply and threshold voltages are normalized to the energy of the corresponding circuit at the nominal voltages (nominal $V_{DD}$ and $V_t$) and room temperature (25°C). Supply voltage scaling lowers both leakage and dynamic switching energy [1], [12], [26], [69], [78], [117]. Alternatively, increasing the device threshold voltage at the nominal supply voltage lowers only the leakage energy while maintaining the switching energy. The energy savings are, therefore, lower with the threshold voltage optimization technique as compared to the supply voltage optimization technique achieving temperature variation insensitive circuit speed [78].
TABLE 5.2
NORMALIZED ENERGY AT THE SUPPLY VOLTAGES PROVIDING TEMPERATURE VARIATION INSENSITIVE SPEED IN 180NM AND 65NM CMOS TECHNOLOGIES

<table>
<thead>
<tr>
<th>CMOS Technology</th>
<th>Temp (°C)</th>
<th>16-Bit Brent-Kung Adder</th>
<th>8-Bit Array Multiplier</th>
<th>16-Bit Ripple Carry Adder</th>
</tr>
</thead>
<tbody>
<tr>
<td>180nm</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VDD</td>
<td>0.96</td>
<td>1.01</td>
<td>0.96</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>0.27</td>
<td>0.27</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>125</td>
<td>0.27</td>
<td>0.28</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>65nm</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VDD</td>
<td>0.32</td>
<td>0.33</td>
<td>0.32</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>0.12</td>
<td>0.19</td>
<td>0.14</td>
<td></td>
</tr>
<tr>
<td>125</td>
<td>0.22</td>
<td>0.57</td>
<td>0.34</td>
<td></td>
</tr>
</tbody>
</table>

TABLE 5.3
NORMALIZED ENERGY WITH THE THRESHOLD VOLTAGES PROVIDING TEMPERATURE VARIATION INSENSITIVE SPEED IN 180NM AND 65NM CMOS TECHNOLOGIES

<table>
<thead>
<tr>
<th>CMOS Technology</th>
<th>Temp (°C)</th>
<th>16-Bit Brent-Kung Adder</th>
<th>8-Bit Array Multiplier</th>
<th>16-Bit Ripple Carry Adder</th>
</tr>
</thead>
<tbody>
<tr>
<td>180nm</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Vt0(T0)</td>
<td></td>
<td>1.26</td>
<td>1.33</td>
</tr>
<tr>
<td>25</td>
<td>0.84</td>
<td>0.77</td>
<td>0.83</td>
<td></td>
</tr>
<tr>
<td>125</td>
<td>0.84</td>
<td>0.78</td>
<td>0.83</td>
<td></td>
</tr>
<tr>
<td>65nm</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Vt0(T0)</td>
<td></td>
<td>0.94</td>
<td>0.92</td>
</tr>
<tr>
<td>25</td>
<td>0.90</td>
<td>0.79</td>
<td>0.85</td>
<td></td>
</tr>
<tr>
<td>125</td>
<td>0.91</td>
<td>0.83</td>
<td>0.86</td>
<td></td>
</tr>
</tbody>
</table>

At the supply and threshold voltages achieving temperature variation insensitive speed, the energy per switching cycle is reduced by up to 4x and 1.3x, respectively, as compared to the energy at the nominal voltages (VDD,nominal = 1.8V and |Vt0,nominal| = 0.46V) in a 180nm CMOS technology. Similarly, the energy consumption at the VDD,insensitive and Vt0,insensitive is lower by up to 8.6x and 1.3x, respectively, as compared to the energy at the nominal voltages (VDD,nominal = 1.0V and |Vt0,nominal| = 0.22V) for circuits
in a 65nm CMOS technology. The energy consumed at the threshold voltages achieving temperature variation insensitive speed is higher than the energy per cycle at the supply voltage providing temperature insensitivity by up to 3.3x and 7.7x for circuits in 180nm and 65nm CMOS technologies, respectively.

Higher supply voltages are preferable in speed critical applications. The performance degradation with the threshold voltage optimization technique diminishes the potential speed gains by employing a higher nominal supply voltage. Furthermore, energy savings achieved by the threshold voltage optimization technique is lower as compared to the supply voltage optimization technique. The supply voltage optimization technique is therefore more effective in simultaneously achieving energy efficiency and temperature variation resilience with a smaller speed penalty as compared to the threshold voltage optimizations in CMOS integrated circuits.

5.4. Chapter Summary

Mobility of the charge carriers is the dominant device parameter that determines the drain current variations when the die temperature fluctuates. In order to compensate for the variation of carrier mobility, the sensitivity of gate overdrive to temperature fluctuations should be enhanced. Supply and threshold voltage optimization techniques are described in this chapter for increasing the gate overdrive variations thereby achieving temperature variation insensitive drain current and circuit performance.

Scaling the supply voltage enhances the variations of the gate-overdrive when the temperature fluctuates. The supply voltages that achieve temperature variation insensitive speed characteristics are 44% to 47% lower than the nominal supply voltage in a 180nm CMOS technology ($V_{DD,nominal} = 1.8V$). Similarly, circuits in a 65nm CMOS technology display temperature variation insensitivity when operated at a voltage 67% to 68% lower than the nominal supply voltage ($V_{DD,nominal} = 1.0V$).

The gate overdrive variations due to temperature fluctuations can also be enhanced by increasing the device threshold voltages. Circuits operating at the nominal supply voltage in a 180nm CMOS technology are insensitive to temperature variations when the threshold voltage is 2.7x to 2.9x higher than the nominal device threshold voltage.
voltage ($|V_{t0,nominal}| = 0.46V$). Similarly, the delay variations of circuits operating at the nominal supply voltage in a 65nm CMOS technology are suppressed when the device threshold voltage is 4.2x to 4.3x higher than the nominal device threshold voltage ($|V_{t0,nominal}| = 0.22V$).

The speed penalty with the threshold voltage optimization technique is higher as compared to the supply voltage optimization technique. Furthermore, energy savings achieved at the threshold voltages achieving temperature variation insensitive speed is lower as compared to the energy savings observed at the supply voltage providing temperature variation insensitivity. The supply voltage optimization technique is therefore more effective in simultaneously achieving energy efficiency and temperature variation tolerance with a smaller speed penalty as compared to the threshold voltage optimizations in CMOS integrated circuits. The energy savings observed with the temperature-aware supply voltage optimization technique are compared with the other low power design methodologies in Chapter 6.
Chapter 6
Temperature-Aware Low Power Design

The primary sources of power dissipation in a full-voltage swing CMOS integrated circuit are the dynamic switching events and the leakage currents, as explained in Chapter 2. The dynamic switching power is consumed during the charging and discharging of parasitic capacitances when the node voltages transition. The dynamic power dissipation is directly related to the switching frequency [1], [117]. The dynamic power can therefore be adjusted to meet power budgets by adjusting the frequency of operation. Supply voltage scaling can also be used to adjust the dynamic power dissipation [1], [12], [117].

Alternatively, the leakage power is consumed when a circuit is idle. The subthreshold and the gate-tunneling leakage currents are the primary sources of leakage power in an idle CMOS circuit, as explained in Chapter 2. Leakage is unavoidable in modern CMOS technologies [7], [22]. Subthreshold leakage current increases exponentially with technology scaling (due to the reduction in the device threshold voltages) [1], [117]. Furthermore, the thinner gate-insulators employed with each new technology generation increase the gate tunneling leakage current [1], [117]. The increase in the leakage currents is perceived as a major roadblock for future CMOS technologies [7]. Design of integrated circuits has thus entered an era of power limited scaling where performance is not the only critical metric [10]-[12].

Power dissipation has become a primary design constraint requiring circuit optimizations to simultaneously consider energy and propagation delay. The effectiveness of the heat removal system poses limits on the power density of integrated systems. Effective heat removal requirements impact the system cost and the maximum attainable performance within a practical power envelope. Power constraints are even more stringent in mobile systems where longer battery lifetime is crucial. The goal of an integrated circuit is, therefore, typically to achieve a high operating frequency while meeting the energy consumption and power density constraints.
Supply voltage scaling is an effective technique to lower all the primary components of power consumption in CMOS circuits [1], [117]. As the supply voltage is reduced, the energy per cycle decreases while the propagation delay increases [102], [103]. The energy-delay product, therefore, has a minimum, as described in [102] and [103]. Furthermore, subthreshold operation minimizes the total energy consumption of a CMOS circuit [66]. In this chapter, the supply voltages that achieve minimum energy-delay product and minimum energy consumption are identified for circuits in 180nm and 65nm CMOS technologies. Results indicate that these supply voltages are lower than the prescribed nominal supply voltage. As shown in Chapter 5, the supply voltages that achieve temperature variation insensitive circuit performance are also lower than the nominal supply voltage in both 180nm and 65nm CMOS technologies. The speed and energy tradeoffs in circuits operating at the supply voltages that provide temperature variation insensitivity, minimum energy consumption, and minimum energy-delay product are compared.

The chapter is organized as follows. Low power designs aimed at minimizing the energy-delay product and energy consumption of CMOS circuits are discussed in Section 6.1. The speed and energy tradeoffs with the supply voltage optimization technique for temperature variation insensitive circuit performance are presented in Section 6.2. The summary of the research results are presented in Section 6.3.

**6.1. Low Power Design with Supply Voltage Scaling**

The results presented in Chapter 5 indicate that there is a supply voltage for which the speed characteristics of an integrated circuit is insensitive to temperature fluctuations. The supply voltage that achieves temperature variation insensitive circuit performance is lower than the nominal supply voltage in both 180nm and 65nm CMOS technologies. Integrated circuits operating at scaled supply voltages consume low power at the cost of reduced speed. The design methodology of optimizing the supply voltage for temperature variation insensitive circuit performance is, therefore, particularly attractive in low power applications with relaxed speed requirements.
Low power designs are typically aimed at reducing the energy consumption or the energy-delay product \cite{66}, \cite{102}, \cite{103}. Energy-delay product metric provides a good compromise between the need to reduce the energy consumption and the requirement to operate the circuits at an appropriate speed \cite{102}. The energy-delay product (EDP) is \cite{1}, \cite{15}, \cite{102}, \cite{117}

\[
EDP \approx \sum_j C_{eff,j} V_{DD}^2 T_g + \sum_j \sum_i I_{leak,j,i} V_{DD} T_g T_c,
\]

\[
I_{leak,j,i} = \frac{\mu_j W_{j,i} C_{OX}}{L_{eff}} V_{T,j,i}^2 e^{\frac{-V_{GS,j,i}}{n V_{T,j,i}}} \left(1 - e^{-V_{T,j,i}}\right),
\]

where $EDP$, $I_{leak}$, $V_{DD}$, $T_g$, $T_c$, $\mu$, $W$, $C_{OX}$, $L_{eff}$, $V_T$, $V_{GS}$, $V_{DS}$, and $n$ are energy-delay product, subthreshold leakage current, supply voltage, propagation delay of the circuit, clock period, carrier mobility, transistor width, oxide capacitance per unit area, effective channel length, threshold voltage with short-channel effects, thermal voltage, gate-to-source voltage, drain-to-source voltage, and subthreshold swing coefficient, respectively. $C_{eff}$ is the average effective switching capacitance of each gate that is extracted to include the average activity factor and the energy consumed due to short circuit current and glitches. The indices $j$ and $i$ cover all of the gates in the circuit and all of the transistors that determine the net subthreshold leakage current of each gate, respectively.

The normalized energy per cycle, delay, and energy-delay product as a function of the supply voltage is shown in Fig. 6.1 at the room temperature (Temp = 25°C) for an inverter in a 180nm CMOS technology. As the supply voltage is reduced, the energy per cycle decreases while the propagation delay increases \cite{66}, \cite{102}, \cite{103}. The energy-delay product given by equation 6.1, therefore, has a minimum, as shown in Fig. 6.1.

Energy per cycle and the propagation delay are also dependent on the die temperature \cite{66}, \cite{89}, \cite{102}, \cite{103}. As the temperature increases, energy consumed by a circuit increases primarily due to the increase in subthreshold leakage current \cite{1}, \cite{117}. Similarly, the propagation delay of circuits in current CMOS technologies increase when the temperature is increased at the nominal supply voltage, primarily due to the
degradation in carrier mobility as explained in Chapter 3. Circuits that operate with the nominal supply voltage, therefore, display the worst case energy-delay product at the maximum temperature for which the circuit is functional.

![Graph showing normalized energy per cycle, delay, and energy-delay product as a function of supply voltage.](image)

Fig. 6.1. Normalized energy per cycle, delay, and energy-delay product as a function of the supply voltage at the room temperature (Temperature = 25°C) for an inverter in the TSMC 180nm CMOS technology.

Speed characteristics of circuits are also dependent on the supply voltage [89], [101]. The temperature fluctuation induced gate overdrive and carrier mobility variations at different supply voltages for devices in a 180nm CMOS technology are presented in Table 6.1. As listed in Table 6.1, scaling the supply voltage enhances the sensitivity of gate-overdrive to temperature variations. At supply voltages below the supply voltage that yields temperature variation insensitive circuit speed, the gate overdrive variations dominate the carrier mobility variations when the temperature fluctuates. The MOSFET drain current and the circuit speed are, therefore, enhanced when the temperature is increased at the supply voltages that satisfy $V_{DD} < V_{DD,insensitive}$.

Energy-delay product at two different temperatures and percent delay variation when the temperature is increased from 25°C to 125°C for an 8-bit array multiplier in a 180nm CMOS technology are shown as a function of the supply voltage in Fig. 6.2. For the supply voltages above the supply voltage achieving temperature variation insensitive
speed, the delay variations are determined primarily by the mobility variations. As listed in Table 6.1, the percent variation in carrier mobility is similar for a specific temperature range at different supply voltages. Alternatively, the sensitivity of gate overdrive to temperature variations is enhanced with the scaling of the supply voltage. As the supply voltage is scaled below the $V_{DD, insensitive}$, the rate of increase in the delay variations (determined primarily by the variations of the gate overdrive for $V_{DD} < V_{DD, insensitive}$) is, therefore, enhanced, as shown in Fig. 6.2.

TABLE 6.1
GATE OVERDRIVE AND CARRIER MOBILITY VARIATIONS AT DIFFERENT SUPPLY VOLTAGES FOR DEVICES IN A 180NM CMOS TECHNOLOGY

<table>
<thead>
<tr>
<th>Supply Voltage (V)</th>
<th>Temperature (°C)</th>
<th>Gate Overdrive (V)</th>
<th>Carrier Mobility ($x10^{-3} \text{ m}^2/\text{Vs}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>PMOS</td>
<td>NMOS</td>
</tr>
<tr>
<td>1.8</td>
<td>25</td>
<td>-1.34</td>
<td>1.33</td>
</tr>
<tr>
<td></td>
<td>125</td>
<td>-1.41</td>
<td>1.39</td>
</tr>
<tr>
<td></td>
<td>Variation (%)</td>
<td>5.37</td>
<td>4.95</td>
</tr>
<tr>
<td>1.1</td>
<td>25</td>
<td>-0.64</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>125</td>
<td>-0.71</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>Variation (%)</td>
<td>11.28</td>
<td>10.48</td>
</tr>
<tr>
<td>0.7</td>
<td>25</td>
<td>-0.24</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td>125</td>
<td>-0.31</td>
<td>0.29</td>
</tr>
<tr>
<td></td>
<td>Variation (%)</td>
<td>30.39</td>
<td>29.01</td>
</tr>
<tr>
<td>0.5</td>
<td>25</td>
<td>-0.04</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>125</td>
<td>-0.11</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>Variation (%)</td>
<td>198.88</td>
<td>249.31</td>
</tr>
</tbody>
</table>

Due to the higher rate of change of delay below $V_{DD, insensitive}$, reduction in the propagation delay dominates the increase in energy in the EDP term as the temperature is increased. The energy-delay product at 125°C is, therefore, lower than the energy-delay product at 25°C, as illustrated in Fig. 6.2. Consequently, the worst case energy-delay product is observed at a lower temperature in ultra-low-voltage circuits operating at supply voltages below the supply voltage that achieves temperature variation insensitive circuit speed.
Fig. 6.2. The energy-delay product at two different temperatures and the percent delay variation as a function of the supply voltage. The temperature is increased from 25°C to 125°C for an 8-bit array multiplier in the TSMC 180nm CMOS technology.

Lowering the energy consumption per switching cycle is the primary goal rather than higher speed in some applications. The energy consumed per cycle is

\[
\text{Energy}_{\text{Total}} \approx \text{Energy}_{\text{Switching}} + \text{Energy}_{\text{Leakage}}
\]

\[
\text{Energy}_{\text{Switching}} \propto V_{DD}^2
\]

\[
\text{Energy}_{\text{Leakage}} = I_{\text{Leakage}} V_{DD} T_c
\]

where \(\text{Energy}_{\text{Total}}\), \(\text{Energy}_{\text{Switching}}\), \(\text{Energy}_{\text{Leakage}}\), and \(I_{\text{Leakage}}\) are the total energy consumed per cycle, total dynamic switching energy per cycle, total leakage energy per cycle, and total leakage current, respectively.

The normalized energy profile of a 16-bit Brent-Kung adder in a 180nm CMOS technology is shown as a function of the supply voltage at 25°C and 125°C in Figs. 6.3 and 6.4, respectively. Similarly, the energy profile of an adder in a 65nm CMOS technology at 25°C and 125°C is shown in Figs. 6.5 and 6.6, respectively. Scaling the
supply voltage reduces the dynamic switching energy, as given by equation 6.4. Scaling the supply voltage, however, also increases the total leakage energy per cycle as given by equation 6.5, due to the increase in the clock period [66]. The total energy consumption, therefore, has a minimum as shown in Figs. 6.3, 6.4, 6.5, and 6.6.

Fig. 6.3. Normalized switching, leakage, and total energy as a function of the supply voltage at 25°C for a 16-bit Brent-Kung adder in the TSMC 180nm CMOS technology.

Fig. 6.4. Normalized switching, leakage, and total energy as a function of the supply voltage at 125°C for a 16-bit Brent-Kung adder in the TSMC 180nm CMOS technology.
Fig. 6.5. Normalized switching, leakage, and total energy as a function of the supply voltage at 25°C for a 16-bit Brent-Kung adder in the predictive 65nm CMOS technology.

Fig. 6.6. Normalized switching, leakage, and total energy as a function of the supply voltage at 125°C for a 16-bit Brent-Kung adder in the predictive 65nm CMOS technology.

The supply voltage that provides minimum energy is determined by the relative significance of dynamic switching and leakage energy components. In a 180nm CMOS technology, the minimum energy consumption at 25°C is observed in the subthreshold
region \((V_{DD} < |V_{t0}| = 0.46V)\), as shown in Fig. 6.3. Leakage current increases at higher temperatures due to the reduction in device threshold voltages and the enhancement of the thermal voltage [1], [117]. The supply voltage that minimizes the energy consumption is higher for circuits with relatively higher leakage currents [66]. The energy consumption at 125°C is therefore minimized when the circuits in a 180nm CMOS technology are operated in the strong inversion region \((V_{DD} > |V_{t0}| = 0.46V)\), as shown in Fig. 6.4.

Supply voltage, threshold voltage, and gate-oxide thickness of MOSFETs are scaled with each new technology generation [1], [117]. Supply voltage scaling reduces the dynamic energy component. Alternatively, the scaling of threshold voltage and gate-oxide thickness increases the leakage energy, as explained in Chapter 2. The dynamic energy to leakage energy ratio is, therefore, reduced with each new technology generation. The increased leakage energy per switching cycle shifts the regime where the energy is minimized in a deeply scaled CMOS technology. For the circuits in this 65nm CMOS technology with significant leakage current, the minimum energy consumption is observed in the strong inversion region \((V_{DD} > |V_{t0}| = 0.22V)\) as shown in Figs. 6.5 and 6.6.

### 6.2. Energy Efficient Temperature Variation Resilient CMOS Circuits

The tradeoffs of attaining temperature variation resilience by operating a circuit at the \(V_{DD,\text{insensitive}}\) are discussed in this section. The energy and propagation delay characteristics at the supply voltages that yield temperature variation insensitive circuit performance, minimum energy-delay product, and minimum energy are compared.

The energy and propagation delay of circuits operating at the nominal supply voltage are presented in Table 6.2. The energy and propagation delay at the supply voltages that achieve temperature variation insensitive circuit performance are listed in Table 6.3 for circuits in both 180nm and 65nm CMOS technologies. Similarly, the energy and propagation delay at the supply voltages achieving minimum energy-delay product at 25°C and 125°C and at the supply voltages achieving minimum energy consumption at
25°C and 125°C are listed in Tables 6.4, 6.5, 6.6, and 6.7 [75], [76], [78]. The energy and the propagation delay listed in Tables 6.3, 6.4, 6.5, 6.6, and 6.7 are normalized to the energy and the propagation delay of the corresponding circuit at the room temperature (Temp = 25°C) and the nominal supply voltage [75], [76], [78].

### TABLE 6.2
DELAY AND ENERGY AT THE NOMINAL SUPPLY VOLTAGE IN 180NM AND 65NM CMOS TECHNOLOGIES

<table>
<thead>
<tr>
<th>CMOS Technology</th>
<th>Nominal Supply Voltage (V)</th>
<th>Temp (°C)</th>
<th>16-Bit Brent-Kung Adder</th>
<th>8-Bit Array Multiplier</th>
<th>16-Bit Ripple Carry Adder</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>25</td>
<td>1264.4</td>
<td>2126.2</td>
<td>2891.8</td>
</tr>
<tr>
<td>180nm</td>
<td></td>
<td>125</td>
<td>1423.2</td>
<td>2465.0</td>
<td>3238.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>25</td>
<td>1934.2</td>
<td>2732.7</td>
<td>1692.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>125</td>
<td>1966.6</td>
<td>2897.4</td>
<td>1722.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>25</td>
<td>633.3</td>
<td>1015.9</td>
<td>1470.7</td>
</tr>
<tr>
<td>65nm</td>
<td></td>
<td>125</td>
<td>978.6</td>
<td>1563.8</td>
<td>2257.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>25</td>
<td>467.7</td>
<td>676.5</td>
<td>363.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>125</td>
<td>506.8</td>
<td>852.5</td>
<td>415.8</td>
</tr>
</tbody>
</table>

### TABLE 6.3
NORMALIZED DELAY AND ENERGY AT THE SUPPLY VOLTAGES PROVIDING TEMPERATURE VARIATION INSENSITIVE CIRCUIT PERFORMANCE IN 180NM AND 65NM CMOS TECHNOLOGIES

<table>
<thead>
<tr>
<th>CMOS Circuits</th>
<th>Temp (°C)</th>
<th>Supply Voltage providing Temperature Variation Insensitive Circuit Performance (V_{DD,insensitive})</th>
<th>180nm CMOS technology</th>
<th>65nm CMOS technology</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>V_{DD} (V)</td>
<td>Delay*</td>
<td>E*</td>
</tr>
<tr>
<td>16-Bit Brent-Kung Adder</td>
<td>25</td>
<td>0.96</td>
<td>2.76</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>125</td>
<td>2.76</td>
<td>0.27</td>
<td></td>
</tr>
<tr>
<td>8-Bit Array Multiplier</td>
<td>25</td>
<td>1.01</td>
<td>2.90</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>125</td>
<td>2.90</td>
<td>0.28</td>
<td></td>
</tr>
<tr>
<td>16-Bit Ripple Carry Adder</td>
<td>25</td>
<td>0.96</td>
<td>2.73</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>125</td>
<td>2.73</td>
<td>0.25</td>
<td></td>
</tr>
</tbody>
</table>
TABLE 6.4
NORMALIZED DELAY AND ENERGY AT THE SUPPLY VOLTAGES PROVIDING MINIMUM ENERGY DELAY PRODUCT AT 25°C IN 180NM AND 65NM CMOS TECHNOLOGIES

<table>
<thead>
<tr>
<th>CMOS Circuits</th>
<th>Temp (°C)</th>
<th>Supply Voltage providing Minimum Energy Delay Product at 25°C</th>
<th>180nm CMOS technology</th>
<th>65nm CMOS technology</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>V_DD (V)</td>
<td>Delay*</td>
<td>E*</td>
</tr>
<tr>
<td>16-Bit Brent-Kung Adder</td>
<td>25</td>
<td>1.07</td>
<td>2.16</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td>125</td>
<td></td>
<td>2.25</td>
<td>0.34</td>
</tr>
<tr>
<td>8-Bit Array Multiplier</td>
<td>25</td>
<td>1.16</td>
<td>2.09</td>
<td>0.37</td>
</tr>
<tr>
<td></td>
<td>125</td>
<td></td>
<td>2.22</td>
<td>0.38</td>
</tr>
<tr>
<td>16-Bit Ripple Carry Adder</td>
<td>25</td>
<td>1.08</td>
<td>2.08</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td>125</td>
<td></td>
<td>2.18</td>
<td>0.33</td>
</tr>
</tbody>
</table>

TABLE 6.5
NORMALIZED DELAY AND ENERGY AT THE SUPPLY VOLTAGES PROVIDING MINIMUM ENERGY DELAY PRODUCT AT 125°C IN 180NM AND 65NM CMOS TECHNOLOGIES

<table>
<thead>
<tr>
<th>CMOS Circuits</th>
<th>Temp (°C)</th>
<th>Supply Voltage providing Minimum Energy Delay Product at 125°C</th>
<th>180nm CMOS technology</th>
<th>65nm CMOS technology</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>V_DD (V)</td>
<td>Delay*</td>
<td>E*</td>
</tr>
<tr>
<td>16-Bit Brent-Kung Adder</td>
<td>25</td>
<td>0.92</td>
<td>3.08</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>125</td>
<td></td>
<td>3.02</td>
<td>0.25</td>
</tr>
<tr>
<td>8-Bit Array Multiplier</td>
<td>25</td>
<td>1.01</td>
<td>2.90</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>125</td>
<td></td>
<td>2.90</td>
<td>0.28</td>
</tr>
<tr>
<td>16-Bit Ripple Carry Adder</td>
<td>25</td>
<td>0.90</td>
<td>3.22</td>
<td>0.22</td>
</tr>
<tr>
<td></td>
<td>125</td>
<td></td>
<td>3.12</td>
<td>0.22</td>
</tr>
</tbody>
</table>

Delay* - Normalized propagation delay. E* - Normalized energy consumption.
### TABLE 6.6
NORMALIZED DELAY AND ENERGY AT THE SUPPLY VOLTAGES PROVIDING MINIMUM ENERGY AT 25°C IN 180NM AND 65NM CMOS TECHNOLOGIES

<table>
<thead>
<tr>
<th>CMOS Circuits</th>
<th>Temp (°C)</th>
<th>V&lt;sub&gt;DD&lt;/sub&gt; (V)</th>
<th>Delay*</th>
<th>E*</th>
<th>V&lt;sub&gt;DD&lt;/sub&gt; (V)</th>
<th>Delay*</th>
<th>E*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>180nm CMOS technology</td>
<td></td>
<td></td>
<td>65nm CMOS technology</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16-Bit Brent-Kung Adder</td>
<td>25</td>
<td>0.26</td>
<td>10167.67</td>
<td>0.02</td>
<td>0.25</td>
<td>45.4</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td>125</td>
<td></td>
<td>742.07</td>
<td>1.06</td>
<td></td>
<td>35.8</td>
<td>0.29</td>
</tr>
<tr>
<td>8-Bit Array Multiplier</td>
<td>25</td>
<td>0.36</td>
<td>1826.80</td>
<td>0.04</td>
<td>0.34</td>
<td>14.8</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td>125</td>
<td></td>
<td>220.10</td>
<td>1.13</td>
<td></td>
<td>15.3</td>
<td>0.55</td>
</tr>
<tr>
<td>16-Bit Ripple Carry Adder</td>
<td>25</td>
<td>0.27</td>
<td>5389.38</td>
<td>0.03</td>
<td>0.31</td>
<td>18.9</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td>125</td>
<td></td>
<td>510.96</td>
<td>0.97</td>
<td></td>
<td>18.5</td>
<td>0.35</td>
</tr>
</tbody>
</table>

### TABLE 6.7
NORMALIZED DELAY AND ENERGY AT THE SUPPLY VOLTAGES PROVIDING MINIMUM ENERGY AT 125°C IN 180NM AND 65NM CMOS TECHNOLOGIES

<table>
<thead>
<tr>
<th>CMOS Circuits</th>
<th>Temp (°C)</th>
<th>V&lt;sub&gt;DD&lt;/sub&gt; (V)</th>
<th>Delay*</th>
<th>E*</th>
<th>V&lt;sub&gt;DD&lt;/sub&gt; (V)</th>
<th>Delay*</th>
<th>E*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>180nm CMOS technology</td>
<td></td>
<td></td>
<td>65nm CMOS technology</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16-Bit Brent-Kung Adder</td>
<td>25</td>
<td>0.47</td>
<td>84.92</td>
<td>0.06</td>
<td>0.34</td>
<td>12.7</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td>125</td>
<td></td>
<td>30.16</td>
<td>0.07</td>
<td></td>
<td>13.8</td>
<td>0.21</td>
</tr>
<tr>
<td>8-Bit Array Multiplier</td>
<td>25</td>
<td>0.59</td>
<td>34.30</td>
<td>0.08</td>
<td>0.47</td>
<td>4.3</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td>125</td>
<td></td>
<td>17.09</td>
<td>0.11</td>
<td></td>
<td>5.8</td>
<td>0.44</td>
</tr>
<tr>
<td>16-Bit Ripple Carry Adder</td>
<td>25</td>
<td>0.50</td>
<td>57.52</td>
<td>0.06</td>
<td>0.41</td>
<td>6.0</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td>125</td>
<td></td>
<td>23.39</td>
<td>0.08</td>
<td></td>
<td>7.6</td>
<td>0.30</td>
</tr>
</tbody>
</table>


As listed in Tables 6.4 and 6.5, the propagation delay of circuits in 180nm and 65nm CMOS technologies varies by -3.06% to +50.23% when the temperature is increased from 25°C to 125°C at the supply voltages providing minimum energy-delay
product. Similarly, as listed in Tables 6.6 and 6.7, the propagation delay varies by -92.7% to +32.8% in circuits of 180nm and 65nm CMOS technologies when the temperature is increased at the supply voltages that yield minimum energy. Therefore, similar to the high-speed integrated circuits operating at the nominal supply voltage, propagation delay of low power integrated circuits with deeply scaled supply voltages are also very sensitive to temperature fluctuations. In circuits optimized for minimum energy and minimum energy-delay product, within die temperature variations due to unbalanced switching activity would typically be small [78]. The primary source of temperature fluctuations in an ultra low-power circuit with a scaled supply voltage would be the variation in the ambient temperature. A change in the ambient temperature affects all the devices on a die similarly.

As listed in Tables 6.3, 6.4, and 6.5, the supply voltages that suppress the delay variations when the temperature fluctuates are similar to the supply voltages providing minimum energy-delay product in a 180nm CMOS technology. Alternatively, the supply voltages that yield minimum energy are lower than the $V_{DD,\text{insensitive}}$, as listed in Tables 6.3, 6.6, and 6.7. The propagation delay, as compared to the delay at the nominal supply voltage, is up to 3.2x and 2.9x longer when circuits in a 180nm CMOS technology are operated at the supply voltages for minimum energy-delay product ($V_{DD}$ optimized for minimum EDP at 125°C) and temperature variation insensitive circuit performance, respectively. At the supply voltages for minimum energy-delay product, the energy per cycle is 63.4% to 78.4% lower than the energy per cycle at the nominal supply voltage. Similarly, the energy per cycle at the supply voltages that yield temperature variation insensitive circuit performance is 72.9% to 75.2% lower than the energy at the nominal supply voltage.

The minimum energy-delay product is 23% to 40% lower than the energy-delay product at the nominal supply voltage. Similarly, the energy-delay product at the supply voltages that yield temperature variation insensitive circuit performance is 22% to 40% lower than the energy-delay product at the nominal supply voltage. The difference of the minimum achievable energy-delay product and the energy-delay product at the supply voltages for temperature variation insensitive circuit performance is less than 3%.
For circuits in a 65nm CMOS technology, the supply voltages that yield temperature variation insensitive circuit performance and minimum energy are lower than the supply voltages providing minimum energy-delay product, as listed in Tables 6.3, 6.4, 6.5, 6.6, and 6.7. When the circuits are operated at the supply voltages for minimum energy, the circuit speed is degraded by up to 45.4x as compared to the speed at the nominal supply voltage. Similarly, the propagation delay at the temperature variation insensitive supply voltages ($V_{DD,\text{insensitive}}$) is up to 17.6x longer than the delay at the nominal supply voltage ($V_{DD,\text{nominal}} = 1.0V$). The minimum achievable energy is 65% to 89% lower than the energy per switching cycle at the nominal supply voltage. Similarly, the energy at the temperature variation insensitive supply voltages ($V_{DD,\text{insensitive}}$) is 55% to 88% lower than the energy at the nominal supply voltage.

As illustrated with the data in Tables 6.4, 6.5, 6.6, and 6.7, low power integrated circuits optimized for minimum energy or minimum energy-delay product are highly sensitive to temperature fluctuations. Alternatively, integrated circuits with supply voltages optimized for temperature fluctuation insensitive speed characteristics also display significantly reduced energy consumption or energy-delay product as compared to the circuits operating with the nominal supply voltage. Energy efficiency and temperature fluctuation tolerance are therefore simultaneously achieved with the proposed supply voltage optimization technique as compared to the traditional margin based designs optimized for functionality at the worst case die temperature.

### 6.3. Chapter Summary

Power dissipation has become a primary design constraint, requiring circuit designs to be optimized considering both energy and delay. Low power designs are typically aimed at reducing the energy consumption or the energy-delay product. Design methodologies to optimize circuits for minimum energy-delay product and minimum energy consumption are described in this chapter. The scaled supply voltages that achieve minimum energy-delay product and minimum energy consumption are identified for circuits in 180nm and 65nm CMOS technologies.
As shown in Chapter 5, the supply voltages that achieve temperature variation insensitive CMOS circuit performance are lower than the nominal supply voltage in both 180nm and 65nm CMOS technologies. The speed and energy tradeoffs in operating the circuits at the supply voltages that provide temperature variation insensitivity, minimum energy consumption, and minimum energy-delay product are compared.

Low power integrated circuits optimized for minimum energy or minimum energy-delay product are highly sensitive to temperature fluctuations. Alternatively, integrated circuits with supply voltages optimized for temperature fluctuation insensitive speed characteristics also display significantly reduced energy consumption or energy-delay product as compared to the circuits operating with the nominal supply voltage. Energy efficiency and temperature fluctuation tolerance are therefore simultaneously achieved with the supply voltage optimization technique as compared to the traditional margin based designs optimized for functionality at the worst case die temperature.

The supply voltages providing minimum energy are observed in the subthreshold region. The switching current in these ultra-low-voltage circuits is the subthreshold leakage current. Subthreshold leakage current is extremely sensitive to temperature fluctuations. A small change in the die temperature exponentially alters the subthreshold leakage current. Furthermore, in these circuits, the circuit speed increases with the die temperature. The reversal in the temperature dependent propagation delay characteristics coupled with the high sensitivity of circuit speed to temperature fluctuations provides opportunities for reducing the energy consumption without degrading the clock frequency at elevated die temperatures in ultra-low-voltage circuits. Temperature adaptive voltage tuning techniques to reduce the high temperature energy consumption of ultra-low-voltage circuits is discussed in Chapter 7.
Chapter 7
High Temperature Energy Reduction in Low-Voltage Circuits

There is a growing interest in ultra-low-power design methodologies due to the increasing market demand for extended battery lifetimes in portable devices and self-sustaining energy-scavenging battery-replacement-free systems [1], [117]. Emerging applications with relatively low throughput requirements, such as distributed sensor networks, are typically aimed at lowering the energy consumption rather than achieving higher clock frequency. Scaling the supply voltage enhances the energy efficiency primarily by reducing the dynamic switching energy. As shown in Chapter 6, the supply voltages that provide minimum energy consumption are typically observed in the subthreshold region [66], [78], [119].

Integrated circuits with ultra-low-voltage power supplies are highly sensitive to process and temperature variations [120]. As the supply voltage is scaled to minimize the energy consumption, the supply voltage to threshold voltage ratio is reduced. The temperature fluctuation induced threshold voltage variations therefore determine the MOSFET drain current variations when the temperature fluctuates in circuits with extremely low power supply voltages, as explained in Chapter 5. Contrary to the standard-higher-voltage circuits designed for high-speed, low-voltage circuits optimized for minimum energy operate faster when the die temperature increases [101], [121].

Variations in the die temperature are caused by the imbalanced switching activity within a die and/or the fluctuations in the environmental temperature. In circuits optimized for minimum energy consumption, on-chip temperature gradients induced by imbalanced switching activity are typically small. Die temperature fluctuations due to the variations in the ambient temperature however can cause significant fluctuations in the speed and power characteristics of ultra-low-voltage circuits [78], [121].

Dynamic supply voltage scaling technique is primarily used for reducing the active mode power consumption of an integrated circuit by exploiting the variations in the computational workload [1], [11], [117], [122], [123]. Alternatively, adaptive body-bias technique reduces both the active and the standby mode power consumption by
dynamically adjusting the device threshold voltages depending on the variations of the workload and the circuit activity [1], [117], [123], [124]. In this chapter, a new temperature-adaptive dynamic supply voltage tuning technique is proposed for reducing the active-mode energy consumption by exploiting the excessive timing slack produced in the clock-period of ultra-low-voltage CMOS circuits at elevated temperatures. The high temperature energy efficiency is enhanced while maintaining a constant-clock-frequency by dynamically scaling the supply voltage of a subthreshold logic circuit. The supply voltages that lower the energy consumption without degrading the circuit speed at increased temperatures are identified for circuits in the TSMC 180nm CMOS technology [71]. An alternative technique based on temperature-adaptive threshold voltage tuning through reverse body-bias is also investigated. The active mode energy consumption characteristics of the two temperature-adaptive voltage tuning techniques are compared. The effectiveness of the proposed temperature-adaptive supply voltage tuning technique for high temperature energy efficiency is also evaluated under process parameter and supply voltage variations.

The chapter is organized as follows. A design methodology to identify the supply voltages providing minimum energy in the standard constant-$V_{DD}$ and constant-frequency systems is presented in Section 7.1. The new temperature-adaptive supply and threshold voltage scaling techniques for dynamically reducing the energy consumed at high die temperatures are described in Section 7.2. The energy characteristics of the temperature-adaptive schemes and the impact of the process parameter and supply voltage variations on the proposed methodologies are evaluated in Section 7.3. A summary of the presented results is provided in Section 7.4.

7.1. Supply Voltage Optimization for Minimum Energy Consumption

The impressive growth of the mobile products industry and the growing interest in the self-sustaining integrated systems with energy-scavenging capability have produced a significant interest in ultra-low-power circuit design. Power consumption of CMOS circuits can be lowered by employing several techniques as described in [1], [117], [122],
In this section, a design methodology for minimizing the energy consumption of CMOS circuits is described.

The two primary sources of power dissipation in CMOS circuits are the leakage power, which results from leakage currents produced by the MOSFETs (subthreshold, gate-tunneling, and junction leakage currents), and the dynamic power, which results from the switching activity. The energy consumed per cycle is

\[
    \text{Energy}_{\text{Total}} \approx \text{Energy}_{\text{Switching}} + \text{Energy}_{\text{Leakage}},
\]

\[
    \text{Energy}_{\text{Switching}} \propto V_{DD}^2,
\]

\[
    \text{Energy}_{\text{Leakage}} = I_{\text{Leakage}} V_{DD} T,
\]

where \(\text{Energy}_{\text{Total}}\), \(\text{Energy}_{\text{Switching}}\), \(\text{Energy}_{\text{Leakage}}\), \(I_{\text{Leakage}}\), \(V_{DD}\), and \(T\) are the total energy consumed per cycle, total dynamic switching energy per cycle, total leakage energy per cycle, total leakage current, supply voltage, and cycle time respectively. The energy efficiency of an integrated circuit (IC) can be enhanced by scaling the power supply voltage [75]. Supply voltage scaling quadratically reduces the dynamic switching energy, as given by (7.2). Scaling the supply voltage, however, also increases the total leakage energy per cycle as given by (7.3), due to the increase in the clock period [66]. The total energy consumed by an IC, therefore, has a minimum as the supply voltage is scaled.

Standard ICs are designed to operate with a constant-supply voltage (constant-\(V_{DD}\)) at a constant-frequency (constant-\(f_s\)) under different environmental conditions. An algorithm that optimizes the supply voltage of a standard constant-\(V_{DD}\) (with no supply voltage scaling capability) and constant-\(f_s\) (with no frequency scaling capability) IC for achieving minimum energy consumption is illustrated in Fig. 7.1, assuming a \(T_1 \rightarrow T_2\) die temperature spectrum. \(V_{DD,\text{nominal}}\), \(V_{DD,\text{min}}\), and \(V_{\text{step}}\) are the nominal supply voltage, supply voltage below which malfunction occurs, and the voltage scaling resolution, respectively. \(V_{DD,\text{nominal}}\) is technology dependent (1.8V for a 180nm CMOS technology). \(V_{\text{step}}\) is 10mV in this study. In the first iterative part of the algorithm, the supply voltage is scaled with a voltage resolution of \(V_{\text{step}}\). The highest constant clock frequency that can be maintained within the entire temperature spectrum is identified for each supply voltage. In the second
part of the algorithm, the energy consumed by the circuit is measured at various temperatures of interest for each pair of supply voltage and the corresponding highest achievable clock frequency. From the measured energy consumption, the constant-supply-voltage that achieves minimum energy at a specific temperature (within the temperature spectrum) is identified assuming a standard constant-$V_{DD}$ and constant-$f_s$ circuit operation.

Fig. 7.1. Flow-chart for identifying the supply voltage that achieves minimum energy consumption at a specific temperature ($T_{specific}$) for a standard constant-$V_{DD}$ and constant-$f_s$ IC.
The above methodology to measure the maximum frequency \( f_{\text{max}} \) achievable with a circuit at a specific supply voltage and temperature is illustrated here using the input/output waveforms shown in Fig. 7.2. Circuits can be either inverting or non-inverting. The input and output waveforms of an inverting circuit are shown in Fig. 7.2. The integrated circuit is initially operated at a low-frequency \( f << f_{\text{max}} \) where \( f_{\text{max}} \) has to be determined. \( \text{Time}_1 \) is the time taken for the falling (rising) output signal to cross 0.1*\( V_{\text{DD}} \) (0.9*\( V_{\text{DD}} \)) after the rising input signal crosses 0.1*\( V_{\text{DD}} \) (0.1*\( V_{\text{DD}} \)) in an inverting (non-inverting) circuit. Similarly, \( \text{Time}_2 \) is the time taken for the rising (falling) output signal to cross 0.9*\( V_{\text{DD}} \) (0.1*\( V_{\text{DD}} \)) after the falling input signal crosses 0.9*\( V_{\text{DD}} \) (0.9*\( V_{\text{DD}} \)) in an inverting (non-inverting) circuit. A 20% margin is added to the maximum of \( \text{Time}_1 \) and \( \text{Time}_2 \) to provide a timing-slack against parameter variations, clock-skew, and clock-jitter. The maximum frequency of a circuit at a specific supply voltage and temperature is

\[
f_{\text{max}} = \frac{1}{2 \times \left[ \max(\text{Time}_1, \text{Time}_2) + 0.2 \times \max(\text{Time}_1, \text{Time}_2) \right]}.
\]  

(7.4)

To find the highest achievable constant-clock-frequency at a particular supply voltage, the maximum achievable frequencies \( f_{\text{max}} \) at the extremes of the die temperature spectrum (\( T_1 \) and \( T_2 \)) are measured using the above procedure. The smaller of the two frequencies is the highest constant-frequency that can be maintained by the circuit within the entire temperature spectrum (\( T_1 \rightarrow T_2 \)) at the particular supply voltage.

The results of the algorithm are listed in Table 7.1 for a 16-bit Brent-Kung adder in a 180nm CMOS technology. The die temperature spectrum is assumed to be from 25°C to 125°C. The standard constant-supply voltages for achieving minimum energy consumption at 25°C and 125°C are reported. As listed in Table 7.1, the temperature that determines the highest operating frequency is also dependent on the supply voltage of the circuit. At the higher supply voltages (such as the \( V_{\text{DD, nominal}} = 1.8V \)), circuits operate slower when the die temperature increases. The maximum achievable (worst-case) frequency is therefore determined by plugging the low-to-high and high-to-low critical path delays observed at the highest temperature into (7.4). Alternatively, as the supply voltage is scaled, the worst-case speed shifts to the lowest operating temperature, as listed
in Table 7.1, due to the determination of the propagation delay characteristics primarily by the gate overdrive variations of the MOSFETs below a specific $V_{DD}$ (0.97V for this Brent-Kung adder). The maximum achievable clock frequency for an entire temperature spectrum is therefore determined by the critical path delays observed at the lowest temperature for $V_{DD} \leq 0.97V$. $V_{DD-25}$ and $V_{DD-125}$ are the constant-supply voltages applied to a standard CMOS circuit (without any voltage tuning capability) for achieving minimum energy consumption at 25°C and 125°C, respectively. The supply voltage that provides minimum energy consumption varies with the operating temperature. The energy consumption of the 16-bit Brent-Kung adder at different temperatures along with the supply voltage that minimizes the energy consumption at each temperature is shown in Fig. 7.3. The supply voltage that provides minimum energy is determined by the relative significance of dynamic switching and leakage energy components [66].

![Fig. 7.2. The input and output waveforms of an inverting circuit.](image-url)
**TABLE 7.1**

**SUPPLY VOLTAGES THAT ACHIEVE MINIMUM ENERGY IN A CONSTANT-$V_{DD}$ AND CONSTANT-$f_s$ BRENT-KUNG ADDER**

<table>
<thead>
<tr>
<th>$V_{DD}$ (V)</th>
<th>Max. frequency at 25°C (MHz)</th>
<th>Max. frequency at 125°C (MHz)</th>
<th>Worst-case frequency (MHz)</th>
<th>Energy consumption at the worst-case frequency and 25°C (pJ)</th>
<th>Energy consumption at the worst-case frequency and 125°C (pJ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.80</td>
<td>296.56</td>
<td>265.14</td>
<td>265.14</td>
<td>2.0000</td>
<td>2.0400</td>
</tr>
<tr>
<td>1.20</td>
<td>173.96</td>
<td>162.68</td>
<td>162.68</td>
<td>0.8370</td>
<td>0.8510</td>
</tr>
<tr>
<td>0.99</td>
<td>117.57</td>
<td>116.84</td>
<td>116.84</td>
<td>0.5531</td>
<td>0.5637</td>
</tr>
<tr>
<td>0.98</td>
<td>115.62</td>
<td>115.23</td>
<td>115.23</td>
<td>0.5464</td>
<td>0.5524</td>
</tr>
<tr>
<td>0.97</td>
<td>110.78</td>
<td>111.99</td>
<td>110.78</td>
<td>0.5344</td>
<td>0.5354</td>
</tr>
</tbody>
</table>

Supply voltage below which the circuit exhibits reverse temperature dependence:

<table>
<thead>
<tr>
<th>$V_{DD}$</th>
<th>Max. frequency at 25°C (MHz)</th>
<th>Max. frequency at 125°C (MHz)</th>
<th>Worst-case frequency (MHz)</th>
<th>Energy consumption at the worst-case frequency and 25°C (pJ)</th>
<th>Energy consumption at the worst-case frequency and 125°C (pJ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.49</td>
<td>3.91</td>
<td>11.00</td>
<td>3.91</td>
<td>0.1200</td>
<td>0.1500</td>
</tr>
<tr>
<td>0.48</td>
<td>3.26</td>
<td>9.94</td>
<td>3.26</td>
<td>0.1160</td>
<td>0.1490</td>
</tr>
<tr>
<td>0.47</td>
<td>2.70</td>
<td>8.87</td>
<td>2.70</td>
<td>0.1110</td>
<td>0.1480</td>
</tr>
<tr>
<td>0.46</td>
<td>2.37</td>
<td>7.90</td>
<td>2.37</td>
<td>0.1060</td>
<td>0.1500</td>
</tr>
<tr>
<td>0.45</td>
<td>1.94</td>
<td>6.86</td>
<td>1.94</td>
<td>0.1020</td>
<td>0.1520</td>
</tr>
</tbody>
</table>

$V_{DD-25}$ $V_{DD-125}$ Minimum energy at 25°C Minimum energy at 125°C

<table>
<thead>
<tr>
<th>$V_{DD-25}$</th>
<th>Minimum energy at 25°C</th>
<th>$V_{DD-125}$</th>
<th>Minimum energy at 125°C</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.28</td>
<td>0.052</td>
<td>0.620</td>
<td>0.052</td>
</tr>
<tr>
<td>0.27</td>
<td>0.041</td>
<td>0.530</td>
<td>0.041</td>
</tr>
<tr>
<td>0.26</td>
<td>0.032</td>
<td>0.45</td>
<td>0.032</td>
</tr>
<tr>
<td>0.25</td>
<td>0.025</td>
<td>0.39</td>
<td>0.025</td>
</tr>
<tr>
<td>0.24</td>
<td>0.020</td>
<td>0.33</td>
<td>0.020</td>
</tr>
</tbody>
</table>

* Results are for a Brent-Kung adder in a 180nm CMOS technology.

Absolute value of the threshold voltage degrades as the temperature increases [73], [74]. Degradation of the threshold voltage coupled with the enhancement of the thermal voltage exponentially increase the subthreshold leakage current at higher temperatures. The supply voltages that minimize the energy consumption are higher for circuits with relatively higher leakage currents [66]. Minimum energy at an elevated temperature is therefore observed at a higher supply voltage, as listed in Table 7.1 and as shown in Fig. 7.3. The algorithm is executed on multiple test circuits and the supply voltages that minimize the energy consumption at 25°C and 125°C for a standard constant-$V_{DD}$ and constant-$f_s$ circuit operation are reported in Table 7.2.
The supply voltages providing minimum energy are observed in the subthreshold region, as listed in Table 7.2 [66], [119]. The switching current in these ultra-low-voltage circuits is the subthreshold leakage current. Subthreshold leakage current is extremely sensitive to temperature fluctuations. A small change in the die temperature exponentially alters the subthreshold leakage current. The reversal in the temperature dependent propagation delay characteristics coupled with the high sensitivity of circuit speed to temperature fluctuations provides opportunities for reducing the energy consumption without degrading the clock frequency at elevated die temperatures in ultra-low supply-voltage circuits.
### TABLE 7.2
SUPPLY VOLTAGES FOR ACHIEVING MINIMUM ENERGY CONSUMPTION IN STANDARD CONSTANT-$V_{DD}$ AND CONSTANT-$f_S$ CIRCUITS

<table>
<thead>
<tr>
<th>Circuits in a 180nm CMOS Technology</th>
<th>Supply Voltages (V)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$V_{DD-25}$</td>
</tr>
<tr>
<td>16-Bit Ripple Carry Adder</td>
<td>0.27</td>
</tr>
<tr>
<td>16-Bit Carry Select Adder</td>
<td>0.32</td>
</tr>
<tr>
<td>16-Bit Brent-Kung Adder</td>
<td>0.26</td>
</tr>
<tr>
<td>8-Bit Array Multiplier</td>
<td>0.36</td>
</tr>
</tbody>
</table>

### 7.2. Techniques for High Temperature Energy Reduction

In this section, the previously proposed conventional voltage scaling and body-bias techniques are briefly discussed. Two new temperature-adaptive dynamic voltage tuning techniques for enhancing the high temperature active mode energy efficiency of circuits operating at ultra-low supply voltages are then introduced.

The operational load for an integrated circuit tends to have peak performance requirements followed by idle periods [1], [117]. Maintaining the full computational capacity at all times, despite the reduction of the throughput requirements with the variations of the workload, wastes significant amount of energy. Dynamic supply voltage scaling technique exploits the variations in the computational workload by dynamically adjusting the supply voltage and the clock frequency of a synchronous system. The primary objective of dynamic supply voltage scaling technique is to provide high throughput during the execution of only the computation-intensive tasks while saving energy during the rest of the time by lowering the supply voltage and the operating clock frequency. The dynamic voltage scaling technique is primarily aimed at reducing the active mode power consumption of an integrated circuit.

Alternatively, the adaptive body-bias techniques utilize the bulk terminal to dynamically modify the threshold voltages of the devices during circuit operation. Depending upon the polarity of the voltage difference between the source and the body
terminals ($V_{SB}$), the threshold voltage can be either increased or decreased as compared to a zero-body-biased transistor. Device threshold voltages can be increased by applying reverse body bias in the standby mode in order to reduce the subthreshold leakage current produced by idle circuits. Furthermore, the dynamic supply voltage scaling and adaptive body-bias techniques can also be used to compensate for the die-to-die and within-die process parameter variations, thereby enhancing the yield [1], [117].

In ultra-low-voltage circuits, temperature gradients due to imbalanced switching activity within a die are typically small. The primary source of die temperature fluctuations in low-voltage circuits are the variations in the ambient temperature. Changes in the ambient temperature tend to affect all the devices in an IC. At elevated die temperatures, the leakage currents as well as the circuit speed are enhanced. Increased leakage power, in turn, further enhances the heat dissipation and elevates the die temperature. This positive feedback between the die temperature, the leakage current, and the total power consumption significantly reduces the battery lifetime in portable devices, accelerates the degradation of the device/circuit reliability due to excessive heating, and can even cause thermal runaway in extreme environments despite the relatively low-supply-voltage. New temperature-adaptive design methodologies are, therefore, highly desirable to enhance the reliability and energy efficiency of ultra-low-voltage circuits operating at environments subject to significant temperature fluctuations.

Integrated circuits are typically designed for guaranteed functionality at the estimated worst-case process and environmental parameter corners. In constant-$V_{DD}$ and constant-$f$, circuits optimized for minimum energy, the worst-case speed is observed at the lowest operating temperature. The lowest temperature therefore determines the achievable maximum clock frequency. As the die temperature increases, the circuits operate faster, thereby producing significant timing slack in the constant clock period.

New temperature-adaptive supply and threshold voltage tuning techniques are proposed in this chapter to dynamically adjust the circuit speed based on the die temperature. The primary objective of the proposed temperature-adaptive schemes is to lower the active-mode energy consumption by exploiting the excessive timing slack produced in the clock period at high die temperatures while maintaining a constant-clock frequency across an entire die temperature spectrum. The objective is achieved by either
dynamically scaling the power supply voltage or dynamically increasing the device threshold voltages through reverse body-bias at elevated temperatures. The temperature-adaptive supply voltage tuning and the temperature-adaptive reverse body-bias techniques are presented in Sections 7.2.1 and 7.2.2, respectively.

### 7.2.1. Temperature-Adaptive Dynamic Supply Voltage Scaling

The temperature-adaptive dynamic supply voltage scaling technique (TA-DVS) is presented in this section. All the primary components of power consumption in a CMOS circuit, namely dynamic switching, short-circuit, and leakage power are significantly reduced by scaling the supply voltage. The propagation delay of a circuit is also strongly dependent on the supply voltage [1], [117]. Scaling the supply voltage reduces the power consumed by a circuit at the cost of degraded circuit performance.

The sum output (SUM[15]) of a 16-bit Brent-Kung adder operating at the constant-$V_{DD-25}$ and various temperatures is shown in Fig. 7.4. $V_{DD-25}$ for the Brent-Kung adder is 0.26V, as listed in Table 7.1. The clock frequency is fixed at 32KHz in a standard CMOS circuit (determined by the lowest operating temperature), the highest constant-$f_s$ that can be maintained at all die temperatures, as listed in Table 7.1. The circuit operates faster at elevated temperatures, thereby producing excessive timing slack in the constant-clock-period, as shown in Fig. 7.4. At elevated temperatures the total energy consumption increases due to the increase in the subthreshold leakage current [1], [117]. The significant timing slack in the clock-period can be exploited to reduce the active-mode energy consumption at elevated temperatures. With the proposed technique, the supply voltage of the circuit is dynamically scaled below $V_{DD-25}$ while maintaining the constant-clock frequency of the circuit as the die temperature increases. The supply voltage of the circuit is tuned until the high temperature circuit performance at the scaled supply voltage matches the circuit performance of the standard constant-$V_{DD}$ and constant-$f_s$ circuit operating at $V_{DD-25}$. Unlike the conventional work-load adaptive dynamic voltage scaling techniques [1], [11], [117], [122], [123], a new die-temperature-adaptive dynamic voltage scaling technique (TA-DVS) is proposed in this chapter for tuning the circuit supply voltage based on the fluctuations of the die temperature and the circuit speed.
The switching current in ultra-low-voltage circuits is the subthreshold leakage current. Supply voltage scaling in the subthreshold regime reduces the $I_{on}/I_{off}$ ratio [125]. Below a certain supply voltage, circuits with reduced $I_{on}/I_{off}$ ratio fail to produce output signals with full rail-to-rail voltage swing [125]. The high temperature signal swing of the sum output (SUM[15]) of a 16-bit Brent-Kung adder at various scaled supply voltages is shown in Fig. 7.5. As the supply voltage is scaled, the signal swing starts to degrade, eventually causing malfunction at 110mV, as shown in Fig. 7.5. The extent of temperature-adaptive dynamic voltage tuning that can be performed with the proposed technique is therefore limited by the acceptable degradation of the output signal voltage swing as well as the circuit speed criterion determined by the lowest operating temperature. Supply voltages that achieve at least a $0.1V_{DD} \rightarrow 0.9V_{DD}$ output voltage swing for the entire temperature range are considered to be fully-functional in this study. Further reduction in the supply voltage lowers the high temperature energy consumption at the cost of unacceptable degradation in the output voltage waveforms and the circuit noise margins.
For the standard constant-$V_{\text{DD}}$ circuits designed for minimum energy at 125°C (supply voltage fixed at $V_{\text{DD-125}}$), the worst-case circuit speed is similarly observed at the lowest temperature, as listed in Table 7.1. The maximum constant operating frequency that can be maintained for the entire die temperature spectrum is, therefore, determined by the lowest temperature in these standard circuits. Similar to the circuits operating at $V_{\text{DD-25}}$, as the die temperature increases, the propagation delays are significantly reduced in a circuit that operates at $V_{\text{DD-125}}$. With the proposed temperature-adaptive supply voltage tuning technique, without violating the constant-clock frequency requirement, the speed of these circuits at elevated temperatures can be dynamically adjusted for exploiting the enhanced timing slack in the clock period. Temperature-adaptive supply voltage tuning technique thereby further reduces the high-temperature energy consumption as compared to even a standard constant-$V_{\text{DD}}$ circuit designed for minimum energy operation at 125°C.

The high temperature energy reduction observed with the proposed temperature-adaptive supply voltage tuning technique in circuits optimized for minimum energy at 125°C is illustrated next for a 16-bit Brent-Kung adder. $V_{\text{DD-125}}$ for the Brent-Kung adder is 0.47V, as listed in Table 7.1. The frequency of the circuit is fixed at 2.7MHz, the
highest constant-$f_s$ that can be maintained at all die temperatures for $V_{DD} = 0.47V$. At 125°C, however, this adder is actually capable of operating at a clock frequency of up to 8.87MHz with this supply voltage. In a standard CMOS circuit, since the $V_{DD}$ and frequency are fixed at 0.47V and 2.7MHz, respectively, the available clock-period is essentially under-utilized at elevated temperatures.

With the proposed technique, the supply voltage of the circuit is further scaled at elevated temperatures ($V_{DD} < V_{DD-125}$) to exploit the excessive slack observed in the clock-period while maintaining the functionality at 2.7MHz. The high-temperature maximum clock frequency and energy consumption of a Brent-Kung adder is listed in Table 7.3 for various power supply voltages. As listed in Table 7.3, at 125°C the supply voltage can be scaled from 0.47V ($V_{DD-125}$) to 0.39V while maintaining the clock frequency at 2.7MHz. Scaling the supply voltage reduces the high-temperature energy consumption from 0.148pJ (the minimum energy achievable with $V_{DD}$ fixed at $V_{DD-125} = 0.47V$) to 0.108pJ ($V_{DD} = 0.39V$), as listed in Table 7.3. Supply voltage scaling at elevated temperatures thereby significantly lowers the energy consumption below the minimum energy achievable with the standard constant-$V_{DD}$ and constant-$f_s$ circuits.

The propagation delays of standard CMOS circuits operating at $V_{DD-25}$ and $V_{DD-125}$ are compared with the high-temperature propagation delays of circuits based on the TA-DVS technique in Tables 7.4 and 7.5 respectively. The scaled supply voltages of the TA-DVS circuits listed in Tables 7.4 and 7.5 are the minimum $V_{DD}$ that achieves at least a $0.1V_{DD} \rightarrow 0.9V_{DD}$ output voltage swing while maintaining the low temperature clock frequency at elevated temperatures. As listed in Table 7.4, the scaled supply voltages that maintain the clock frequency at high temperatures are 29.6% (ripple carry adder) to 38.9% (array multiplier) lower as compared to the supply voltages required by the standard constant-$V_{DD}$ circuits for achieving minimum energy at 25°C ($V_{DD-25}$). Similarly, with the proposed technique, the supply voltage at elevated temperatures can be scaled by up to 17% (Brent-Kung adder) as compared to the supply voltage required by the standard constant-$V_{DD}$ circuits for achieving minimum energy at 125°C ($V_{DD-125}$), as listed in Table 7.5.
### TABLE 7.3
MAXIMUM-\(f_s\) AND ENERGY CONSUMPTION FOR A BRENT-KUNG ADDER AT DIFFERENT SUPPLY VOLTAGES (TEMPERATURE = 125°C)

<table>
<thead>
<tr>
<th>(V_{DD}) (V)</th>
<th>Max. frequency at 125°C (MHz)</th>
<th>Energy consumption at 125°C (pJ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.47 ((V_{DD-125}))</td>
<td>8.87</td>
<td>0.148</td>
</tr>
<tr>
<td>0.45</td>
<td>6.86</td>
<td>0.138</td>
</tr>
<tr>
<td>0.43</td>
<td>5.45</td>
<td>0.127</td>
</tr>
<tr>
<td>0.41</td>
<td>4.18</td>
<td>0.117</td>
</tr>
<tr>
<td>0.39</td>
<td>3.20</td>
<td>0.108</td>
</tr>
<tr>
<td>0.37</td>
<td>2.40</td>
<td>0.098</td>
</tr>
</tbody>
</table>

* The target clock frequency at \(V_{DD-125}\) is 2.7MHz. At elevated temperatures, the supply voltage can be scaled while maintaining the clock frequency.

### TABLE 7.4
PROPAGATION DELAY COMPARISON OF CONSTANT-\(V_{DD}\) CIRCUITS AT \(V_{DD-25}\) AND CIRCUITS WITH TA-DVS CAPABILITY

<table>
<thead>
<tr>
<th>Circuits in a 180nm CMOS Technology</th>
<th>Standard constant voltage operation at (V_{DD-25})</th>
<th>TA-DVS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(V_{DD}) (V)</td>
<td>PD (25°C, 125°C)</td>
</tr>
<tr>
<td>16-Bit Ripple Carry Adder</td>
<td>0.27</td>
<td>15.59, 1.48</td>
</tr>
<tr>
<td>16-Bit Carry Select Adder</td>
<td>0.32</td>
<td>3.40, 0.32</td>
</tr>
<tr>
<td>16-Bit Brent Kung Adder</td>
<td>0.26</td>
<td>10.10, 0.81</td>
</tr>
<tr>
<td>8-Bit Array Multiplier</td>
<td>0.36</td>
<td>3.89, 0.47</td>
</tr>
</tbody>
</table>

The normalized high temperature energy consumption of the standard constant-\(V_{DD}\) circuits and the circuits based on the TA-DVS technique are listed in Table 7.6. The high temperature energy consumption is reduced by up to 40% (carry select adder) and 28% (Brent-Kung adder) with the temperature-adaptive dynamic supply voltage tuning technique as compared to the standard constant-\(V_{DD}\) circuits providing minimum energy at 25°C (\(V_{DD-25}\)) and 125°C (\(V_{DD-125}\)), respectively.
TABLE 7.5
PROPAGATION DELAY COMPARISON OF CONSTANT-$V_{DD}$ CIRCUITS AT
$V_{DD-125}$ AND CIRCUITS WITH TA-DVS CAPABILITY

<table>
<thead>
<tr>
<th>Circuits in a 180nm CMOS Technology</th>
<th>Standard constant voltage operation at $V_{DD-125}$</th>
<th>TA-DVS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$V_{DD}$ (V)</td>
<td>PD</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$25^\circ C$</td>
</tr>
<tr>
<td>16-Bit Ripple Carry Adder</td>
<td>0.50</td>
<td>0.166</td>
</tr>
<tr>
<td>16-Bit Carry Select Adder</td>
<td>0.53</td>
<td>0.056</td>
</tr>
<tr>
<td>16-Bit Brent Kung Adder</td>
<td>0.47</td>
<td>0.127</td>
</tr>
<tr>
<td>8-Bit Array Multiplier</td>
<td>0.59</td>
<td>0.065</td>
</tr>
</tbody>
</table>


TABLE 7.6
NORMALIZED ENERGY SAVINGS WITH THE TEMPERATURE-ADAPTIVE
VOLTAGE SCALING SCHEME

<table>
<thead>
<tr>
<th>Circuits in a 180nm CMOS Technology</th>
<th>High Temperature (125°C) Energy Consumption</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$V_{DD-25}$</td>
</tr>
<tr>
<td>16-Bit Ripple Carry Adder</td>
<td>1.00</td>
</tr>
<tr>
<td>16-Bit Carry Select Adder</td>
<td>1.00</td>
</tr>
<tr>
<td>16-Bit Brent-Kung Adder</td>
<td>1.00</td>
</tr>
<tr>
<td>8-Bit Array Multiplier</td>
<td>1.00</td>
</tr>
</tbody>
</table>

7.2.2. Temperature-Adaptive Body Bias

An alternative voltage tuning technique based on temperature-adaptive body-bias (TA-BB) is presented in this section. Similar to the dependence on the supply voltage, the propagation delay of a circuit is also strongly dependent on the device threshold voltages.
The absolute value of the threshold voltages degrade as the temperature increases, thereby simultaneously enhancing the circuit speed and the subthreshold leakage currents at elevated temperatures [1], [75]-[78], [117]. In ultra-low-voltage circuits exhibiting reversed temperature dependence, the threshold voltage of devices is dynamically increased through reverse body-bias at elevated temperatures to exponentially reduce the leakage current without degrading the clock frequency. The device threshold voltages are increased until the high temperature circuit performance of the TA-BB circuit matches the worst-case circuit performance of a standard-zero-body-biased circuit operating at a constant-$V_{DD}$. Unlike the conventional body-bias techniques aimed at altering the device threshold voltages based on the variations of the workload and the circuit activity, the proposed temperature-adaptive body-bias technique alters the threshold voltages of the devices based on the fluctuations of the die temperature and the circuit speed.

The temperature-adaptive body-bias technique (TA-BB) is illustrated in Fig. 7.6. Integrated circuits with the TA-BB technique operate with a constant supply voltage. The supply voltage and the target operating frequency ($f_{target}$) are determined according to the algorithm illustrated in Fig. 7.1. For minimum energy consumption at 25°C (125°C), the supply voltage of the circuit is fixed at $V_{DD-25}$ ($V_{DD-125}$). A ring oscillator providing a replica of the critical path of the entire integrated circuit translates the die temperature to a specific clock frequency ($f_{clock}$) for a specific set of body-bias voltages produced by the PMOS and NMOS body-bias generators. Note that a relatively uniform temperature is assumed across the die with this technique. The ring oscillator frequency ($f_{clock}$) is compared with the target operating frequency ($f_{target}$) and a frequency error signal ($f_{error}$) is generated. Using this error signal, the body-bias generators either modify or maintain the body-bias voltages applied to the devices in the integrated circuit. The device threshold voltages are, thereby, dynamically tuned based on the die temperature variations for maintaining a constant circuit speed across the entire die temperature spectrum.
Fig. 7.6. Temperature-adaptive body-bias technique. $V_{DD}$: standard constant-supply voltage providing minimum energy ($V_{DD-25}$ or $V_{DD-125}$).

The propagation delays of standard zero-body-biased circuits are compared with the high-temperature propagation delays of circuits based on the TA-BB technique operating at $V_{DD-25}$ and $V_{DD-125}$ in Tables 7.7 and 7.8, respectively. As listed in Table 7.7, the applicable high temperature reverse body-bias voltages are 0.40V (array multiplier) to 0.47V (Brent-Kung adder) with the TA-BB technique while maintaining the same clock-frequency as compared to the standard-zero-body-biased circuits operating at $V_{DD-25}$. Similarly, the applicable high temperature reverse body-bias voltages are 0.21V (array multiplier) to 0.25V (Brent-Kung adder) with the TA-BB technique for maintaining the same clock frequency as compared to the standard-zero-body-biased circuits operating at $V_{DD-125}$, as listed in Table 7.8.
### TABLE 7.7
PROPAGATION DELAY COMPARISON OF ZERO-BODY-BIASED CIRCUITS OPERATING AT $V_{DD-25}$ AND CIRCUITS WITH TA-BB CAPABILITY

<table>
<thead>
<tr>
<th>Circuits in a 180nm CMOS Technology</th>
<th>Standard constant voltage operation at $V_{DD-25}$</th>
<th>TA-BB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$</td>
<td>V_{SB}</td>
</tr>
<tr>
<td></td>
<td>25°C</td>
<td>125°C</td>
</tr>
<tr>
<td>16-Bit Ripple Carry Adder</td>
<td>0.00</td>
<td>15.59</td>
</tr>
<tr>
<td>16-Bit Carry Select Adder</td>
<td>0.00</td>
<td>3.40</td>
</tr>
<tr>
<td>16-Bit Brent Kung Adder</td>
<td>0.00</td>
<td>10.10</td>
</tr>
<tr>
<td>8-Bit Array Multiplier</td>
<td>0.00</td>
<td>3.89</td>
</tr>
</tbody>
</table>

### TABLE 7.8
PROPAGATION DELAY COMPARISON OF ZERO-BODY-BIASED CIRCUITS OPERATING AT $V_{DD-125}$ AND CIRCUITS WITH TA-BB CAPABILITY

<table>
<thead>
<tr>
<th>Circuits in a 180nm CMOS Technology</th>
<th>Standard constant voltage operation at $V_{DD-125}$</th>
<th>TA-BB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$</td>
<td>V_{SB}</td>
</tr>
<tr>
<td></td>
<td>25°C</td>
<td>125°C</td>
</tr>
<tr>
<td>16-Bit Ripple Carry Adder</td>
<td>0.00</td>
<td>0.166</td>
</tr>
<tr>
<td>16-Bit Carry Select Adder</td>
<td>0.00</td>
<td>0.056</td>
</tr>
<tr>
<td>16-Bit Brent Kung Adder</td>
<td>0.00</td>
<td>0.127</td>
</tr>
<tr>
<td>8-Bit Array Multiplier</td>
<td>0.00</td>
<td>0.065</td>
</tr>
</tbody>
</table>

* $V_{RBB}$: Reverse body-bias voltages that match the high temperature performance of the TA-BB circuits with the low-temperature performance of the standard-zero-body-biased.

PD: propagation delay in micro-seconds.

The normalized high temperature energy consumption of the standard-zero-body-biased circuits and the circuits based on the TA-BB technique are listed in Table 7.9. The high temperature energy consumption is increased by up to 6x (ripple carry adder) and 1.2x (ripple carry and Brent-Kung adder) with the temperature-adaptive reverse-body-
bias technique as compared to the standard-zero-body-biased circuits providing minimum energy at 25°C ($V_{DD-25}$ and $|V_{SB}| = 0V$) and 125°C ($V_{DD-125}$ and $|V_{SB}| = 0V$), respectively.

### TABLE 7.9
NORMALIZED ENERGY REDUCTION WITH THE TEMPERATURE-ADAPTIVE VOLTAGE SCALING SCHEME

<table>
<thead>
<tr>
<th>Circuits in a 180nm CMOS Technology</th>
<th>High Temperature (125°C) Energy Consumption</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$V_{DD-25}$</td>
</tr>
<tr>
<td>16-Bit Ripple Carry Adder</td>
<td>1.00</td>
</tr>
<tr>
<td>16-Bit Carry Select Adder</td>
<td>1.00</td>
</tr>
<tr>
<td>16-Bit Brent-Kung Adder</td>
<td>1.00</td>
</tr>
<tr>
<td>8-Bit Array Multiplier</td>
<td>1.00</td>
</tr>
</tbody>
</table>

The reason for the higher energy consumption at elevated temperatures in circuits with the TA-BB technique is illustrated here with a p-channel MOSFET in a 180nm CMOS technology. The switching current at ultra-low power-supply-voltages is the subthreshold leakage current. A p-channel MOSFET operating in the subthreshold regime with the drain biased at 0V and the gate and source terminals biased at 0.27V (the supply voltage providing minimum energy for a ripple-carry adder at 25°C, as listed in Table 7.2) is shown in Fig. 7.7. The currents observed at the different terminals of this PMOS transistor for various body-bias voltages ($V_{BB}$) are listed in Table 7.10 along with the total power consumption of the device.

Applying reverse body-bias increases the device threshold voltage thereby reducing the subthreshold leakage current [1], [117]. Applying reverse body bias, however, also increases the junction leakage currents due to the enhanced band-to-band tunneling [1], [117]. As listed in Table 7.10, even for a small reverse body-bias voltage ($|V_{SB}| = 0.05V$), the leakage current through the body-diodes increases by up to 76.7% (from 309.63pA to 547.24pA) as compared to a zero-body-biased transistor. For the relatively high reverse body-bias voltages required in circuits with the TA-BB technique ($|V_{SB}|$ ranging from 0.21V to 0.47V), the increase in the substrate current dominates the
reduction in the subthreshold leakage current, thereby increasing the total power consumed by the device, as listed in Table 7.10.

Fig. 7.7. A PMOS device in the TSMC 180nm CMOS technology. The gate and source terminals are biased at 0.27V. Temperature = 125°C. The device is reverse-body-biased by applying a voltage higher than 0.27V to the body terminal.

TABLE 7.10
POST-LAYOUT CURRENT MEASURED AT THE DIFFERENT TERMINALS OF THE PMOS DEVICE FOR VARIOUS BODY-BIAS VOLTAGES

| VBB (V) | |V_SB| (V) | ID (pA) | IS (pA) | IG (pA) | IB (pA) | Device Total Power Consumption (pW) |
|---|---|---|---|---|---|---|---|
| 0.27 | 0.00 | -397.55 | -87.92 | 0.00 | 309.63 | 107.34 |
| 0.32 | 0.05 | -368.15 | 179.09 | 0.00 | 547.24 | 126.76 |
| 0.37 | 0.10 | -348.95 | 253.70 | 0.00 | 602.64 | 154.48 |
| 0.42 | 0.15 | -336.31 | 279.31 | 0.00 | 615.62 | 183.15 |
| 0.47 | 0.20 | -327.95 | 290.77 | 0.00 | 618.72 | 212.29 |
| 0.52 | 0.25 | -322.38 | 297.14 | 0.00 | 619.52 | 241.92 |
| 0.57 | 0.30 | -318.64 | 301.14 | 0.00 | 619.78 | 271.97 |
| 0.62 | 0.35 | -316.13 | 303.79 | 0.00 | 619.92 | 302.33 |
| 0.67 | 0.40 | -314.43 | 305.60 | 0.00 | 620.03 | 332.91 |
7.3. Effectiveness of the Temperature-Adaptive Voltage Tuning Schemes

The effectiveness of the proposed temperature-adaptive schemes for lowering the active mode energy consumption is evaluated in this section. The energy consumption with the TA-DVS and TA-BB design methodologies are compared in Section 7.3.1. The impact of process parameter and environmental variations on the reliability of the proposed schemes is evaluated in Section 7.3.2.

7.3.1. Characteristics of the Temperature-Adaptive Schemes

The tradeoffs in the implementation of the temperature-adaptive voltage tuning schemes in circuits optimized for minimum energy consumption are presented in this section. The percent energy reduction provided with the two temperature-adaptive schemes is compared to the standard CMOS circuits operating at V_{DD-25} in Table 7.11. The high temperature energy efficiency is significantly enhanced by up 40% with the temperature-adaptive supply voltage tuning technique, as listed in Table 7.11. Alternatively, the energy consumption increases by up to 6x with the temperature-adaptive reverse body-bias technique as compared to the standard-zero-body-bias circuits operating at V_{DD-25}. TA-BB technique is therefore not effective for enhancing the high-temperature energy efficiency in ultra-low-voltage subthreshold logic circuits.

<table>
<thead>
<tr>
<th>Circuits in a 180nm CMOS Technology</th>
<th>Percent Energy Reduction (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TA-DVS</td>
</tr>
<tr>
<td>16-Bit Ripple Carry Adder</td>
<td>31</td>
</tr>
<tr>
<td>16-Bit Carry Select Adder</td>
<td>40</td>
</tr>
<tr>
<td>16-Bit Brent-Kung Adder</td>
<td>36</td>
</tr>
<tr>
<td>8-Bit Array Multiplier</td>
<td>38</td>
</tr>
</tbody>
</table>
7.3.2. Impact of the Process-Parameter and Supply Voltage Variations

Subthreshold logic circuits are highly sensitive to variations in the process parameters, the supply voltage, and the operating temperature [1], [66], [117], [120]. Both the performance and the energy consumption of integrated circuits are altered due to the fluctuations of the circuit parameters [1], [117], [126], [127]. The impact of parameter variations on the proposed temperature-adaptive voltage tuning technique is evaluated in this section. As described in Section 7.3.1, only the TA-DVS technique is effective in enhancing the high-temperature energy efficiency in ultra-low-voltage subthreshold logic circuits. Therefore, only the TA-DVS technique is evaluated in this section under parameter variations.

Random and systematic fluctuations in the channel length ($L_{\text{GATE}}$), the doping concentration ($N_{\text{CH}}$), and the gate-oxide thickness ($T_{\text{OX}}$) cause variations in the threshold voltage of a MOSFET. Fluctuation in the threshold voltage alters the performance and the power consumption (both dynamic and leakage power consumption) of a circuit. In this section, the variations in the performance and the energy consumption due to the process variations in the channel length ($L_{\text{GATE}}$), the doping concentration ($N_{\text{CH}}$), and the gate-oxide thickness ($T_{\text{OX}}$) are evaluated. Each parameter is assumed to have an independent normal Gaussian statistical distribution with a three-sigma variation of 10% [127].

Another important source of noise in CMOS integrated circuits is the power supply noise [1], [117]. Integrated circuits are typically designed to meet the performance specifications at a voltage 10% lower than the nominal supply voltage to account for the supply voltage variations [1], [117]. The supply voltage is assumed to have an independent normal Gaussian statistical distribution with a three-sigma variation of 10%.

Monte-Carlo simulations (30 simulations) are run to evaluate the performance and the energy consumption fluctuations in circuits with the TA-DVS technique. The delay versus energy consumption plots for the 16-bit ripple-carry adders operating at $V_{\text{DD-25}}$ and the optimized high temperature supply voltage with the TA-DVS technique are shown in Fig. 7.8. $V_{\text{DD-25}}$ for the ripple-carry adder is 0.27V (as listed in Table 7.2). In the presence of variations, the high temperature (125°C) energy consumption of standard constant-$V_{\text{DD}}$ ripple carry adders operating at $V_{\text{DD-25}}$ ranges from 1.54pJ to 1.84pJ, as
shown in Fig. 7.8. Alternatively, ripple carry adders with TA-DVS have a lower energy consumption at 125°C (energy consumption ranges from 1.17pJ to 1.28pJ) in the presence of process parameter and supply voltage variations, as shown in Fig. 7.8. Furthermore, the high-temperature propagation delay of ripple carry adders with TA-DVS (propagation delay range is from 4.85µs to 5.41µs, as shown in Fig. 7.8) is smaller as compared to the low-temperature performance of the standard constant-supply-voltage ripple carry adder (propagation delay at $V_{DD} = V_{DD-25}$ and temperature = 25°C is 15.59µs, as listed in Table 7.4). The smaller high-temperature propagation delay in circuits with the TA-DVS technique indicates that there is sufficient timing slack in the constant-clock-period even in the presence of variations.

The mean and the standard deviation of the high-temperature (125°C) energy consumption of the standard constant-$V_{DD}$ ripple-carry adder circuits operating at $V_{DD-25}$ are 1.67pJ and 84.9fJ, respectively. The 3-sigma offset of the lowest energy consumption (mean - 3*standard deviation) in these circuits is 1.41pJ. Alternatively, the mean and the standard deviation of the high-temperature (125°C) energy consumption of the ripple-carry adders with TA-DVS are 1.22pJ and 25.4fJ, respectively. The 3-sigma offset of the highest energy consumption (mean + 3*standard deviation) in the circuits with the TA-DVS technique is 1.29pJ. These results indicate that the highest possible energy consumption of a circuit with TA-DVS is still lower than the lowest possible energy consumption of a standard constant-$V_{DD}$ circuit when the parameter fluctuations are considered. The effectiveness of the proposed temperature-adaptive dynamic voltage scaling technique for enhancing the high temperature energy efficiency is therefore maintained in the presence of process-parameter and supply-voltage variations.
Fig. 7.8. Delay versus energy consumption plots for the 16-bit ripple carry adders operating at $V_{DD}$ and at the optimized high temperature supply voltage with the TA-DVS in the presence of process parameter and supply voltage variations. $N_{CH}$, $L_{GATE}$, $T_{OX}$, and $V_{DD}$ are assumed to have independent normal Gaussian statistical distributions with a three-sigma variation of 10%.

7.4. Chapter Summary

Gate overdrive variation with temperature dominates the speed characteristics of circuits operating at ultra-low-voltages. In ultra-low power-supply-voltage CMOS circuits, the circuit speed is enhanced with the increased temperature. The excessive timing slack observed in the clock period at elevated temperatures provides new opportunities to lower the active mode energy consumption without violating the constant-clock-frequency requirement. Temperature-adaptive dynamic supply voltage tuning technique is proposed in this chapter to reduce the high temperature energy consumption of ultra-low-voltage subthreshold logic circuits.

The temperature-adaptive supply voltage scaling technique dynamically adjusts the power supply voltage of a circuit based on the die-temperature fluctuations. The high temperature energy consumed with the temperature-adaptive voltage scaling technique is reduced by up to 40% as compared to the minimum energy achievable with the standard
constant-$V_{DD}$ and constant-frequency circuits. An alternative technique based on temperature-adaptive reverse body-bias that dynamically tunes the threshold voltages of the devices based on the fluctuations of the die temperature and the circuit speed is also evaluated. Temperature-adaptive dynamic supply voltage tuning technique is shown to be very effective to reduce the high temperature energy consumption without degrading the clock frequency in subthreshold logic circuits operating at ultra-low power-supply-voltages. The temperature-adaptive voltage scaling schemes are applied to the subthreshold memory banks for providing enhanced energy efficiency at elevated temperatures in Chapter 8.
Chapter 8
Temperature-Adaptive Low-Voltage Memory Banks

Scaling is the primary thrust behind the advancement of CMOS technology. The rapid scaling of device dimensions has enabled a dramatic increase in functionality and complexity of integrated circuits, as explained in Chapter 1. The performance of digital circuits can be further enhanced by increasing the memory capacity. Larger memories improve the effective memory bandwidth by reducing the average memory access time [128]. In the current state-of-the-art high performance integrated circuits, embedded memories occupy a majority of the total chip area [129]. The increasing trend of the on-die memory usage for the Intel Xeon microprocessor is shown in Fig. 8.1 [81]. The amount of on-chip memory is expected to continue to increase for enhancing the performance of future generations of portable devices and high-performance processors, as indicated in the ITRS roadmap [35]. Larger on-chip memories increase the total power consumption of integrated circuits [107].

Fig. 8.1. On-die memory capacity in Intel Xeon microprocessors [81].
Similar to high performance microprocessors, static random access memory (SRAM) arrays occupy a significant percent of the total chip area in low power mobile processors [130], [131]. The large switching capacitance in the bit-lines and the word-lines of an SRAM array contribute to the high dynamic power consumption during a memory access. Furthermore, the leakage currents of large memory banks can dominate the total power consumption in ultra-low-voltage circuits [107]. To achieve higher reliability and longer battery life-time in portable applications, the power consumed by the memory array should be reduced. Scaling the supply voltage enhances the energy efficiency of CMOS circuits [1], [117]. The optimum supply voltage that provides minimum energy consumption is typically observed in the subthreshold region [66], [78].

Circuits employed in the logic core exhibit reversed temperature dependence when optimized for minimum energy consumption, as explained in Chapters 6 and 7. Similar to the combinational circuits, the read and write propagation delay of subthreshold memory banks are enhanced at elevated temperatures. The reduction in the high temperature propagation delay provides new opportunities to reduce the energy consumption of subthreshold SRAM arrays at elevated temperatures. Temperature-adaptive voltage scaling techniques are presented in Chapter 7 to reduce the high temperature energy consumption of circuits exhibiting reversed temperature dependence. The effectiveness of the temperature-adaptive schemes for ultra-low-voltage memory banks is evaluated in this chapter. The high temperature energy consumption is reduced by exploiting the excessive timing slack produced in the constant clock-period of subthreshold memory circuits at elevated temperatures. The supply voltage and the reverse-body-bias voltage that lowers the energy consumption without degrading the circuit speed at increased temperatures are identified for an SRAM array in the TSMC 180nm CMOS technology [71]. The effectiveness of the TA-DVS technique under process parameter and supply voltage variations is also explored.

The chapter is organized as follows. The sizing constraints in a subthreshold SRAM cell and the design of a 64-bit x 64-bit ultra-low-voltage SRAM array are discussed in Section 8.1. The supply voltage that provides minimum energy consumption for an SRAM array in a 180nm CMOS technology is identified in Section 8.2. The efficiency of the temperature-adaptive supply and threshold voltage scaling techniques to
reduce the energy consumption at elevated temperatures is presented in Section 8.3. The influence of the temperature-adaptive dynamic supply voltage scaling scheme on the data stability of the memory cells is examined in Section 8.4. The effectiveness of the TA-DVS scheme under process parameter and supply voltage variations is also evaluated. The chapter is summarized in Section 8.5.

8.1. Sizing of a Subthreshold SRAM Bit Cell

In this section, the sizing constraints for the stability and the functionality of a conventional super-threshold ($V_{DD} > |V_t|$) SRAM cell are reviewed. The sizing constraints for the robust operation of the subthreshold SRAM circuits ($V_{DD} < |V_t|$) are then distinguished. The simulation setup for a 64-bit x 64-bit SRAM array is also presented.

A commonly used six transistor (6T) SRAM bit-cell is shown in Fig. 8.2. A single word-line (WORD) and both true and complementary bit-lines (BL and BLB) are utilized with the standard 6T-SRAM bit-cell. The bit-cell is composed of a pair of cross coupled inverters (P1/N1 and P2/N2) and two access transistors for the bit-lines (A1 and A2), as shown in Fig. 8.2. Both the true and the complementary versions of the data are stored in the cross coupled inverters (on node1 and node2).

![Six transistor (6T) SRAM bit-cell.](image)

Fig. 8.2. Six transistor (6T) SRAM bit-cell.
Regardless of a read or a write cycle, both bit-lines (BL and BLB) are periodically pre-charged to $V_{DD}$. Without loss of generality, in the following discussion it is assumed that node1 and node2 are initially at 0V and $V_{DD}$, respectively, as shown in Fig. 8.2. The word-line is raised for a read operation. BL is pulled down through A1 and N1. The current flowing through N1 raises the voltage on node1 while BL is being discharged. If the node1 voltage rises above the switching threshold voltage of the P2/N2 inverter, the data stored in the SRAM cell is flipped. To prevent the loss of data during a read operation, the pull-down transistors in the cross-coupled inverters (N1 and N2) must be stronger as compared to the access transistors (A1 and A2) [44]. Alternatively, when a ‘0’ is to be written to node2, BLB is pulled low by the write driver. After BLB is discharged, the word-line is asserted. A ‘0’ is forced onto node2 through A2. P2 opposes the high-to-low transition of node2. A1 and A2 must be stronger as compared to P1 and P2 to be able to flip the state of a memory cell with brute-force through the bit-line access transistors [44], [104]-[107].

Static Noise Margin (SNM) is the maximum amount of noise that is tolerated at the data storage nodes of an SRAM cell [106], [107]. The voltage transfer characteristics (VTC) of two cross-coupled inverters are shown in Fig. 8.3. The resulting curve is called the “butterfly curve”. The SNM is the length of the largest square that can be embedded inside the lobes of a butterfly curve, as illustrated in Fig. 8.3 [104]-[107].

As described in [104], [105], [106], and [107], for the memory circuits operating in the super-threshold (strong inversion) region, the read static noise margin increases with an increase in the memory cell ratio. The memory cell ratio is the ratio of the pull-down NMOS transistor width to the access transistor width ($W_{N1}/W_{A1}$ or $W_{N2}/W_{A2}$ where W is the width of the corresponding device in the 6T-SRAM bit-cell). The SNM of a 6T-SRAM bit-cell measured during a read operation (read static noise margin) is plotted in Fig. 8.4 for different supply voltages and memory cell ratios. As shown in Fig. 8.4, at lower supply voltages that are attractive for reduced energy consumption, the sensitivity of the read static noise margin to the SRAM cell ratio is diminished. As the supply voltage is scaled to the subthreshold region, the dependence of the data stability on the SRAM cell ratio becomes negligible.
Fig. 8.3. Voltage transfer characteristics (VTC) of the cross-coupled inverters to measure the static noise margin. The length of the side of the largest embedded square in the butterfly curve is the SNM.

Fig. 8.4. Read static noise margins of the 6T-SRAM cells for different supply voltages and cell ratios. \( W_{A1} = W_{A2} = W_{P1} = W_{P2} = 360\text{nm} \). \( W_{N1} = W_{N2} = (\text{Memory Cell Ratio}) \times 360\text{nm} \). TSMC 180nm CMOS technology.
The reason for the diminishing sensitivity of the read static noise margin to the transistor sizes at scaled supply voltages is identified next. The switching current at ultra-low-voltages is the subthreshold leakage current. For devices operating in the weak inversion region, the switching current is exponentially dependent on the voltage levels. Alternatively, increasing the device width produces only a linear increase in the switching current. A linear change in the subthreshold switching current has a relatively small impact on the voltage transfer characteristics. The sensitivity of the read noise margin to the memory cell ratio is therefore negligible in an ultra-low supply voltage subthreshold memory circuit. In this study, all the devices in the subthreshold SRAM cells are sized identical ($W_{N1} = W_{A1} = W_{P1} = W_{N2} = W_{A2} = W_{P2}$) since increasing the memory cell ratio does not provide a significant enhancement in data stability.

The carrier mobility of a PMOS device is lower as compared to an NMOS device since a hole is heavier as compared to an electron. The total weak inversion current produced by a PMOS device is, therefore, smaller as compared to an NMOS device with similar physical dimensions (width, length, and $t_{ox}$) and similar voltage difference across the device terminals. Therefore, even when the bit-cell devices are sized equal, the access transistors ($A1$ and $A2$) are stronger than the pull-up devices ($P1$ and $P2$), thereby satisfying the necessary condition for write ability.

The performance and the power consumption of the subthreshold logic circuits are affected significantly by the process parameter variations [1], [107], [117]. To suppress the fluctuations of the threshold voltage due to the process parameter variations, devices can be sized larger than the minimum width allowed in a given technology [107]. The minimum device width allowed in this TSMC 180nm CMOS technology is 220nm. In this study, the widths of all the devices in the SRAM bit-cells are sized 360nm for robustness against process variations.

The layout of the 6T-SRAM bit-cell drawn using the design rules of the TSMC 180nm CMOS technology [44] is shown in Fig. 8.5. The dimensions of the layout of a single bit-cell are $49\lambda \times 16\lambda$ where $\lambda$ is 90nm. The width of each device is $4\lambda$. The bit-lines and the word-lines are routed using the metal layers 2 and 3, respectively. The per cell word-line resistance ($R_{\text{WORD}}$) and capacitance ($C_{\text{WORD}}$) estimated using the sheet resistance and the capacitance values provided in [44] are $0.6891\Omega$ and $1.023\, \text{fF}$,
respectively. Similarly, the per cell bit-line (BL/BLB) resistance ($R_{BL/BLB}$) and capacitance ($C_{BL/BLB}$) are 0.225Ω and 0.334fF, respectively. The bit-lines and the word-lines are modeled as $\pi$–networks. The simulation setup of the 64-bit x 64-bit SRAM array is shown in Fig. 8.6. The read and write propagation delays are measured for the SRAM bit-cell farthest from the word-line driver and the read driver, as shown in Fig. 8.6. The read and write circuitry used with this SRAM array are shown in Fig. 8.7 [44].

![Fig. 8.5. Layout of the 6T-SRAM bit-cell.](image)

### 8.2. Supply Voltage for Minimum Energy Consumption

An algorithm to optimize the logic circuits for minimum energy consumption is presented in Chapter 7. In this section, the presented algorithm is employed to minimize the energy consumption of the 64-bit x 64-bit SRAM array in a 180nm CMOS technology. The supply voltage that achieves minimum energy consumption for the memory bank at 25°C is identified.
Fig. 8.6. The simulation setup of a 64-bit x 64-bit SRAM array. The read and the write propagation delays are measured with respect to the shaded SRAM cell.

In the first part of the algorithm presented in Chapter 7 (Fig. 7.1), the highest constant clock frequency that can be maintained within the entire temperature spectrum is identified for each supply voltage. The methodology employed to measure the maximum frequency \( f_{\text{max}} \) achievable with the SRAM array at a specific supply voltage and temperature is illustrated in Fig. 8.8. The signals for the read and the write operations in an SRAM bit-cell are shown in Fig. 8.8. The initial voltages of node1 and node2 in the SRAM bit-cell are assumed to be 0V and \( V_{DD} \), respectively, as shown in Fig. 8.2. First, a 0V is written into the SRAM bit-cell. The BLB is discharged by applying 0V on DATA_IN during the write operation. Node1 transitions from 0V \( \rightarrow V_{DD} \), as shown in Fig. 8.8. The contents of the SRAM bit-cell are flipped. The write operation is followed by a read operation. Following this sequence, another write operation is performed.
During the second write operation, DATA_IN is maintained at $V_{dd}$. BL is discharged. Node1 transitions from 0V → $V_{dd}$ during the write operation, as shown in Fig. 8.8. Finally, the most recent data is read from the selected bit-cell of the memory array. The voltage at DATA_OUT rises during the read operation indicating that the data stored in the SRAM bit-cell corresponds to logic 1.

Fig. 8.7. The write and read drivers of the SRAM array. (a) The write driver. (b) The read driver.
Similar to the logic circuits, the SRAM array is initially operated at a low clock frequency ($f << f_{max}$, where $f_{max}$ has to be determined). $Time_1$ ($Time_2$) is the time taken for the rising (falling) node1 voltage to cross $0.9*V_{DD}$ ($0.1*V_{DD}$) after the rising WORD signal crosses $0.1*V_{DD}$ ($0.1*V_{DD}$). Similarly, $Time_3$ is the time taken for the rising DATA_OUT signal to cross $0.9*V_{DD}$ after the rising WORD signal crosses $0.1*V_{DD}$. A 20% margin is added to the maximum of $Time_1$, $Time_2$, and $Time_3$ to provide a timing slack against parameter variations, clock jitter, and clock-skew. The maximum clock frequency achievable with a synchronous memory circuit operating at a specific supply voltage and temperature is

$$f_{max} = \left\{ 2 \times \max(Time_1, Time_2, Time_3) \right\}^{-1}.$$  \hspace{1cm} (8.1)

To find the highest achievable constant-clock-frequency at a particular supply voltage, the maximum achievable frequencies ($f_{max}$) at the extremes of the die temperature spectrum ($T_1$ and $T_2$) are measured using the procedure described above. The
smaller of the two frequencies is the highest constant-frequency that can be maintained by the circuit within the entire temperature spectrum \( (T_1 \rightarrow T_2) \) at the particular supply voltage. After the \( f_{max} \) is calculated, the energy consumption at a specific supply voltage and temperature is measured to identify the supply voltage that provides minimum energy consumption, as illustrated in Chapter 7.

The results of the algorithm applied to a 64-bit x 64-bit SRAM array are listed in Table 8.1. The die temperature spectrum is assumed to be from 25°C to 125°C. The standard constant-supply voltage for achieving minimum energy consumption at 25°C is reported. As listed in Table 8.1, the temperature that determines the highest operating frequency is also dependent on the supply voltage of the circuit. At the higher supply voltages (such as the \( V_{DD,\text{nominal}} = 1.8V \)), the SRAM circuit operates slower when the die temperature increases. The maximum achievable (worst-case) frequency is therefore determined by the \( f_{max} \) at the highest temperature. Alternatively, as the supply voltage is scaled, the worst-case clock frequency shifts to the lowest operating temperature, as listed in Table 8.1. The maximum achievable clock frequency for the entire temperature spectrum is therefore determined by the propagation delays (read and write) observed at the lowest temperature for \( V_{DD} \leq 0.85V \). \( V_{DD-25} \) is the constant-supply-voltage applied to a standard CMOS circuit (without any voltage tuning capability) for achieving minimum energy consumption at 25°C.

The long channel threshold voltages of the n-channel and the p-channel devices in this 180nm CMOS technology are 0.48V and -0.45V, respectively [71]. The supply voltage providing minimum energy (\( V_{DD-25} \)) is lower than the threshold voltage of the devices in this technology (the SRAM array, therefore, operates in the subthreshold regime), as listed in Table 8.1 [66], [107]. The switching current at these ultra-low supply voltages is the subthreshold leakage current. As described in Chapter 2, the subthreshold leakage current is extremely sensitive to temperature fluctuations. Absolute value of the threshold voltage degrades and the thermal voltage is enhanced as the temperature increases [73], [74]. A small change in the die temperature exponentially alters the subthreshold leakage current [1], [117]. The reversal in the temperature dependent propagation delay characteristics coupled with the high sensitivity of the circuit speed to the temperature fluctuations provides opportunities for reducing the energy consumption
without degrading the clock frequency at elevated die temperatures in an ultra-low supply-voltage memory array.

### TABLE 8.1
SUPPLY VOLTAGES THAT ACHIEVE MINIMUM ENERGY IN A CONSTANT-$V_{DD}$ AND CONSTANT-$f_S$ 64-BIT x 64-BIT SRAM ARRAY

<table>
<thead>
<tr>
<th>$V_{DD}$ (V)</th>
<th>Max. frequency at 25°C (MHz)</th>
<th>Max. frequency at 125°C (MHz)</th>
<th>Worst-case frequency (MHz)</th>
<th>Energy consumption at the worst-case frequency and 25°C (pJ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.80</td>
<td>521.7</td>
<td>392.7</td>
<td>392.7</td>
<td>22.60</td>
</tr>
<tr>
<td>1.70</td>
<td>486.2</td>
<td>388.7</td>
<td>388.7</td>
<td>20.30</td>
</tr>
<tr>
<td>1.60</td>
<td>465.3</td>
<td>374.7</td>
<td>374.7</td>
<td>17.91</td>
</tr>
<tr>
<td>0.86</td>
<td>180.2</td>
<td>176.2</td>
<td>176.2</td>
<td>4.84</td>
</tr>
<tr>
<td>0.85</td>
<td>173.3</td>
<td>173.6</td>
<td>173.3</td>
<td>4.72</td>
</tr>
<tr>
<td>0.84</td>
<td>167.9</td>
<td>168.1</td>
<td>167.9</td>
<td>4.59</td>
</tr>
</tbody>
</table>

Supply voltage below which the circuit exhibits reverse temperature dependence

<table>
<thead>
<tr>
<th>$V_{DD}$ (V)</th>
<th>Max. frequency at 25°C (MHz)</th>
<th>Max. frequency at 125°C (MHz)</th>
<th>Worst-case frequency (MHz)</th>
<th>Energy consumption at the worst-case frequency and 25°C (pJ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.30</td>
<td>157.2</td>
<td>2020.4</td>
<td>157.2</td>
<td>0.776</td>
</tr>
<tr>
<td>0.29</td>
<td>131.1</td>
<td>1798.6</td>
<td>131.1</td>
<td>0.728</td>
</tr>
<tr>
<td><strong>0.28</strong></td>
<td><strong>101.9</strong></td>
<td><strong>1568.2</strong></td>
<td><strong>101.9</strong></td>
<td><strong>0.711</strong></td>
</tr>
<tr>
<td>0.27</td>
<td>86.7</td>
<td>1331.9</td>
<td>81.1</td>
<td>0.716</td>
</tr>
<tr>
<td>0.26</td>
<td>64.8</td>
<td>1116.4</td>
<td>64.8</td>
<td>0.729</td>
</tr>
</tbody>
</table>

* Results are for a 64-bit x 64-bit SRAM array in a 180nm CMOS technology.

### 8.3. High Temperature Energy Reduction in Memory Banks

The temperature-adaptive supply and threshold voltage tuning techniques are presented in Chapter 7. The effectiveness of the temperature-adaptive schemes to reduce the high temperature energy consumption in subthreshold memory arrays is evaluated in this section. The energy reduction observed with the temperature-adaptive supply voltage
scaling and the temperature-adaptive reverse-body-bias techniques are examined in Sections 8.3.1 and 8.3.2, respectively.

**8.3.1. Temperature-Adaptive Supply Voltage Scaling**

The effectiveness of the temperature-adaptive dynamic supply voltage scaling technique (TA-DVS) is presented in this section. All the primary components of power consumption in a CMOS circuit, namely dynamic switching, short-circuit, and leakage power are significantly reduced by scaling the supply voltage. The read and the write propagation delays of an SRAM array are also strongly dependent on the supply voltage [1], [117]. Scaling the supply voltage reduces the power consumed by a circuit at the cost of degraded circuit performance.

The CLK and DATA_OUT of the 64th column in a 64-bit x 64-bit SRAM array operating at the constant-VDD-25 with various temperatures are shown in Fig. 8.9. VDD-25 for the SRAM array is 0.28V, as listed in Table 8.1. The clock frequency is fixed at 101.9MHz in a standard CMOS circuit (determined by the lowest operating temperature), the highest constant-\(f_c\) that can be maintained at all die temperatures, as listed in Table 8.1. The circuit operates faster at elevated temperatures, thereby producing excessive timing slack in the constant-clock-period, as shown in Fig. 8.9. At elevated temperatures the total energy consumption also significantly increases due to the increase in the subthreshold leakage current [1], [117]. The significant timing slack in the clock-period can be exploited to reduce the active-mode energy consumption at elevated temperatures with the temperature-adaptive supply voltage scaling technique presented in Chapter 7.
Fig. 8.9. CLK and DATA_OUT of the 64th column in a 64-bit x 64-bit SRAM array operating at $V_{DD-25} (0.28V)$ and various temperatures.

The propagation delays (read and write) and the energy consumption of the SRAM array at various power supply voltages and temperatures are listed in Table 8.2. For all the supply voltages listed in Table 8.2, the outputs achieve at least a $0.1V_{DD} \rightarrow 0.9V_{DD}$ voltage swing (condition for functionality). $Max\_delay$ is the maximum of the read and the write propagation delays at a given supply voltage and temperature. At the supply voltage providing minimum energy ($V_{DD} = 0.28$), the $max\_delay$ reduces from 3.72µs to 0.22µs when the temperature increases from 25°C to 125°C, as listed in Table 8.2. To exploit this slack in the clock period, the supply voltage can be scaled with the TA-DVS technique while maintaining constant frequency. The supply voltage can be scaled down to 0.21V at the highest temperature while maintaining the circuit delay lower than the lowest temperature propagation delay ($max\_delay$), as listed in Table 8.2. Supply voltage scaling reduces the energy consumption of a circuit. With the proposed TA-DVS technique, the high temperature energy consumption is reduced by up to 32.8% (from 12.93pJ to 8.69pJ) without degrading the clock frequency of the SRAM array in this 180nm CMOS technology.
TABLE 8.2
READ DELAY, WRITE DELAY, AND ENERGY CONSUMPTION OF THE SRAM ARRAY AT VARIOUS TEMPERATURES AND POWER SUPPLY VOLTAGES

<table>
<thead>
<tr>
<th>Temp (°C)</th>
<th>V_{DD} (V)</th>
<th>Read Delay (µs)</th>
<th>Write Delay (µs)</th>
<th>Max_delay (µs)</th>
<th>Energy consumption (pJ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>0.28</td>
<td>3.72</td>
<td>2.95</td>
<td>3.72</td>
<td>0.711</td>
</tr>
<tr>
<td></td>
<td>0.28</td>
<td>0.22</td>
<td>0.20</td>
<td>0.22</td>
<td>12.93</td>
</tr>
<tr>
<td></td>
<td>0.27</td>
<td>0.23</td>
<td>0.25</td>
<td>0.23</td>
<td>12.33</td>
</tr>
<tr>
<td></td>
<td>0.26</td>
<td>0.26</td>
<td>0.30</td>
<td>0.30</td>
<td>11.63</td>
</tr>
<tr>
<td></td>
<td>0.25</td>
<td>0.27</td>
<td>0.34</td>
<td>0.34</td>
<td>11.05</td>
</tr>
<tr>
<td></td>
<td>0.24</td>
<td>0.28</td>
<td>0.42</td>
<td>0.42</td>
<td>10.46</td>
</tr>
<tr>
<td></td>
<td>0.23</td>
<td>0.30</td>
<td>0.50</td>
<td>0.50</td>
<td>9.91</td>
</tr>
<tr>
<td></td>
<td>0.22</td>
<td>0.32</td>
<td>0.59</td>
<td>0.59</td>
<td>9.30</td>
</tr>
<tr>
<td>125</td>
<td>0.21</td>
<td>0.33</td>
<td>0.72</td>
<td>0.72</td>
<td>8.69</td>
</tr>
<tr>
<td></td>
<td>0.20</td>
<td>0.35</td>
<td>4.97</td>
<td>4.97</td>
<td>8.14</td>
</tr>
</tbody>
</table>

*Max_delay is the maximum of the read and the write propagation delays at a specific supply voltage and temperature.

8.3.2. Temperature-Adaptive Body Bias

The effectiveness of the alternative voltage tuning technique based on temperature-adaptive body-bias (TA-BB) is presented in this section. In ultra-low-voltage subthreshold memory circuits exhibiting reversed temperature dependence, the threshold voltages of devices are dynamically increased through reverse-body-bias at elevated temperatures to exponentially reduce the leakage current without degrading the clock frequency. The device threshold voltages are increased until the high temperature performance of the SRAM array with TA-BB matches the worst-case performance of a standard-zero-body-biased SRAM array operating at a constant-$V_{DD}$.

The propagation delays (read and write) and the energy consumption of the SRAM array at various body-bias voltages and temperatures are listed in Table 8.3. At $V_{DD-25}$ ($V_{DD} = 0.28$), the max_delay of the memory circuit is reduced when the
temperature increases. To exploit the available timing slack, the proposed TA-BB technique is applied to the SRAM array. For the memory circuits in this 180nm CMOS technology, the applicable high temperature reverse-body-bias voltage without degrading the clock frequency is up to 0.37V with the TA-BB technique, as listed in Table 8.3. The energy consumption of the SRAM array for the various body-bias voltages are also listed in Table 8.3. Results indicate that the high temperature energy consumption with this TA-BB technique increases by up to 1.47x (from 12.93pJ to 19.02pJ).

### Table 8.3

READ DELAY, WRITE DELAY, AND ENERGY CONSUMPTION OF THE SRAM ARRAY AT VARIOUS TEMPERATURES AND BODY-BIAS VOLTAGES.

| Temp (°C) | \(|V_{SB}|\) (V) | Read Delay (µs) | Write Delay (µs) | Max_delay (µs) | Energy consumption (pJ) |
|-----------|-----------------|-----------------|------------------|----------------|------------------------|
| 25        | 0.0             | 3.72            | 2.95             | 3.72           | 0.711                  |
| 0.0       | 0.0             | 0.22            | 0.20             | 0.22           | 12.93                  |
| 0.17      | 0.77            | 0.65            | 0.77             | 13.30          |
| 0.21      | 1.03            | 0.84            | 1.03             | 14.23          |
| 0.25      | 1.38            | 1.08            | 1.38             | 15.30          |
| 0.29      | 1.86            | 1.38            | 1.86             | 16.47          |
| 0.33      | 2.52            | 1.77            | 2.52             | 17.72          |
| 0.37      | 3.45            | 2.25            | 3.45             | 19.02          |
| 0.41      | 4.74            | 2.85            | 4.74             | 20.42          |

*Max_delay* is the maximum of the read and the write propagation delays at a specific supply voltage and temperature.

The reason for the higher energy consumption in a TA-BB SRAM array is illustrated here with a p-channel MOSFET in a 180nm CMOS technology. The switching current at ultra-low power-supply-voltages is the subthreshold leakage current. A p-channel MOSFET operating in the subthreshold regime with the drain biased at 0V and the gate and source terminals biased at 0.28V (\(V_{DD_{25}}\) for the SRAM array, as listed in Table 8.1) is shown in Fig. 8.10. The currents observed at the different terminals of this
PMOS transistor for various body-bias voltages (V_{BB}) are listed in Table 8.4 along with the total power consumption of the device.

Fig. 8.10. A PMOS device in the TSMC 180nm CMOS technology. The gate and source terminals are biased at 0.28V. Temperature = 125°C. The device is reverse-body-biased by applying a voltage higher than 0.28V to the body terminal.

### TABLE 8.4
POST-LAYOUT CURRENT MEASURED AT THE DIFFERENT TERMINALS OF THE PMOS DEVICE FOR VARIOUS BODY-BIAS VOLTAGES

<table>
<thead>
<tr>
<th>V_{BB} (V)</th>
<th></th>
<th>V_{SB} (V)</th>
<th>ID (pA)</th>
<th>IG (pA)</th>
<th>IS (pA)</th>
<th>IB (pA)</th>
<th>Device Total Power Consumption (pW)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.28</td>
<td></td>
<td>0.00</td>
<td>-228.1</td>
<td>0.0</td>
<td>-67.6</td>
<td>160.4</td>
<td>26.0</td>
</tr>
<tr>
<td>0.32</td>
<td></td>
<td>0.04</td>
<td>-209.2</td>
<td>0.0</td>
<td>61.7</td>
<td>270.8</td>
<td>103.9</td>
</tr>
<tr>
<td>0.36</td>
<td></td>
<td>0.08</td>
<td>-195.8</td>
<td>0.0</td>
<td>109.5</td>
<td>305.3</td>
<td>140.6</td>
</tr>
<tr>
<td>0.40</td>
<td></td>
<td>0.12</td>
<td>-186.3</td>
<td>0.0</td>
<td>129.8</td>
<td>316.1</td>
<td>162.8</td>
</tr>
<tr>
<td>0.44</td>
<td></td>
<td>0.16</td>
<td>-179.5</td>
<td>0.0</td>
<td>140.1</td>
<td>319.5</td>
<td>179.8</td>
</tr>
<tr>
<td>0.48</td>
<td></td>
<td>0.20</td>
<td>-174.6</td>
<td>0.0</td>
<td>146.1</td>
<td>320.6</td>
<td>194.8</td>
</tr>
<tr>
<td>0.52</td>
<td></td>
<td>0.24</td>
<td>-171.0</td>
<td>0.0</td>
<td>150.0</td>
<td>321.0</td>
<td>208.9</td>
</tr>
<tr>
<td>0.56</td>
<td></td>
<td>0.28</td>
<td>-168.5</td>
<td>0.0</td>
<td>152.8</td>
<td>321.2</td>
<td>222.7</td>
</tr>
<tr>
<td>0.60</td>
<td></td>
<td>0.32</td>
<td>-166.6</td>
<td>0.0</td>
<td>154.7</td>
<td>321.3</td>
<td>236.1</td>
</tr>
</tbody>
</table>
Applying reverse body-bias increases the device threshold voltage thereby reducing the subthreshold leakage current [1], [117]. Applying reverse body bias, however, also increases the junction leakage currents due to the enhanced band-to-band tunneling [1], [117]. As listed in Table 8.4, even for a small reverse body-bias voltage (|V_{SB}| = 0.04V), the leakage current through the body-diodes increases by up to 68.8% (from 160.4pA to 270.8pA) as compared to a zero-body-biased transistor. For the relatively higher reverse-body-bias voltage applied to the TA-BB SRAM array (|V_{SB}| = 0.37V), the increase in the substrate current dominates the reduction in the weak inversion current, thereby increasing the total power consumed by the individual devices, as listed in Table 8.4.

8.4. Effectiveness of the TA-DVS Technique

In this section, the reliability of the subthreshold memory circuits with the TA-DVS technique is evaluated. The influence of the temperature adaptive dynamic supply voltage tuning technique on the data stability of the subthreshold memory cells is presented in Section 8.4.1. The effectiveness of the proposed TA-DVS scheme under the presence of process parameter and environmental variations is evaluated in Section 8.4.2.

8.4.1. Influence of the TA-DVS Technique on the Noise Margins

In this section, the influence of the proposed temperature adaptive dynamic supply voltage scaling scheme on the noise margins of the SRAM cell is evaluated. The hold static noise margin and the read static noise margin of the SRAM bit-cells (all the devices in the bit-cells are sized the same) at different supply voltages and temperatures are listed in Table 8.5. The read static noise margin of the SRAM bit-cells operating at V_{DD-25} (V_{DD} = 280mV) at both 25°C and 125°C is 35mV, as listed in Table 8.5. With the proposed dynamic supply voltage tuning technique, the power supply voltage of the SRAM array can be scaled down to 210mV when the temperature increases to 125°C. Both the hold static noise margin and the read static noise margin are reduced when the supply voltage is scaled. With the proposed technique, the read static margin of the SRAM bit-cell is reduced by up to 19mV (from 35mV to 16mV) at high temperatures, as listed in Table 8.5. Provided that this degradation in the read static noise
margin can be tolerated, the proposed temperature adaptive dynamic voltage scaling scheme can be employed to reduce the high temperature power consumption by up to 32.8%, as discussed in Section 8.3.1.

<table>
<thead>
<tr>
<th>$V_{DD}$ (mV)</th>
<th>Hold Noise Margin (mV)</th>
<th>Read Noise Margin (mV)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>25°C</td>
<td>125°C</td>
</tr>
<tr>
<td>280</td>
<td>91</td>
<td>90</td>
</tr>
<tr>
<td>270</td>
<td>86</td>
<td>85</td>
</tr>
<tr>
<td>260</td>
<td>81</td>
<td>80</td>
</tr>
<tr>
<td>250</td>
<td>76</td>
<td>75</td>
</tr>
<tr>
<td>240</td>
<td>71</td>
<td>70</td>
</tr>
<tr>
<td>230</td>
<td>66</td>
<td>65</td>
</tr>
<tr>
<td>220</td>
<td>61</td>
<td>60</td>
</tr>
<tr>
<td><strong>210</strong></td>
<td><strong>56</strong></td>
<td><strong>55</strong></td>
</tr>
<tr>
<td>200</td>
<td>51</td>
<td>50</td>
</tr>
</tbody>
</table>

8.4.2. Impact of the Process-Parameter and Supply Voltage Variations

Subthreshold logic circuits are highly sensitive to variations in the process parameters, the supply voltage, and the operating temperature [1], [66], [107], [117]. Both the performance and the energy consumption of integrated circuits are altered due to the fluctuations of the circuit parameters [1], [117], [126], [127]. The impact of the parameter variations on the proposed TA-DVS technique is evaluated in this section.

Random and systematic fluctuations in the channel length ($L_{GATE}$), the doping concentration ($N_{CH}$), and the gate-oxide thickness ($T_{OX}$) cause variations in the threshold voltage of a MOSFET. Fluctuation in the threshold voltage alters the performance and the power consumption (both dynamic and leakage power consumption) of a circuit. In this
study, the variations in the performance and the energy consumption due to the process variations in the channel length \(L_{GATE}\), the doping concentration \(N_{CH}\), and the gate-oxide thickness \(T_{OX}\) are evaluated. Each parameter is assumed to have an independent normal Gaussian statistical distribution with a three-sigma variation of 10% [127].

Another important source of noise in CMOS integrated circuits is the power supply noise [1], [117]. Integrated circuits are typically designed to meet the performance specifications at a voltage 10% lower than the nominal supply voltage to account for the supply voltage variations [1], [117]. The supply voltage is assumed to have an independent normal Gaussian statistical distribution with a three-sigma variation of 10%.

Monte-Carlo simulations are run to evaluate the performance and the energy consumption fluctuations in circuits with the TA-DVS technique. The delay versus energy consumption plots for the 64-bit x 64-bit SRAM arrays operating at \(V_{DD,25}\) and the optimized high temperature supply voltage with the TA-DVS technique are shown in Fig. 8.11. \(V_{DD,25}\) for the SRAM array is 0.28V (as listed in Table 8.1). In the presence of variations, the high temperature \(125^\circ C\) energy consumption of the standard constant-\(V_{DD}\) memory circuits operating at \(V_{DD,25}\) ranges from 12pJ to 14.2pJ, as shown in Fig. 8.11. Alternatively, the SRAM arrays with TA-DVS have a lower energy consumption at \(125^\circ C\) (energy consumption ranges from 8.1pJ to 9.7pJ) in the presence of process parameter and supply voltage variations, as shown in Fig. 8.11. Furthermore, the high temperature delay of the TA-DVS SRAM circuits (propagation delay range is from 543ns to 819ns, as shown in Fig. 8.11) is smaller as compared to the low-temperature delay of the standard constant-supply-voltage SRAM arrays (propagation delay at \(V_{DD} = V_{DD,25}\) and temperature = \(25^\circ C\) is 3719ns, as listed in Table 8.2). The smaller high-temperature propagation delay in the TA-DVS memory arrays indicate that there is sufficient timing slack in the constant-clock-period even in the presence of parameter variations.

The mean and the standard deviation of the high-temperature \(125^\circ C\) energy consumption of the standard constant-\(V_{DD}\) memory circuits operating at \(V_{DD,25}\) are 12.98pJ and 555fJ, respectively. The 3-sigma offset of the lowest energy consumption (mean - 3*standard deviation) in these circuits is 11.32pJ. Alternatively, the mean and the standard deviation of the high-temperature \(125^\circ C\) energy consumption of the SRAM arrays with TA-DVS are 8.85pJ and 425fJ, respectively. The 3-sigma offset of the highest
energy consumption (mean + 3*standard deviation) in the circuits with the TA-DVS technique is 10.12pJ. These results indicate that the highest possible energy consumption of a circuit with TA-DVS is still lower than the lowest possible energy consumption of a standard constant-$V_{DD}$ circuit when the parameter fluctuations are considered. The effectiveness of the proposed temperature-adaptive dynamic voltage scaling technique for enhancing the high temperature energy efficiency in subthreshold memory arrays is therefore maintained in the presence of process-parameter and supply-voltage variations.

![Fig. 8.11. Delay versus energy consumption plots for the 64-bit x 64-bit SRAM arrays operating at $V_{DD,25}$ and at the optimized high temperature supply voltage with the TA-DVS in the presence of process parameter and supply voltage variations. $N_{CH}$, $L_{GATE}$, $T_{OX}$, and $V_{DD}$ are assumed to have independent normal Gaussian statistical distributions with a three-sigma variation of 10%.

8.5. Chapter Summary

In ultra-low power-supply-voltage subthreshold memory arrays, the circuit performance is enhanced with the increased temperature. The excessive timing slack observed in the clock period at elevated temperatures provides new opportunities to lower the active mode energy consumption without violating the constant-clock-frequency
requirement. The effectiveness of the temperature-adaptive dynamic supply voltage tuning and temperature-adaptive body-bias technique to reduce the high temperature energy consumption of ultra-low-voltage subthreshold memory circuits is evaluated in this chapter.

The temperature-adaptive supply voltage scaling technique dynamically adjusts the power supply voltage of a circuit based on the die-temperature fluctuations. The high temperature energy consumed with the temperature-adaptive voltage scaling technique is reduced by up to 32.8% as compared to the minimum energy achievable with a standard constant-$V_{DD}$ and constant-frequency 64-bit x 64-bit SRAM array in a 180nm CMOS technology. The alternative temperature-adaptive reverse body-bias technique is also evaluated. Temperature-adaptive dynamic supply voltage tuning technique is shown to be very effective to reduce the high temperature energy consumption without degrading the clock frequency in subthreshold memory circuits operating at ultra-low power-supply-voltages.
Chapter 9
Conclusions

Scaling is the primary thrust behind the advancement of CMOS technology. The IC industry has been consistently scaling the design rules, increasing the chip areas, and manufacturing larger wafers with each new technology generation. Consequently, the semiconductor industry has enjoyed phenomenal enhancement in circuit speed and functionality combined with a steady decline in the cost per function. In addition to the higher circuit speed with technology scaling, the performance of integrated circuits have also been enhanced by utilizing the growing number of transistors to develop novel circuit techniques and microarchitectures. Technology scaling related enhancements coupled with advances in circuit structures and microarchitectures have significantly increased the performance of integrated circuits. The primary side effect of enhanced circuit performance and functionality is typically an increase in the power consumption.

High-end processors represent one end of the semiconductor market. Increasing the clock frequency has traditionally been the primary objective in the design of high performance systems. Using power hungry circuit techniques and microarchitectures to enhance the clock frequency has increased the power consumption of the high performance circuits many folds over the years. Alternatively, a second family of ICs is utilized in systems that target miniaturization and portability. Typically, portable applications are aimed at reducing the power consumption. However, particularly since the late 1990s, customer demand has been growing for higher performance and wider variety of applications in mobile systems. The techniques employed for enhanced performance increase the power consumption of portable applications as well.

The power consumed by a circuit is dissipated as heat through the substrate. The heat removal is traditionally achieved with inexpensive packages, passive heat sinks, and air flow fans. With power consumption rising well above 100W in the current integrated systems, however, more expensive packaging and cooling technologies will soon be required. Another important consequence of increasing power consumption is the increasing die temperature gradients and the formation of local hot-spots. Fluctuations of
the ambient temperature can also vary the die temperature. The variation of temperature alters the speed and power consumption characteristics of CMOS circuits. Several techniques to reduce the sensitivity of circuit speed to the fluctuations of die temperature are described in this dissertation. The goal of the proposed techniques is to simultaneously achieve temperature variation resilience and enhanced energy efficiency in CMOS circuits. Furthermore, temperature-adaptive design methodologies for enhanced energy efficiency in circuits operating at elevated temperatures are also presented. The speed and energy tradeoffs with the different techniques are provided.

An analysis of the power and heat dissipation related problems faced by the semiconductor industry starts with the identification of the sources of power consumption. The primary sources of power consumption in CMOS integrated circuits are described in this dissertation. The dominant component of the power consumption of a typical CMOS circuit is the switching power. The dynamic switching power is consumed during the charging and discharging of parasitic capacitances when the node voltages transition. Other significant sources of power consumption in digital circuits are the leakage currents. The subthreshold, the gate-oxide tunneling, and the junction leakage currents cause power consumption even in circuits that are idle. Furthermore, the finite slopes of the input signals temporarily produce direct current conduction paths between the power supply and the ground distribution network, thereby causing short-circuit power consumption.

Fluctuations in the die temperature alter the number of charge carriers in a semiconductor device. The variation of carrier concentrations changes the behavior of semiconductor devices with temperature. The four primary MOSFET device parameters that are affected by the temperature fluctuations are the threshold voltage, the carrier mobility, the saturation velocity, and the parasitic drain/source resistances. The threshold voltage, the carrier mobility, and the saturation velocity degrade when the die temperature increases. Alternatively, the source/drain resistances increase at elevated temperatures. Effective variation of MOSFET drain current is determined by the variation of the dominant device parameter when the temperature fluctuates.
Influence of temperature fluctuations on the device and circuit characteristics are evaluated in Chapter 3 for the TSMC 180nm and the Berkeley Predictive 65nm CMOS technologies. At the nominal supply voltage, the carrier mobility variations are dominant for the devices in the two different CMOS technologies when the temperature is increased from 25°C to 125°C. The drain current of the devices operating at the nominal supply voltage, therefore, degrade at elevated temperatures. Variation of the current produced by the active devices alters the propagation delay and the energy consumption characteristics of CMOS circuits. When operating at the nominal supply voltage, the speed of circuits degrades by up to 15.9% and 54.5% as the temperature is increased from 25°C to 125°C in 180nm and 65nm CMOS technologies, respectively. Alternatively, the energy consumption increases by up to 26% at 125°C as compared to 25°C.

Computer-aided design (CAD) tools are used for the pre-fabrication characterization of integrated circuits. Design objectives such as speed, area, reliability, and power consumption can be verified with the aid of CAD tools. Reducing the power consumption is a primary objective in the design of digital integrated circuits. Techniques to lower the power consumption require precise power measurements in CMOS circuits. Accurate power estimation with the circuit simulators is critical to be able to correctly identify the most effective power reduction techniques that satisfy the design objectives. Furthermore, the fluctuations of the power consumption with temperature necessitate a comprehensive power measurement methodology to accurately characterize the total power consumption of nanoscale circuits at different die temperatures.

A generic methodology to accurately measure the power and energy consumption with the circuit simulators is presented in this dissertation. The currents drawn from the power supplies at the different terminals of the device contribute to the total device power consumption. The instantaneous power consumption of the device can be accurately measured by algebraically summing the power drawn from the independent sources connected to the different device terminals. A generic expression for accurately measuring the total power consumption of a complex circuit considering all of the circuit terminals is also derived.

The power consumption of devices and circuits measured with the generic measurement methodology are compared with the power estimated using the built-in
power estimation commands of two commercial circuit simulators: HSPICE and CADENCE-SPECTRE. Results indicate that the gate-oxide tunneling leakage currents are completely ignored in the power calculations with the built-in commands. This inappropriate assumption drastically underestimates the power measurements by up to 8540x with the built-in power estimation functions. For the deeply scaled CMOS circuits in the nano-meter regime, power and energy measurement with the proposed explicit methodology is strongly recommended for an accurate pre-fabrication circuit characterization. The actual power consumption and the power savings provided with the techniques explored in this dissertation are measured using the generic methodology proposed in Chapter 4.

Variations in the drain current with temperature cause significant fluctuations in the speed characteristics of CMOS circuits as discussed in Chapter 3 with various examples. The propagation delay variation with temperature necessitates the verification of the timing constraints at different die temperatures. Decreasing the sensitivity of the circuit speed to the die temperature fluctuations is desirable for reducing the uncertainty in the propagation delay characteristics of CMOS circuits. Design methodologies for suppressing the variations of the drain current and the propagation delay due to temperature fluctuations are described in Chapter 5. Supply and threshold voltage optimization techniques are proposed in this dissertation to achieve temperature variation insensitive constant circuit speed.

The mobility of charge carriers is the dominant device parameter that determines the drain current variations when the die temperature fluctuates. In order to compensate for the variation of carrier mobility, the sensitivity of gate overdrive to temperature fluctuations should be enhanced. With the supply voltage optimization technique, the supply voltage of the circuits is scaled to increase the sensitivity of the gate overdrive. For a particular lower supply voltage, the temperature fluctuation induced gate overdrive variation completely counterbalances the carrier mobility variation, thereby providing temperature variation insensitive constant MOSFET drain current. Similar to the MOSFET drain currents, the speed characteristics of CMOS circuits can also be made insensitive to temperature variations with the supply voltage optimization technique. Circuits in a 180nm CMOS technology display a temperature variation insensitive
behavior when operated at a supply voltage 44% to 47% lower than the nominal supply voltage ($V_{DD,\text{nominal}} = 1.8V$). Similarly, the supply voltages providing temperature variation insensitivities are 67% to 68% lower than the nominal supply voltage ($V_{DD,\text{nominal}} = 1.0V$) for the circuits in a 65nm CMOS technology.

An alternative design methodology based on optimizing the threshold voltages to suppress the propagation delay variations when the temperature fluctuates is also evaluated in this dissertation. For circuits operating at the nominal supply voltage, the gate overdrive sensitivity to temperature fluctuations can be enhanced by increasing the device threshold voltage. For a particular higher threshold voltage, the temperature fluctuation induced gate overdrive variation completely counterbalances the carrier mobility variation, thereby providing temperature variation insensitive constant MOSFET drain current. The threshold voltage optimization technique also achieves temperature variation insensitive propagation delay characteristics in CMOS circuits. Results indicate that the circuits operating at the nominal supply voltage in a 180nm CMOS technology are insensitive to temperature variations when the threshold voltage is 2.7x to 2.9x higher than the nominal device threshold voltage ($|V_{t0,\text{nominal}}| = 0.46V$). Similarly, the delay variations of circuits operating at the nominal supply voltage in a 65nm CMOS technology are suppressed when the device threshold voltage is 4.2x to 4.3x higher than the nominal device threshold voltage ($|V_{t0,\text{nominal}}| = 0.22V$).

The speed and the energy trade-offs with the proposed supply and threshold voltage optimization techniques are compared in this dissertation. The propagation delay is up to 25.6x higher when circuits in a 65nm CMOS technology are operated at the threshold voltages that achieve temperature variation insensitive speed ($V_{t0,\text{insensitive}}$) as compared to the circuits that are operated at the supply voltage that provides temperature variation insensitivity ($V_{DD,\text{insensitive}}$). Furthermore, the energy consumption at $V_{t0,\text{insensitive}}$ is higher than the energy per cycle at $V_{DD,\text{insensitive}}$ by up to 3.3x and 7.7x for the circuits in 180nm and 65nm CMOS technologies, respectively. Higher supply voltages are preferable in speed critical applications. The performance degradation with the threshold voltage optimization technique diminishes the potential speed gains by employing a higher nominal supply voltage. The energy savings achieved by the threshold voltage optimization technique is also lower as compared to the supply voltage optimization
technique. The supply voltage optimization technique is therefore more effective in simultaneously achieving energy efficiency and temperature variation resilience with a smaller speed penalty as compared to the threshold voltage optimization approach in CMOS integrated circuits.

Integrated circuits operating at scaled supply voltages consume lower power at the cost of reduced speed. The design methodology of optimizing the supply voltage for temperature variation insensitive circuit performance is particularly attractive in low power applications with relaxed speed requirements. The speed and energy consumption with the supply voltage optimization technique are compared with the other low power techniques based on supply voltage scaling in Chapter 6 of this dissertation. Supply voltage scaling is an effective technique to lower all the primary components of power consumption in CMOS circuits. As the supply voltage is reduced, the energy per cycle decreases while the propagation delay increases. The energy-delay product, therefore, has a minimum. Furthermore, the subthreshold operation minimizes the total energy consumption of a CMOS circuit. The supply voltages that achieve minimum energy-delay product and minimum energy consumption are identified in this dissertation for the circuits in 180nm and 65nm CMOS technologies. Results indicate that the propagation delay of low-power integrated circuits with deeply scaled supply voltages is highly sensitive to temperature fluctuations.

At the supply voltages for minimum energy-delay product in a 180nm CMOS technology, the energy per cycle is 63.4% to 78.4% lower than the energy per cycle at the nominal supply voltage. Similarly, the energy per cycle at the supply voltages that yield temperature variation insensitive circuit performance is 72.9% to 75.2% lower than the energy at the nominal supply voltage. Furthermore, the difference of the minimum achievable energy-delay product and the energy-delay product at the supply voltages for temperature variation insensitive circuit performance is less than 3%. For circuits in a 65nm CMOS technology, the minimum achievable energy is 65% to 89% lower than the energy per switching cycle at the nominal supply voltage. The energy consumption at $V_{DD, insensitive}$ is about 12% higher as compared to the minimum achievable energy consumption. Integrated circuits with supply voltages optimized for temperature fluctuation insensitive speed characteristics display significantly reduced energy
consumption and energy-delay product similar to the circuits optimized for minimum energy-delay product and minimum energy consumption. Energy efficiency and temperature fluctuation tolerance can therefore be simultaneously achieved with the proposed supply voltage optimization technique as compared to the traditional margin based designs optimized for functionality at the worst case die temperature.

The supply voltages providing minimum energy are observed in the subthreshold region. The switching current in these ultra-low-voltage circuits is the subthreshold leakage current. Subthreshold leakage current is extremely sensitive to temperature fluctuations. A small change in the die temperature exponentially alters the subthreshold leakage current. Furthermore, in these circuits, the circuit speed increases with the die temperature. The reversal in the temperature dependent propagation delay characteristics coupled with the high sensitivity of circuit speed to temperature fluctuations provides opportunities for reducing the energy consumption without degrading the clock frequency at elevated die temperatures in ultra-low-voltage circuits. New temperature-adaptive dynamic voltage tuning techniques are proposed in this dissertation for reducing the active-mode energy consumption by exploiting the excessive timing slack produced in the clock-period of ultra-low-voltage CMOS circuits at elevated temperatures. An algorithm to identify the supply voltages that provide minimum energy consumption is also presented.

The constant clock frequency of a standard circuit is determined by the circuit speed along the critical delay paths at the worst case die temperature. In circuits optimized for minimum energy consumption, the worst case circuit speed is observed at the lowest operating temperature. The circuit operates faster at elevated temperatures, thereby producing excessive timing slack in the constant-clock-period. At elevated temperatures the total energy consumption also increases due to the exponential increase in the subthreshold leakage current. With the proposed temperature-adaptive supply voltage tuning technique (TA-DVS), the supply voltage of the circuit is dynamically scaled below the supply voltage providing minimum energy consumption at 25°C ($V_{DD,25}$) while maintaining the constant-clock frequency of the circuit as the die temperature increases. The supply voltage of the circuit is tuned until the high temperature circuit performance at the scaled supply voltage matches the circuit performance of the standard constant-$V_{DD}$
and constant-\(f/s\) circuit operating at \(V_{DD-25}\). An alternative voltage tuning technique based on temperature-adaptive body-bias (TA-BB) is also presented in this dissertation. In ultra-low-voltage circuits exhibiting reversed temperature dependence, the threshold voltage of devices is dynamically increased through reverse body-bias at elevated temperatures to exponentially reduce the leakage current without degrading the clock frequency.

The energy savings achieved with the two different temperature-adaptive voltage tuning techniques are compared. The high temperature energy efficiency is enhanced by up to 40% with the temperature-adaptive dynamic supply voltage tuning technique. Alternatively, the energy consumption increases by up to 6x with the temperature-adaptive reverse body-bias technique as compared to the standard-zero-body-bias circuits operating at \(V_{DD-25}\). The increase in the band-to-band tunneling junction leakage currents at higher reverse-body-bias voltages increase the energy consumption of circuits with the TA-BB.

The performance of digital circuits can be enhanced by increasing the capacity of the integrated on-chip memory. Larger embedded memories enhance the effective memory bandwidth by reducing the average memory access time. Similar to high performance microprocessors, static random access memory (SRAM) arrays occupy a significant percent of the total chip area in low-power mobile processors. The large switching capacitance in the bit-lines and the word-lines of an SRAM array contribute to the high dynamic switching power consumption during a memory access. Furthermore, the leakage currents of large memory banks can dominate the total power consumption in ultra-low-voltage circuits. To achieve enhanced reliability and longer battery life-time in portable applications, the power consumed by the memory arrays should be reduced.

Similar to the circuits employed in the logic core, memory arrays optimized for minimum energy consumption exhibit reversed temperature dependence. Both the read and the write propagation delays of subthreshold memory banks are reduced at elevated temperatures. The reduction in the high temperature propagation delay provides new opportunities to reduce the energy consumption of subthreshold SRAM arrays at elevated temperatures. The effectiveness of the temperature-adaptive schemes proposed in Chapter
7 is evaluated for ultra-low-voltage memory banks. The sizing constraints for data stability in a subthreshold SRAM bit cell are also presented.

Results indicate that the high temperature energy consumption is reduced by up to 32.8% with the proposed TA-DVS technique without degrading the clock frequency of an SRAM array in a 180nm CMOS technology. The high temperature energy consumption with the TA-BB technique, however, increases by up to 1.47x. The influence of the temperature-adaptive voltage tuning techniques on the data stability of the subthreshold memory cells is also presented in this dissertation. The effectiveness of the proposed TA-DVS scheme under the presence of process parameter and environmental variations is evaluated.

Process and environmental variations are enhanced with each new technology generation. Enhancing the reliability of circuits in the presence of variations will be a primary requirement in the design of future digital integrated circuits. Temperature effects cannot be ignored in nano-CMOS designs due to the enhanced sensitivity of the device and circuit characteristics to the variations in the die temperature. Developing temperature variation tolerant and temperature-adaptive circuit techniques are necessary for enhancing the reliability of circuits that operate in environments with significant temperature fluctuations. Several new techniques that simultaneously enhance the energy efficiency and the reliability of digital CMOS circuits when the temperature fluctuates are proposed in this dissertation.
Chapter 10

Future Research

Some interesting research activities for the near future are briefly described in this chapter. The advantages of the temperature-adaptive voltage scaling techniques for high-temperature energy efficiency are presented in Chapters 7 and 8. The challenges of developing a temperature-adaptive dynamic voltage scaling power supply are presented in Section 10.1. The reduction in the supply voltage to threshold voltage ratio with each new technology generation enhances the sensitivity of the gate-overdrive. Circuits in the future CMOS technologies will exhibit reversed temperature dependence even at the nominal supply voltages [101]. Research plans to develop high-speed signaling techniques with the future nanoscale wire technologies taking advantage of the enhanced circuit speed at elevated temperatures is described in Section 10.2.

There is an increasing market demand for low-power mobile systems [1], [117]. Reducing the data retention power of DRAMs can significantly enhance the battery lifetime of portable applications. Future research plans to explore temperature-adaptive power reduction techniques for enhanced energy efficiency in DRAMs are discussed in Section 10.3.

The scaling of planar silicon devices has been continuing for approximately five decades. However, due to the fabrication related difficulties and the degradation in device electrical characteristics caused by enhanced short channel effects, scaling of standard planar MOSFETs is expected to slow down within the next decade [7]. Today, the integrated circuit technologies shift from the single-gate planar MOSFETs to the multiple-gate three dimensional devices [104], [105], [110]. FinFETs are the most attractive choice among the multiple-gate device architectures due to the self alignment of the two vertical gates and the fabrication compatibility of the FinFETs with the existing standard CMOS fabrication process [105]. Future research directions to develop temperature-variation tolerant CMOS ICs with the emerging FinFET technology are described in Section 10.4.
Performance of nano-CMOS circuits becomes increasingly dominated by wire delays due to the decreasing wire pitch and the increasing die size in each new technology generation [132]. Furthermore, the heterogeneous integration of different technologies on a single die becomes increasingly desirable, for which planar (two-dimensional) ICs may not be suitable. Novel three dimensional (3-D) chip design strategies are being investigated for developing future systems-on-chip [132], [136]-[140]. Interesting research topics related to 3-D integration are presented in Section 10.5.

10.1. Temperature-Adaptive Voltage Scaling Power Supplies

Temperature-adaptive dynamic supply voltage scaling (TA-DVS) schemes are presented in Chapters 7 and 8 for enhancing the high temperature energy efficiency of ultra-low-voltage subthreshold logic circuits. Research ideas to implement a high energy-efficiency and compact temperature-adaptive voltage scaling power supply for monolithic integration are presented in this section.

A system with temperature-adaptive supply voltage tuning capability is illustrated in Fig. 10.1. The low-temperature-$V_{DD}$ and the target operating frequency ($f_{\text{target}}$) of the integrated circuit for achieving minimum energy consumption are determined at the lowest operating temperature according to the algorithm presented in Chapter 7 (Fig. 7.1). A ring oscillator providing a replica of the critical path of the entire integrated circuit can be employed to track the fluctuations of the critical-path propagation delay with the variations of the ambient temperature at a specific supply voltage. A relatively uniform temperature is assumed across the die with this technique. Note that the uniform die temperature assumption is typically satisfied with the ultra-low-voltage subthreshold logic circuits where TA-DVS is attractive. The ring oscillator translates the variations in the die temperature to a specific clock frequency ($f_{\text{clock}}$) for a specific power supply voltage generated by the DC-DC converter. As the die temperature increases, the ring oscillator frequency ($f_{\text{clock}}$) also increases due to the enhanced gate overdrive voltages of the MOSFETs. The ring oscillator frequency is compared to the target clock frequency ($f_{\text{target}}$), generating a frequency error signal ($f_{\text{error}}$). The pulse width modulator using this error signal generates the control signals for the DC-DC converter to either modify or
maintain the output voltage. The power supply voltage is, thereby, dynamically tuned based on the variations of the die temperature using the closed loop feedback circuitry shown in Fig. 10.1.

Integrated circuits functioning with ultra-low power-supply-voltages are reported in [123], [125], and [141]. The multiply and accumulate unit designed in [123] exhibits functionality at 175mV. However, the generation and the distribution mechanisms of such ultra-low voltages are not clearly described in [123], [125], and [141]. A dynamic supply voltage scaling system for operating the circuits in the subthreshold region during periods of low-throughput requirements is presented in [142]. The design presented in [142] however requires multiple external supply voltages to efficiently reduce the power consumption. Designing a high-precision and energy efficient ultra-low-voltage power supply with dynamic voltage tuning capability for the temperature-adaptive schemes proposed in this dissertation is crucial for effective energy reduction at the system level. Developing novel techniques to lower the energy losses in the voltage scaling system will be an important future research topic.

Fig. 10.1. Temperature-adaptive dynamic supply voltage scaling technique.
Switching DC-DC converters are preferred over linear regulators due to the typically higher energy conversion efficiency [143], [146]. A primary factor that determines the quality of a DC-DC converter is the stability of the output voltage over a wide range of input voltages [1], [117], [122], [144]-[146]. The output stability is characterized by the peak-to-peak output voltage ripple under changing conditions of the input voltage. A simplified buck converter is shown in Fig. 10.2 [143]. Increasing the filter capacitance (C_f) reduces the ripples in the output voltage (V_{Out}) and improves the low-voltage conversion efficiency [1], [122]. Alternatively, the transition time of the DC-DC converter can be reduced by using a smaller filter capacitance. The transition time is the time required to transition from one voltage level to another at the output of the voltage regulator. Reducing both the transition time and the output voltage ripple is critical for providing effective power savings with a dynamic voltage scaling technique.

The increase in the energy consumption of the DC-DC converter with the output ripple voltage is shown in Fig. 10.3 [122]. The energy loss is less than 5% for higher supply voltages. The energy overhead with the DC-DC converter however increases at a lower supply voltage, as shown in Fig. 10.3 [122]. The value of the filter capacitance should therefore be carefully chosen in the design of an ultra-low-voltage DC-DC converter. A large filter capacitor reduces the output voltage ripple, thereby improving the conversion efficiency of the voltage regulator. Alternatively, the transition time increases with a large filter capacitance. Exploring the optimum filter capacitance value that effectively improves the energy savings with the temperature-adaptive supply voltage scaling technique is an important research area.

Estimating the impact of temperature fluctuations on the conversion efficiency of a voltage regulator will be a primary research goal. Developing techniques for implementing low-voltage DC-DC converters with higher voltage conversion efficiency, smaller area penalty, and smaller voltage settling time will be an important research objective. Characterizing the efficiency and the quality of the voltage converter under the impact of process-parameter and supply voltage variations will be another research topic.
10.2. Thermal Variation Aware Interconnect Design in Nano-CMOS Technologies

Critical dimensions of a transistor are scaled with each new technology generation [1], [117]. The reduction in the size and the cost of transistors enables the production of
integrated circuits with enhanced functionality. A complex integrated circuit with many functional blocks is shown in Fig. 10.4. Global interconnects are employed to connect the different functional blocks across an IC, as shown in Fig. 10.4.

The reduction in the defect density due to the maturing fabrication technology enables the manufacturing of integrated circuits with larger dies, as explained in Chapter 1. The increasing die area with each new technology generation increases the total interconnect length in an integrated circuit. The propagation delay of a wire is a function of the interconnect length [44]. Both the wire resistance \( R \) and the wire capacitance \( C \) increase with the increasing interconnect length. The \( RC \) delay of an interconnect therefore increases quadratically with the increasing wire length \( l \) [44].

The propagation delay of interconnect can be reduced using the repeater insertion technique described in [44]. A complex integrated circuit with repeaters is shown in Fig. 10.5. Each wire is split into \( N \) segments with the repeater insertion technique. An inverter (repeater) actively drives each of the \( N \) segments, as shown in Fig. 10.6. With repeaters, the interconnect \( RC \) delay tends to increase linearly with the wire length, as explained in [44]. The optimum number of repeaters for a long interconnect and the optimum size of each repeater are [26], [44].
\[ S_{opt} = \sqrt{\frac{R_d C_w}{R_w C_d}}, \quad (10.1) \]

\[ N_{opt} = \frac{t_{p-wire}}{t_{p-repeater}}, \quad (10.2) \]

where \( N_{opt} \), \( t_{p-wire} \), and \( t_{p-repeater} \) are the optimum number of repeaters, the propagation delay of the unbuffered wire, and the propagation delay of the repeater, respectively. \( R_w \) and \( C_w \) are the resistance and the capacitance per unit length, respectively, of the unbuffered wire while \( R_d \) and \( C_d \) are the resistance and the input capacitance, respectively, of a minimum sized repeater. \( S_{opt} \) is the size of the repeater as compared to a minimum sized repeater.

Variations in the die temperature alter the speed characteristics of the wires as well as the devices, as explained in Chapter 3. The temperature gradients on a typical integrated circuit with imbalanced local utilization and activity factor are shown in Fig. 10.7. Increase in the die temperature increases the propagation delay of interconnects in the current CMOS technologies. Both the repeater delay (due to the dominant carrier mobility variations in the devices) and the wire delay of each interconnect segment (due
to the increase in the wire resistance) increase at elevated temperatures. The worst-case interconnect propagation delay is therefore observed at the highest temperature along a typical repeater based interconnect line.

Fig. 10.6. Interconnects. (a) Without repeaters. (b) With repeaters.

Fig. 10.7. Typical integrated circuit with on-die temperature gradients.
The supply and threshold voltages are scaled with each new technology generation [35]. The supply and threshold voltage scaling trends are shown in Fig. 10.8 [108]. The supply voltage is scaled primarily based on the maximum clock frequency requirement in a new technology generation. The speed of a circuit can be further enhanced by scaling the threshold voltages. The threshold voltages are, however, scaled at a much slower rate as compared to the supply voltages due to the subthreshold leakage power constraints. The supply voltage to threshold voltage ratio is reduced with each new technology generation, as shown in Fig. 10.8. The variation of the gate overdrive (\(V_{DD} - V_t\)) with the temperature, therefore, plays an increasingly important role in determining the speed characteristics of circuits in scaled future CMOS technologies.

![Fig. 10.8. The supply and threshold voltages in different CMOS technology generations.](image)

Circuits in future CMOS technologies will exhibit reversed temperature dependence at the nominal supply voltage due to the enhanced gate-overdrive variations [101]. Contrary to the older technologies, in a 45nm CMOS technology, the speed of the repeaters tends to increase while the wire delays become longer at elevated temperatures. The enhancement of the buffer speed with the increased temperature adds a new dimension to the design of repeater driven interconnects for the future integrated systems. Exploring this intrinsic counterbalancing effect of the opposite thermal behaviors of the
nano-gates and nano-wires in deeply scaled CMOS technologies will be a primary research focus. Developing temperature-aware repeater insertion techniques to reduce the interconnect power consumption by exploiting the enhanced high temperature speed of CMOS gates with reverse temperature dependence will be an important research topic.

**10.3. Temperature-Adaptive Dynamic Memory**

The customer demand for portable applications has been continuously increasing since the late 1990s. Dynamic random access memory (DRAM) circuits are widely employed in mobile applications to enhance the performance of battery operated devices [133]. In a typical DRAM cell, data is stored using the MOSFET junction capacitance. The stored data is lost over time due to the leakage currents produced by a MOSFET. DRAMs employ a refresh circuitry to maintain the data stored in each DRAM cell. A DRAM refresh is necessary to prevent the loss of data. However, continuous refreshing of the DRAM cells causes significant power consumption. Reducing the power consumed during a DRAM refresh cycle can substantially increase the battery lifetime in portable applications.

The data retention capability of DRAMs is altered due to the fluctuations in the die temperature. The required data refresh rate for a DRAM cell at various temperatures is shown in Fig. 10.9. At time $t_0$, a logic high voltage is written into the DRAM cell, as shown in Fig. 10.9. The stored voltage leaks with time due to the MOSFET leakage currents. The DRAM cell can be refreshed after the storage node voltage falls below the refresh threshold voltage, as shown in Fig. 10.9. The refresh rate of a DRAM cell is determined by the total leakage current produced by a MOSFET at different temperatures. Subthreshold leakage current increases exponentially at elevated temperature [78]. The data stored in a DRAM cell is therefore lost at a faster rate at high die temperatures. In order to prevent the loss of data, the required data refresh period is smaller at elevated temperatures as compared to the refresh period at low temperatures, as shown in Fig. 10.9. The DRAMs however typically employ fixed refresh periods [134]. For maintaining data at all die temperatures, the data refresh period is fixed based on the refresh period required by the worst-case leakage condition observed at the highest
temperature. Fixing the refresh period to cater the needs for a high temperature circuit operation, however, wastes significant recharging power when the DRAM operates at a lower temperature. Dynamically adjusting the data refresh period of a DRAM based on the die temperature is desirable for reducing the recharging power consumed by DRAMs [134].

A temperature-adaptive self-recharging circuit for dynamically altering the data refresh period of a DRAM based on the die temperature is proposed in [134]. The comparator circuitry proposed in [134] to trigger the recharging cycle based on the temperature fluctuations is shown in Fig. 10.10. $V_{ref}$ is the refresh threshold voltage. When the cell voltage falls below $V_{ref}$, trigger is activated to signal the need for a DRAM refresh. The refresh threshold voltage ($V_{ref}$) increases with the temperature [134]. The DRAM refresh period is therefore small at elevated temperatures. Alternatively, the smaller magnitude of $V_{ref}$ reduces the rate of DRAM refresh at low temperatures. The effectiveness of the temperature-adaptive self-recharging circuit is however determined by the bias voltage ($V_{bias}$) under which the circuitry operates [134].

![Fig. 10.9. DRAM refresh cycles at different temperatures.](image-url)
Fig. 10.10. Comparator circuitry for temperature-adaptive DRAM refresh cycle [134].

Trigger is activated to refresh the DRAM cells.

According to [134], $V_{bias}$ is an externally supplied analog voltage. The generation and distribution of the $V_{bias}$ is not clearly described in [134]. Furthermore, the effect of parameter variations and the power/area overhead of the comparator circuitry are not presented in [134]. The comparator circuit must be process-variation tolerant with a small power and area overhead. Developing a novel low-power and robust control technique for the reliable and temperature-adaptive refreshing of the data in DRAMs will be another future research goal.

10.4. Thermal-Aware FinFET Device Optimizations

Scaling is the primary thrust behind the advancement of CMOS technology, as emphasized throughout this dissertation. However, the increased subthreshold and gate-oxide leakage currents coupled with the enhanced device sensitivity to process parameter fluctuations have reduced the pace of technology scaling [7]. Multi-gate three dimensional devices offer distinct advantages as compared to the standard single-gate planar MOSFETs in suppressing the leakage currents in addition to providing significant speed enhancement at scaled device dimensions [109].
FinFETs are the most attractive choice among the multi-gate device architectures due to the self-alignment of the two vertical gates and the fabrication compatibility of the FinFETs with the existing standard CMOS fabrication process [105]. Furthermore, the short-channel effects and the drain induced barrier lowering (DIBL) are significantly suppressed in FinFETs as compared to single-gate planar MOSFETs. In this section, future research directions to develop temperature variation tolerant CMOS ICs with the emerging FinFET technologies are described. New device architectures will be explored to achieve temperature variation insensitive constant drain current characteristics with the FinFETs.

The 3-D architectures of tied-gate and independent-gate FinFETs are shown in Fig. 10.11. The cross-sectional top view of a FinFET is shown in Fig. 10.12. $H_{\text{fin}}$ is the height of the fin. The body-thickness (fin-width) is $t_{\text{si}}$. A fabrication process is described in [113] for implementing tied-gate and independent-gate FinFETs on the same die. In [110], [111], and [112] independent gate FinFETs are utilized to reduce the number of transistors required for implementing specific logic functions as compared to the standard circuits with tied-gate FinFETs. 2-input NAND gates implemented using standard single gate planar MOSFETs and independent gate FinFETs are shown in Fig. 10.13.

![Fig. 10.11. FinFET architectures. (a) Tied-gate FinFET. (b) Independent-gate FinFET.](image-url)
The power consumption and the power density of high performance ICs are expected to continue to increase with each new technology generation. The increasing power consumption causes higher heat dissipation. The hot-spots in future technology generations are therefore expected to be hotter as compared to today’s high performance ICs. Circuits with the emerging technologies, such as the FinFET technology, are therefore expected to experience higher on-chip die temperature variations as compared to the current advanced ICs based on the traditional single-gate planar MOSFETs.

The fluctuation of the temperature alters the speed and power consumption characteristics of FinFET based circuits. The supply and threshold voltages that provide
temperature variation tolerant circuit performance with the FinFET technology will be identified. The development of new technology guidelines to enhance the temperature variation resilience of FinFET integrated circuits will be an important research objective. The short channel effects, drain-induced barrier lowering, and subthreshold slope observed with different device architectures will be compared. The speed and power characteristics of the FinFET circuits optimized for temperature variation insensitive performance will be evaluated.

10.5. 3-D Stacked Integrated Circuits

Continuous scaling of device dimensions reduces the gate delays. Technology scaling, however, increases the interconnect delay due to the increase in the wire resistance and wire length [132], [136]-[139]. The gate and wire delays in different CMOS technology generations are compared in Fig. 10.14 [132]. Repeater insertion and wire sizing techniques have been proposed to reduce the delay of interconnects [137]. The increase in the clock frequency with technology scaling, however, reduces the benefits of repeater insertion [137]. Furthermore, interconnect loading effects increase the power consumption of high performance chips [132]. To sustain the performance improvements in future technology generations, unconventional design methodologies are required.
Three-dimensional (3-D) integration is a promising alternative which offers the opportunity to relieve the deleterious effects of long interconnects [132], [136]-[138]. Furthermore, 3-D chip design technology can be exploited to build a system-on-chip (SoC) by placing circuits with different voltage and performance requirements in different layers of a silicon stack. The different layers of a 3-D IC are connected with inter-layer interconnects [132], [136], [137]. The schematic diagrams of an SoC using the standard planar 2-D and the emerging 3-D technologies are shown in Fig. 10.15. Three dimensional integration can reduce the interconnect length, thereby reducing the wiring capacitance, the power dissipation, and the chip area [132]. Furthermore, digital and analog components in mixed signal systems can be placed in different silicon layers thereby achieving better noise immunity [132].
Heat is generated in an integrated circuit due to the switching activity of the devices and the wires. The generated heat flows through the substrate and the package. The ambient temperature on a die is raised due to the heat flow [44]. The heat flow path in a 2-D IC and a 3-D chip are compared in Fig. 10.16. A 2-D planar IC has only one layer of silicon, as shown in Fig. 10.16a. The heat generated in the substrate of the 2-D planar chip flows through the package and the heat sink. Alternatively, a 3-D IC has many stacked silicon layers, as shown in Fig. 10.16b. The heat generated in the nth silicon layer has to pass through (n-1) silicon layers before reaching the package. This increases the thermal resistance of the heat flow path. Furthermore, the glue layers have lesser thermal conductivity [132]. Due to the increased thermal resistance in the heat flow path of a 3-D chip, the temperature and the power density of the 3-D IC increase with the increasing number of silicon layers, as shown in Fig. 10.17 [132]. In addition to the temperature variations within a single plane (due to the imbalance switching activity), significant temperature gradients are expected between the different silicon layers of a 3-D design [132].
Fig. 10.16. Heat flow path in typical ICs. (a) Schematic view of heat flow in a 2-D planar IC. (b) Schematic of an ‘n’ layer 3-D IC with heat sink at the bottom.

Fig. 10.17. Temperature profile of a 3-D chip with increasing number of silicon layers. The temperature and the power density increase as more silicon layers are packed in the 3-D IC [132].

Temperature fluctuations alter the speed characteristics of CMOS gates and interconnects [75]-[78], [91]. Furthermore, the maximum temperature observed in each silicon layer is determined by the distance of the silicon layer from the package [132]. Many techniques have been proposed to reduce the increase in the temperature at the top
layers of a 3-D IC. A thermal via placement technique is proposed in [136] to reduce the thermal resistance of a 3-D chip. However, placing thermal vias in areas of high temperature such as the upper layer has little impact in reducing the thermal problems [136]. Furthermore, thermal vias occupy valuable routing space.

Improving interconnect performance can significantly enhance the speed of the integrated circuit in a 3-D chip. Design methodologies to reduce the propagation delay of inter-layer interconnects is proposed in [137] and [140]. The proposed techniques, however, do not consider the temperature profile of a 3-D chip. New methodologies will be investigated to accurately characterize and reduce to the interconnect propagation delays considering the unconventional temperature gradients and the heat dissipation challenges in 3-D integrated circuits. The realistic thermal modeling of 3-D integrated circuits will be an important research area. Reducing the maximum temperature observed particularly in the upper layers of a 3-D IC will be a primary research objective.


[81] Intel Corporation. *www.intel.com*


Appendix A Publications

JOURNAL PUBLICATIONS


CONFERENCE PUBLICATIONS


