Linköping Studies in Science and Technology
Dissertation No. 1197

Low-Power Clocking and Circuit Techniques for Leakage and Process Variation Compensation

Martin Hansson

Linköping University
INSTITUTE OF TECHNOLOGY

Electronic Devices
Department of Electrical Engineering
Linköpings universitet, SE-581 83 Linköping, Sweden
Linköping 2008

ISBN 978-91-7393-847-1
ISSN 0345-7524

Low-Power Clocking and Circuit Techniques for Leakage and Process Variation Compensation

Martin Hansson

Copyright © Martin Hansson, 2008

ISBN: 978-91-7393-847-1

Linköping Studies in Science and Technology
Dissertation No. 1197
ISSN: 0345-7524

Electronic Devices
Department of Electrical Engineering
Linköping University
SE-581 83 Linköping
SWEDEN

Cover image:
Circuit model for the low-power LC-tank energy recovering clocking technique presented with chip measurements in Chapter 6 of this thesis. The LC-tank drives the flip-flops in the data paths directly without intermediate buffering. The capacitance *C* in the figure models the total clock load in the flip-flops and the clock distribution, while the resistance *R* models the parasitic losses in the inductance and interconnects.

Printed by LiU-Tryck, Linköping University
Linköping, Sweden, 2008

Abstract

Over the last four decades the integrated circuit industry has evolved in a tremendous pace. This success has been driven by the scaling of device sizes leading to higher and higher integration capability, which have enabled more functionality and higher performance. The impressive evolution of modern high-performance microprocessors have resulted in chips with over a billion transistors as well as multi-GHz clock frequencies. As the silicon integrated circuit industry moves further into the nanometer regime, scaling of device sizes is still predicted to continue at least into the near future. However, there are a number of challenges to overcome to be able to continue the increase of integration at the same pace. Three of the major challenges are increasing power dissipation due to clocking of synchronous circuit, increasing leakage currents causing growing static power dissipation and reduced circuit robustness, and finally increasing spread in circuit parameters due to physical limitations in the manufacturing process. This thesis presents a number of circuit techniques that aims to help in all three of the mentioned challenges.

Power dissipation related to the clock generation and distribution is identified as the dominating contributor of the total active power dissipation for multi-GHz systems. As the complexity and size of synchronous systems continues to increase, clock power will also increase. This makes novel power reduction techniques absolutely crucial in future VLSI design. In this thesis an energy recovering clocking technique aimed at reducing the total chip clock power is presented. Based on theoretical analysis the technique is shown to enable considerable clock power savings. Moreover, the impact of the proposed technique on conventional flip-flop topologies is studied. Measurements on an experimental chip design proves the technique, and shows more than 56% lower

clock power compared to conventional clock distribution techniques at clock frequencies up to 1.76 GHz.

Static leakage power dissipation is a considerable contributor to the total power dissipation. This power is dissipated even for circuits that are idle and not contributing to the operation. Hence, with increasing number of transistors on each chip, circuit techniques which reduce the static leakage currents are necessary. In this thesis a technique is discussed which reduces the static leakage current in a microcode ROM resulting in 30% reduction of the leakage power with no area or performance penalty.

Apart from increasing static power dissipation the increasing leakage currents also impact the robustness constraints of the circuits. This is important for regenerative circuits like flip-flops and latches where a changed state due to leakage will lead to loss of functionality. This is a serious issue especially for high-performance dynamic circuits, which are attractive in order to limit the clock load in the design. However, with the increasing leakage the robustness of dynamic circuits reduces dramatically. To improve the leakage robustness for sub-90 nm low clock load dynamic flip-flops, a novel keeper technique is proposed. The proposed keeper utilizes a scalable and simple leakage compensation technique, which is implemented on a reconfigurable flip-flop. At normal clock frequencies the flip-flop is configured in dynamic mode, and reduces the clock power by 25% due to the lower clock load. During any low-frequency operation, the flip-flop is configured as a static flip-flop retaining full functional robustness.

As scaling continues further towards the fundamental atomistic limits, several challenges arise for continuing industrial device integration. Large inaccuracies in lithography process, impurities in manufacturing, and reduced control of dopant levels during implantation all cause increasing statistical spread of performance, power, and robustness of the devices. In order to compensate the impact of the increasingly large process variations on latches and flip-flops, a reconfigurable keeper technique is presented in this thesis. In contrast to the traditional design for worst-case process corners, a variable keeper circuit is utilized. The proposed reconfigurable keeper preserves the robustness of storage nodes across the process corners without degrading the overall chip performance.

Populärvetenskaplig sammanfattning

Utvecklingen inom halvledarelektronik har de senaste fyra årtiondena utvecklats från sin tidiga barndom under 60-talet, då hundratals transistorer på ett chip var ansett som science fiction, tills idag när mikroprocessorer till dagens datorer innehåller miljarder av transistorer, som utför beräkningar åtskilliga miljarder gånger per sekund. Denna fantastiska framgång har möjliggjorts med hjälp av nedskalning av transistorernas storlekar, vilket har medfört att mer funktionallitet och fler kretsar har kunnat integreras i avancerade system på samma chip. Flertalet teknologiska innovationer de senaste årtiondena, som internet, bärbara datorer och mobiltelefoner, hade inte varit möjligt utan denna utveckling. Priset för denna höga beräkningshastighet och det stora antalet transistorer är dock en ökad effektförbrukning.

I stort sett alla microprocessorer idag ordnar alla operationer genom att synkronisera dem med en eller flera klocksignaler. Dessa signaler behöver ofta distribueras över hela chippet och driva alla synkroniseringskretsar med klockfrekvenser på åtskilliga miljarder svängningar per sekund. Detta förbrukar en betydande och växande andel av den effekt en mikroprocessor använder. För att minska denna effektförbrukning krävs nya kretstekniker, som minskar belastingen på distributionsnätet, men även nya och innovativa metoder för att distribuera signalerna på ett energieffektivt sätt. I denna avhandling presenteras en klockningsteknik för att distribuera klocksignaler i digitala system. Principen bakom tekniken är att återanvända den energi som åtgår till att ladda upp klocklasten. Teoretiska resonemang har visat att stora energibesparingar är möjliga, och praktiska mätningar på tillverkade experimentchip har visat att effektförbrukingen kan mer än halveras vid klockfrekvenser på upp till 1.76 GHz.

Idealt har alla transistorer i digitala system betraktats som enkla switchar som leder ström när de är på och inte leder ström när de är av. Under de tidiga åren av integrerade kretsar var detta ett fullt tillräckligt sätt att se på digitala kretsar. Men med de konstant minskande storlekarna har fysikaliska fenomen gjort att denna ideala syn på transistorer inom digitalteknik förändrats. Transistorerna har istället blivit svårare att stänga av helt, vilket har lett till att en växande andel av den totala effektförbrukningen kommer från så kallat läckage genom transistorer som egentligen skulle vara avstängda. Med ett antal hundra miljoner transistorer på ett chip så kan denna effekt uppgå till en stor andel av den totala effektförbrukingen. Ytterliggare en konsekvens av de minskande storlekarna är att tillverkningsprocessen blir mer och mer komplex. Begränsningar i bland annat de optiska system som används vid tillverkningen gör att precisionen av de geometriska storlekarna påverkas. Detta leder i sin tur till att både prestandan och effektförbrukningen i kretsarna varierar mer och mer från de typiska värdena som man vanligtvis använder under design av kretsarna.

I denna avhandling presenteras ett antal kretstekniker vilka syftar till att kompensera för det ökande läckaget för kretsar där ett visst tillstånd måste behållas i minnet. Dessutom föreslås en metod för att reducera den totala läckageströmmen för avancerade läsminnen i mikroprocessorer. Slutligen föreslås en teknik för att i efterhand kompensera för de ökande variationerna i prestanda och tillförlitlighet på grund av tillverkningsosäkerheter.

Preface

This dissertation presents the research I have been involved in during the period June 2003 through December 2007 at the Electronic Devices group, Department of Electrical Engineering, Linköping University, Sweden. This work has been supported by Intel Corporation and the Swedish Foundation for Strategic Research (SSF).

I began my research working on low clock load techniques for flip-flops, but the research topic have grown to include global low-power clocking techniques for multi-GHz VLSI designs, process variation tolerant circuit techniques, and circuit techniques for low leakage and leakage tolerance. My research has resulted in a number of papers published in international conferences and journals. The following papers are included in the thesis:

- **Paper 1: Martin Hansson** and Atila Alvandpour, "A Low Clock Load Conditional Flip-Flop," in *Proceedings of IEEE International System-on-Chip Conference*, pp. 169-170, Santa Clara, California, USA, September 2004.

- **Paper 2: Martin Hansson** and Atila Alvandpour, "Power-Performance Analysis of Sinusoidally Clocked Flip-Flops," in *Proceedings of 23rd IEEE NORCHIP Conference,* pp. 153-156, Oulu, Finland, November 2005.

- **Paper 3: Martin Hansson**, Behzad Mesgarzadeh, and Atila Alvandpour, "1.56 GHz On-chip Resonant Clocking in 130nm CMOS," in *Proceedings of the IEEE Custom Integrated Circuit Conference*, pp. 241-244, San Jose, California, USA, September 2006.

- **Paper 4:** Behzad Mesgarzadeh**, Martin Hansson**, and Atila Alvandpour, "Jitter Characteristic in Resonant Clock Distribution," in *Proceedings of the 32nd European Solid-State Circuit Conference*, pp. 464-467, Montreux, Switzerland, September 2006.

- **Paper 5: Martin Hansson** and Atila Alvandpour, "A Leakage Compensation Technique for Dynamic Latches and Flip-flops in Nano-scale CMOS," in *Proceedings of IEEE International System-on-Chip Conference*, pp. 83-84, Austin, Texas, USA, September 2006.

- **Paper 6: Martin Hansson** and Atila Alvandpour, "Comparative Analysis of Process Variation Impact on Flip-Flop Power-Performance," in *IEEE International Symposium on Circuits and Systems,* pp. 3744-3747, New Orleans, Louisiana, USA, May 2007.

- **Paper 7:** Behzad Mesgarzadeh, **Martin Hansson**, and Atila Alvandpour, "Jitter Characteristic in Charge Recovery Resonant Clock Distribution," in *IEEE Journal of Solid-State Circuits*, vol. 42, no. 7, pp. 1618-1625, July 2007.

- **Paper 8:** Behzad Mesgarzadeh, **Martin Hansson**, and Atila Alvandpour, "Low-Power Bufferless Resonant Clock Distribution Networks," in *50th IEEE International Midwest Symposium on Circuits and Systems*, pp. 960-963, Montreal, Canada, August 2007 *(This paper was awarded the best student paper award)*.

During the period June to October 2004, I participated in an internship at Intel Circuit Research Laboratory in Hillsboro, Oregon, USA. There I was involved in research on low leakage and process tolerant circuit techniques. This work is presented in the two following papers also included in this thesis:

- **Paper 9: Martin Hansson**, Atila Alvandpour, Steven K. Hsu, and Ram K. Krishnamurthy, "A Process Variation Tolerant Technique for sub-70 nm Latches and Flip-Flops," in *Proceedings of the 23rd IEEE NORCHIP Conference*, pp. 149-152, Oulu, Finland, November 2005.

- **Paper 10:** Steven K. Hsu**, Martin Hansson**, Amit Agarwal, Sanu K. Mathew, Atila Alvandpour and Ram K. Krishnamurthy, "A 9GHz 320x80bit Low Leakage Microcode Read Only Memory in 65nm CMOS," in *Proceedings of the 32nd European Solid-State Circuit Conference*, pp.299-302, Montreux, Switzerland, September 2006.

During the course of my doctoral studies I have also been involved in other research projects that have generated the following papers falling outside the scope of this thesis:

- **Martin Hansson**, and Atila Alvandpour, "Crosstalk Analysis Considering Power and Delay on Interconnects," in *Proceedings of the 21st IEEE NORCHIP Conference*, pp. 196-199, Riga, Latvia, November, 2003.

- Robert Malmqvist, and **Martin Hansson**, "SiGe BiCMOS LNA's and Tunable Active Filter for Future Wide-Band Multi-Purpose Array Antennas," in *Proceeding of the national conference GigaHertz*, Linköping, Sweden, November, 2003.

- Peter Caputa, Henrik Fredriksson, **Martin Hansson**, Stefan Andersson, Atila Alvandpour, and Christer Svensson, "An Extended Transition Energy Cost Model for Buses in Deep Submicron Technologies," in *Proceeding of the 14th International Workshop on Power and Timing Modeling, Optimization and Simulation*, pp. 849-858, Santorini, Greece, September, 2004.

- Robert Malmqvist, **Martin Hansson**, Carl Samuelsson, Mattias Alfredson, "Some Important Aspects on the Design of Active Microwave Filters using Standard RF Silicon Process Technologies," in *Proceeding of the 34th European Microwave Conference*, pp. 941-944, Amsterdam, The Netherlands, October, 2004.

- Nasir Mehmood, **Martin Hansson**, Atila Alvandpour, "An Energy-Efficient 32-bit Multiplier Architecture in 90-nm CMOS," in *Proceedings of the 24th IEEE NORCHIP Conference*, pp. 35-38, Linköping, Sweden, November 2006.

Finally, during the period January to March 2008, I joined the Circuit Research Laboratory at Intel Corporation in Hillsboro, Oregon, USA, for a second graduate technical internship. At Intel I was involved in a research project on energy-efficient, high-bandwidth network-on-chip circuit techniques, which resulted in the following paper, falling outside the scope of this thesis:

- Mark Anders, Himanshu Kaul, **Martin Hansson**, Ram Krishnamurthy, Shekhar Borkar, "A 2.9Tb/s 8W 64-Core Circuit-switched Network-on-Chip in 45nm CMOS," to be presented at *the 34th European Solid-State Circuit Conference*, Edinburgh, UK, September 2008.

Contributions

The main contributions of this dissertation are as follows:

- Development, design, and analysis of an energy recovering resonant clocking technique for multi-GHz synchronous VLSI systems. Successful CMOS implementation proving the proposed resonant clocking technique. Analysis and comparisons of power-performance impact on conventional flip-flops when used in resonant clocked systems.
- An analysis and design of a leakage compensation keeper used for low clock load dynamic latches and flip-flops, including successful CMOS implementation proving the proposed leakage compensating keeper on a reconfigurable flip-flop.
- Analysis and comparisons of process variation impact on conventional flip-flop topologies.
- Studies and analysis on a reconfigurable process variation compensation technique for high-performance static latches and flip-flops.
- Power and performance comparison of a low-leakage technique for high-performance ROM circuits.

Abbreviations

AC	Alternating Current
ALU	Arithmetic-Logic Unit
AND	Logic AND function
ASIC	Application Specific Integrated Circuit
BIST	Built-In Self-Test
CMOS	Complementary Metal-Oxide-Semiconductor
DC	Direct Current
DIBL	Drain-Induced Barrier Lowering
DSP	Digital Signal Processor
FO	Fan-Out
GBL	Global Bitline
IC	Integrated Circuit
IEEE	The Institute of Electrical and Electronics Engineers
ITRS	International Technology Roadmap for Semiconductors
Kb	Kilobit (here 10^3 bits)
LBL	Local Bitline
LC	Inductance-Capacitance
MOS	Metal-Oxide-Semiconductor
MOSFET	Metal-Oxide-Semiconductor Field Effect Transistor
MSFF	Master-Slave Flip-Flop
MUX	Multiplexer
NAND	Logic not-AND function
NMOS	Negative-channel Metal-Oxide-Semiconductor

NOR	Logic not-OR function
OR	Logic OR function
PCB	Printed Circuit Board
PDP	Power-Delay-Product
PMOS	Positive-channel Metal-Oxide-Semiconductor
PRBS	Pseudorandom Sequence
RC	Resistance-Capacitance
RF	Radio Frequency
RLC	Resistance- Inductance-Capacitance
RMS	Root-Mean-Square
ROM	Read-only memory
SAFF	Sense-Amplifier Flip-Flop
SDL	Set-Dominant Latch
SOC	System-on-Chip
SR	Set-Reset
TG-MSFF	Transmission-gate Master-Slave Flip-Flop
VCO	Voltage-Controlled Oscillator
VLSI	Very-Large Scale Integration
XOR	Logic exclusive-OR gate

Acknowledgments

During the course of the last five years as a PhD. student, I have met and worked together with a large number of people. Without the help, support, and encouragement of all these persons it would have been considerably harder to complete this thesis. I would like to express my gratitude and thank the following people for what they have done for me and this thesis:

- First and foremost I would like to express my deepest gratitude to my supervisor, advisor, and guide into the world of integrated circuit research, Professor Atila Alvandpour. Without your guidance, patience, and support this thesis would not exist. Thanks for giving me the opportunity to pursue a career as Ph.D. student and for letting me get a chance to try the American lifestyle not only once but twice!

- I would also like to thank Professor Christer Svensson for all interesting discussions and valuable comments throughout the course of my doctoral studies.

- I thank Tek. Lic. Behzad Mesgarzadeh for the excellent collaboration during the course of our joint resonant clocking project. I have greatly appreciated your comments, help, cooperation, and patience during all the long hours of layout and tape-out work, chip measurements, and paper writing close to deadlines. Thanks!

- Dr. Stefan Andersson deserves many thanks for starting this adventure by letting me know about the free Ph.D. position some five years ago. You have also been a big help during a large part of my Ph.D. studies and a great company during the Intel summer of 2004!

- Dr. Henrik Fredriksson deserves thanks for all help and discussions about all kinds of stuff. Your expertise in various program language related problems and invaluable tips in numerous chip design issues have been greatly appreciated.

- I want to thank M.Sc. Timmy Sundström for excellent collaboration during chip tape-outs, student labs, graduate courses, and for being an outstanding friend. You also deserve extra credit for proof reading this thesis.

- M.Sc. Jonas Fritzin deserves a great deal of thanks for being a great friend and colleague. But also for being a reliable company this spring during all long hours at the office, any day of the week. You work too much! Also, many thanks for finding time to proof read this thesis.

- I would like to thank Dr. Peter Caputa for invaluable assistance before my first US trip and for all interesting discussions during my first three years at Electronic Devices.

- I thank our secretary Anna Folkeson for keeping track of the group and your invaluable support in everything from travel arrangements, course registration, and all other administrative tasks.

- Research engineer Arta Alvandpour deserves thanks for taking care of all PCB designs and layouts, solving all computer and tool related problems, and for instructing me in the basics of the Persian language.

- All the past and present members of the Division of Electronic Devices, especially M.Sc. Rashad Ramzan, M.Sc. Naveed Ahsan, M.Sc. Shakeel Ahmad, Ass. Prof. Jerzy Dabrowski, Dr. Kalle Folkesson, Dr. Håkan Bengtsson, Dr. Darius Jakonis, Adj. Prof. Aziz Ouacha, M.Sc. Joacim Olsson, Dr. Ingemar Söderquist, and also Professor Dake Liu, Dr. Anders Nilsson, and Dr. Daniel Wiklund, who are present and former members of the Division of Computer Engineering. Thanks for creating such a great research environment.

- A great deal of thanks goes to all the people at the Circuit Research Lab, Intel in Hillsboro, Oregon, USA, especially M.Sc. Mark Anders, Dr. Himanshu Kaul, Dr. Steven Hsu, Dr. Amit Agarwal, Dr. Sanu Matthew, Dr. Ram Krishnamurthy, M.Sc. Matthew Haycock, and M.Sc. Shekhar Borkar. Thanks for making both of my internships with you such great experiences!

- I would also like to thank Dr. Sriram Vangal for being a really good friend, for your hospitality, and for always offering me help with all things and during my last stay in Oregon.

- Thanks to all friends and family who have encourage me during the years, but who I could not fit in here.

- Last by certainly not least, I would like to thank my parents Ellinor and Anders Hansson for all encouragement and love, and my brother Andreas Hansson for all discussions about other things not related to science and technology. Your support has been greatly appreciated, especially when being on the other side of the planet. *Tack för allt!*

Martin Hansson
Linköping, Sweden
August, 2008

Contents

Abstract iii

Preface vii

Contributions xi

Abbreviations xiii

Acknowledgments xv

Part I Background 1

Chapter 1 Introduction 3

1.1 Historical Perspective 3
1.2 Future Trends and Challenges 5
1.3 Dissertation Motivation and Scope 6
1.3.1 Low-Power Clocking 6
1.3.2 Leakage Tolerant Design 6
1.3.3 Process Variation Aware Design 7
1.4 Dissertation Overview 7
1.5 Bibliography 9

Chapter 2 Background to CMOS Technology 11

2.1 Introduction 11
2.2 The MOS Device 11
2.2.1 Threshold Voltage 12

2.2.2 Static Current-Voltage Characteristics 13
2.2.3 Subthreshold Conduction 15
2.2.4 Scaling and Small Geometry Effects 16
2.3 Power Dissipation in CMOS 19
2.3.1 Switching Power 19
2.3.2 Short-Circuit Power 20
2.3.3 Leakage Power 20
2.4 Basics of Integrated Circuit Manufacturing 21
2.4.1 Lithography 22
2.4.2 Etching 22
2.4.3 Implantation, Oxidation, and Deposition 23
2.5 Process Variation 24
2.5.1 Geometry Variations 25
2.5.2 Material Variations 26
2.5.3 Modeling of Process Variation 27
2.6 Bibliography 29

Chapter 3 Clocking and Synchronization **35**

3.1 Introduction 35
3.2 Synchronization Circuits 36
3.2.1 Level-Sensitive Latches 36
3.2.2 Edge-Triggered Flip-flops 37
3.3 Characterizing Synchronization Circuits 39
3.3.1 Characterizing Timing for Latches and Flip-Flops 39
3.3.2 Power-Delay Design Space 40
3.3.3 A Flip-Flop Optimization Approach 41
3.4 Clock Signal Integrity 42
3.4.1 Clock Jitter 42
3.4.2 Clock Skew 43
3.5 Synchronization Approaches 44
3.5.1 Edge-Triggered Clocking 44
3.5.2 Level-Sensitive Clocking 45
3.6 Common Flip-Flop Topologies 46
3.6.1 Master-Slave Latch Pairs 46
3.6.2 Pulsed Latches 47
3.6.3 Sense-Amplifier Based Flip-flops 48
3.7 Conventional Clock Distribution Techniques 49
3.7.1 Tapered Clock Buffer Chain 49
3.7.2 Clock Trees 50
3.7.3 Grid Clock Distribution 50
3.7.4 Length-Matched Serpentines 50
3.8 Bibliography 52

Part II Low-Power Clocking 55

Chapter 4 Background 57

4.1 Introduction 57
4.2 Power Analysis of Conventional Buffered Clocking 58
4.3 Conventional Low-Power Clocking Techniques 59
4.3.1 Frequency and Voltage Reductions 60
4.3.2 Low-Swing Clocking 60
4.3.3 Clock Gating 61
4.3.4 Clock Load Reduction 61
4.3.5 Summary 61
4.4 Energy Recovery Clocking Techniques 62
4.4.1 Adiabatic Switching 62
4.4.2 Oscillator Driven Global Clock Networks 63
4.4.3 Bufferless LC-tank Resonant Clocking 63
4.5 Power Analysis of LC-tank Resonant Clocking 64
4.6 Issues Concerning Tank Q-value 66
4.7 Bibliography 67

Chapter 5 Resonant Clocking - Impact on Flip-Flops 71

5.1 Introduction 71
5.2 Analyzed Flip-Flop Topologies 73
5.3 Comparison and Discussion 75
5.3.1 Simulation Setup 75
5.3.2 Power-Delay Comparison 75
5.4 Bibliography 80

Chapter 6 Chip Implementations and Evaluation 83

6.1 Introduction 83
6.2 Resonant Clocking Evaluation Test Chip 83
6.2.1 Top Level Chip Organization 84
6.2.2 Conventional Clock Drivers 85
6.2.3 Implemented Oscillator Topology 86
6.2.4 Clock Distribution Network 86
6.2.5 Inductor Implementation 87
6.2.6 Implemented Flip-Flops 88
6.2.7 Organization of the Data-Path Blocks 90
6.3 Power Measurement Results 91
6.3.1 On-Chip Resonant Core Power Comparison 91
6.3.2 Influence of Inductor Q-value 93

6.4 Clock Signal Integrity ... 95
6.4.1 Oscillator Power Supply Sensitivity ... 95
6.4.2 Data Dependent Phase Noise ... 95
6.4.3 Implemented Jitter Suppression Technique ... 96
6.5 Frequency Tunability ... 98
6.5.1 Tunability Using Injection Locking ... 98
6.5.2 Capacitive Tuning on a Oscillator Test Chip ... 98
6.5.3 Switchable Inductance ... 99
6.5.4 Chip Measurement Results ... 101
6.6 Bibliography ... 102

Chapter 7 Conclusions and Future Work **103**

7.1 Conclusions ... 103
7.1.1 Low-Power Resonant Clocking ... 103
7.1.2 Flip-Flop Behavior in Resonant Clocking Systems ... 104
7.2 Future Work ... 105
7.2.1 Low-Power Resonant Clocking ... 105
7.2.2 Flip-Flops for Resonant Clocking Systems ... 106
7.3 Bibliography ... 106

Part III Leakage Tolerant Circuit Design **107**

Chapter 8 Background **109**

8.1 Introduction ... 109
8.2 Leakage Reduction Techniques ... 110
8.2.1 Power Gating and Multiple-V_{th} Techniques ... 110
8.2.2 Selective Long-Channel Insertion ... 111
8.2.3 Threshold Voltage Modulation ... 111
8.3 Dynamic Circuits ... 112
8.3.1 Low-Power Dynamic Flip-Flop ... 112
8.3.2 Leakage Robustness Issues ... 113
8.3.3 Static Weak-Keeper Flip-Flops ... 114
8.4 Bibliography ... 115

Chapter 9 Leakage Compensation Keeper **119**

9.1 Introduction ... 119
9.2 Reconfigurable Leakage Compensation Keeper ... 120
9.2.1 Principle of Operation ... 120
9.2.2 Reconfigurable Dynamic Flip-Flop ... 123
9.3 Simulation Results ... 124
9.3.1 Robustness using Leakage Compensation ... 124

9.3.2 Performance Impact of Leakage Compensation Keeper ... 127
9.3.3 Leakage Compensation Keeper for Low Clock Power ... 128
9.4 Experimental Chip Results ... 130
9.4.1 Chip Implementation ... 130
9.4.2 Measurement Results ... 131
9.5 Bibliography ... 133

Chapter 10 Low-Leakage Microcode ROM **135**

10.1 Introduction ... 135
10.2 ROM Organization ... 137
10.3 Microcode Heuristics ... 138
10.4 Programmable Logic Technique ... 140
10.4.1 Removal of Unused Devices ... 140
10.4.2 Optimization of Driver Strength ... 141
10.5 Comparison Results and Discussion ... 143
10.6 Bibliography ... 144

Chapter 11 Conclusions and Future Work **147**

11.1 Conclusions ... 147
11.1.1 Leakage Compensation Keeper ... 147
11.1.2 Low-Leakage High-Speed ROM ... 148
11.2 Future Work ... 148
11.3 Bibliography ... 148

Part IV Process Variation Aware Design **151**

Chapter 12 Background **153**

12.1 Introduction ... 153
12.2 Impact of Process Variation ... 153
12.3 Process Variation Compensation Techniques ... 155
12.3.1 Power Supply and Body Bias Adjustments ... 155
12.3.2 Reconfigurable Designs ... 156
12.3.3 Device Sizing ... 156
12.4 Bibliography ... 157

Chapter 13 Impact of Process Variation on Flip-Flops **159**

13.1 Introduction ... 159
13.2 Flip-Flop Topologies and Optimization ... 161
13.2.1 Flip-Flop Topologies ... 161
13.2.2 Optimization Approach ... 161

13.3 Process Variation Impact on Flip-Flop Timing ... 162
13.3.1 Setup Time Margin ... 162
13.3.2 Statistical Simulation Approach ... 163
13.4 Proces Variation Simulation Results ... 165
13.5 Summary and Discussion ... 168
13.6 Bibliography ... 170

Chapter 14 Process Variation Compensation Keeper 173

14.1 Introduction ... 173
14.2 Reconfigurable Keeper for Latches and Flip-Flops ... 174
14.2.1 Circuit Concept ... 174
14.2.2 Reconfigurable Keeper for Static MUX-Latches ... 175
14.2.3 Reconfigurable Keeper for Static MSFFs ... 177
14.3 Simulation Results ... 178
14.3.1 Reconfigurable Static 5-to-1 MUX-Latch ... 179
14.3.2 Reconfigurable Uninterrupted Keeper for Static Flip-Flops ... 181
14.4 Bibliography ... 182

Chapter 15 Conclusions and Future Work 185

15.1 Conclusions ... 185
15.1.1 Process Variation Impact on Flip-Flop Power-Performance ... 185
15.1.2 Reconfigurable Process Variation Tolerant Keeper ... 185
15.2 Future Work ... 186
15.3 Bibliography ... 186

Part I

Background

Chapter 1

Introduction

1.1 Historical Perspective

During the last five decades the electronics industry has evolved tremendously, and the last ten years of aggressive scaling have moved integrated circuits from the micrometer regime down to the nanoscale regime [1], [2]. In the late 1950s, putting more than one transistor on a piece of semiconductor device was considered cutting edge. The concept of integrated circuits with even as little as tens of devices was unheard of. To obtain a 50% probability of functionality for a 20-transistor circuit, the probability of individual device functionality had to be $(0.5)^{1/20}$ = 96.6%, which was considered optimistic well beyond anything imaginable [3]. Nevertheless, ongoing innovations in technology and integration have continued to overcome the predicted limits [2], [4], and today transistors are manufactured with gate lengths well below 100 nm, and integrated circuits contains over a billion transistors per chip [5].

In 1965 Gordon Moore published his famous paper [6], in which he predicted that the number of components per integrated circuit for minimum cost would increase by two every year. This prediction was updated ten years later, predicting that the number of devices should double every second year from then on, which is popularly refereed to as “Moore’s Law” [7]. These predictions have since then inspired the microelectronic industry to strive for increased complexity and lower fabrication costs of integrated circuits. Up until now

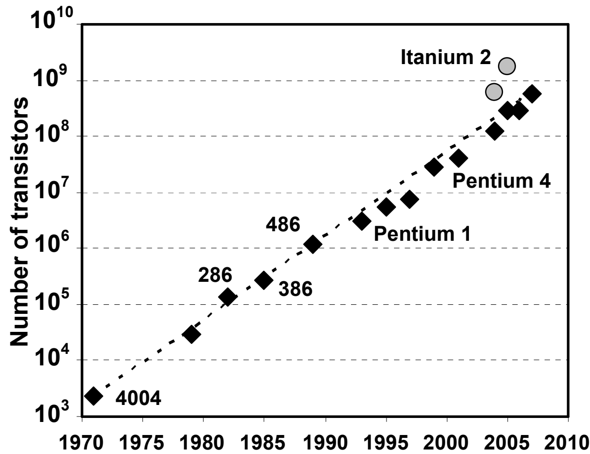
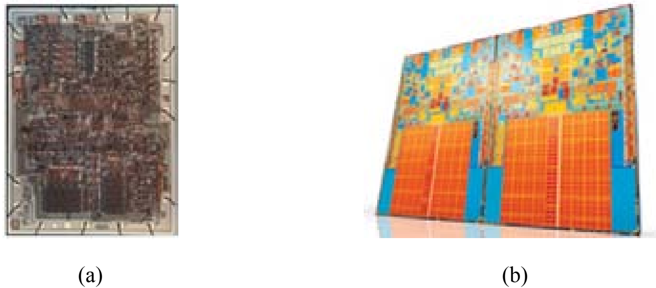


Figure 1.1: 40 years of evolution in Intel® microprocessors [5].



(a) (b)

Figure 1.2: (a) Intel® 4004 in 10 µm (1971), (b) Intel® Core 2 Quad™ in 45 nm (2008) (reprinted with permission from Intel) [5].

Moore's predictions have been quite accurate, as a result of vast improvements in circuit capabilities, enabled by dimensional scaling. This can be illustrated in the form of the microprocessor evolution in the last four decades seen in Figure 1.1. From the first Intel® 4004 microprocessor (Figure 1.2(a)) with 2300 transistors clocked at a frequency of 108 kHz to the present Core™ 2 Quad (Figure 1.2(b)) with 820 million transistors and clocked at frequencies above 3 GHz, the number of transistors has roughly doubled every two years [5].

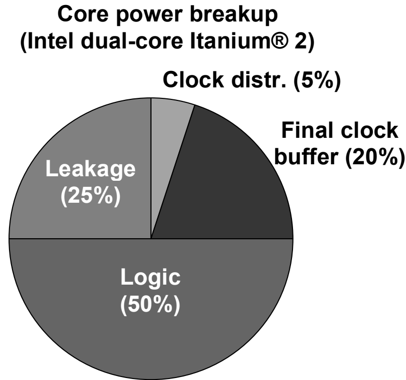


Figure 1.3: Power breakout of a high-performance microprocessor [8].

1.2 Future Trends and Challenges

Without the incredible progress in silicon technology and device integration, many high-technology achievements such as Internet, portable computers, and mobile phones, would never been able to be realized [2]. As the silicon integrated circuit industry moves further into the nanometer regime, scaling of device sizes is still predicted to continue into at least the near future [9], with gate lengths approaching and passing 10 nm within the next ten years [10]. Certainly, as the race for more mobility and accessibility of electronic devices increases, the evolution of integrated circuits will continue to increase its importance in the high-technology society [2], [11].

However, continuing scaling is facing a number of challenging problems, which need to be treated. These challenges are both related to the difficulties in process and manufacturing, which due to fundamental physical limits results in growing costs to continue the integration [1], [2], [9], as well as in circuits and architectures. The rapid increase in the number of transistors on each chip has enabled a dramatic increase in the performance of computing systems. Consequently, the extreme speeds and amounts of transistors integrated in high-performance VLSI systems have led to escalating power dissipation [12] - [16]. The increasing power dissipation has largely been caused by active power, especially in the clock network [8]. Moreover, in the last years more and more of the power dissipation has been due to static leakage currents in the transistors,

which are caused by the continuing scaling of feature sizes and voltages [8], [17]. This is illustrated in Figure 1.3, which shows that only half of the power dissipated in a high-performance microprocessor is from actual computations, while the other half is due to either clock power or leakage power [8]. Furthermore, the diminishing sizes make it considerably more challenging to manufacture integrated circuits with good accuracy. This has lead to increasing statistical variations around the expected circuit performance, which is projected to become even worse in future CMOS technologies [18] - [21].

1.3 Dissertation Motivation and Scope

This thesis covers some of the main challenges in future CMOS technologies. The main focus of the thesis is on techniques for reducing clock power, which will be a common theme throughout the thesis. However, the research presented in this thesis can be largely divided into the three following topics.

1.3.1 Low-Power Clocking

Power dissipation related to the clock generation and distribution is identified as the dominating contributor of the total active power dissipation for digital multi-GHz systems [8]. With the increasing complexity and the growing number of devices in synchronous systems, clock power will continue to increase, and threatens to become a limiter for the continuing integration of more functionality [9], [16]. This makes novel power reduction techniques crucial in future VLSI design.

In this thesis, an energy recovering clocking technique, aimed at reducing the total chip clock power, is presented. The technique enables considerable savings of the clock power dissipation (over 56%) compared to conventional clock distribution techniques at clock frequencies up to 1.76 GHz.

1.3.2 Leakage Tolerant Design

Leakage power contributes to a considerable part of the total power dissipation, and has become one of the primary design constraints in VLSI systems [8], [17]. This limits the amount of integration and thereby the functionality in all from battery power mobile processors to high-performance server processors where the cooling cost is limiting the power envelop requires power constrained design. Therefore, circuit techniques that reduce the leakage are needed. Apart from increasing power dissipation, the increasing leakage currents also impact the robustness constraints for the circuits [15]. This is an issue especially for

low-power, high-performance dynamic circuits, which require higher and higher refresh frequencies in order to maintain the stored charges on the floating nodes.

In this thesis a technique is discussed, which reduces the static leakage current part in a microcode ROM, resulting in 30% reduction of the leakage power. In order to improve the leakage robustness for sub-90 nm low clock load dynamic flip-flops, a novel keeper technique is proposed. The proposed keeper is implemented on a dynamic reconfigurable flip-flop, which utilizes a scalable and simple leakage compensation technique. During any low-frequency operation, the flip-flop is configured as a static flip-flop for increased functional robustness.

1.3.3 Process Variation Aware Design

As scaling continues further towards the fundamental atomistic limits, several challenges arise for continuing industrial device integration. Large inaccuracies in lithography process, impurities in manufacturing, and reduced control of dopant levels during implantation, all cause increasing statistical spread of performance and power in the devices [18] - [21].

In this thesis an analysis of the process variation impact on a number of conventional flip-flops are presented. The statistical spread in performance and power dissipation is discussed. In order to compensate for the impact of the increasingly large process variations on latches and flip-flops, a reconfigurable keeper technique is presented. In contrast to traditional worst-case design, a variable keeper circuit is utilized, which preserves the robustness of the storage nodes across process corners, without degrading the overall chip performance.

1.4 Dissertation Overview

This thesis is divided into four main parts, which treats the three above mentioned topics. **Part I** begins with **Chapter 1,** which provides a brief discussion on the history of CMOS technology scaling and a future outlook. Also, the motivation of the work presented in this thesis is given together with this outline. **Chapter 2** aims to give an introduction into the world of integrated circuits and particularly CMOS VLSI technology and the issues and challenges of today, such as power dissipation, leakage, and process variation. This is followed by an introduction to synchronous digital circuit design given in **Chapter 3**. Design and characteristics of timing circuits and clocking are discussed in order to provide the background to the discussions in the three following parts of the thesis.

In **Part II** a low-power resonant clocking technique is presented and discussed through theoretical reasoning but also by experimental chip measurement results. The discussion, analysis, measurements, and results in this part are largely based on the previously presented publications in **Paper 2**, **Paper 3**, **Paper 4**, **Paper 7**, and **Paper 8**. Part gives an introduction to conventional low-power techniques for clock networks, given in **Chapter 4**. Furthermore, the power dissipation in a general clock network is analyzed, and the concept and the theoretical aspects of energy recovering clocking are discussed. **Chapter 5** presents an analysis and comparison of some common flip-flop topologies under the impact of sinusoidal clock from the resonant clock driver. This is then followed in **Chapter 6** by a thorough presentation and discussion of the implemented test chips on which the proposed resonant clocking technique is implemented. **Chapter 7** concludes the second part of the thesis and provides a short discussion on future research issues related to low-power clocking.

Part III treats the research on low-leakage and leakage compensations techniques. The discussion and results presented in this part are based on the publications in **Paper 1**, **Paper 5**, and **Paper 9**. A background into common leakage reduction techniques and a brief introduction to the issues with robustness in dynamic circuits is given in **Chapter 8**. A proposed low clock load dynamic flip-flop, incorporating a novel leakage compensating keeper, is presented in **Chapter 9**. Both simulation results showing the concept and chip measurement results proving the functionality are provided. **Chapter 10** presents the design and implementation of a layout programmable leakage reduction technique for high-performance ROM circuits. Conclusions on the third part of the thesis are given in **Chapter 11** followed by a short discussion on future research issues related to leakage problems.

In **Part IV** the impact of process variation on timing circuits is analyzed and a compensation technique to increase process variation tolerance is discussed. These discussion and results are based on the results presented in **Paper 6**, and **Paper 10**. In **Chapter 12** a short introduction on process variation impact on digital circuits is given together with a brief summary of some common process variation compensation techniques. This is followed by a more focused study in **Chapter 13** on the impact of statistical process parameter variation on some common flip-flop circuits. **Chapter 14** then presents an implementation of a reconfigurable and scalable keeper circuit solution aimed to reduce the process variation induced spread in robustness, performance, and power for static flip-flops and latches. Finally part four is concluded in **Chapter 15** and a short discussion on future research issues related to process variation problems are given.

1.5 Bibliography

[1]. G. Moore, "No Exponential is Forever: But "Forever" Can Be Delayed!" in *Digest of Technical Papers IEEE Solid-State Circuits Conference*, pp. 20-23, 2003.

[2]. S. Chou, "Integration and Innovation in the Nanoelectronics Era," in *Digest of Technical Papers IEEE Solid-State Circuits Conference*, pp. 36-41, 2005.

[3]. S.A. Campell, *The Science and Engineering of Microelectronic Fabrication*, Oxford University Press, 1996, ISBN: 0-19-510508-7.

[4]. C. Svensson, "Forty Years of Feature-Size Predictions (1962-2002)," in *Digest of Technical Papers IEEE Solid-State Circuits Conference*, pp. S28-S29, 2003.

[5]. http://www.intel.com, accessed: June 2008.

[6]. G.E. Moore, "Cramming more components onto integrated circuits," in *Electronics*, vol. 38, no. 8, 1965.

[7]. G.E. Moore, "Progress on Digital Integrated Electronics," in *Technical Digest of International Electron Device Meeting*, pp. 11-13, 1975.

[8]. S. Naffziger, B. Stackhouse, T. Grutkowski, "The Implementation of a 2-core Multi-Threaded Itanium®-Family Processor," in *Digest of Technical Papers IEEE International Solid-State Circuit Conference,* pp. 182-183, 2005.

[9]. T.-C. Chen, "Where CMOS is going: trendy hype vs. real technology," in *Digest of Technical Papers IEEE Solid-State Circuits Conference*, pp. 1-18, 2006.

[10]. http://www.itrs.net, June 2008.

[11]. M. Muller, "Embedded Processing at the Heart of Life and Style," in *Digest of Technical Papers IEEE Solid-State Circuits Conference*, pp. 32-37, 2008.

[12]. S. Borkar, "Design challenges of technology scaling," in *IEEE Micro*, Volume 19, Issue 4, pp. 23-29, 1999

[13]. V. De and S. Borkar, "Technology and Design Challenges for Low Power and High-Performance," in *Proceedings of International Symposium on Low Power Electronics and Design*, pp. 163-168, 1999.

[14]. S. Rusu, "Trends and challenges in VLSI technology scaling towards 100nm," in *Proceedings of the 27th European Solid-State Circuits Conference*, pp. 194-196, 2001.

[15]. R. Krishnamurthy, A. Alvandpour, V. De, and S. Borkar, "High performance and low-power challenges for sub-70-nm microprocessor circuits," in *Proceedings of the Custom Integrated Circuits Conference*, pp. 125-128, 2002.

[16]. P.P. Gelsinger, "Microprocessors for the New Millennium: Challenges, Opportunities, and New Frontiers," in *Digest of Technical Papers IEEE Solid-State Circuits Conference*, pp. 22-25, 2001.

[17]. K. Roy, S. Mukhopadhyay, and H. Mahmoodi-Meimand, "Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicrometer CMOS Circuits," in *Proceeding of the IEEE*, vol. 91, no. 2, pp. 305-327, 2003.

[18]. S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, V. De, "Parameter Variations and Impact on Circuits and Microarchitecture", in *Proceedings of Design Automation Conference*, pp. 338-342, 2003.

[19]. M.T. Bohr, "Nanotechnology Goals and Challenges for Electronic Applications," in *IEEE Transactions on Nanotechnology*, vol. 1, no. 1, pp. 56-62, 2002.

[20]. K.J. Kuhn, "Reducing Variation in Advanced Logic Technologies: Approaches to Process and Design for Manufacturability of Nanoscale CMOS," in *IEEE International Electron Device Meeting*, pp. 471-474, 2007.

[21]. S. Nassif, K. Bernstein, D.J. Frank, A. Gattiker, W. Haensch, B.L. Ji, E. Nowak, D. Pearson, N.J. Rohrer, "High Performance CMOS Variability in the 65nm Regime and Beyond," in *IEEE International Electron Device Meeting*, pp. 569-571, 2007.

Chapter 2

Background to CMOS Technology

2.1 Introduction

The extraordinary evolution in microelectronics would not have been as impressive were it not for the invention of MOS devices, and lately CMOS circuits. CMOS devices have grown to become without comparison the most commonly used devices in VLSI circuits and still remain the workhorse of the entire digital electronics industry [1]. This thesis seeks to introduce and present design and circuit techniques aimed to combat escalating problems in VLSI systems subjected to extreme scaling in future nanoscale CMOS technologies. In order to do this, some insight into the world of CMOS technology is required. This chapter aims to provide that insight.

2.2 The MOS Device

A MOSFET is a voltage controlled device with the four terminals, drain, source, gate, and bulk. There are two types of MOSFET devices used in CMOS circuits, negative-channel MOS (NMOS) and positive-channel MOS (PMOS), which have complementary properties [1]. Figure 2.1 shows a schematic and a cross section view of a NMOS transistor and in the discussion that follows the physical and electrical properties of the NMOS device will be treated. However, the properties of the PMOS can be treated in a similar fashion.

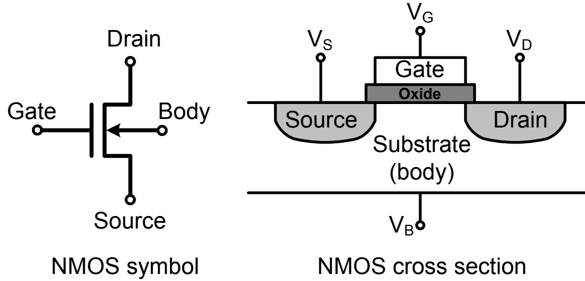


Figure 2.1: Schematic and cross section views of an NMOS transistor.

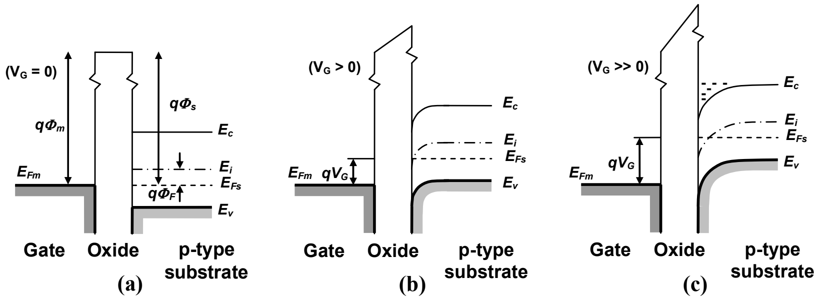


Figure 2.2: Energy band diagram of the channel region for an ideal NMOS device ($\Phi_m = \Phi_s$) at (a) equilibrium, (b) depletion, and (c) inversion.

2.2.1 Threshold Voltage

The gate terminal of the NMOS device in Figure 2.1 is separated from the positively doped (P-doped) substrate or body region by a thin insulating layer called gate oxide. The body represents the doped silicon substrate in which the transistors are manufactured. The region between the source and drain is usually called the channel region although there is only a physical channel under certain bias conditions.

The physical properties of the interface between the gate and the substrate at any point along the channel region can be described using an energy band diagram. Figure 2.2 shows the energy band of the channel region of an ideal

MOS device[1] biased so that source, drain, and body terminals are connected to ground [2]. If there is no potential voltage difference between the gate terminal and the substrate, the Fermi level for the gate (E_{FM}) and for the P-type substrate (E_{Fs}) will align as shown in Figure 2.2(a), and the channel region is at equilibrium. When a positive voltage ($V_G > 0$), relative to the substrate, is asserted to the gate terminal an electric field is created between the gate and the substrate. This will attract electrons to gather in the P-doped region between the negatively doped (N-doped) source and drain, which shifts the Fermi level (E_{Fs}) towards the conduction band (E_C). This is shown in Figure 2.2(b) as a bending of the energy bands and results in that the substrate region closest to the oxide becomes depleted. When the applied voltage difference becomes even larger the concentration of electrons will increase and the Fermi level shift even closer towards the conduction band, effectively making the channel N-type instead. At a certain voltage difference between the gate and the substrate the channel region has become as strongly N-type as the substrate is P-type and this is defined as strong inversion[2] [2], as shown in Figure 2.2(c). The voltage required to make the channel strongly inverted is defined as the threshold voltage of the device [2] denoted V_{th0} when the bulk voltage is asserted to ground potential. Once the gate-channel voltage have reached the threshold voltage an N-type channel is formed between the drain and source terminals. The voltage asserted on the body terminal (V_B in Figure 2.1), will modulate the threshold voltage according to

$$V_{th} = V_{th0} + \gamma\left(\sqrt{|-2\Phi_F + V_{SB}|} - \sqrt{|2\Phi_F|}\right), \tag{2.1}$$

where γ is defined as the body-bias coefficient, which depends on physical parameters of the device, and Φ_F is the Fermi potential of the substrate [1], [2].

2.2.2 Static Current-Voltage Characteristics

If the gate terminal of the NMOS transistors is asserted a voltage so that the gate-source voltage is larger than the threshold voltage ($V_{GS} > V_{th}$), then a small voltage difference between the drain and source (V_{DS}) results in a current flow in the channel as shown in Figure 2.3. The property of being able to control the conductivity between the two terminals with the gate voltage is what makes the MOS transistors attractive as a switch in digital circuits [1], [2]. The voltage at any point along the channel is denoted $V(x)$ and if the gate-to-channel voltage is larger than the threshold voltage at every point between the drain and source

[1] Metal gate and identical work functions $\Phi_m = \Phi_s$.

[2] $qV > q|2\Phi_F|$

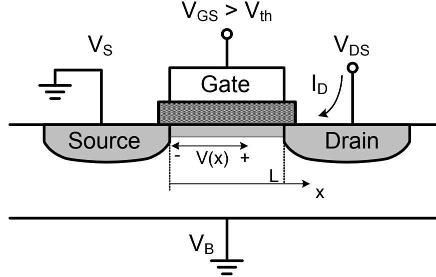


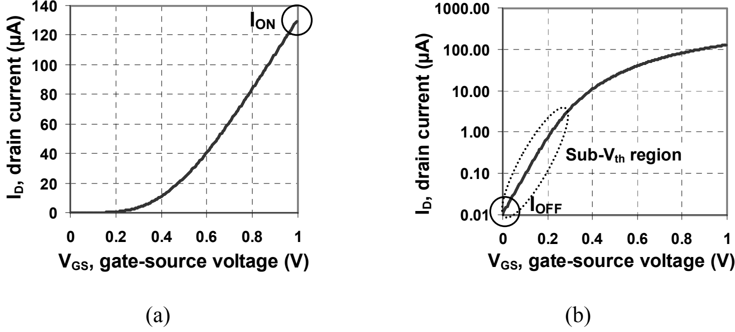
Figure 2.3: Cross section view of NMOS with channel formed between source and drain.

(V_{GS}-$V(x)$ > V_{th}) then strong inversion is achieved throughout the channel and the transistor is defined to be in the linear region [1], [2]. The drain current is then approximately linearly proportional to the drain-source potential. As the drain-source voltage is increased the strong-inversion condition ceases to exist at some point in the channel, which occur when V_{GS}-$V(x)$ < V_{th}. At this point the conduction channel is pinched-off, and the voltage difference in the remaining channel is fixed at V_{GS}-V_{th}, which makes the drain current constant regardless of V_{DS} to a first order [1], [2]. However, in reality V_{DS} still modulates the efficient channel length, hence still possesses some modulating ability on the saturation current [1], [2]. Furthermore, with the small transistor sizes in modern CMOS transistors, the electric field strength, between the source and drain terminals, is high during normal voltage operation. The speed, with which the charge carriers can propagate inside the channel, is for weak electric fields linearly depending on the field strength. But, when the field strength increases above a certain value the velocity of the charge carriers will saturate due to scattering effects in the channel [1], [2]. The transistor is then said to be in velocity saturation mode. The drain current in velocity saturation can be expressed as

$$I_{D,vsat} = \mu_n C_{ox} \frac{W}{L}\left((V_{GS} - V_{th})V_{D,vsat} - \frac{V_{D,vsat}^2}{2} \right)(1 + \lambda V_{DS}) \ , \quad (2.2)$$

where W and L are the width and length of the channel, μ_n is the carrier mobility, C_{ox} is the gate-oxide capacitance, and $V_{D,vsat}$ is the drain-source velocity saturation voltage [1].

In a digital gate the drain-source voltage of the transistors are usually either zero or equal to the power supply voltage. This means that during normal



(a) (b)

Figure 2.4: Drain current characteristic for gate-source voltage ($V_{DS} > V_{DS,vsat}$) (a) linear scale, (b) semi-log scale.

operation the current through the transistor will vary between two distinct values usually refereed to as I_{ON} and I_{OFF}, as the gate voltage is changed between zero and the power supply. Here I_{ON} is defined as the maximum drain current of the device when V_{GS} and V_{DS} are both equal to the power supply voltage, which is modeled by the expression in equation (2.2) if the transistor is in velocity saturation. Figure 2.4(a) shows the drain current and I_{ON} of an NMOS transistor in a modern CMOS technology.

2.2.3 Subthreshold Conduction

Figure 2.4(b) shows the drain current for an NMOS transistor in a semi-log scale. Noticeable is that the drain current will not go down to zero directly below the threshold voltage, but will instead follow an exponential relationship in the subthreshold region. This region of the transistor curve is referred to as weak-inversion conduction or subthreshold conduction. The current transport between the drain and source terminals is due to diffusion of carriers along the channel surface, which yields an exponential relation to the gate voltage [1] - [3]. The weak inversion conduction of can be modeled as

$$I_{D,subth} = \mu_0 C'_{ox} \frac{W}{L_{eff}} v_T^2 e^{1.8} \cdot e^{1/mv_T (V_{GS} - V_{th0} - \gamma V_S + \eta V_D)} \left(1 - e^{-V_{DS}/v_T}\right), \tag{2.3}$$

where V_{th0} is the zero bias threshold voltage, v_T is the thermal voltage, γ is the body effect coefficient, and m is the subthreshold swing coefficient [3].

Another contributing factor to the off-state current in the subthreshold mode is the reverse-biased diodes that are formed between the drain/source areas and the substrate. Minority carrier diffusion and drift near the depletion region edge together with electron-hole-pair generation in the depletion region of the reversed PN-junction cause a current to flow from drain/source to the substrate. The resulting leakage current from both of these effects depends on the junction area and the doping concentration of the diffusion regions [3], [4]. An additional junction leakage effect called band-to-band tunneling can occur if both the N-region and P-region in the MOS device are heavily doped. If the reverse-biased junction is asserted a high electric field electrons are able to tunnel from the valence band in the P-region to the conduction band in the N-region. In order for this tunneling to take place the voltage drop over the junction needs to be larger then the bandgap [4]. The weak conduction region and the junction leakage currents cause the transistor characteristic to deviate from the switch-like behavior that is desired in digital circuits, by causing the off-state current to be higher than zero (I_{OFF}).

2.2.4 Scaling and Small Geometry Effects

The progress in the electronics industry the last four decades have largely been contributed by the scaling of the CMOS technologies. In the last 40 years the channel length of state-of-the-art CMOS technologies have scaled to roughly half every fourth year with a new technology node released once every second year, as shown for the MOS transistor gate length in Figure 2.5 [5]. The principle of scaling was introduced by Dennard et al. in 1974 [6], which quickly became the guide for the industry on the road to continue Moore's law [7].

The scaling principle advocated by Dennard et al. [6] are known as constant

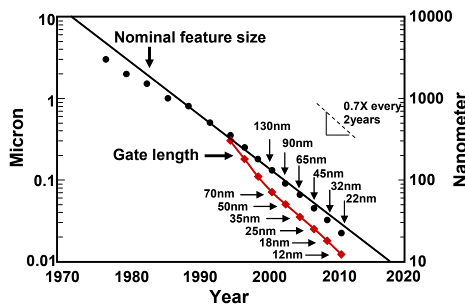
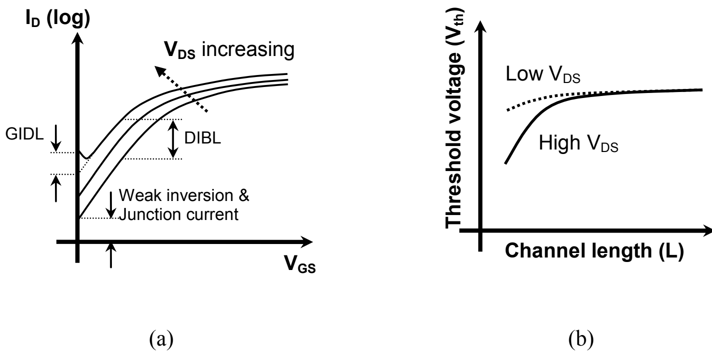


Figure 2.5: Scaling of transistor gate length over the last four decades [5].

field scaling, where the magnitude of the internal electric fields in the MOSFET devises are preserved by scaling all physical dimensions and all voltages by the same factor S to achieve a constant field. As the device dimensions and voltages are scaled, the gate oxide capacitance will reduce with the scaling factor together with the saturation drain currents. This result in that the delays of CMOS circuits will be reduced by S and the power dissipation will reduce by S^2 [1]. However, intrinsic device voltages such as bandgaps and built-in junction potentials cannot be scaled due to physical limitations. Furthermore, threshold voltages can not be scaled down arbitrarily because of increasing subthreshold conduction. Therefore, for the last 10 years power supply and threshold voltages have not scaled as fast as the process, as suggested by the scaling theory [1].

Moreover, for the small feature sizes of today's modern CMOS technologies there exists a number of physical phenomenas that impact the characteristics of the transistors. One of these effects is drain induced barrier lowering (DIBL), which is due to electrostatic interaction between the source and drain for short channel devices. The effect is due to that the drain depletion region is extended deeper into the substrate as the drain voltage increases, and for a short channel device this extension of the depletion region also influences the source depletion region. This electrostatic interaction between the depletion region cause the potential barrier between the source and drain, to reduce as the drain voltage is increased, which results in a reduction of the transistors threshold voltage [3], [4]. In equation (2.3) the DIBL effect is modeled with the parameter η [3]. Figure 2.6(a) shows the drain current versus the gate-source voltage at three different drain-source bias voltages, where the increased drain-source voltage



(a) (b)

Figure 2.6: (a) I_D versus V_{GS} at different V_{DS} showing different impact on I_{OFF}. (b) Threshold voltage roll-off with changing channel length.

results in a vertical shift of the drain current causing an increase in the weak-inversion current and I_{OFF}, and a reduction of V_{th}.

Another small geometry effect occurs when the potential difference between the gate and drain terminals of an NMOS transistor becomes high. Then a narrow depletion region forms in the heavily-doped N-type drain region underneath the gate. If the resulting band bending exceeds the bandgap band-to-band tunneling occurs, which creates electron hole-pairs. The electrons go to the drain causing an increase in the drain current [2], [4]. This effect is called gate-induced drain leakage (GIDL) and result in an increase in the off-state current as depicted in Figure 2.6(a).

In order to reach the strong inversion the gate voltage need to invert the charge in the depletion region in the channel [2], [4]. When the gate length reduces, the overlap regions between the gate and the source/drain region cause a charge sharing between the gate depletion region and the source/drain depletion regions, which leads to that the gate terminal need to invert less charge in order to reach strong inversion. This is commonly referred to as short-channel effect and is illustrated in Figure 2.6(b), where a shorter transistor leads to a reduction of the threshold voltage, so called V_{th} roll-off [2], [4].

In order to scale the performance of the transistors the gate oxide thickness has been reduced for each generation. However, as the physical gate oxide thickness is reduced, the field strength between the gate and the substrate increases. This enables electrons or holes in the substrate to directly tunnel through the gate oxide [8] - [10]. As the gate tunneling current is inversely proportional to the oxide thickness the scaling of the gate oxide have lead to larger and larger gate leakage current for each CMOS technology generation.

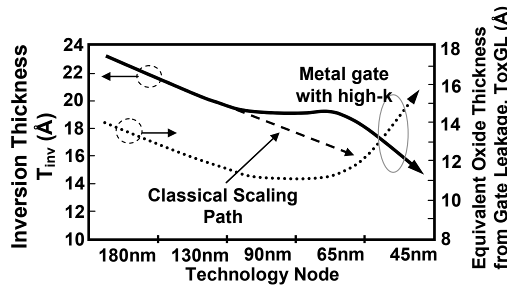


Figure 2.7: Gate leakage and gate-oxide thickness for nanoscale CMOS [12].

This led to that the oxide thickness stopped scaling for one CMOS generation in order to keep the leakage under control as shown in Figure 2.7 [11], [12]. To combat the gate leakage, while still obtaining improving performance by scaling, high-k dielectric materials have been proposed [11] - [13]. High-k materials make it possible to manufacture physically thicker oxide layers, while improving the electrical properties of the gate-oxide as shown in Figure 2.7. From being a research topic for many years, with the introduction of their 45-nm technology node, Intel announced a Hafnium-based high-k, metal-gate transistor, which shows gate leakage reductions in the order of 20X or higher. Hence, the scaling of electrical equivalent oxide thickness has been able to continue in the historical rate [11] - [13].

2.3 Power Dissipation in CMOS

Generally the power dissipation of a simple CMOS inverter (seen in Figure 2.8) can be divided into dynamic power and static power [1], [14]. The two main sources of dynamic power are switching power and short-circuit power, while the static power dissipation emanates from leakage currents in various forms.

2.3.1 Switching Power

Switching power is the power dissipated when the capacitive load is charged and discharged. If an input voltage of zero is assumed the PMOS transistor in Figure 2.8 will start to charge the capacitor C, which require the energy CV_{dd}^2 from the power supply and the capacitor is charged with the charge CV_{dd}. Once the input

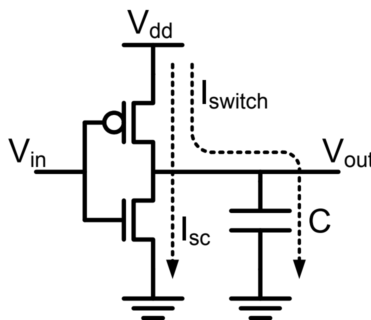


Figure 2.8: Schematic of a basic CMOS inverter, including dynamic currents (switching and short-circuit currents).

changes to V_{dd} the PMOS turns off and the NMOS starts to conduct, which will discharge the capacitor to ground. This requires no additional energy from the power supply, which means that only rising outputs dissipates power. Switching power is therefore described by the relation in equation (2.4), where C is the load capacitor that is charged, f_{clk} is the clock frequency with which the gate switches, V_{dd} is the power supply voltage, and α is the switching activity ratio, which determines how frequently the output switches from low-to-high per clock cycle [1].

$$P_{switch} = \alpha \cdot f_{clk} \cdot C \cdot V_{dd}^2 \quad (2.4)$$

2.3.2 Short-Circuit Power

The short-circuit power is due to the direct path between the power supply and ground that forms briefly when both PMOS and NMOS transistors conduct current simultaneously. Equation (2.5) describes the short-circuit power dissipation for the CMOS inverter in Figure 2.8, where β is the gain factor of the transistors, τ is the input rise/fall time, and V_{th} is the threshold voltage of the transistors [15]. The short-circuit power will increase in cases where the input rise/fall times to the gates are large compared to the output rise/fall times. For a well sized static CMOS gate the short-circuit power can be kept below 10% of the switching component [14].

$$P_{sc} = \frac{\beta}{12}\left(V_{dd} - 2V_{th}\right)^3 \cdot \tau \cdot f_{clk} \quad (2.5)$$

2.3.3 Leakage Power

Historically dynamic power has been the dominating contributor the power dissipation in digital CMOS design. However, with the continuing scaling of transistor sizes and voltages, leakage currents have grown to become large contributors of the overall power dissipation as shown in Figure 2.9 [3], [4], [16]. There are several leakage mechanisms that contribute to the total leakage for CMOS circuits both in active mode and in stand-by. Figure 2.10(a) shows the main leakage mechanisms that are present in a MOS transistor under normal operating conditions. These are subthreshold leakage current (I_{subth}), gate leakage currents (I_{gso}, I_{gc}, I_{gb}, I_{gdo}), and reversed junction leakage current (I_{junc}) [3], which all were introduced in section 2.2. High-performance VLSI designs usually operates at elevated temperatures, which is due to the heat generated by

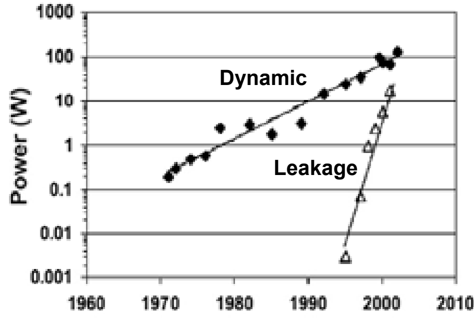


Figure 2.9: Dynamic versus leakage power for microprocessors [16].

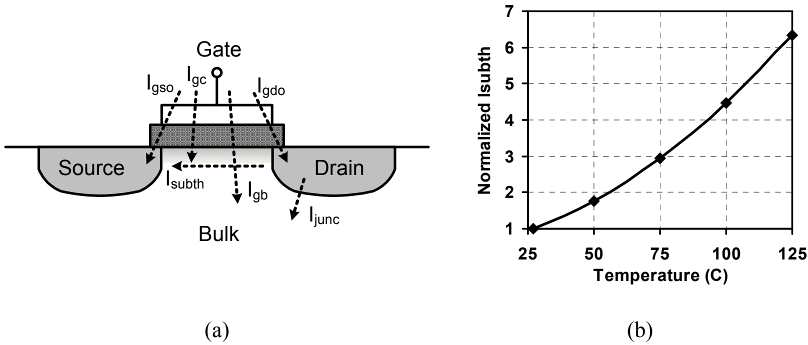


Figure 2.10: (a) Main leakage components for a MOS transistor.
(b) I_{subth} versus temperature for a 130 nm CMOS process.

the large power dissipation. Figure 2.10(b) shows how the subthreshold leakage current (I_{subth}) increases with increased temperature. This greatly affects the leakage power dissipation and robustness of VLSI designs during operation.

2.4 Basics of Integrated Circuit Manufacturing

The fabrication of an integrated circuit involves several complex and expensive processing steps. All processing steps need to be done in certain orders depending on the final product. The collection and ordering of the different

processing steps for making a certain product (like a microprocessor) is called a technology, which can be thought of as the product recipe [17], [18]. An integrated circuit starts with an identified need for an electronic system, which is translated into a circuit design with the number of devices needed to implement the design, including the sizes of each device and how to interconnect them. The interface between the processing engineers and the circuit designers are through a set of design rules. If the designer commit to these rules it is guaranteed that the circuit can be manufactured in the given technology. The designer therefore hands over a set of layers, which compose of the blueprint on how the transistors and other devices should be placed and connected.

2.4.1 Lithography

The blueprint provided by the circuit designer is transferred to the final silicon wafers using a technique called lithography. Each manufacturing stage involves applying a certain processing step to a selective area of the chip. The selectivity is accomplished by first applying a photoresist material on the surface of the chip. This material is either hardened or softened by the exposure of light. A photomask is used to block the light from the areas which should be processed. Photomasks are fabricated on various types of high-quality glass, where one side is filled with a patterned opaque layer formed usually by Chromium. The quality of the mask is essential because the same mask could be used for manufacturing of millions of chips [17].

2.4.2 Etching

Once the photoresist have been exposed and formed on the surface of the wafer, the non-exposed resist is rinsed away. The remaining image then needs to be transferred onto a layer underneath the resist. This is done using a technique called etching, which can be conducted either using wet etching or dry etching [1], [17]. Wet etching is a chemical process where the wafers are exposed to various types of chemical solvents that react with the wafer surface not covered with photoresist. Dry etching utilizes an electrically charged plasma, where the electrical field in the etching chamber causes the plasma to align in the vertical direction forming a chemical and physical sandblasting on the exposed surfaces. An important property of an etching technique is the etch anisotropy, which relates to the ratio between the lateral and vertical etching. Anisotropy equal to one is perfectly anisotropic, thus only etches in the vertical direction. Due to the vertical alignment, plasma etching has anisotropy close to one, which results in sharp and well defined patterns needed for the small feature sizes in modern CMOS technologies making dry etching the predominant etching method [1], [17].

2.4.3 Implantation, Oxidation, and Deposition

The starting material for an intergraded circuit process is usually a lightly-doped single-crystalline silicon wafer. To form the various active and passive devices the doping levels on the wafer need to be altered selectively. An example is the source and drain regions in the NMOS transistor in Figure 2.1, which are heavily N-doped regions, formed inside the lightly P-doped substrate. The predominant method of selective dopant insertion is ion implantation. Today advanced CMOS technologies require more than twenty different implants for which ion implantation is utilized [18]. Dopant ions are accelerated in an electrical field and are then brought to the wafer where they collide and penetrate the wafer surface. The penetration depth can be controlled with high precision by adjusting the strength of the electric field. The dosage of the impurity atoms introduced is tightly controlled by adjusting the ion current. However, ion implantation is a violent process because when the ions impact the silicon surface they cause displacement of the Si-atoms due to the high impact energy. In order to repair this damage and also to align the ions in the Si-crystal structure the wafer is heated to a high temperature and then slowly cooled down in a process called annealing [17], [18].

Oxides are extensively used in integrated circuits as field insulation between devices and of course as oxides in MOSFETs. The popularity of silicon as the

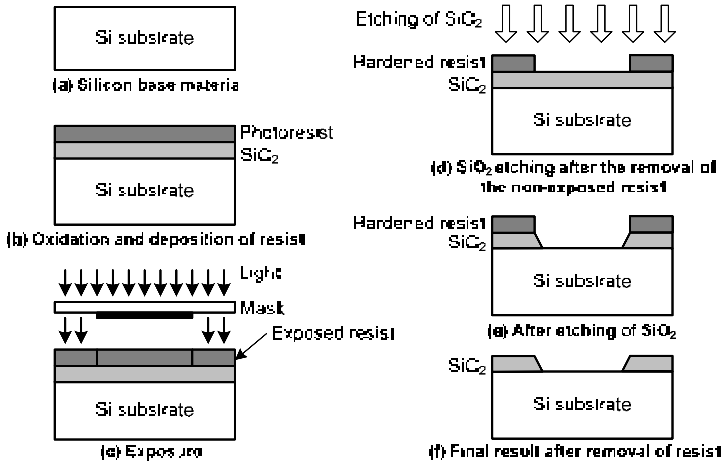


Figure 2.11: A simple example of the basic processing steps for patterning oxide.

semiconductor of choice in integrated circuit manufacturing is partly because of its ability to form excellent oxides, relatively free of impurities. In a manufacturing process the wafers are put in an oxygen rich environment where an elevated temperature can control the thickness of the oxide with high precision [17], [19]. To illustrate the basic steps described so far, a simple example of the patterning, oxidation, and etching is shown in Figure 2.11.

All structures that are needed to connect the devices together on the IC need to be built up on top of the active areas. This is done by depositing material on the wafer surface. There are several techniques that are utilized for this purpose both physical and chemical. Mainly two techniques are used in modern CMOS processes. These are sputtering and chemical vapor deposition and the choice of method depend mainly on the precision of the applied process step. In sputtering, a target is bombarded with high energy ions inside a plasma vacuum chamber. The bombarding releases ions from the target material. The target and the wafers are charged with opposing polarity creating an electric field between them. Once the target ions have been released they travel in the direction of the field and are deposed onto the wafer surface. Instead of bombarding the solid form of the material to be deposited, chemical vapor deposition (CVD) uses a molecular gas form of the material. The wafers are inserted in a reactor chamber where they are heated and a molecular gas of the material to be deposed is inserted in the chamber reacting with the heated wafer surfaces. This technique is extensively used for contacts between interconnecting layers in integrated circuits, due to superior geometric precision [1], [17].

2.5 Process Variation

Process variations are deviations from the intended circuit or device parameters. Both environmental variations and physical process variations will impact integrated circuits. Environmental variations occur during operation of the circuit, changing the device parameters due to temperature, power supply, and activity differences across the chip. Due to local hot-spots, such as an ALU in a microprocessor, temperature variations on different sites of the chip can be rather high. However, variations of power supply and temperature gradients can be treated as systematic fluctuations using worst-case guard-banding techniques [20], [21].

In nanoscale CMOS a number of limitations in the fabrication process creates variability in the transistor parameters, such as threshold voltage and channel dimensions [20] - [23]. Physical process variability is divided into mainly two levels depending on where the variations are introduced or caused. Fabrication

inaccuracies that impact the entire die or wafer are usually referred to as die-to-die or inter-die variations. This means that all devices are affected in the same way across the entire chip, which results in a shift in the mean value of some parameters [24]. Inaccuracies that impact different regions spatially within the die are defined as within-die or intra-die variation. Intra-die variation can be categorized as either systematic variation, with a smooth and continuous gradient across the die, or as uncorrelated random variation, changing discontinuously across the die [20], [21]. Intra-die variation leads to fluctuations within each die resulting in different parameters values for transistors within close proximity [25]. Historically fabrication induced process variation has been treated solely as a global systematic die-to-die variation. However, with the continuing reduction of device dimensions, within-die variation has grown [24]. The causes of process variation can mainly be categorized as variation in the geometry of devices and interconnects, and variation in the material.

2.5.1 Geometry Variations

The geometric structures of the devices and interconnects on the dies are crucial to the functionality. An essential parameter is the control of the film thickness during manufacturing. Especially gate oxide thickness is important as it directly impacts the threshold voltage of the transistor. Other structures that depend on the thickness are the deposited interconnect layers. The thicknesses of the films on the dies are in general relatively well controlled across the wafers. Variations in the film thickness are therefore usually only considered from wafer to wafer. These variations come from changes in the fabrication environments such as processing temperatures, equipment properties, and wafer polishing [22].

The lateral dimensions of devices and interconnects are typically more prone to manufacturing induced variation. Lateral dimensions are for instance channel length and width, and variations in the width and spacing of interconnect layers. Both have a profound impact on the circuit characteristics. Limitations in the photolithography process are large contributors to the geometric variations. Lithography has in the past relied on aggressively scaling the wavelength of light to enable the diminishing feature sizes following classical CMOS scaling. The resolution of a lithography system is usually expressed in terms of its wavelength and numerical aperture (*NA*) and a process dependent constant *k* according to the expression given in

$$\text{Resolution} = k\frac{\lambda}{NA}. \tag{2.6}$$

Typical values of k and NA have historically been in the range 0.5 to 0.8 [26]. Thus, the resolution of a given process has been largely determined by the wavelength of the light (λ) used to expose the wafer. However, with the introduction of the 0.18-µm CMOS technology node, this relationship changed and today transistors with gate lengths as small as 35 nm are fabricated using light sources with wavelength of 193 nm [11], [27] - [29]. This progress has been enabled by advances in image resist materials and image improvement techniques such as phase shift masks, optical proximity correction, inversion lithography, etc. [26], [27]. These advances have increased the NA and reduced the k constant, resulting in dramatic improvement in resolution. Nevertheless, even with the advances made in the lithography process there is a growing difference between the layout viewed in the design stage and the manufactured process. This leads to reduced control of relative critical dimensions. Lithography related inter-die variations can be accounted to impurities and defects of the mask and lenses, wafer placement, and photo system deviations. Systematic within-die variation can be accounted to wafer alignment, lens aberrations, stepper non-uniformities, and wafer topology. Random within-die variations related to lithography can be caused by shifts in depth of focus and line-edge roughness [22], [26], [30]. Moreover, inaccuracies in the etching process and the chemical-mechanical polishing of the wafers are further aggravating the variations in the geometric dimensions [22].

2.5.2 Material Variations

Variation in the materials is related to the implantation of dopant and the growth of new layers on the wafer. As advanced channel engineering is extensively used in modern CMOS technologies to limit the small geometry effects, dopant dosage and variation in the implant precision have profound impact on the variation [22]. Deviation arising due to implant dose, energy, or angle variation of the ion beam during ion implantation can affect junction depth and dopant profiles of the devices. Another effect is charge interaction between the ions in the ion beam, which also results in variation of the implant dose. The high temperatures generated during the beam exposure can also damage the photo-resist, severely altering the critical dimension geometries. Furthermore, small differences in beam angles can create variation in doping dosage due to shadowing. Another issue is contamination that can occur from particles sputtered by the ion beam from various objects in the implant chambers [18]. Ion implantation is utilized as the predominate tool to adjust the threshold voltage levels of the transistors during manufacturing. With shrinking channel dimensions the number of impurity dopant atoms needed to adjust the threshold voltage reduces. Therefore random fluctuations in the number of dopant atoms

inserted during fabrication leads to larger and larger variation of the threshold voltage of the transistors. Random dopant fluctuation is assumed to be the major cause of device mismatch in modern CMOS technologies [31]. Other material variation sources include variations and impurities in the crystal structures during deposition and annealing, such as crystallographic grain or phase effects. These effects can cause substantial variation in the conductivity of interconnect layers and contacts between metal layers [22].

2.5.3 Modeling of Process Variation

In order to model die-to-die variations the process foundries provides the designer with different set of parameters that covers the larger part of the variation spread. The statistical distribution of the process parameter variations can be approximated with a Gaussian distribution [23], [32], as shown in Figure 2.12 for the threshold voltage. The mean process values are usually refereed to as typical values and represent the average value for the given parameter. Apart from the typical process parameters process corners related to specific points on the distribution curve are also provided. Deviations from the mean values that results in an increase of the circuit performance is refereed to as fast process corners, and vice versa for slow process corners. In order to cover the worst-case spread the fast and slow corners are referred to as a shift of all parameters from the mean values corresponding to a number of standard deviations. Usually the corners covers ±3 times the standard deviation (σ), which results in more than 99% coverage of the die-to-die process parameter spread. Combinations of fast

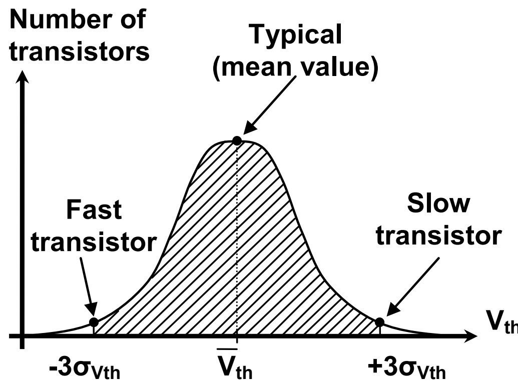


Figure 2.12: Distribution of threshold voltage for an NMOS transistor.

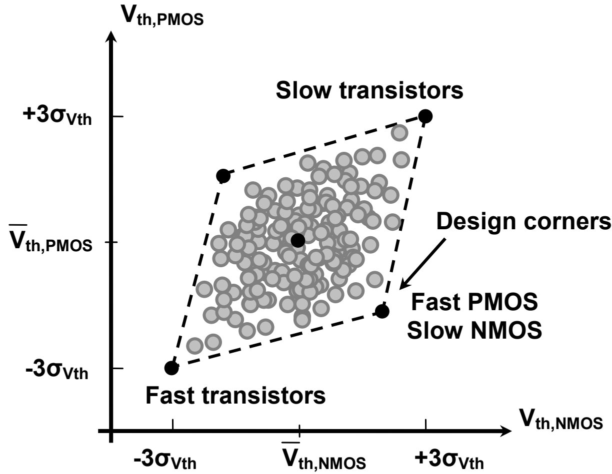


Figure 2.13: Example of threshold voltage design corners for an NMOS- PMOS transistor pair.

NMOS transistors and slow PMOS transistors and vice versa are also provided to analyze circuits for worst-case NMOS-PMOS matching variations [1], [22]. An example of different design corners for a NMOS-PMOS pair is shown in Figure 2.13.

Historically, corner analysis has been the predominant tool in digital design when only die-to-die variations were considered. However, with the growing within-die variation, specifically random variations, a more statistical approach is required in the design in contrast to the determinist corner analysis. For analysis of single gates and small systems usually Monte-Carlo simulations are used [34] - [36]. The circuits are then analyzed over a large number of Monte-Carlo simulations runs, where the different transistor parameters are randomly varied in the entire design space. This treats both the global systematic variation as well as local systematic and random variations. The foundry provides the circuit designer with model files containing information on the distribution and standard deviations of a number of transistor parameters. The local random variation is analyzed by using random process parameters for each transistor chosen independently, which enables the analysis of mismatch effects between the transistors in a circuit. By using this approach also close proximity effects such as random dopant fluctuations and gate-length variations can be modeled, analyzing the effect of local mismatch between transistors.

2.6 Bibliography

[1]. J.M. Rabaey, A. Chandrakasan, B. Nikolic, *Digital Integrated Circuits – A Design Perspective*, Prentice-Hall, 2003, ISBN: 0-13-597444-5.

[2]. B.G. Streetman and S. Banerjee, *Solid State Electronic Devices,* Prentice Hall, 2000, ISBN: 0-13-025538-6.

[3]. V. De, Y. Ye, A. Keshavarzi, S. Narendra, J. Kao, D. Somasekhar, R. Nair, S. Borkar, "Techniques for Leakage Power Reduction," in A. Chandrakasan, W.J. Bowhill, F. Fox, *Design of High-Performance Microprocessor Circuits*, IEEE Press, 2001, ISBN: 0-7803-6001-X.

[4]. K. Roy, S. Mukhopadhyay, and H. Mahmoodi-Meimand, "Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicron CMOS Circuits," in *Proceedings of the IEEE*, vol. 91, no. 2, pp. 305-327, 2003.

[5]. S. Chou, "Integration and Innovation in the Nanoelectronics Era," in *Digest of Technical Papers IEEE Solid-State Circuits Conference*, pp. 36-41, 2005.

[6]. R.H. Dennard, F.H. Gaensslen, H.-N. Yu, V.L. Rideout, E. Bassous, and A.R. LeBlanc, "Design of Ion-Implanted MOSFET's with Very Small Physical Dimensions," in *IEEE Journal of Solid State Circuits*, vol. SC-9, no. 5, 1974, pp. 256-268.

[7]. G.E. Moore, "Cramming more components onto integrated circuits," in *Electronics*, vol. 38, no. 8, 1965.

[8]. S. Mukhopadhyay, C. Neau, R.T. Cakici, A. Agarwal, C.H. Kim, and K. Roy, "Gate leakage Reduction for Scaled Devices Using Transistor Stacking," in *IEEE Transaction on VLSI Systems*, vol. 11, no. 4, pp. 716-730, 2003.

[9]. K.M. Cao, W.-C. Lee, W. Liu, X. Jin, P. Su, S.K.H. Fung, J.X. An, B. Yu, and C. Hu, "BSIM4 Gate Leakage Model Including Source-Drain Partition," in *Technical Digest of International Electron Devices Meeting*, pp. 815-818, 2000.

[10]. M. Drazdziulis and P. Larsson-Edefors, "A Gate Leakage Reduction Strategy for Future CMOS Circuits," in *Proceedings of the 29th European Solid-State Circuit Conference,* pp. 317-320, 2003.

[11]. K. Mistry, C. Allen, C. Auth, B. Beattie, D. Bergstrom, M. Bosi, M. Brazier, M. Buehler, A. Cappellani, R. Chau, C.-H. Choi, G. Ding, K. Fisher, T. Ghani, R. Grover, W. Han, D. Hanken, M. Hattendorf, J. He, J. Hicks, R. Huessner, D. Ingerly, R. Jain, R: James, L. Jong, S. Joshi, C. Kenyon, K. Kuhn, K. Lee, H. Lin, J. Maiz, B. McIntyre, P. Moon, J. Neirynck, S. Pae, C. Parker, D, Parsons, C. Prasad, L. Pipes, M. Prince, R. Ranade, T. Reynolds, J. Sandford, L. Shifren, J. Sebastian, J. Seiple, D. Simon, S. Sivakumar, P. Smith, C. Thomas, T. Troeger, P. Vandervoorn, S. Williams, K. Zawadzki, "A 45nm Logic Technology with High-k+Metal Gate Transistors, Strained Silicon, 9 Cu Interconnect Layer, 193nm Dry Patterning, and 100% Pb-free Packaging," in *IEEE International Electron Device Meeting*, December 2007, pp. 247-250.

[12]. T.-C. Chen, "Where CMOS is going: trendy hype vs. real technology," in *Digest of Technical Papers IEEE Solid-State Circuits Conference*, pp. 1-18, 2006.

[13]. http://www.intel.com, accessed June 2008.

[14]. A.P. Chandrakasan and R.W. Brodersen, "Minimizing Power Consumption in Digital CMOS Circuits," in *Proceeding of the IEEE*, vol. 83, no. 4, pp. 498-523, 1995.

[15]. H.J. Veendrick, "Short-Circuit Dissipation of Static CMOS Circuitry and Its Impact on the Design of Buffer Circuits," in *IEEE Journal of Solid-State Circuits*, vol. SC-19, no. 4, pp. 468-473, 1984.

[16]. G. Moore, "No Exponential is Forever: But "Forever" Can Be Delayed!" in *Digest of Technical Papers 2003 IEEE Solid-State Circuits Conference*, pp. 20-23, 2003.

[17]. S.A. Campell, *The Science and Engineering of Microelectronic Fabrication*, Oxford University Press, 1996, ISBN: 0-19-510508-7.

[18]. M. Ameen, I. Berry, W. Class, H.-J. Gossmann, L. Rubin, "Ion Implantation," in R. Doering, Y. Nishi (ed.), *Handbook of Semiconductor Manufacturing Technology*, CRC Press, 2008, ISBN: 978-1-57444-675-3.

[19]. C.R. Cleavelin, L. Colombo, H. Niimi, S. Pas, E.M. Vogel, "Oxidation and Gate Dielectrics," in R. Doering, Y. Nishi (ed.), *Handbook of Semiconductor Manufacturing Technology*, CRC Press, 2008, ISBN: 978-1-57444-675-3.

[20]. S. Nassif, K. Bernstein, D.J. Frank, A. Gattiker, W. Haensch, B.L. Ji, E. Nowak, D. Pearson, N.J. Rohrer, "High Performance CMOS Variability in the 65nm Regime and Beyond," in *IEEE International Electron Device Meeting*, pp. 569-571, 2007.

[21]. S.B. Samaan, "The Impact of Device Parameter Variations in the Frequency and Performance of VLSI Chips," in *IEEE International Conference on Computer Aided Design*, pp. 343-346, 2004.

[22]. D. Boning and S. Nassif, "Models of Process Variations in Device and Interconnects," in A. Chandrakasan, W.J. Bowhill, F. Fox, *Design of High-Performance Microprocessor Circuits*, IEEE Press, 2001, ISBN: 0-7803-6001-X.

[23]. S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, V. De, "Parameter Variations and Impact on Circuits and Microarchitecture," in *Proceedings of Design Automation Conference*, pp. 338-342, 2003.

[24]. K.A. Bowman, S.G. Duvall, and J.D. Meindl, "Impact of Die-to-Die and Within-Die Parameter Fluctuations on the Maximum Clock Frequency Distribution for Gigascale Integration," in *IEEE Journal of Solid-State Circuits*, vol. 37, no. 2, pp. 183-190, 2002.

[25]. S. Bhunia, S. Mukhopadhyay, and K. Roy, "Process Variation and Process-Tolerant Design," in *International Conference on VLSI Design*, pp. 699-704, 2007.

[26]. L.R. Harriott, "Limits of Lithography," in *Proceeding of the IEEE*, vol. 89, no. 3, March 2001, pp. 366-274.

[27]. G.E. Fuller, "Optical Lithography," in R. Doering, Y. Nishi (ed.), *Handbook of Semiconductor Manufacturing Technology*, CRC Press, 2008, ISBN: 978-1-57444-675-3.

[28]. K.-L. Cheng, C.C. Wu, Y.P. Wang, D.W. Lin, C.M. Chu, Y.Y. Tarng, S.Y. Lu, S.J. Yang, M.H. Hsieh, C.M. Liu, S.P. Fu, J.H. Chen, C.T. Lin, W.Y. Lien, H.Y. Huang, P.W. Wang, H.H. Lin, D.Y. Lee, M.J. Huang, C.F. Nieh, L.T. Lin, C.C. Chen, W. Chang, Y.H. Chiu, M.Y. Wang, C.H. Yeh, F.C. Chen, C.M. Wu, Y.H. Chang, S.C. Wang, H.C. Hsieh, M.D. Lei, K. Goto, H.J. Tao, M. Cao, H.C. Tuan, C.H. Diaz, and Y.J. Mii, "A Highly Scaled, High Performance 45nm Bulk Logic CMOS Technology

with 0.242μm^2 SRAM Cell," in *IEEE International Electron Device Meeting*, pp. 243-246, 2007.

[29]. T. Miyashita, K. Ikeda, Y.S. Kim, T. Yamamoto, Y. Sambonsugi, H. Ochimizu, T. Sakoda, M. Okuno, H. Minakata, H. Ohta, Y. Hayami, K. Ookoshi, Y. Shimamune, M. Fukuda, A. Hatada, K. Okabe, T. Kubo, M. Tajima, T. Yamamoto, E. Motoh, T. Owada, M. Nakamura, H. Kudo, T. Sawada, J. Nagayama, A. Satoh, T. Mori, A. Hasegawa, H. Kurata, K. Sukegawa, A. Tsukune, S. Yamaguchi, K. Ikeda, M. Kase, T. Futatsugi, S. Satoh, and T. Sugii, "High-Performance and Low-Power Bulk Logic Platform Utilizing FET Specific Multiple-Stressors with Highly Enhanced Strain and Full-Porus Low-k Interconnects for 45-nm CMOS Technology," in *IEEE International Electron Device Meeting*, pp. 251-254, 2007.

[30]. B.H. Calhoun, Y. Cao, X. Li, K. Mai, L.T. Pileggi, R.A. Rutenbar, and K.L. Shepard, "Digital Circuit Design Challenges and Opportunities in the Era of Nanoscale CMOS," in *Proceeding of the IEEE*, vol. 96, no. 2, Feb. 2008.

[31]. K.J. Kuhn, "Reducing Variation in Advanced Logic Technologies: Approaches to Process and Design for Manufacturability of Nanoscale CMOS," in *IEEE International Electron Device Meeting*, pp. 471-474, 2007.

[32]. H. Mahmoodi, S. Mukhopadhyay, and K. Roy, "Estimation of Delay and Variations due to Random-Dopant Fluctuations in Nanoscale CMOS Circuits," in *IEEE Journal of Solid-State Circuits*, vol. 40, no. 9, pp. 1787-1796, 2005.

[33]. K. Bernstein, K.M. Carrig, C.M. Durham, P.R. Hansen, D. Hogenmiller, E.J. Novak, N.J. Rohrer, *High Speed CMOS Design Styles*, Kluwer Academic Publishers, 1999, ISBN: 0-7923-8220-X.

[34]. S. Zanella, A. Nardi, A. Neviani, M. Quarantelli, S. Saxena, and C. Guardiani, "Analysis of the Impact of Process Variation on Clock Skew," in *IEEE Transaction on Semiconductor Manufacturing*, vol. 13, no. 4, pp. 401-406, 2000.

[35]. R. Rao, A. Srivastava, D. Blaauw, D. Sylvester, "Statistical Estimation of Leakage Current Considering Inter- and Intra-Die Process Variation, " in

Proceedings of the International Symposium on Low Power Electronics and Design, pp. 84-89, 2003.

[36]. D. Sylvester, K. Agarwal, S. Shah, “Variability in nanometer CMOS: Impact, analysis, and minimization,” in *Integration the VLSI Journal*, 2007, doi:10.1016/j.vlsi.2007.09.001.

[37]. International Technology Roadmap for Semiconductors (ITRS) – 2007 ed., http://www.itrs.net, accessed June 2008.

Chapter 3

Clocking and Synchronization

3.1 Introduction

The absolute majority of high-performance digital designs today utilize a synchronous clock to order events [1]. Although the principle of synchronization is easy in the system design perspective, ordering all events in a high-performance design in a synchronous fashion requires generation and distribution of clock signals at multi-GHz clock frequencies, which is extremely challenging. Moreover, synchronization circuits such as latches and flip-flops constitute the clocked registers that synchronize the data flow in a VLSI circuit. Hence, flip-flops and latches are among the most important circuit blocks in a digital synchronous chip design. Ideally, timing circuits like flip-flops and latches should add as little latency as possible, and have low power dissipation. In practice however, clocked registers can actually consume a substantial fraction of the clock-cycle period, and dissipates a considerable portion of the total power. It is therefore of highest interest to design both power-efficient clock distribution circuits and synchronization circuits, that are optimized for their desired task.

This chapter seeks to introduce the concept of synchronization and the circuits and designs techniques that are commonly used in order to construct a synchronous VLSI design. Most of the definitions introduced in this chapter will be extensively used in the following parts of the thesis.

3.2 Synchronization Circuits

3.2.1 Level-Sensitive Latches

Latches are the simplest kind of synchronizing circuit in a sequential design. A latch is a level sensitive device that is either transparent or opaque, depending on the signal level of the clock input. A simple schematic of a transmission-gate latch is shown in Figure 3.1. When the clock signal (*clk*) is high the latch lets the input (*D*) pass to the output (*Q*), while if the clock is low the output (*Q*) will hold the previous input data on the output.

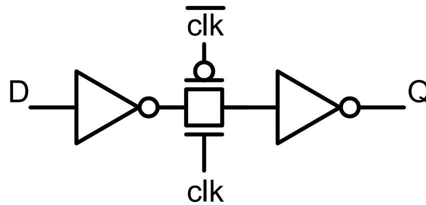


Figure 3.1: Schematic example of a simple level-sensitive latch.

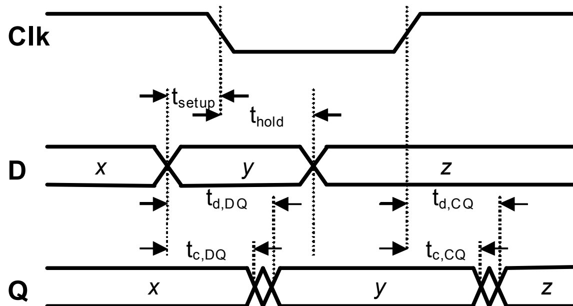


Figure 3.2: Timing definitions for a level-sensitive latch.

A level-sensitive latch that is transparent for a high clock signal samples the data on the falling clock-edge [1], [4], and stores it at the latch output during the low clock phase. The timing diagram for a level-sensitive latch is shown in Figure 3.2 and the different delays are defined as [4]:

- Data-to-output delay ($t_{d,DQ}$) is defined as the delay from the data edge to the output transition during the transparent phase of the sampling clock.
- Clock-to-output delay ($t_{d,CQ}$) is defined as the delay from the beginning of the transparent phase of the clock to the output of the latch.
- Setup time (t_{setup}) is defined as the time that the input (*D*) needs to be stable before the sampling clock edge in order to capture the correct data.
- Hold time (t_{hold}) is defined as the minimum time that the input (*D*) needs to be stable after the sampling clock edge in order for the latch to have time to capture the correct data.

3.2.2 Edge-Triggered Flip-flops

An edge-triggered flip-flop samples the data input on one edge of the clock, but in contrast to a level-sensitive latch, keeps the sampled data on the output during the remainder of the clock period. A simple master-slave flip-flop can be constructed from two cascaded level-sensitive latches, as shown in Figure 3.3.

When the clock signal (*clk*) is low the first latch, called master latch, is transparent and the input is transferred to the intermediate node (*X*). The second latch, called the slave latch, is opaque so the output (*Q*) is held at its previous state. When the clock signal (*clk*) makes a low-to-high transition the master latch becomes opaque, thus latching the data on the input on the intermediate data node (*X*). The slave latch becomes transparent, and the intermediate data at (*X*) is transferred to the output (*Q*). The data on the output is valid for the remainder of the clock period [1], [4].

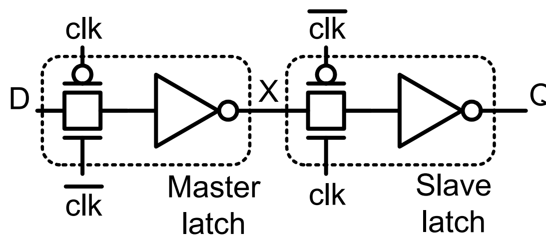


Figure 3.3: Schematic example of an edge-triggered flip-flop.

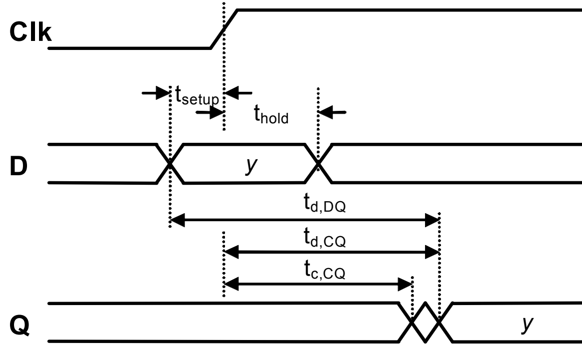


Figure 3.4: Timing definitions for a positively edge-triggered flip-flop.

A timing diagram of a positive edge-triggered flip-flop is shown in Figure 3.4 [4]. All timing relations for the edge-triggered flip-flop are referred only to the sampling clock edge[1]. The timing relations for an edge-triggered flip-flop are defined by essentially five different delays, which are [1]:

- Setup time (t_{setup}) is defined as the time that the input (*D*) must be stable before the sampling clock edge in order for the flip-flop to capture the correct data.
- Hold time (t_{hold}) is defined as the time that the input (*D*) must be stable after the sampling clock edge in order for the flip-flop to capture the correct data.
- Contamination delay ($t_{c,CQ}$) is the minimum delay from the active clock edge to the output (*Q*) of the flip-flop.
- Propagation delay or clock-to-output delay ($t_{d,CQ}$) is defined as the maximum delay from the active clock edge to the output (*Q*) of the flip-flop.
- Data-to-output delay or total delay ($t_{d,DQ}$) is the sum of the setup time and the propagation delay, which represents the total latency of the flip-flop.

[1] Rising edge in Figure 3.4

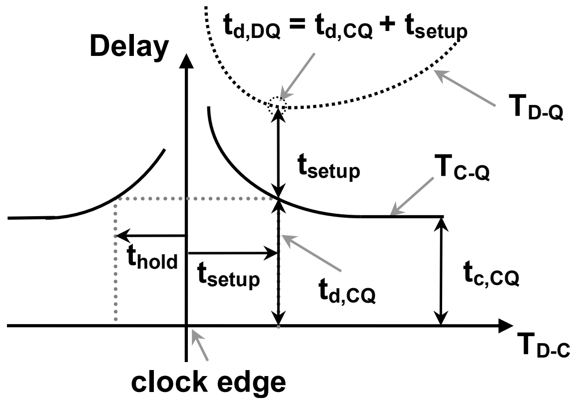


Figure 3.5: Delay metrics for flip-flops.

3.3 Characterizing Synchronization Circuits

3.3.1 Characterizing Timing for Latches and Flip-Flops

The timing characteristics of latches and flip-flops are extremely important, because it influence the performance of a synchronous system directly. Both for edge-triggered flip-flops and level-sensitive latches the delay metrics are highly dependent on the arrival of the incoming data relative the sampling clock edge. This section describes a method to extract all delay metrics for edge-triggered flip-flop. However, the timing metrics for level-sensitive latches can be treated in a similar way just by keeping track of which clock edge to relate to. Figure 3.5 shows a method to extract the delays for a given edge-triggered flip-flop. The figure shows the delay plotted against the total data-to-clock delay ($T_{D\text{-}C}$), which is the time difference between a change of data and the latching clock edge. When the data changes early, the clock-to-output delay ($T_{C\text{-}Q}$) will have its minimum value. This is defined as the contamination delay, and is denoted by $t_{c,CQ}$. As the data edge arrives closer to the latching clock edge, the clock-to-output delay will start to increase monotonically. The total delay will follow the expression in equation (3.1), calculated from a data change on the input of a flip-flop the previous cycle, to a valid data on the output of the flip-flop. The total flip-flop delay is shown as the dashed plot in Figure 3.5, here denoted as $T_{D\text{-}Q}$. At a certain data-to-clock delay, the data-to-output delay ($T_{D\text{-}Q}$) will have an

optimum value, which is denoted $t_{d,DQ}$. It is obvious from Figure 3.5 that if the data changes later than this optimum point, the total delay will increase dramatically as the flip-flop becomes more metastable. The data-to-output delay will increase until the flip-flop fails to capture the input data. The data-to-clock delay that results in the minimum data-to-output delay is defined as the setup time t_{setup}. The clock-to-output delay or propagation delay, $t_{d,CQ}$, is defined as the clock-to-output delay at the setup time according to Figure 3.5. Equation (3.2) defines the optimal data-to-output delay or propagation delay ($t_{d,DQ}$), which is the total minimum latency for the flip-flop from a data change on the input to a corresponding change of the output [5], [6].

$$T_{D-Q} = T_{D-C} + T_{C-Q} \tag{3.1}$$

$$t_{d,DQ} = \min\left(T_{D-Q}\right) = t_{setup} + t_{d,CQ} \tag{3.2}$$

The negative side of the T_{D-C}-axis in Figure 3.5 corresponds to the case where data changes after the latching clock edge. If the data changes long after the latching edge, the output delay would correspond to the contamination delay ($t_{c,CQ}$). However, if data changes close after the latching edge, the output delay will increase monotonically. There is a certain point on the negative T_{D-C} axis where the output delay equals the maximum clock-to-output delay $t_{d,CQ}$. The negative value of the data-to-clock delay at this point is defined as the hold time t_{hold} as shown in Figure 3.5.

When cascading two flip-flops the output of the first flip-flop is driving the input of the second one. The first flip-flop can change output value as early as $t_{c,CQ}$ after the clock edge, while the second flip-flop needs to have the input stable as long as t_{hold}, in order to capture the correct data. If $t_{c,CQ}$ is shorter than t_{hold}, the second flip-flop can fail to capture the data correctly. This is referred to as internal race violation, and to assure that this does not occur the timing metric in equation (3.3) need to be fulfilled [6].

$$t_{race} = t_{c,CQ} - t_{hold} > 0 \tag{3.3}$$

3.3.2 Power-Delay Design Space

When optimizing flip-flop circuits, trade-offs between power and delay can be made as for all logic design. A power-efficient flip-flop is one that for a certain delay has the minimal power dissipation and vice versa. This can be illustrated in a design-space graph shown in Figure 3.6, which shows the total power for

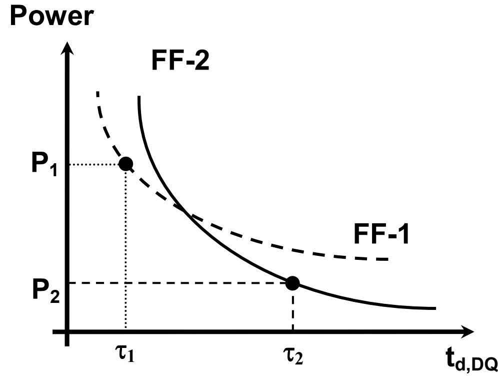


Figure 3.6: Power-delay design space for two different flip-flop topologies.

two flip-flops plotted versus the total minimum latency ($t_{d,DQ}$). If a fair and accurate comparison between different flip-flop topologies should be done, a power-delay plot like Figure 3.6 is needed. As an example comparing flip-flop FF-1 with FF-2 only at one point will yield that one of the topologies is better then the other in general. However, the truth might be that they are the better choice in different parts of the design space. For instance, a low-latency flip-flop that dissipates more power (FF-1 at τ_1) could be used in critical parts of a design, while using a slower less power-consuming flip-flop (FF-2 at τ_2) in non-critical parts.

3.3.3 A Flip-Flop Optimization Approach

Figure 3.7 shows a schematic picture of an optimization approach, which is utilized for finding power-delay-optimal transistor sizing of flip-flops throughout the work presented in this thesis. The flip-flop to be optimized is initially sized for a few different power and delay points in the design space. An automated simulation script then finds all delays (especially the total latency, $t_{d,DQ}$) and the total power for each sizing point. The initial simulation results in a number of points in the power-delay design space, shown as the first iteration in Figure 3.7. Out of these points the ones with minimum power for a certain delay are chosen, as shown in the figure. These sizing points are then feed back to the evaluation script, which tweaks the size of the transistors around their given values. This results in a number of new sizing points, which are simulated for power and delay, and the result is once again analyzed. After a number of

iterations the optimization loop is stopped and the output is the optimal power-delay curve, which is used for comparison to other flip-flop topologies.

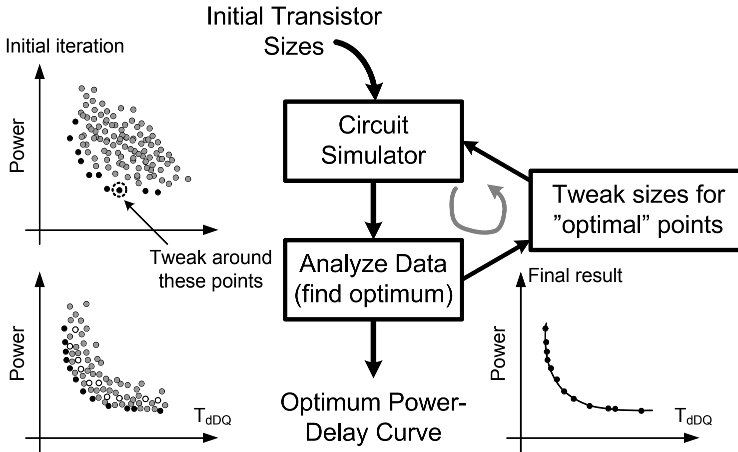


Figure 3.7: An iterative flip-flop optimization approach.

3.4 Clock Signal Integrity

Ideally all clocked registers in a synchronous design should receive the clock signal simultaneously. However, limitations and parasitic effects in the distribution of the clock signals across the chip will usually result in non-idealities that cause phase differences between the delivered clocks signals.

3.4.1 Clock Jitter

Any temporal variation in the arrival of the clock edges between different clock-cycles related to the nominal clock edge are referred to as clock jitter [1], [2]. This jitter arises from noise and inaccuracy in the clock generation circuitry and in noise generated in the clock distribution network [2]. Clock jitter is considered a random variable with zero-mean value around the ideal clock edge arrival time, as shown in Figure 3.8 [1]. Jitter on the clock signal impacts the performance of sequential system negatively by limiting the amount of time available for logic evaluation out of the nominal clock cycle, thereby reducing the maximum clock frequency of the system [1].

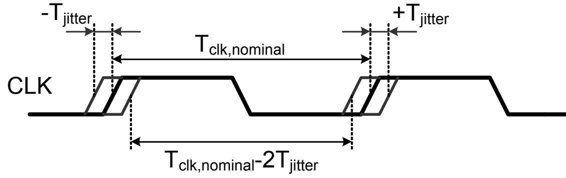


Figure 3.8: Timing diagram showing jitter influence on the clock period.

3.4.2 Clock Skew

In a VLSI circuit design, clock signals are usually tapped from a global clock network at different locations. Any imbalance or difference in the distribution paths between different tap locations causes delay difference between the clock signals [1], [2]. This results in clock skew, which is shown in Figure 3.9 between clock signals CLK1 and CLK2. Skew between the clocks are caused not only by imbalance in the network, but also by process variation that is locally changing the strength of the drivers and the RC-delay of the wires. Clock skew can directly impact the performance and functionality of a sequential system. In contrast to the random behavior of clock jitter, clock skew is random in size and polarity between different data paths, but can be considered as fixed or slowly changing for a particular data path. A positive skew between two clock signals is defined as when the launching clock, CLK1, leads the capturing clock signal, CLK2 (as shown in Figure 3.9). Consequently, a negative skew is defined as the launching clock lagging the capturing clock ($T_{skew} < 0$ in Figure 3.9) [1].

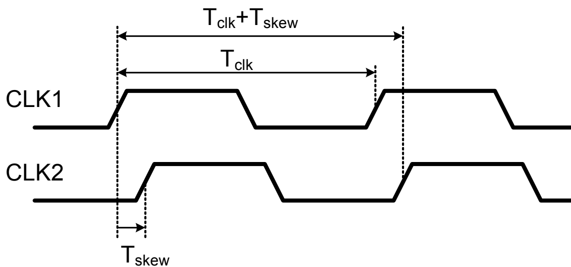


Figure 3.9: Timing diagram showing clock skew between two clock signals.

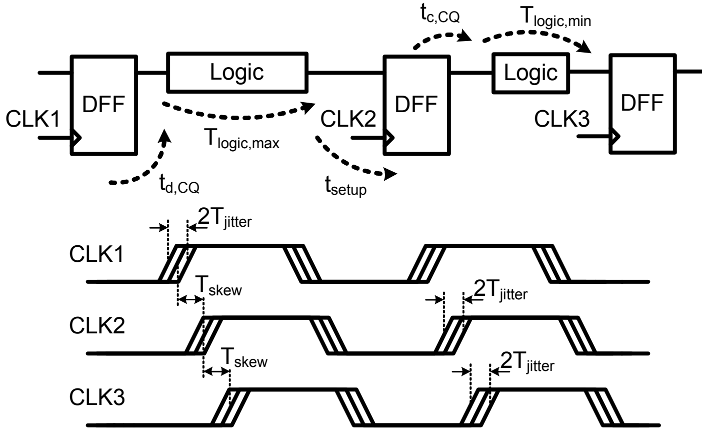


Figure 3.10: General sequential circuit showing delay metrics.

3.5 Synchronization Approaches

3.5.1 Edge-Triggered Clocking

Edge-triggered clocking is a robust and simple clocking approach, where static logic is placed between edge-triggered flip-flops. All events are triggered on the active clock edge, and the logic has the entire clock cycle to evaluate [1], [2]. Figure 3.10 shows a simple example of a pipeline stage using an edge-triggered clocking approach. Here the clock signals CLK1, CLK2, and CLK3 are all derived from the same source. However, both skew and jitter introduced in the clock distribution cause a phase difference between the signals according to the discussion in the previous section. At the rising edge of CLK1 the output of the first flip-flop is updated after a worst-case delay of $t_{d,CQ}$. This triggers the static logic to evaluate, which requires an evaluation time of $T_{logic,max}$. The second flip-flop requires the input to be valid at least the setup time t_{setup} before the active edge of CLK2. The timing relation for an edge-triggered system is therefore given by

$$T_{cycle,avail} \geq t_{d,CQ} + t_{setup} + T_{logic,max} \,, \tag{3.4}$$

where $T_{cycle,avail}$ is the clock cycle time available for completing the flip-flop delays and logic evaluation [1]. The available cycle is equal to the ideal clock period accounting for any skew and jitter between CLK1 and CLK2 according to

$$T_{cycle,avail} = T_{clock} + T_{skew} - 2T_{jitter}\,, \tag{3.5}$$

where $2T_{jitter}$ is the worst-case peak-to-peak jitter, as shown in Figure 3.8, and T_{skew} is defined as either positive or negative as shown in Figure 3.9 [1].

The logic path with delay $T_{logic,max}$ denotes the worst-case delay path in the system, which is the performance limiter for the maximum clock frequency. However, there is also a minimum delay path in a sequential system. In Figure 3.10 the minimum logic delay path is between the second and the third flip-flop, and is denoted by the delay $T_{logic,min}$. If the delay between the two flip-flops is short enough, the data updated on the output of flip-flop two can reach the input of flip-flop three before its master latch closes. This creates a race condition, which is a serious issue because it can not be mitigated by any reduction of the operational frequency. In order to guarantee that this condition does not occur, the following min-delay or race condition need to be fulfilled

$$t_{hold} + T_{skew} + 2T_{jitter} < t_{c,CQ} + T_{logic,min}\,, \tag{3.6}$$

where t_{hold} is the hold time for flip-flop three and $t_{c,CQ}$ is the contamination delay of flip-flop two [1]. From equation (3.6) it is evident that the worst-case situation arises when there are no logic in between two cascaded flip-flops, which reduces the equation to the same as equation (3.3) but including jitter and skew. Hence, both skew and jitter further worsens the race concerns for edge-triggered clocking. However, according to equation (3.6) there is a certain amount of minimum logic that can be inserted between two consecutive flip-flops, in order to alleviate the problems with race in the pipeline stages.

3.5.2 Level-Sensitive Clocking

Another clocking approach that is capable of incorporating a wider selection of logic styles is level-sensitive clocking. In contrast to the edge-triggered clocking approach, logically adjacent latches are separated by a half-cycle in level-sensitive clocking. The logic is distributed between latches in either half-cycle, and can be thought of as adding logic between the latches of a master-slave flip-flop [1] - [3]. Figure 3.11 shows a pipeline stage using level-sensitive clocking, where every other latch is clocked with the inverted clock signal. An advantage with level-sensitive clocking is that because the latches are transparent during a

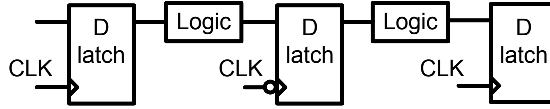


Figure 3.11: General sequential circuit showing delay metrics.

half-phase of the clock, there is a possibility to give away and borrow time between the clock phases [3]. This ability also makes the level-sensitive clocking technique less sensitive to clock skew, which has made it a popular choice in high-performance designs [1] - [3].

3.6 Common Flip-Flop Topologies

In the literature there are a large number of flip-flop circuits proposed, which can be classified into mainly three categories. These are master-slave latch pairs, pulsed latches, and sense-amplifier based flip-flops [1].

3.6.1 Master-Slave Latch Pairs

The most common approach to build an edge-triggered flip-flop is to combine two level-sensitive latches, which are clocked on opposite clock phases. An example of a common static TG-MSFF is shown in Figure 3.12. The advantages

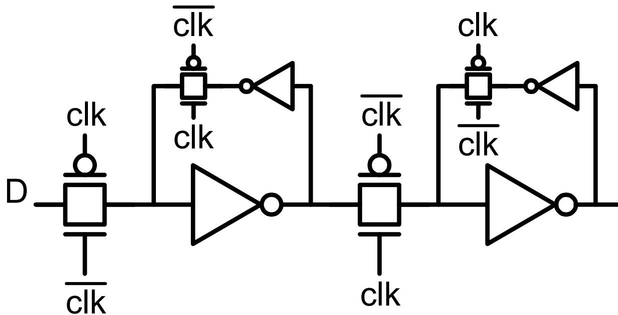


Figure 3.12: Conventional master-slave transmission-gate flip-flop.

with this flip-flop are simplicity and excellent race immunity [5]. However, because the setup time of the flip-flop is mainly determined by the propagation delay of the master latch, and the output latency is determined by the propagation delay through the slave latch, the total flip-flop latency will be quite large. Moreover, the hard edge property of a MSFF renders any time borrowing between clock phases impossible, and increases the clock skew sensitivity [7], [8]. Therefore, this flip-flop is frequently used in non-critical data paths, where the larger latency and clock skew sensitivity is not impacting the performance of the system.

3.6.2 Pulsed Latches

To counteract the hard edge property of master-slave edge-triggered flip-flops, several pulsed-latch approaches have been presented. The principle of a pulsed latch is to create a short pulse on the latching edge of the clock, and then clock the latch with that pulse, thereby obtaining an edge-triggering behavior. A simple example of a pulsed-latch using a clocked-CMOS (C^2MOS) latch is shown in Figure 3.13. The pulse generator could be an external circuit or integrated in the latch design [1]. However, an external clock-pulse generator could be shared with a number of other latches in order to reduce the total clock power. For a rising edge of the clock (*clk*) the output of the pulse-generator (clk_{pulse}) will go high, making the latch transparent. After a delay (t_{delay}) the output of the pulse-generator will go low, thus making the latch opaque. During the high pulse the latch transfers any change of data on the input. This property

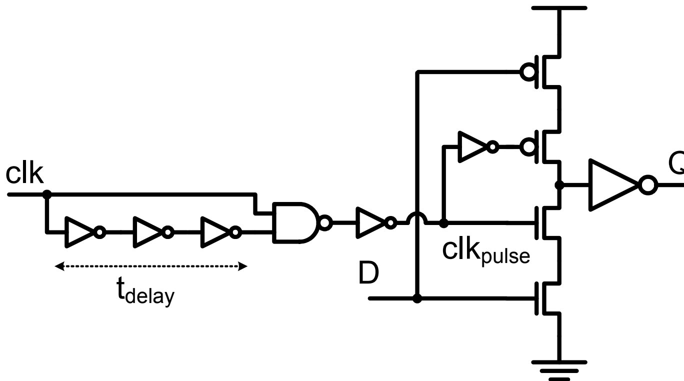


Figure 3.13: Pulsed C^2MOS latch with external pulse-generator.

is referred to as negative setup time, because data will be correctly latched even though arriving after the rising edge of the main clock signal. This property can be utilized to borrow time from neighboring clock cycles. This soft-edge property can also be used to trade off time borrowing for clock skew absorption [8]. However, the soft-edge property exists at the expense of hold time, because during the duration of the latching pulse, the input can not be allowed to change data erroneously in order not to corrupt the output value. Hence, pulsed flip-flops with negative setup time usually have large positive hold times, and limited internal race robustness. However, several pulsed latches have been described in the literature, and some of them are utilized as low-latency flip-flops in critical pipeline stages in high-performance microprocessors where the logic depth mitigates the reduced internal race margin. Some of the most popular topologies are the pulsed hybrid-latch flip-flop (HLFF), and the semi-dynamic pulsed-latch flip-flop (SDFF), presented in [10] and [11], respectively.

3.6.3 Sense-Amplifier Based Flip-flops

A third technique to implement an edge-triggered flip-flop is to utilize a sense-amplifier to sample the data [1], [12]. A typical sense-amplifier based flip-flop is shown in Figure 3.14 [13], where a pre-charged sense-amplifier front-end is

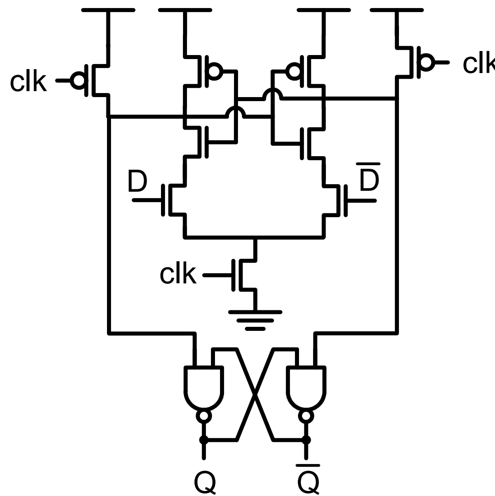


Figure 3.14: Example of a sense-amplifier-based flip-flop.

used to sample the complementary data inputs when the clock makes a rising transition. A NAND-based SR-latch captures the sampled data and holds it until next rising clock-edge. Due to the amplification provided by the feedback in the cross-coupled inverters, the flip-flop can sample input signals with small amplitude difference. Sense-amplifier flip-flops could therefore be utilized as synchronous level-converters between different power-supply regions [14]. Another advantage with the sense-amplifier flip-flop is the low number of clocked transistors, which gives low clock load. Moreover, the flip-flop in Figure 3.14 can at most do one transition per clock cycle, making it attractive as an interface between static and pre-charged dynamic CMOS circuits. One of the largest drawbacks with the sense-amplifier flip-flops is the pre-charged behavior of the sample-stage, which is power-consuming especially when the data activity on the inputs is low.

3.7 Conventional Clock Distribution Techniques

There are a number of conventional clock distribution techniques that could be considered for high-speed synchronous designs, each with specific advantages and disadvantages. Clock distribution can be divided into largely two parts; final stage driver, and pre-driver network. The final stage driver feeds the clock to the final load, which consists of the clocked circuits such as flip-flops and latches. To distribute the clock signal from the central clock source to the final stage drivers a pre-driver network is used [2], [15].

3.7.1 Tapered Clock Buffer Chain

The simplest pre-driver technique used for clock distributions is to utilize a large buffer chain. The clock signal is buffered through a number of gain stages to the final clock drivers. Usually the technique is used to distribute the clock in a wide wire across the chip, from where the clock is distributed using a more local clock distribution technique. The skew from the clock source to the output of the

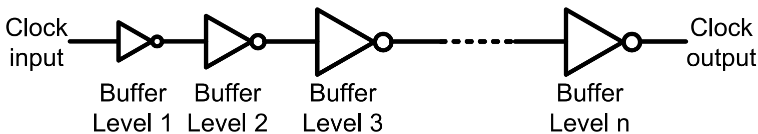


Figure 3.15: Example of n-stage clock buffer chain.

buffer chain is not of any concern, because the output is only one point, which means that the skew at the main clock wire on the chip is considered to be zero. However, if the clock signals are tapped at different distances from the main clock wire the skew will increase, and the skew follows the contour of the clock wire [2], [15]. Figure 3.15 shows an n-stage tapered buffer chain where each inverter is upsized with a certain tapering factor compared to the preceding one, thus providing a large final gain capable of driving the large clock load.

3.7.2 Clock Trees

If the final clock drivers are distributed across the chip, the pre-driver network must also be able to distribute the clock. A technique to do this is to utilize so called binary trees, also known as H-trees. Instead of buffering the clock through a large buffer chain placed on one spot of the chip, the pre-driver buffers can be distributed across the H-tree network. To reduce the skew in the tree all branches need to be balanced in order for their delays to be matched. This balancing is done by matching the RLC delay of each wire segment, which can be a quite cumbersome task. However, there are tools available that can do this automatically. Figure 3.16 shows an example of an H-tree with 16-leaf nodes [2] [15].

3.7.3 Grid Clock Distribution

Clock grids are common as final driver networks because they require no matching of delays from the source to the final loads. The clock is also accessible at many more points compared to a binary tree, which makes the grid approach appealing to designers as it is forgiving to late design changes. Clock skew is usually not considered an issue for a clock grid, because it is usually implemented either dense or small, so that the delay between different taps is limited. The principle is shown in Figure 3.17 where the drivers feed the clock signal to the grid. An obvious disadvantage with the grid approach is that it leads to a large amount of wiring, which results in excessive loading of the clock drivers, thus causing additional clock power [1], [2], [15].

3.7.4 Length-Matched Serpentines

Length-matched serpentines are a clocking technique that is similar to the clock tree, but the length of the branches is adjusted to balance for any load mismatch. This way the skew compared to the other serpentines can be minimized. This approach is also appealing for the cases where the clock load is asymmetrically distributed. It is compatible with the buffer chain in that it can distribute the clocks from the main clock wire to the final loads with minimum skew [1], [2], [15].

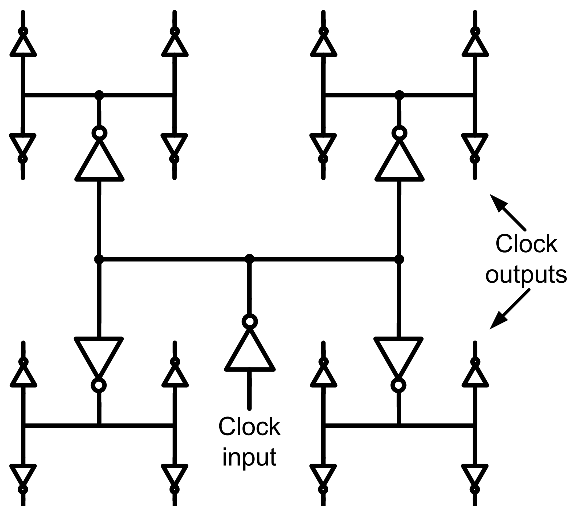


Figure 3.16: Example of a RLC-balanced 16-leaf H-tree clock-network.

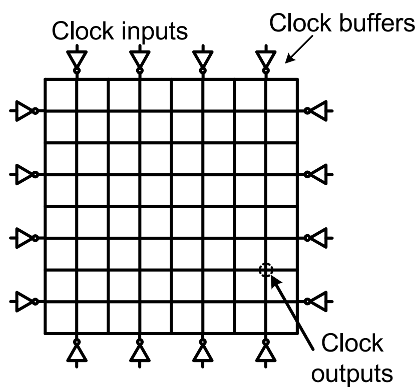


Figure 3.17: Simple example of a clock grid.

3.8 Bibliography

[1]. J.M. Rabaey, A. Chandrakasan, B. Nikolic, *Digital Integrated Circuits – A Design Perspective*, Prentice-Hall, 2003, ISBN: 0-13-597444-5.

[2]. K. Bernstein, K.M. Carrig, C.M. Durham, P.R. Hansen, D. Hogenmiller, E.J. Novak, N.J. Rohrer, *High Speed CMOS Design Styles*, Kluwer Academic Publishers, 1999, ISBN: 0-7923-8220-X.

[3]. D. Harris, *Skew-Tolerant Circuit Design*, Morgan Kaufmann Publishers, 2001, ISBN: 1-55860-636-X.

[4]. W.J. Dally, J.W. Poulton, *Digital System Engineering*, Cambridge University Press, 1998, ISBN: 0-521-59292 2.

[5]. V. Stojanovic and V. Oklobdzija, "Comparative Analysis of Master-Slave Latches and Flip-Flops for High-Performance and Low-Power Systems," in *IEEE Journal of Solid-State Circuits,* vol. 34, no. 4, pp. 536-548, 1999.

[6]. D. Markovic, B. Nikolic, and R.W. Brodersen, "Analysis and Design of Low-Energy Flip-Flops," in *Proceedings of the 2001 International Symposium on Low Power Electronics and Design*, pp. 52-55, 2001.

[7]. V.G. Oklobdzija, "Clocking Multi-GHz Systems," in C. Piguet, *Low-Power Electronics Design*, CRC Press, 2005, ISBN: 0-8493-1941-2.

[8]. N. Nedovic, V.G. Oklobdzija, W.W. Walker, "A Clock Skew Absorbing Flip-Flop," in *Digest of Technical Papers IEEE International Solid-State Circuits Conference*, pp. 342-343, 2003.

[9]. J. Yuan and C. Svensson, "High-speed CMOS Circuits Techniques," in *IEEE Journal of Solid-State Circuits*, vol. 26, no. 1, pp. 62-70, 1989.

[10]. H. Partovi, R. Burd, U. Salim, F. Weber, L. DiGregorio, D. Draper, "Flow-Through Latch and Edge-Triggered Flip-flop Hybrid Elements," in *Digest of Technical Papers IEEE International Solid-State Circuits Conference,* pp. 138-139, 1996.

[11]. F. Klass, C. Amir, A. Das, K. Aingaran, C. Truong, R. Wang, A. Mehta, R. Heald, and G. Yee, "A New Family of Semidynamic and Dynamic Flip-Flops with Embedded Logic for High-Performance Processors," in *IEEE Journal of Solid-State Circuits*, vol. 34, no. 5, pp. 712-716, 1999.

[12]. B. Nicolic, V. Stojanovic, V.G. Oklobdzija, W. Jia, J. Chiu, M. Leung, "Sense Amplifier-Based Flip-Flop," in *Digest of Technical Papers 1999 IEEE International Solid-State Circuits Conference,* pp. 282-283, 1999.

[13]. J. Montanaro, R.T. Witek, K. Anne, A.J. Black, E.M. Cooper, D.W. Dobberpuhl, P.M. Donahue, J. Eno, W. Hoeppner, D. Kruckemyer, T.H. Lee, P.C.M Lin, L. Madden, D. Murray, M.H. Pearce, S. Santhanam, K.J. Snyder, R. Stehpany, S.C. Thierauf, "A 160-MHz, 32-b, 0.5-W CMOS RISC microprocessor," in *IEEE Journal of Solid-State Circuits*, vol. 31, no. 11, pp. 1703-1714, 1996.

[14]. F. Ishihara, and B. Nikolic, "Level-Conversion for Dual-Supply System," in *IEEE Transactions on VLSI Systems*, vol. 12, no. 2, pp. 185-195, 2004.

[15]. D.W. Bailey, "Clock Distribution," in A. Chandrakasan, W.J. Bowhill, F. Fox, *Design of High-Performance Microprocessor Circuits*, IEEE Press, 2001, ISBN: 0-7803-6001-X.

Part II

Low-Power Clocking

Chapter 4

Background

4.1 Introduction

With the continuing evolution of integrated circuits, and the constant thrive for more functionality in advanced VLSI systems, circuit complexity is constantly increasing. In today's high-performance VLSI circuits, power dissipation is one of the major design challenges. A major contributor to the total power in modern microprocessors is from the clock distribution network [1] - [4], which can dissipate as much as 70% of the total power for certain high-performance applications [3].

The increased complexity and performance requirements in high-speed VLSI systems, such as microprocessors, rely either on increased pipelining and higher clock frequencies, or increased parallelism in order to reach the tight performance demands [5]. Both approaches increase the clock load because of the growing number of clocked elements, thereby further aggravating the clock power problem. This is setting the limit on the amount of functionality that can be integrated and on the achievable performance. This is forcing designers and researchers to find novel low-power solutions for the clock distribution, in order to enable further increase in functionality. Conventional high performance clocking techniques, which were described in Chapter 3, are mature and robust. Still, they are based on a relatively rigid and traditional philosophy, which

enforces a power-hungry clock distribution network, leaving almost no room for low power.

This chapter begins with a power analysis of conventional clock networks. This is followed by a brief discussion on some conventional and previously presented techniques, utilized to achieve low power dissipation in conventional buffer driven clock networks. A background is also given to energy recovering clocking, which is proposed as a technique to considerably reduce the clock power dissipation.

4.2 Power Analysis of Conventional Buffered Clocking

In order to be able to compare the power dissipation of a conventional clocking network to other clocking techniques, a power analysis is needed. In section 3.7 conventional clocking and clock distribution techniques were discussed. Regardless of clock network style, all of these techniques can be modeled by a simplified buffer chain [6]. For instance, the loads on the branches in a balanced H-tree are effectively driven in parallel, which can be modeled by a single lumped capacitive load. Furthermore, the grid capacitance can also be modeled as a lumped capacitor assuming that a local clock grid is used for the final distribution to the clock elements (flip-flops and latches). A simplified model of a clock distribution network, with a four-level clock tree and local grid distribution, is shown in Figure 4.1. The capacitance C_{grid} is the lumped model of the local distribution grid, while C_{FF} denotes the equivalent capacitive load due to the flip-flops and latches. The final clock load comprises the sum of C_{grid}

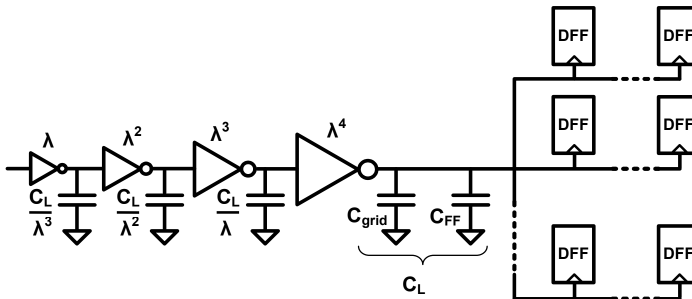


Figure 4.1: Simplified model of a conventional clock distribution with local clock grid and clock loads.

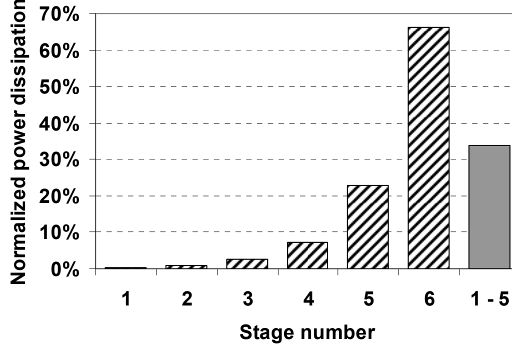


Figure 4.2: Normalized power dissipation for a 6-stage inverter chain with tapering factor of 3.

and C_{FF} and is denoted by C_L. The four-stage buffer chain is sized with a tapering factor of λ, which in a well-designed buffer gives the intermediate loads in the chain according to Figure 4.1 [6], [7]. The total switching power dissipation can be expressed as the relationship in equation (4.1), where N is the number of stages in the buffer. If the clock buffers are designed with a tapering factor of three ($\lambda = 3$), the relation in equation (4.1) yields that more than two thirds of the power is dissipated in the last buffer stage [6], [7]. This is also illustrated by the simulation shown in Figure 4.2, for a six-stage buffer chain with tapering factor equal to three. Consequently, with the overall objective to reduce the total clock power, primary concern must be to limit the clock power in the last driver stage (local drivers) in order to make significant impact. Hence, solutions requiring final clock buffers can only reduce the minor part of the total clock power.

$$P_{clock} = C_L V_{dd}^2 f_{clk} + \sum_{n=1}^{N-1} \left(\frac{1}{\lambda^n} \right) C_L V_{dd}^2 f \tag{4.1}$$

4.3 Conventional Low-Power Clocking Techniques

Because of the escalating clock power dissipation, a number of techniques have been proposed in order to reduce the power dissipation in conventional clock

networks. If most of the power dissipated in the clock distribution networks is assumed to be due to active switching power, there exist a limited number of possibilities to reduce the power dissipation according to the expression in equation (4.1).

4.3.1 Frequency and Voltage Reductions

The clock frequency (f_{clk}) has a direct linear relationship on the switching power dissipation, which means that if the clock frequency can be reduced the clock power will decrease. However, with an unchanged amount of hardware, the amount of operations that can be done each second naturally reduces. This leads to a decrease in the throughput of the circuit. In order to reduce this performance penalty, an increased amount of parallelism can be incorporated. This leads to an area penalty and an increase in the clock load because more circuits are needed, which also applies to the clocked circuits. Nevertheless, frequency is a powerful tool in order to limit the average clock power in a system. Especially for systems where the need for the highest performance is only required temporarily. During the more normal operation modes, which require lower computing speed, the clock frequency can be lowered, and the average clock power dissipation can thereby be reduced [8].

Moreover, the switching power varies as the square of the power supply voltage (V_{dd}). Hence considerable power savings is possible by scaling down the power supply. As the delay of the digital gates increases with lower power supply voltage, this will also lead to a reduction of the throughput. However, similarly to the case with temporarily changing the frequency, the power supply voltage can also be reduced when the required computation speed is low. Dynamic voltage and frequency scaling during operation is something that is commonly used for microprocessors, which can incorporate a number of different power-saving modes. When the computation need is low, both power supply voltage and clock frequency of the chip is reduced, leading to considerable average clock and total power reductions [8], [9].

4.3.2 Low-Swing Clocking

An alternative to scaling down the power supply for all circuits, including the clock network, would be to scale down the power supply voltage for only the clock drivers. Using low-swing clock signals on the global clock grid and then level-convert and use a high-voltage clock locally have been proposed in [10]. However, as discussed in section 4.2 the major part of the clock power is dissipated in the final driver, thus a global low-swing clock technique can only reduce a minor part of the clock power. Instead, techniques using low-swing clock signal all the way to the clock load have been proposed, and proved

feasible showing substantial clock power savings [11] - [15]. However, as the driving strength of the clock buffers in the clock network is reduced at lower power supply, to maintain driving performance and robustness the clock buffers will need to be upsized [11], which can mitigate some of the positive effects.

4.3.3 Clock Gating

The clock signal is the only signal in synchronous design which has an activity ratio of one, meaning that it switches all the time. A common technique to limit the activity ratio at least locally in a system is to incorporate so called clock gating. Clock gating means that the clock is masked using a control signal, which blocks the clock signal coming from the clock source, to circuits further downstream in the network [7], [16], [17]. Clock gating can be performed at many different levels of granularity. At the local level the final clock driver is gated, reducing both the power in the final driver and the switching power in the clocked elements. The gating can also be done higher in the clock network hierarchy, by gating the clock to larger blocks. Different architectural methods to find the granularity of the gating and the potential blocks to be gated have been proposed [18] - [20].

4.3.4 Clock Load Reduction

According to equation (4.1) the clock switching power changes linearly with the clock load, $C_{L,}$ of the clock network. Reducing the clock load can be accomplished for instance by designing clocked elements with reduced loading on the clock network, which could be accomplished by down sizing of the transistors in the flip-flops. However, this leads to a negative impact on the performance of the flip-flops, and therefore to a global performance penalty of the design.

4.3.5 Summary

Even though conventional methods exist to combat the escalating clock power, many of these techniques are depending on the application and architecture. Aggressive clock gating and dynamic voltage and frequency scaling reduce the clock power during periods of low computing intensity. Still the clock power at full-speed modes, which accounts for substantial part of the total power in microprocessors, is not reduced with these techniques [2], [3]. Therefore there is still a large need for more general clock power saving techniques, which also limits the power dissipation during the high-performance modes.

4.4 Energy Recovery Clocking Techniques

As a potential low-power clocking solution, a number of energy recovering clocking techniques have emerged, and gained increasing research interest due to the ability for large clock power savings. The main advantage, compared to the conventional clocking, is that resonant clocking enable recycling of the energy needed to charge the clock load when the clock network is discharged. Ideally this can lead to zero power dissipation, but even considering practical limitations in integrated electronics, such as interconnect parasitics, large power saving is feasible resulting in clock power dissipation well below fCV^2.

4.4.1 Adiabatic Switching

The ultimate goal for an energy recovering clock driver is to recycle all the energy used to charge the clock capacitance. The adiabatic principle of charging a capacitance has the potential to reduce the power dissipation well below the conventional fCV^2 relationship [21]. The concept of adiabatic switching can be ideally described by the switch circuit in Figure 4.3. The charge that the capacitance C is charged with is $Q = CV_{dd}$, and the voltage drop across the resistive switch is $V_R = IR$. The energy consumed in the resistor is therefore $E = V_R Q = IR\ CV$. The time it takes to charge the capacitance between 0 and V_{dd} is denoted T, and the charging current equals $I = CV_{dd}/T$, which gives the energy according to $E = RC^2 V_{dd}^2 / T$. This relation shows that the energy consumption related to charging a capacitance with a constant current can get arbitrarily small, provided that the charge time is long enough [21]. Although successful implementations of modified adiabatic circuits have been implemented and run above 100 MHz [22], a pure adiabatic technique relies on slow transition times, which makes it less suitable for multi-GHz clocked systems.

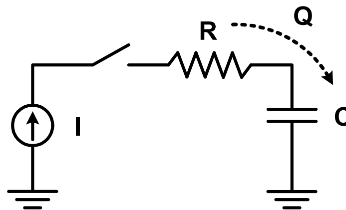


Figure 4.3: Constant current charging with an ideal resistive switch.

4.4.2 Oscillator Driven Global Clock Networks

Instead of using fully adiabatic switching, partly adiabatic techniques based on high-Q resonator circuits have been investigated for implementation of an energy recovering clocking. Low-power clocking based on sinusoidal clocks generated by resonating LC-tanks have been treated in [23], where a four-phase clock generator have been used to drive a transmission-gate based flip-flop. The four phases are used in order to reduce the overlapping transparent window in the flip-flop when using only two phases. However, the proposed four-phase clock requires large routing resources. To increase the freedom in the choice of clocked elements, a sine-to-square buffer is proposed, which will block the clock driver from resonating the largest part of the energy. This will therefore limit the amount of power that could be saved [23].

Clock distribution based on resonating drivers has also displayed several advantageous properties, compared to the conventional clocking, when it comes to clock signal integrity. Global clock distribution based on standing-wave oscillators have shown good numbers on clock skew and clock jitter [24]. However, the technique requires clock signal amplification and buffering due to the limited amplitude of the standing-wave clock signal across the grid. This is therefore largely aimed for the reduction of the clock skew across the chip, while any reduction of the power dissipation will be limited to the global network. Furthermore, the technique has still to show any power benefit even at the global network level. Another solution, aimed at reducing the jitter of the global clock signal, uses distributed LC-tank oscillators that are injection locked to an external clock signal [25], [26]. By utilizing distributed resonators in the clock network, low jitter has been displayed together with large reduction of the global clock power dissipation. The technique although promising, does not consider the impact which the resonant clocking will have on a conventionally designed digital synchronous system.

4.4.3 Bufferless LC-tank Resonant Clocking

As discussed in section 4.2, in order to be able to reduce the major part of the clock power, the power dissipation in the final driver need to be addressed. An energy recovery clocking technique that uses an LC-tank oscillator to drive the entire clock load in a design without intermediate buffers has been presented in [6]. The general idea of a resonator driven clock system is to use a high-Q LC-tank to form an oscillator directly driving the entire final clock load. In order to make the resonant clocking become a feasible alternative to the conventional buffer driven clock distribution approach, a fair comparison between comparable designs using both clocking solutions is required. Recently several

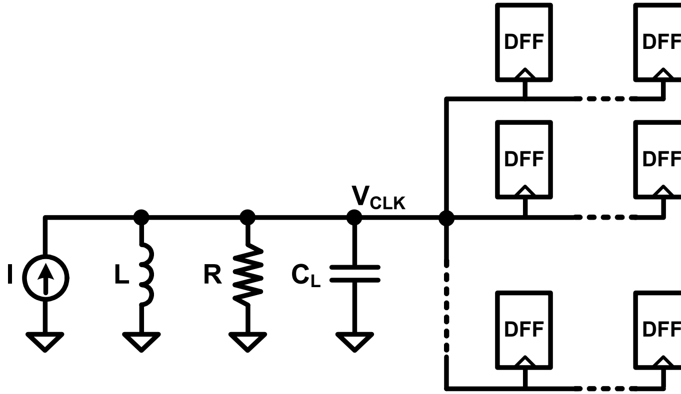


Figure 4.4: Simplified schematic model of a resonator driven clock network.

working energy recovering solutions of bufferless LC-tank resonant clock drivers, running at GHz-range frequencies have been shown on fabricated silicon [27] - [31].

4.5 Power Analysis of LC-tank Resonant Clocking

Using the same clock load as in Figure 4.1, a simplified schematic model of a LC-tank oscillator driven system can be modeled as in Figure 4.4. At resonance, the impedance for the parallel inductance L and capacitance C_L will be infinite. Hence, power will only be dissipated in the shunted resistance R, which comprises the total equivalent parasitic resistance in the LC-tank including the parasitic resistance in both the inductor, and in the parasitic resistance in the distributed capacitance from the resonator to the timing circuits. A current source, used as a negative resistance, is required in order to compensate for the energy loss in the parasitic elements [6].

The output voltage of the resonator circuit V_{CLK} will be a sinusoidal signal in the form of $V_0\cos(2\pi f_0 t+\phi)+V_0$, where f_0 to a first order can be calculated according to equation (4.2). The driving strength of the compensating current source sets the magnitude of V_0, and the average power dissipated by the resonator circuit is determined by the expression in equation (4.3). The quality factor (Q-value) of the parallel RLC-tank can be expressed as $Q_{tank} = 2\pi f_0 RC_L$, and if the magnitude and DC-level are both $V_0 = V_{dd}/2$, a full swing clock signal

is obtained. Equation (4.3) can then be expressed as equation (4.4). This shows that a better Q-value of the tank reduces the power dissipation of the resonator circuit. By using the expression in equation (4.1) using a buffer with four stages ($N = 4$) and a tapering factor equal to three ($\lambda = 3$), the power can be compared to the power dissipated by the bufferless LC-tank resonator driven clock network. This comparison yields the expression given in equation (4.5), which implies that the Q-value of the LC-tank needs to be larger than $\pi/2$ (≈ 1.6), in order to obtain power saving compared to the entire conventional clock buffer [6].

$$f_0 = \frac{1}{2\pi\sqrt{LC}} \tag{4.2}$$

$$P_{resonant} = \frac{\frac{1}{T_0}\int_0^{T_0} V_{Clk}^2\,dt}{R} = \frac{3V_0^2}{2R} \tag{4.3}$$

$$P_{resonant} = \frac{3}{4Q_{tank}}\pi V_{dd}^2 f_0 C_L \tag{4.4}$$

$$\frac{P_{resonator}}{P_{clock}} = \frac{3}{4Q_{tank}}\pi \Bigg/ \sum_{n=0}^{N-1}\left(\frac{1}{\lambda^n}\right) = \frac{3}{4Q_{tank}}\pi \Bigg/ \frac{40}{27} \approx \frac{\pi}{2Q_{tank}} \tag{4.5}$$

Although the power efficiency of a resonant clocking scheme ideally can approach 100%, in today’s standard CMOS technologies, parasitic resistances in the inductance and interconnect network result in losses that can not be neglected, thus leading to clock power dissipation. However, a substantial clock power saving can still be achieved using resonant clocking, compared to conventional clocking schemes. As an example, according to equation (4.5), for a tapering factor of three in the conventional clock buffers, a Q-value of $Q_{tank} > \pi$ would result in more than 50% clock power saving in the resonant clocking network, compared to the conventional scheme. Such power savings have also been reported in previous published works [27] - [31].

4.6 Issues Concerning Tank Q-value

As shown in equation (4.5), the efficiency of the LC-tank energy recovering clocking technique relies entirely on what quality factor of the LC-tank that can be obtained. The two contributing components on the Q-value of the LC-tank are the distributed capacitance in the clock network and clock elements, and the Q-value of the inductance in the LC-tank. The definition of the Q-value is given by equation (4.6) [32], where ω_0 is the resonance angular frequency for the LC-tank given by $\omega_0 = 1/\sqrt{LC}$. The stored energy at resonance for the circuit in Figure 4.4 is $E_{tot} = \frac{1}{2} C_L V^2 = \frac{1}{2} C_L (RI)^2$, while the average power dissipation is given by $P_{avg} = \frac{1}{2} I^2 R$. This yields an expression for the total LC-tank Q-value given in equation (4.7) [32]. Let R_L and R_C denote the equivalent parallel parasitic resistance in the inductor and capacitor, respectively. The quality factor for each component can then be expressed as $Q_L = R_L / \omega_0 L$ and $Q_C = \omega_0 C_L R_C$ for the inductance and capacitance, respectively. The total LC-tank Q-value can, by applying equation (4.7), be expressed as a parallel connection of the Q-value of the two lossy reactive components, which yields the expression given in equation (4.8) [6], [32]. This means that the total Q-value is limited by the smallest of the Q-values for the reactive components.

$$Q_{tank} = \omega_0 \frac{\textit{energy stored}}{\textit{average power dissipated}} = \omega_0 \frac{E_{tot}}{P_{avg}} \tag{4.6}$$

$$Q_{tank} = \frac{1}{\sqrt{LC}} \frac{\frac{1}{2} C_L (IR)^2}{\frac{1}{2} I^2 R} = \frac{R}{\sqrt{L/C_L}} = \frac{R}{\omega_0 L} = \omega_0 C_L R \tag{4.7}$$

$$Q_{tank} = \omega_0 C_L (R_L // R_C) = \frac{Q_L Q_C}{Q_L + Q_C} = Q_L // Q_C \tag{4.8}$$

In conventional RF design the Q-value of on-chip inductance is usually considered the limiter of the total Q-value of an LC-tank oscillator. Nevertheless, obtaining Q-values reaching above ten for inductance values of a few nano-Henrys has been shown to be possible in CMOS technologies [33], [34]. Several techniques have been proposed in order to improve and optimize the Q-values for planar integrated inductors, such as using shunted metal layers [33] and providing good substrate shielding [34]. Using high-resistive substrates underneath the inductor could further improve the Q-value

[33]. Other more exotic techniques like wafer-level packaging [35] or micro-machined devices [36] have also been proposed to improve the inductor Q-value considerably. Hence, it is clear that the limiting factor for the total LC-tank Q-value is not the inductance but instead the losses for the distributed capacitance.

In order to achieve a low-loss clock network to the flip-flops and latches in the design, all presently used techniques to obtain low-skew clock grids should be applied. This includes using wide metal wires in the top metal layers in order to reduce the wire resistance. However, for a conventionally driven clock grid a tradeoff between power dissipation and grid resistance have to be made. This tradeoff is not present in the same degree for the case of resonant clocking [6]. Although techniques are available to limit the losses in the clock grid it is hard to obtain equivalent parasitic series resistance values less than a few Ohms, which is still high enough to make the Q-value of the distributed capacitance the limiting factor for the overall LC-tank Q-value.

4.7 Bibliography

[1]. S.D. Naffziger, G. Colon-Bonet, T. Fischer, R. Riedlinger, T.J. Sullivan, and T. Grutkowski, "The Implementation of the Itanium 2 Microprocessor," in *IEEE Journal of Solid State Circuits*, vol. 37, no. 11, pp. 1448-1460, 2002.

[2]. S. Naffziger, B. Stackhouse, T. Grutkowski, "The Implementation of a 2-core Multi-Threaded Itanium®-Family Processor," in *Digest of Technical Papers IEEE International Solid-State Circuit Conference,* pp. 182-183, 2005.

[3]. C.J. Anderson, J. Petrovick, J.M. Keaty, J. Warnock, G. Nussbaum, J.M. Tendler, C. Carter, S. Chu, J. Clabes, J. DiLullo, P. Dudley, P. Harvey, B. Krauter, J. LeBlanc, P.-F. Lu, B. McCredie, G. Plum, P. J. Restle, S. Runyon, M. Scheuermann, S. Schmidt, J. Wagoner, R. Weiss, S. Weitzel, B. Zoric, "Physical Design of a Fourth-Generation POWER GHz Microprocessor," in *Digest of Technical Papers IEEE International Solid-State Circuit Conference,* pp. 232-233, 2005.

[4]. E.G. Friedman, "Clock Distribution Networks in Synchronous Digital Integrated Circuits," in *Proceeding of the IEEE*, vol. 89, no. 5, pp. 665-692, 2001.

[5]. http://www.intel.com, accessed: June 2008.

[6]. A.J. Drake, K.J. Nowka, T.Y. Nguyen, J.L. Burns, and R.B. Brown, “Resonant Clocking Using Distributed Parasitic Capacitance,” in *IEEE Journal of Solid-State Circuits*, vol. 39, no. 9, pp. 1520-1528, 2004.

[7]. D.W. Bailey, “Clock Distribution,” in A. Chandrakasan, W.J. Bowhill, F. Fox, *Design of High-Performance Microprocessor Circuits*, IEEE Press, 2001, ISBN 0-7803-6001-X.

[8]. T.D. Burd, T.A. Pering, A.J. Stratakos, and R.W. Brodersen, “A Dynamic Voltage Scaled Microprocessor System,” in *IEEE Journal of Solid State Circuits*, vol. 35, no. 11, pp. 1571-1580, 2000.

[9]. R.K. Krishnamurthy, S.K. Mathew, M.A. Anders, S.K. Hsu, H. Kaul, S. Borkar, “High-performance and Low-voltage Challenges for Sub-45nm Microprocessor Circuits,” in *The 6th International Conference On ASIC*, vol. 1, pp. 283-286, 2005.

[10]. J. Pangjun and S.S. Sapatnekar, “Low-Power Clock Distribution Using Multiple Voltages and Reduced Swings,” in *IEEE Transaction on Very Large Scale Integration Systems*, vol. 10, no. 3, pp. 309-318, 2002.

[11]. D. Markovic, J. Tschanz, and V. De, “Feasability Study of Low-Swing Clocking,” in *Proceeding of the 24th International Conference on Microelectronics*, vol. 2, pp. 547-550, 2004.

[12]. H. Kawaguchi and T. Sakurai, “A Reduced Clock-Swing Flip-Flop (RCSFF) for 63% Power Reduction,” in *IEEE Journal of Solid-State Circuits*, vol. 33, no. 5, pp. 807-811, 1998.

[13]. C. Kim and S.-M. Kang, “A Low-Swing Clock Double-Edge Triggered Flip-Flop,” in *IEEE Journal of Solid-State Circuits*, vol. 37, no. 5, pp. 648-652, 2002.

[14]. R.K. Krishnamurthy, S. Hsu, M. Anders, B. Bloechel, B. Chatterjee, M. Sachdev, S. Borkar, “Dual Supply Voltage Clocking for 5GHz 130nm Integer Execution Core,” in *Symposium On VLSI Circuits Digest of Technical Papers*, pp. 128-129, 2002.

[15]. H. Kojima, S. Tanaka, and K. Sasaki, “Half-Swing Clocking Scheme for 75% Power Saving in Clocking Circuitry,” in *IEEE Journal of Solid-State Circuits*, vol. 30, no. 4, pp. 432-435, 1995.

[16]. S. Rusu, and G. Singer, "The First IA-64 Microprocessor," in *IEEE Journal of Solid State Circuits*, vol. 35, no. 11, pp. 1539-1544, 2000.

[17]. S. Tam, S. Rusu, U.N. Desai, R. Kim, J. Zhang, and I. Young, "Clock Generation and Distribution for the First IA-64 Microprocessor," in *IEEE Journal of Solid State Circuits*, vol. 35, no. 11, pp. 1545-1552, 2000.

[18]. Q. Wu, M. Pedram, and X. Wu, "Clock-Gating and Its Application to Low Power Design of Sequential Circuits," in *IEEE Transaction on Circuits and Systems-I: Fundamental Theory and Applications*, vol. 47, no. 103, pp. 415-420, 2000.

[19]. J. Oh and M. Pedram, "Gated Clock Routing for Low-Power Microprocessor Design," in *IEEE Transaction on Computer-Aided Design of Integrated Circuits and Systems*, vol. 20, no. 6, pp. 715-722, 2001.

[20]. H. Jacobson, P. Bose, Z. Hu, A. Buyuktosunoglo, V. Zyuban, R. Eickemeyer, L. Eisen, J. Griswell, D. Logan, B. Sinharoy, J. Tendel, "Stretching the Limits of Clock-Gating Efficiency in Server-Class Processors," in *Proceedings of the 11th International Symposium on High-Performance Computer Architecture*, pp. 238-242, 2005.

[21]. L. Svensson, "Adiabatic and Clock-Powered Circuits," in C. Piguet, *Low-Power Electronics Design*, CRC Press, 2005, ISBN: 0-8493-1941-2.

[22]. W.C. Athas, N. Tzartzanis, L. Svensson, and L. Peterson, "A Low-Power Microprocessor Based on Resonant Energy," in *IEEE Journal of Solid-State Circuits*, vol. 32, no. 11, pp. 1693-1701, 1997.

[23]. B. Voss and M. Glesner, "A Low Power Sinusoidal Clock," in *IEEE International Symposium on Circuits and Systems*, vol. 4, pp. 108-111, 2001.

[24]. F. O'Mahony, C.P. Yue, M.A. Horowitz, and S.S. Wong, "A 10-GHz Global Clock Distribution Using Coupled Standing-Wave Oscillators," in *IEEE Journal of Solid-State Circuits*, vol. 38, no. 11, pp. 1813-1820, 2003.

[25]. S.C. Chan, P.J. Restle, K.L. Shepard, N.K. James, R.L. Franch, "A 4.6GHz Resonant Global Clock Distribution Network," in *Digest of Technical Papers IEEE International Solid-State Circuit Conference*, pp. 342-343, 2004.

[26]. S.C. Chan, K.L. Shepard, P.J. Restle, "1.1 to 1.6 GHz Distributed Differential Oscillator Clock Network," in *Digest of Technical Papers IEEE International Solid-State Circuit Conference*, pp. 518-519, 2005.

[27]. M. Hansson, B. Mesgarzadeh, and A. Alvandpour, "1.56 GHz On-chip Resonant Clocking in 130nm CMOS," in *Proceedings of the IEEE Custom Integrated Circuit Conference*, pp. 241-244, 2006.

[28]. B. Mesgarzadeh, M. Hansson, and A. Alvandpour, "Jitter Characteristic in Charge Recovery Resonant Clock Distribution," in *IEEE Journal of Solid-State Circuits*, vol. 42, no. 7, pp. 1618-1625, 2007.

[29]. B. Mesgarzadeh, M. Hansson, and A. Alvandpour, "Low-Power Bufferless Resonant Clock Distribution Networks," in *IEEE International Midwest Symposium on Circuits and Systems*, pp. 960-963, 2007.

[30]. J. Kim, C.H. Ziesler, and M.C. Papaefthymiou, "Charge-Recovery Computing on Silicon," in *IEEE Transactions on Computers*, vol. 54, no. 6, pp. 651-659, 2005.

[31]. V.S. Sathe, J.C. Kao, and M.C. Papaefthymiou, "Resonant-Clock Latched-Based Design," in *IEEE Journal of Solid State Circuits*, vol. 43, no. 4, pp. 864-873, 2008.

[32]. T.H. Lee, *The Design of CMOS Radio Frequency Integrated Circuits*, Cambridge University Press, 1998, ISBN: 0-521-63922-0.

[33]. J.N. Burghartz and B. Rejaei, "On the Design of RF Spiral Inductors on Silicon," in *IEEE Transaction on Electron Devices*, vol. 50, no. 3, pp. 718-729, 2003.

[34]. C.P. Yue and S.S. Wong, "On-Chip Spiral Inductors with Patterned Ground Shields for Si-Based RF IC's," in *IEEE Journal of Solid-State Circuits*, vol. 33, no. 5, pp. 743-752, 1998.

[35]. S.-W. Yoon, S. Pinel, and J. Laskar, "A 0.35-μm CMOS 2-GHz VCO in Wafer-Level Package," in *IEEE Microwave and Wireless Components Letter*, vol. 15, no. 4, pp. 229-231, 2005.

[36]. C.L. Chua, D.K. Fork, K.V. Schuylenbergh, and J.-P. Lu, "Out-of-Plane High-Q Inductors on Low-Resistance Silicon," in *IEEE Journal of Microelectromechanical Systems*, vol. 12, no. 6, pp. 989-995, 2003.

Chapter 5

Resonant Clocking - Impact on Flip-Flops

5.1 Introduction

As discussed in Chapter 4, in order to take the largest advantage of the resonant clocking approach, the LC-tank oscillator need to be driving the clock load without any intermediate buffering. The clocked elements, like flip-flops and latches, will therefore be clocked with the output of the oscillator directly. The clock signal generated by the LC-tank oscillator will, to a first order, be a sinusoidal clock signal, which will not have the fast-edge property that a high-performance system usually relies on. Flip-flops and latches optimized and adapted to a sinusoidal clock are therefore needed.

There have been a few suggestions on alternative flip-flop and latch implementations in the resonant clocking literature. Suggestions using four-phase clocking of transmission-gate based flip-flops have shown functionality [1] but considerable delay overhead [2]. Adiabatic flip-flops where the clock provides both power supply voltage and timing information to the flip-flop have been proposed [3]. Although no additional buffers were needed and the energy in the flip-flop was recycled, this flip-flop implementation consumed a large part of the clock cycle, leaving less time for logic evaluations [3], [4]. Furthermore, the clocked transistors were connected to the clock network through the source of the transistor. Hence, introduced additional resistive loss in the LC-tank, and

therefore leads to a reduction of the tank efficiency [4]. A number of differential structures such as sense-amplifier flip-flops, and pulsed differential latches with conditional precharge have been presented in [2]. A recent proposal has been presented, where a latch based design methodology using both one-phase and two-phase clocking [4] was used in a resonant clocking environment. This has shown good power efficiency in the resonant clocking network as no buffers are inserted in between the clock driver and the latches.

In this chapter, a number of conventional flip-flops, usually designed for clock signals with fast edge rate, are investigated as potential register candidates for a resonant clocking system running in the GHz-range. As the targeted clock frequency is relatively high the edge rates for the sinusoidal clock signal, although reduced compared to conventional clocks, is still assumed to be relatively sharp. Therefore the benefit of keeping traditional flip-flop topologies potentially outweighs some of the negative effects.

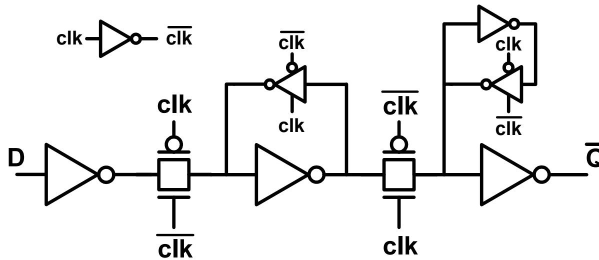


Figure 5.1: Conventional TG-MSFF with local clock-inversion.

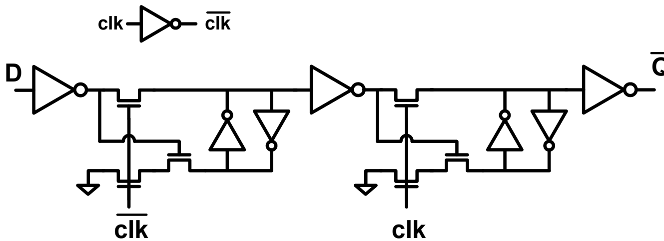


Figure 5.2: Read-port master-slave flip-flop (RP-MSFF) with local clock-inversion.

5.2 Analyzed Flip-Flop Topologies

Figure 5.1 shows the first of the studied flip-flops, which is the well-known TG-MSFF. This flip-flop requires two clock phases, and usually the second clock phase is generated locally in each flip-flop. However, the slow edges of the sinusoidal clock will lead to large short-circuit power dissipation in the local clock driver. An alternative would be to use a differential oscillator and distribute two sinusoidal clock phases. Therefore, a power-performance comparison of the TG-MSFF clocked with two sinusoidal clock phases is included in this study. Of course, clocking the flip-flops with two phases requires that both phases are distributed over the chip.

As discussed in the previous chapter, low-swing clocking technique is a potential candidate for low-power clocking. The issue with poor driving capability of a low-swing clock is sort of similar to the case with reduced edge rate in resonant clocking. For low-swing clocking there have been some

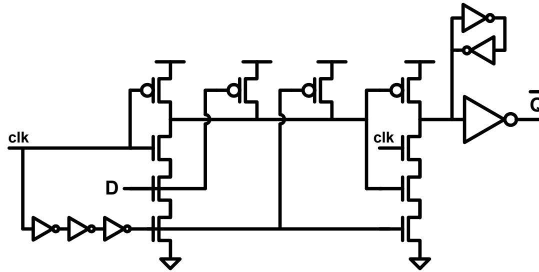


Figure 5.3: Pulsed hybrid latch flip-flop (HLFF) [7].

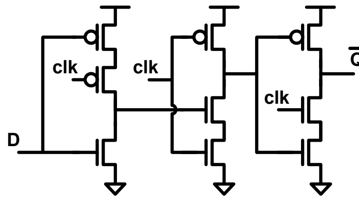


Figure 5.4: True-single phase clocked flip-flop (TSPC-FF) [8], [9].

proposals of using pass-gate based flip-flop topologies [5], [6]. An example is the read-port MSFF (RP-MSFF) shown in Figure 5.2, also included in this flip-flop study.

Pulsed latches are in general considered high-performance flip-flops. The advantage of the pulsed flip-flops is their short setup time, which results in short synchronization overhead needed for performance demanding applications. The drawback is the implicit pulse generation included in the latch, which dissipates considerable power. Furthermore, the hold time of pulsed latches are in general low as discussed in section 3.6.2, thus requiring extensive timing verification. However, as pulsed latch approaches have been commonly used in speed-critical data paths for high-performance VLSI designs, it is important to include them in this analysis in order to evaluate the impact of sinusoidal clock signals. To represent pulsed latches in this study the pulsed hybrid-latch flip-flop (HLFF) [7] is chosen. Figure 5.3 shows the schematic of the HLFF.

To remove the requirement of two clock phases, flip-flops that only require a single-phase clock, is needed. One type of flip-flops that have the single-phase property is the true single phase clock flip-flop (TSPC-FF) presented in [8], [9]. Figure 5.4 shows the implementation of the TSPC-FF. This flip-flop is known to require a sharp edge rates to function properly, thereby making it interesting to investigate how the slow edge of the sinusoidal clock impacts the flip-flop characteristics. A second type of flip-flops that only requires one clock-phase is

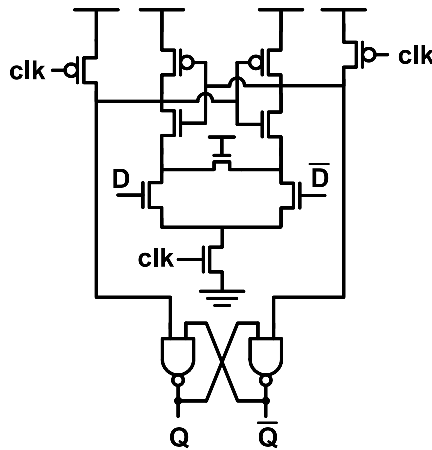


Figure 5.5: Sense Amplifier Flip-Flop (SAFF).

sense-amplifier-based flip-flops. Therefore, a conventional SAFF [10], as shown in Figure 5.5, is included in the flip-flop study.

5.3 Comparison and Discussion

5.3.1 Simulation Setup

In order to compare the power and performance impact due to sinusoidal clocks on each flip-flop, the power-performance design space for each topology is investigated. The total flip-flop power is optimized against the total flip-flop latency ($t_{d,DQ}$), and each flip-flop is simulated using a testbench, where the flip-flop output is loaded with a FO-4 inverter load, and the input buffers are designed for a FO-4 driving strength. The optimization is done using the algorithm described in section 3.3.3, which results in a power-delay curve for each flip-flop and clocking strategy, respectively. Simulations are carried out in a commercially available 130-nm CMOS technology. All simulations are performed using typical process parameters, a temperature of 110 ºC, and a power supply voltage of 1.2 V. The aim for the LC-oscillator clock strategy is to reduce power for a high performance system where a relatively high clock frequency is chosen. For a reasonable clock frequency during the comparison, a clock signal with frequency of 1 GHz is used. This gives a rise/fall time[1] roughly 10X slower for the sinusoidal clock compared to the conventional buffered clock. The clock buffers used for the conventional clock are sized so that a constant edge rate of 30 ps is obtained. The total power dissipation for each flip-flop is defined according to the expression $P_{total} = V_{dd}\,(I_{Vdd}+|I_{Data}|+|I_{Clk}|)$, where I_{Vdd} is the current drawn from the power supply to the flip-flop, I_{Data} is the current drawn by the data input driver(s) of the flip-flop, and I_{Clk} is the current drawn from the clock driver. In the power definition the current flowing into the clock node for the sinusoidally clocked system will be considered consumed. Although a resonant clock system is able to recover clock power, the efficiency of the LC-oscillator will impact the amount of the current that can be recovered. Therefore no energy recovering is considered in this chapter, unless otherwise stated, and the focus will be entirely on the effects in the flip-flop.

5.3.2 Power-Delay Comparison

Figure 5.6 shows the simulation results of the three flip-flops shown in Figure 5.1, Figure 5.2, and Figure 5.4, using both buffered and sinusoidal clocks. The additional clock phase needed for the TG-MSFF and the RP-MSFF is here

[1] Rise/fall times are here defined as the time between the 10% and 90% voltages.

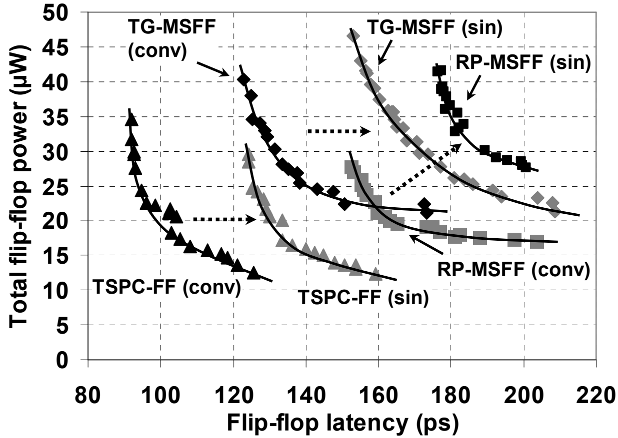


Figure 5.6: Power-delay comparison between TG-MSFF, RP-MSFF, and TSPC-FF, using both conventional and sinusoidal clocks.

generated internally. As can be seen the power-delay curve for the TG-MSFF and the TSPC-FF are shifted mostly in the delay direction as indicated by the horizontal arrows in the plot. For the TSPC-FF the delay penalty compared at the same power level (iso-power) is 27% at the low-power end (≈ 12 µW) of the curve and increases to 33% at the high-performance end of the power-delay curve (≈ 30 µW). The comparison for the TG-MSFF at equal power gives that at the low-power end of the curve (≈ 21 µW) the sinusoidal clock results in about 20% delay penalty, which increases to 30% as the delay is reduced towards the high-performance end of the design space (≈ 40 µW). For the RP-MSFF the sinusoidal clock results in a shift in both power and delay, indicating that this flip-flop suffers more due to the slow-edge clock signal.

Figure 5.7 shows the power-delay comparison between the HLFF clocked with conventional and sinusoidal clocks. The result of the TG-MSFF is included for comparison. The results for the HLFF at iso-power show a delay penalty of 42% at a power dissipation level of 80 µW, while the delay penalty is 50% at the low-power end of the curve (≈ 62 µW). Hence, the sinusoidal clock signal has a relatively large negative impact on the power-performance of the HLFF.

Figure 5.8 shows the power-delay comparison for the SAFF (Figure 5.5), with the TG-MSFF included for comparison. For the SAFF the delay penalty varies between 34%, compared at an iso-power level of 21 µW, and 18% compared at

the high-performance end (≈ 45 μW) of the curve. This is comparable to the results shown for the TG-MSFF.

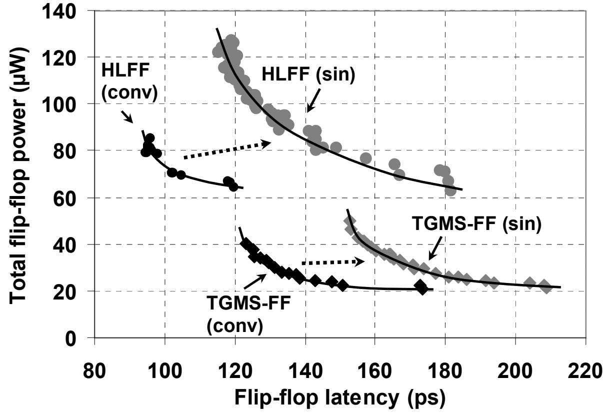


Figure 5.7: Power-delay comparison between TG-MSFF and HLFF using both conventional and sinusoidal clocks.

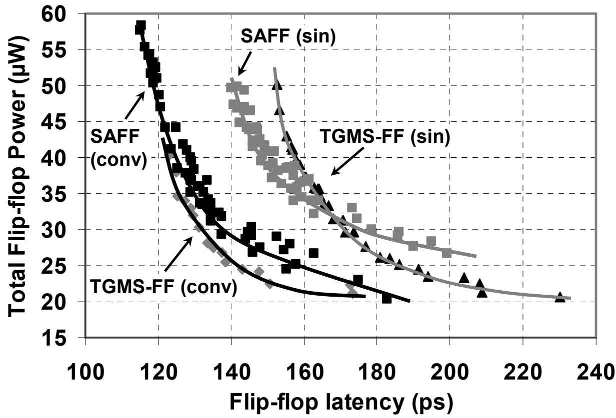


Figure 5.8: Power-delay comparison between TG-MSFF and SAFF using both conventional and sinusoidal clock signal.

TABLE 5-1: COMPARISON OF THE STUDIED FLIP-FLOP TOPOLOGIES AT ISO-DELAY

Flip-flop topology	Clocking strategy	D-Q delay ps	Hold ps	Total Power µW	Min Race ps	PDP fJ
TG-MSFF	Conv.	173.4	-21.4	21.1	120.8	3.66
	1 ph. sinus	173.7	-25.0	29.6	100.6	5.15
	2 ph. sinus	172.9	-50.8	27.5	121.4	4.76
RP-MSFF	Conv.	181.0	-31.0	17.8	125.5	3.22
	1 ph. sinus	181.3	-21.8	36. 7	118.2	6.65
TSPC-FF	Conv.	125.5	36.3	12.5	38.6	1.56
	1 ph. sinus	125.2	77.0	24.7	-41.7	3.09
HLFF	Conv.	119.7	45.8	64.3	24.5	7.70
	1 ph. sinus	118.7	48.1	111.3	11.7	13.2
SAFF	Conv.	144.1	59.3	28.7	51.7	4.13
	1 ph. sinus	146.0	29.2	41.5	50.7	6.06

Table 5-1 summarizes the simulated results and comparison at equal delay (iso-delay) for each flip-flop, respectively. The power-delay points compared at iso-delay are increased 40% and 30%, for the TG-MSFF for one-phase and two-phase sinusoidal clocking, respectively. Noticeable is that the power dissipation for the master-slave flip-flops is increasing considerably in the one-phase clock case. The reason for this is shown in the power breakout in Figure 5.9, where the power consumption in the local clock inverter (Clock inv) increases 3X for the sinusoidally clocked case due to increased short-circuit currents. If two sinusoidal clock phases are distributed to the flip-flops there is no need for a power hungry local clock buffer in the TG-MSFF flip-flop as discussed before. Figure 5.10 shows power-delay plots of the TG-MSFF in Figure 5.1, for conventional clock as well as for one-phase and two-phase sinusoidal clocks. For the high-performance design points, the delay penalty is around 30% for both one and two phase clocking. However, for low-performance design points the power-benefit from the two phase clocking is obvious. The RP-MSFF clocked with a one-phase sinusoidal clock shows a very large power penalty in Table 5-1 resulting in a 2X increase in the PDP, suggesting that this flip-flop is not a suitable choice in a resonant clock system. Table 5-1 also shows that the sinusoidal 1-GHz clock, will have limited impact on hold times and race margins for the master-slave flip-flops. This is a large advantage as the flip-flop will remain robust against hold time violations. Both the TSPC-FF and the HLFF suffer a serious hold time penalty, which reduces the race-margin considerably. For the TSPC-FF the race margin becomes negative, which creates

serious timing issues. This is because the slow edge rates of the clock increase the transparency period of the flip-flop. Considerable design efforts need to be spent on verifying the timing if the TSPC-FF or HLFF are used in a LC-oscillator clocked system.

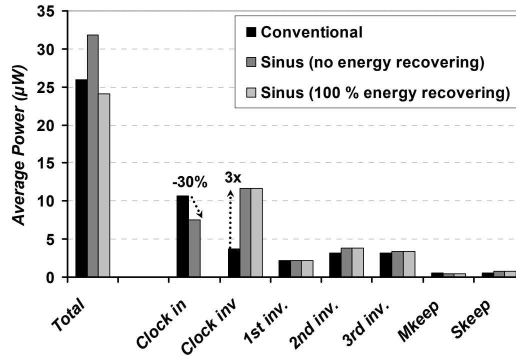


Figure 5.9: Power consumption break-up of a PowerPC flip-flop.

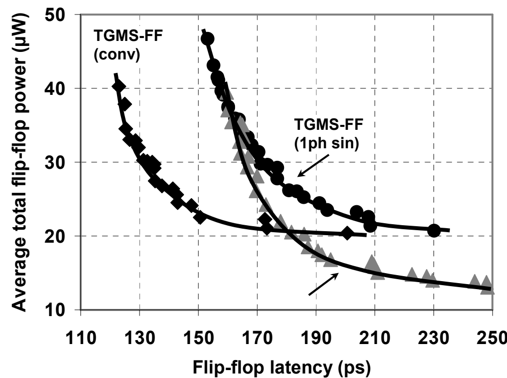


Figure 5.10: Power-delay comparison between TG-MSFF using conventional, one-phase, and two-phase sinusoidal clocking respectively.

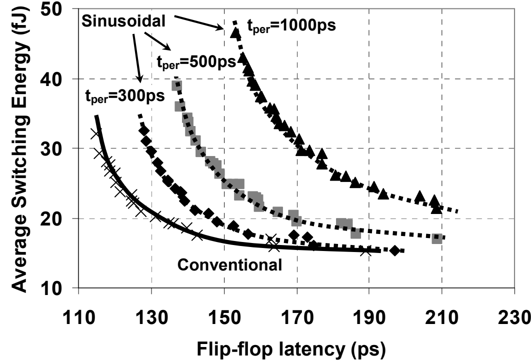


Figure 5.11: Energy-delay comparison between conventional and sinusoidally clocked TG-MSFF at different clock periods.

The SAFF behaves rather robust against the sinusoidal clock. The sinusoidal clock does not cause any large degradation of the race margins, and the power-delay penalty is comparable with the TG-MSFF clocked with two-phase. The advantage with the SAFF is of course the single clock-phase property. The power-delay-points compared at equal delay are increased 47% for the SAFF.

As a consequence of the fact that there are no buffers between the clock driver and the clock load, the edge rate of the sinusoidal clock signal will also vary when the clock frequency changes. Obviously, the performance of the flip-flops will improve when the frequency of the clock signal increases, because the rise/fall times reduce. This effect can be seen in Figure 5.11, for the TG-MSFF simulated with a conventional clock and a sine-wave clock with frequencies 1 GHz, 2 GHz, and 3.3 GHz. The result clearly shows the negative impact of the reduced edge rate at the slower clock frequencies.

5.4 Bibliography

[1]. B. Voss and M. Glesner, “A Low Power Sinusoidal Clock,” in *IEEE International Symposium on Circuits and Systems*, vol. 4, pp. 108-111, 2001.

[2]. M. Cooke, H. Mahmoodi-Meimand, K. Roy, “Energy Recovery Clocking Scheme and Flip-Flops for Ultra Low-Energy Applications,” in

Proceeding of the International Symposium on Low-Power Electronics and Design, pp. 54-59, 2003.

[3]. C.H. Zeisler, J. Kim, V.S. Sathe, and M.C. Papaefthymiou, “A 225 MHz resonant clocked ASIC chip,” in *Proceeding of the International Symposium on Low-Power Electronics and Design*, pp. 48-53, 2003.

[4]. V.S. Sathe, J.C. Kao, and M.C. Papaefthymiou, “Resonant-Clock Latched-Based Design,” in *IEEE Journal of Solid State Circuits*, vol. 43, no. 4, pp. 864-873, 2008.

[5]. D. Markovic, J. Tschanz, and V. De, “Feasability Study of Low-Swing Clocking,” in *Proceeding of the 24th International Conference on Microelectronics*, vol. 2, pp. 547-550, 2004.

[6]. R.K. Krishnamurthy, S. Hsu, M. Anders, B. Bloechel, B. CHatterjee, M. Sachdev, S. Borkar, “Dual Supply Voltage Clocking for 5GHz 130nm Integer Execution Core,” in *Symposium On VLSI Circuits Digest of Technical Papers*, pp. 128-129, 2002.

[7]. H. Partovi, R. Burd, U. Salim, F. Weber, L. DiGregorio, and D. Draper, ”Flow-Through Latch and Edge-Triggered Flip-flop Hybrid Elements,” in *Digest of Technical Papers IEEE International Solid-State Circuit Conference*, pp. 138-139, 1996.

[8]. Y. Ji-Ren, I. Karlsson, and C. Svensson, “A True Single-Phase-Clock Dynamic CMOS Circuit Technique,” in *IEEE Journal of Solid-State Circuits*, vol. SC-22, no.5, pp- 899-901, 1987.

[9]. J. Yuan and C. Svensson, “High-speed CMOS Circuits Techniques,” in *Journal of Solid-State Circuits*, vol. 26, no. 8, pp. 62-70, 1989.

[10]. D. Dobberpuhl, “The Design of a High Performance Low Power Microprocessor,” in *Proceeding of the International Symposium on Low-Power Electronics and Design*, pp. 11-16, 1996.

[11]. T. Sakurai, H. Kawaguchi, and T. Kuroda, “Low-Power CMOS Design through V_{TH} Control and Low-Swing Circuits,” in *Proceeding of the International Symposium on Low-Power Electronics and Design*, pp. 1-6, 1997.

[12]. J. Frenkil, “A multi-level approach to low-power IC design,” in *IEEE Spectrum*, pp. 54-60, Feb. 1998.

[13]. W. C. Athas, N. Tzartzanis, L. Svensson, and L. Peterson, “A Low-Power Microprocessor Based on Resonant Energy,” in *IEEE Journal of Solid-State Circuits*, vol. 32, no. 11, pp. 1693-1701, 1997.

[14]. S. C. Chan, P. J. Restle, K. L. Shepard, N. K. James, and R. L. Franch, “A 4.6GHz Resonant Global Clock Distribution Network,” in *Digest of Technical Papers IEEE International Solid-State Circuit Conference*, pp. 342-343, 2004.

[15]. A. J. Drake, K. J. Nowka, T. Y. Nguyen, J. L. Burns, and R. B. Brown, “Resonant Clocking Using Distributed Parasitic Capacitance,” in *Journal of Solid-State Circuits*, vol. 39, no. 9, pp. 1520-1528, 2004.

[16]. D. Markovic, B. Nikolic, and R.W. Brodersen, “Analysis and Design of Low-Energy Flip-Flops,” in *Proceeding of the International Symposium on Low-Power Electronics and Design*, pp. 52-55, 2001.

[17]. G. Gerosa, S. Gary, C. Dietz, D. Pham, K. Hoover, J. Alvarez, H. Sanches, P. Ippolito, T. Ngo, S. Litch, J. Eno, J. Golab, N. Vanderschaaf, and J. Kahle, “A 2.2W, 80 MHz Superscalar RISC Microprocessor,” in *Journal of Solid-State Circuits*, vol. 29, no. 12, pp. 1440-1454, 1994.

Chapter 6

Chip Implementations and Evaluation

6.1 Introduction

The discussion in section 4.4 shows that resonant clocking has large potential for considerable power reductions in the clock distribution network. However, in order to prove and evaluate the concept, and to identify possibilities and challenges two experimental chips are manufactured and evaluated. This chapter will discuss the implementation and evaluations of one experimental test chip incorporating both resonant clock drivers and conventional clock drivers to enable true and accurate power comparisons. The test chip also incorporates the possibility of using off-chip high-Q inductors to evaluate the impact of the Q-value of the inductors, compared to the Q-value of the clock network itself. A second test chip is also fabricated, which enables evaluation of some simple tuning and gating possibilities using the proposed resonant clocking technique.

6.2 Resonant Clocking Evaluation Test Chip

The proposed resonant clocking technique, discussed in section 4.4.3, is incorporated in an experimental test chip, which is fabricated in a 130-nm, multiple-V_{th} CMOS process with six Cu-metal layers. The implemented test chip includes comparable designs for proposed resonant LC-tank clocking and conventional clocking. The following section will describe the implemented

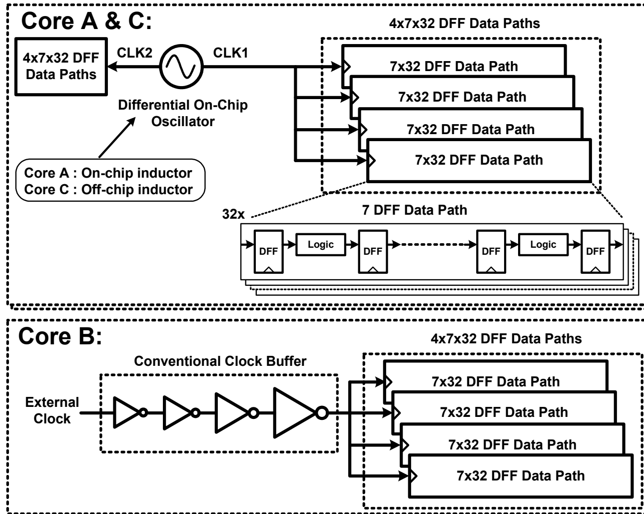
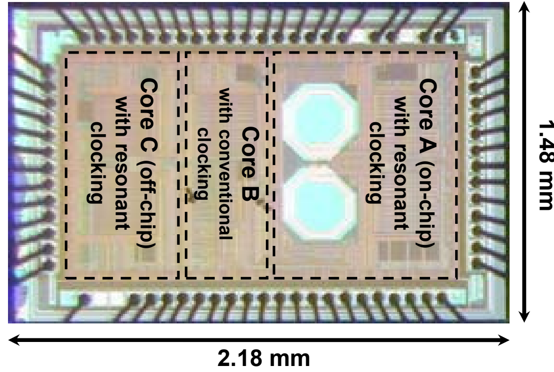


Figure 6.1: Block level chip organization.

parts of the chip including oscillator, clock load, evaluation circuitry, and flip-flops.

6.2.1 Top Level Chip Organization

Figure 6.1 shows the block level organization of the test-chip. The chip consists of three separate cores, which are used for different evaluations of the resonant clocking technique. Core-A incorporates the differential resonant clock driver, which drives two synchronous data path blocks, one for each clock phase. The clock load consists of the flip-flops in the synchronous data path block, and these flip-flops are driven directly from the oscillator without intermediate buffers. The inductor in the LC-tank oscillator for Core-A is implemented as a fully integrated on-chip inductance. The conventional clock buffer in Core-B is implemented to enable true and accurate power comparisons between the two different clocking techniques. Identical clock load is used for the conventional driver in the form of the previously mentioned synchronous data path block. Finally, Core-C consists of the same oscillator and clock load as in Core-A. However, in Core-C the integrated inductance for the LC-tank oscillator is removed, and the differential clock outputs are instead connected to two bondpads to enable attachment of an off-chip high-Q inductance during the chip



Technology	130-nm CMOS
Power supply voltage	1.2 V
Power supply regions	26 separate
Interconnect layers	6 Cu layers
Total chip area	3.22 mm^2
Chip core area	2.08 mm^2
Number of pads	76 (pitch 80 μm)
Amount of transistors	~ 200 000
Total no. of flip-flops	~ 6000

Figure 6.2: Chip photograph of resonant clocking experimental chip in 130-nm CMOS technology and chip data.

measurement. Figure 6.2 shows the micrograph and the chip data for the fabricated test chip displaying the three separate cores. For fair and accurate power comparisons, 26 separate power supply regions with individual bondpads are used.

6.2.2 Conventional Clock Drivers

Figure 6.3 shows the implementation of the conventional clock driver used in Core-B. The conventional clock driver is realized as a four-stage inverter chain with a tapering factor of three. The conventional driver is designed to achieve an output clock signal with a fast edge rate, comparable to state-of-the-art high-performance microprocessors. The clock buffer drives one data path block providing a single phase clock to the implemented flip-flops, while the second phase is generated locally in the flip-flops.

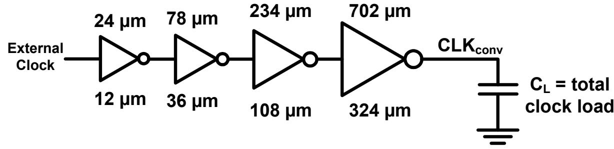


Figure 6.3: Conventional 4-stage clock-buffer chain.

6.2.3 Implemented Oscillator Topology

The resonant clock driver is implemented as a complementary differential oscillator topology, as shown in Figure 6.4. The complementary structure is known to be able to provide larger gain for the same power dissipation compared to a NMOS-only solution [1]. Furthermore, the symmetry of the differential design has been shown to limit the up conversion of *1/f* noise, hence improving the phase noise performance of the oscillator [1] - [3]. The cross-coupled inverters use PMOS and NMOS devices with minimum channel lengths and transistor widths of 400 µm and 200 µm, respectively. The current source, implemented as a 400 µm wide NMOS transistor, is utilized primarily to control the voltage swing and secondly to control the oscillation frequency of the oscillator. The biasing voltage for the current source is controlled by the external bias voltage V_{bias} via the current mirror circuitry, as shown in Figure 6.4.

6.2.4 Clock Distribution Network

According to equation (4.5) on page 65, for large clock power saving, the total Q-value of the LC-tank should be as high as possible, where the LC-tank Q-value equals the parallel equivalent of the Q-value of the inductance and the capacitance ($Q_L \mid\mid Q_C$). The Q-value of the capacitance is mainly determined by the losses in the wires distributing the clock to the flip-flops and latches. Therefore, a dense low-resistive wire grid in the top-metal layer (metal 6) is used to distribute the clock signals across the data path blocks. The local clocks are tapped from the grid to the flip-flops through metal via contacts. Seen from the LC-tank, these via contacts are connected in parallel, resulting in tolerable losses for the distributed capacitance. A wide metal-6 wire (width ≈ 30 µm) is used to distribute the output clock from the LC-tank oscillator, or the conventional clock buffer, to a 14x14-wire clock grid built up from 5 µm wide

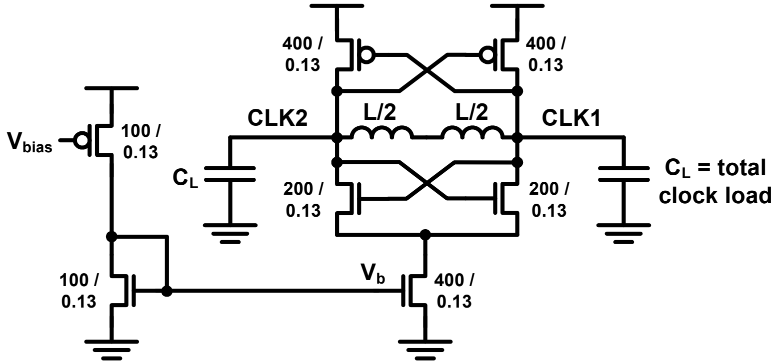


Figure 6.4: Differential complementary oscillator used as resonant clock-driver.

metal-6 wires. The implemented clock grid yields an extracted worst-case parasitic resistance of around one Ohm from one corner of the grid to the other corner.

6.2.5 Inductor Implementation

A key component required for the LC-tank oscillator is of course an inductor with sufficiently high Q-value. The 1.2-nH on-chip inductor, used in Core-A of the implemented test chip, is formed by using two serially connected octagonal single-turn spiral inductors. Using two serially connected inductors instead of one single inductor has been previously reported to give a symmetrical oscillator layout improving the phase noise performance [1] - [3]. Each inductor is realized as a single-turn spiral in parallel 30 µm wide metal-5 and metal-6 wires, with an outer diameter of 300 µm. The total inductance value per inductor is estimated to 600 pH. A layout view of the implemented oscillator is shown in Figure 6.5, and as can be seen the inductors occupy a considerable amount of the total oscillator area. However, the large area consumption is traded-off in this experiment in order to increase the rate of success of the entire clock experiment. Each output phase of the oscillator drives the common data path block, which consists of the clock network and clocked registers.

To enable comparison between inductors with different Q-values, Core-C incorporates the possibility to utilize external inductors in the resonant LC-tank oscillator. The off-chip inductor in Core-C is formed by using externally attached gold bondwires, and is implemented in two separate versions. The first

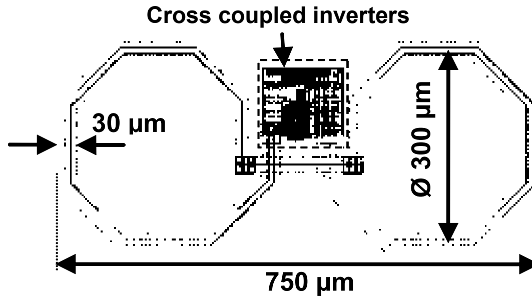


Figure 6.5: Layout of oscillator with 2x0.6-nH on-chip single-turn coils.

version uses a bondwire which just stretches across the chip from edge to edge. This gives a bondwire length of approximately 1.5 mm and a diameter of 33 μm. The other inductor version is implemented with a bondwire roughly 1.9X longer. Obviously, using bondwires as inductors is only for test purpose enabling a first order study of the impact on the resonator performance and efficiency, by using low-Q on-chip inductors compared to high-Q off-chip inductors.

6.2.6 Implemented Flip-Flops

The experimental chip incorporates conventional master-slave flip-flops, as shown in Figure 6.6. The choice of using master-slave flip-flops is done in order to increase the chance of success for the chip, because of the race robustness of these flip-flops, and due to its relatively low performance impact from sinusoidal clocks, as discussed in the previous chapter.

The chosen master-slave flip-flop requires two clock phases, and in the implemented chip the second clock phase is generated locally in each individual flip-flop. This is despite the fact that the differential LC-tank oscillator provides both clock phases. The reason is to simplify the clock routing to the data paths via the clock grid. Clearly, this does not affect the clock power comparisons, as the main clock distribution networks in both cores have exactly equal clock loads. However, it does lead to increased power dissipation in the flip-flops clocked with the oscillator, which is expected as the short-circuit current through the local clock inverter will increase. Nevertheless, one of the experiments in the test chip is to determine the impact of the resonant clock strategy on conventional flip-flops. Moreover, identical transistor sizing is used in the flip-flops for both conventional clocking and resonant clocking, which could result

in that the additional power-performance tradeoff is larger than it needs to be, compared to if the flip-flops would have been optimized for the specific clocking strategy, as described in Chapter 5. However, the chosen strategy is used in order to obtain exactly the same clock load for the conventional and the resonant clocking in the experimental chip. Based on simulations the delay tradeoff at iso-power for the specific flip-flop and transistor sizing is 23%.

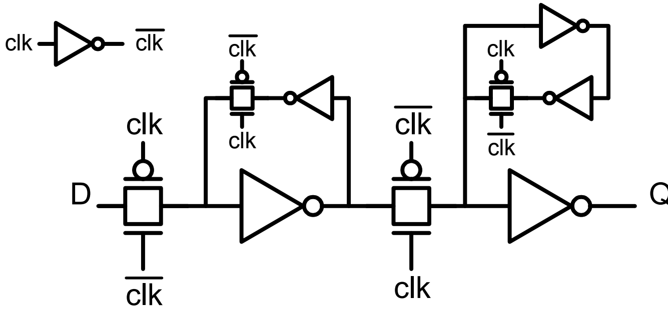


Figure 6.6: Conventional TGMS flip-flop with interrupted keepers.

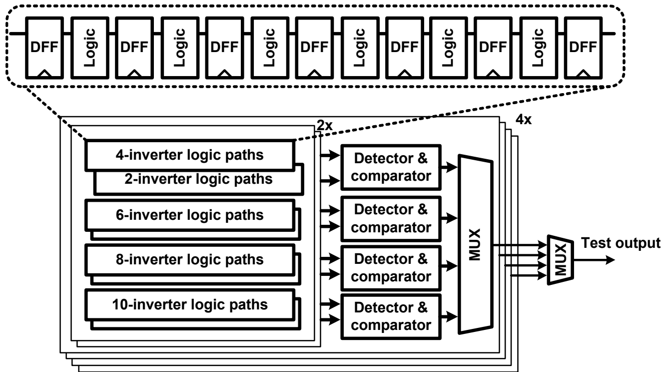


Figure 6.7: Detailed block level schematic of the data path block.

6.2.7 Organization of the Data-Path Blocks

The data path block is organized in four sub-blocks, where each sub-block consists of 32 pipeline paths that include seven flip-flops per path. This yields a total of 896 flip-flops per data path block. The clock load capacitance for all flip-flops in the 4x32 pipeline paths including the clock distribution network (metal grid) is extracted to 9 pF. Sixteen out of the 32 pipelines in the sub-blocks have a constant logic depth of two FO-2 inverters between the flip-flops. The remaining sixteen pipeline paths are realized with increasing logic depth from four FO-2 inverters up to as much as ten FO-2 inverters. Figure 6.7 shows a detailed block level schematic of the implemented synchronous data path blocks.

In order to verify the functionality of the pipeline path a simple evaluation circuitry is added. Pipeline paths with logic depth between four and ten FO-2 inverters, are paired together with pipeline paths with logic depth of two FO-2 inverters, as shown in Figure 6.8. The same input is feed into the two paths, and the outputs are compared using an XOR gate. The result is finally captured by an SR latch. If the outputs are not identical the output of the SR-latch will go high, hence indicating an error. The proposed validation circuitry is able to capture setup time violations in the more critical pipeline stages with a granularity of two FO-2 inverter delays. The test circuit relies on that no hold time violations occur. However, the good internal race robustness of the master-slave flip-flop structure minimizes the risk of hold time violations in the implemented test chip. With the implemented test circuitry it is possible to monitor the flip-flop functionality online during the entire measurement.

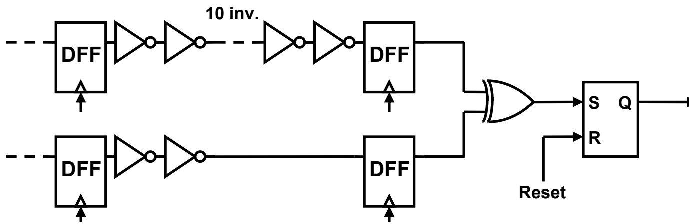


Figure 6.8: Test circuitry functionality evaluation of the pipeline paths.

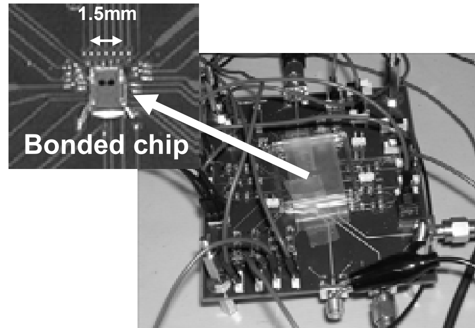


Figure 6.9: Test chip during measurements with chip directly bonded to the PCB.

6.3 Power Measurement Results

This section presents the measured power comparisons of the experimental 130-nm CMOS test chip. The conventional core is compared both to the resonant clocking technique using on-chip inductor, and off-chip inductor implementations of the oscillator. The experimental test chip is bonded directly to a PCB to enable high-performance measurements with minimal interference from off-chip parasitic components. Three different PCB versions are utilized corresponding to the three different cores on the die. This is done in order to be able to optimize the PCB footprint, so that the bondwire length can be minimized, thus reducing parasitic effects. Figure 6.9 shows one of the PCBs during measurements and the bonded chip.

6.3.1 On-Chip Resonant Core Power Comparison

Figure 6.10 shows the measured power saving in the resonant clocking Core-A compared to the conventional clocking Core-B, at a clock frequency of 1.56 GHz across a data activity from 0% to 80%, using a power supply of 1.2 V. The most important result is that the core with the resonant clocking dissipates on average 57% less clock power, which indicates a total LC-tank Q-value higher than 3.7. Total chip power saving for this test chip is between 14% and 29% depending on the input data activity. Figure 6.11 shows the measured power break-up at a data activity of 30% and at a clock frequency of 1.56 GHz. The relatively slower edge rate of the sinusoidal clock, delivered from the resonant LC-tank clock driver, compared to the conventional clock increases the

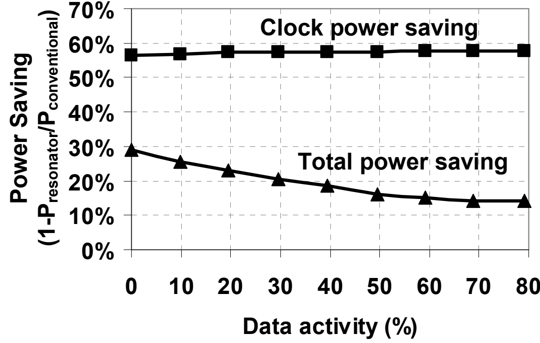


Figure 6.10: Power saving using proposed resonant clocking technique compared to a conventional buffer-driven clock driver.

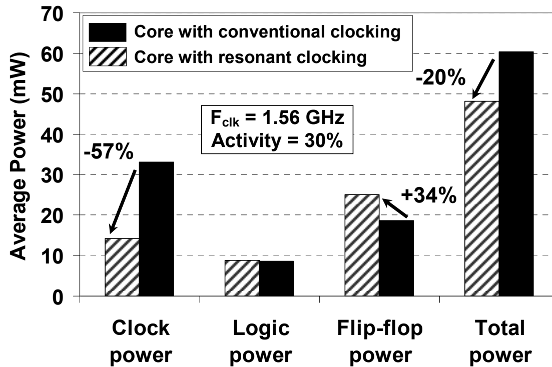


Figure 6.11: Power comparison between resonant clocking using on-chip inductance oscillator and conventional buffer-driven clocking

power consumption in the flip-flops with 34%. This is mainly due to increased short-circuit currents in the local clock inverter in the flip-flops, as is expected from the analysis discussed in Chapter 5. This suggests that the resonant clocking implementation would benefit from more customized flip-flops or by using two-phase clocking to mitigate the increase in power. Despite the higher

power consumption in the data paths, the 57% clock power saving still results in a total power reduction of 20% compared to the conventional core implemented on the same chip (Core-B). This proves the feasibility of the resonator clock strategy as a technique to substantially reduce the total chip clock power, even when utilizing a conventional logic design approach.

6.3.2 Influence of Inductor Q-value

The comparison between the three resonant clocking cores, using both the on-chip inductance oscillator and the two off-chip oscillators, is shown in Figure 6.12. All measurements are done for a nominal power supply voltage. The measured clock frequencies are 1.56 GHz for Core-A with the on-chip inductor, 1.08 GHz for Core-C with the longer off-chip inductor, and 1.76 GHz for the same core with the shorter off-chip inductor. Core-B with the conventional clock distribution is clocked by an external signal generator, where the clock frequency is matched to that of the other cores during the one-to-one power comparisons. Furthermore, as in the case for the conventional and resonant comparison described earlier, the flip-flops and data path functionality is continuously monitored during the measurements. Figure 6.12 summarizes the measured energy breakup per cycle at 30% data activity for the entire resonant clocking test chip including the three different resonate cases compared to the conventional clock buffer. The 1.56-GHz LC-resonator with on-chip inductance results in about 57% lower clock power and 20% lower total core power. The short bondwire version of Core-C oscillates at the frequency of 1.76 GHz

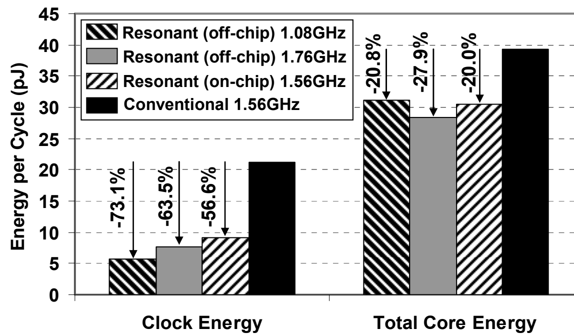


Figure 6.12: Detailed energy per cycle comparison between the resonant clocking technique using both on-chip and off-chip inductors.

enabling a 64% reduction of the clock power and a 28% reduction of the total power. Finally the long bondwire version of Core-C achieves a 73% reduction of the clock power at the oscillation frequency of 1.08 GHz, which resulted in a total power reduction of 21% on the implemented test chip. This clearly demonstrates the excellent potential of such a charge-recovery clocking for significant clock and total chip power savings.

As indicated by the measurements the power efficiency of the resonant clocking technique benefits from the relatively higher Q-value of the off-chip inductors. However, the relative power numbers and the effective Q-values of the LC tanks suggest that the off-chip inductances with higher Q_L (>10) do not increase the Q-value of the tank proportionally. The reason is that the quality factor of the tank equals to $Q_L||Q_C$, as discussed in section 4.6, and Q_C is limited by the resistive loss in the clock distribution network. To gain more insight into this fact, a numerical example can be used. If the network resistance is about 1 Ω, for 1.5-GHz oscillation frequency in a resonant clocking network with a total clock load of 18 pF, the quality factor of the capacitor will be about six. It means that utilizing an inductor with quality factor of twelve will result in a tank quality factor of only about four. Thus highly energy-efficient LC tank clock resonators require not only high-Q inductors but also low-loss on-chip clock network.

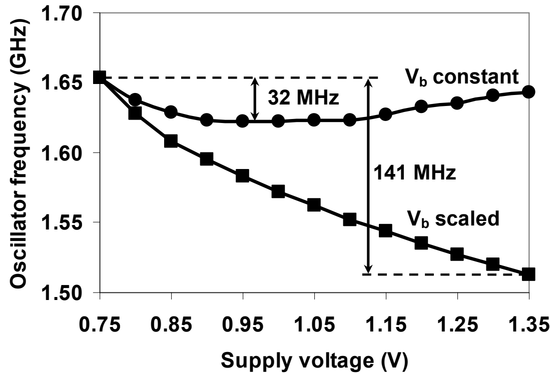


Figure 6.13: Measured oscillator frequency over a range of power-supply voltages.

6.4 Clock Signal Integrity

6.4.1 Oscillator Power Supply Sensitivity

The oscillation frequency characteristics for a V_{dd} range of 0.75-1.35 V is shown in Figure 6.13, using both constant and scaled bias voltage. Here the oscillator current source bias voltage (V_b in Figure 6.4) is held constant by adjusting the bias voltage to the PMOS transistor connected to the current mirror (V_{bias} in Figure 6.4). Therefore the oscillation current should ideally be held constant. For the scaled case, the PMOS bias voltage is held at ground, giving a constant maximum current for the given power-supply value. For a constant bias voltage, the oscillation frequency remains close to constant across the V_{dd} range of 0.75-1.35 V. With a scaled bias voltage, the oscillation frequency is reduced from the maximum of 1.65 GHz at 0.75 V, down to 1.52 GHz for a power supply voltage of 1.35 V. An interesting observation is that for the power supply voltage of 1.35 V, the frequency can be tuned around 130 MHz by adjusting V_{bias}. Although, by changing the bias to the current source in the oscillator, the clock amplitude is also changed, which can impact the functionality of the data paths.

6.4.2 Data Dependent Phase Noise

Since the oscillator directly drives the clocked devices without intermediate buffers, there is no decoupling between the clock generator and the clock load. Hence, any change of the clock load in the clocked devices due to the data change, will be directly transferred to the clock generator. Special attention is

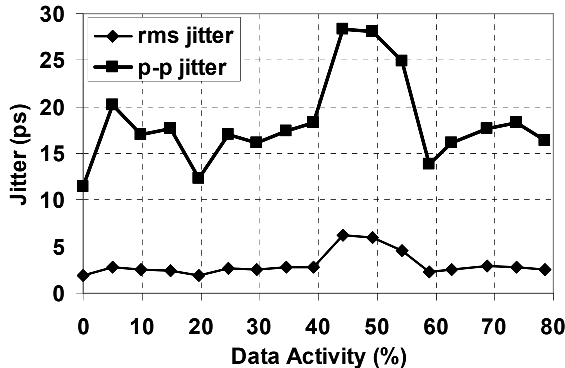


Figure 6.14: Oscillator clock jitter vs. data activity.

therefore needed to determine the impact of data activity on the clock signal integrity.

In order to evaluate the impact of the data activity in the flip-flops, the jitter is measured and observed for one of the clock phases of the resonant clocking cores. The clock signal is buffered in a source-follower before being fed off-chip through a separate bondpad. The data signal applied to all of the flip-flops is applied externally through a data buffer, and all flip-flop inputs are changing at the same time and at the same direction. This results in the worst-case jitter scenario, and Figure 6.14 shows the peak-to-peak and RMS jitter measurements for data activities from 0% to 80%, with an average clock frequency of 1.56 GHz. Note that the data frequency equal to half the clock frequency (α = 0.5) is here defined as 100% data activity. As can be seen in Figure 6.14, at a data activity of 45%, when the data frequency is close to one fourth of the oscillation frequency, the worst-case peak-to-peak jitter is measured to 28.4 ps and the RMS jitter to 6.3 ps. This is identified to be caused by (i) variation of the capacitive loading of the clock network, due to data dependent changes of the gate capacitance of the transmission-gates in the flip-flops, and more importantly, (ii) due to mixing between the data input signal and the clock signal due to the direct capacitive coupling. This mixing results in an up-conversion of several frequency tones, including the fourth harmonic of the data signal. At sertain data frequencies some of these tones appear close to fundamental oscillation tone of the clock driver. The combination of these tones increases the noise power rapidly, resulting in so called jitter peaking in the generated clock [6].

$$L_{tot} = \frac{{\omega_L}^2 \cdot \cos^2\theta}{(\Delta\omega)^2 + {\omega_L}^2 \cdot \cos^2\theta} L_{ext} + \frac{(\Delta\omega)^2}{(\Delta\omega)^2 + {\omega_L}^2 \cdot \cos^2\theta} L_{free} \tag{6.1}$$

6.4.3 Implemented Jitter Suppression Technique

Clearly, 28.4 ps of the clock period in peak-to-peak jitter at 1.56 GHz will result in a direct performance penalty for the synchronous system according to the discussion in section 3.5. To mitigate some of the data dependent jitter, the implemented oscillator incorporates an optional jitter suppression technique based on injection locking. Although unintended injection locking can cause severe problems in certain circuit applications, this phenomenon can be utilized in proper way for different purposes [4] - [7]. The impact of injection locking in phase noise reduction in on-chip applications have previously been studied [5], [7], and it can be shown that if the oscillator is injection-locked to a source

with phase noise of L_{ext} the total phase noise after injection locking (L_{tot}) can be calculated by equation (6.1), where θ is the phase difference between the output signal and the injection source after lock, $\Delta\omega$ is the frequency offset, and ω_L is one-sided lock range [6]. According to equation (6.1) the phase noise over the lock range is dominated by the externally injected phase noise.

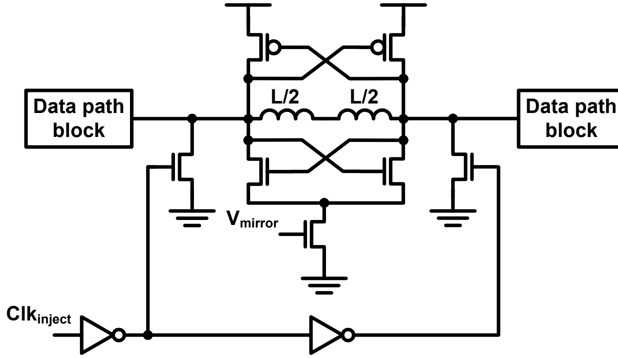


Figure 6.15: Schematic of resonant clock driver with injection-locking capability.

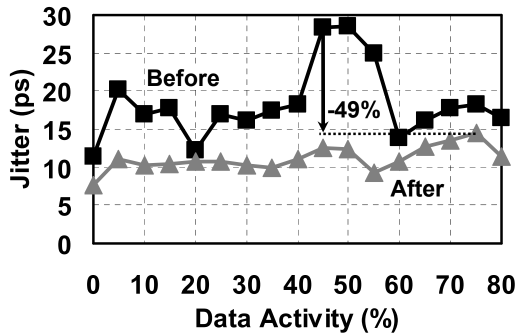


Figure 6.16: Peak-to-peak jitter before and after applying the injection-locking technique.

Figure 6.15 shows the two additional NMOS transistors that are incorporated in the oscillator implementation. A low-phase-noise external signal (with 7-ps peak-to-peak and 1-ps RMS jitter) is injected into the oscillator using the two NMOS pull-down transistors. Utilizing injection locking, the worst-case peak-to-peak jitter is reduced about 50%, from 28.4 ps down to 14.5 ps as shown in Figure 6.16. The worst-case RMS jitter is reduced from 6.2 ps down to 2.0 ps.

6.5 Frequency Tunability

One of the main challenges to solve with the resonant clocking technique is the tunability issue. Even if the clock load and the inductance can be extracted with high accuracy, variations in the process will still lead to shifts in the capacitance and inductance values. This will cause the natural oscillation frequency of the resonant clock driver to drift away from the intended target frequency. Moreover, the resonant clock driver works well for high clock frequencies as was shown in section 6.3. But if the clock frequency is reduced to a low frequency, which is commonly used during test of the ICs, then the robustness, power dissipation, and performance of the flip-flops would be seriously degraded. Hence, in order to be feasible, the resonant clock driver requires tunability in order to compensate for process spread, and the ability to be run in a low-frequency test mode, which should not impact the clocked elements drastically.

6.5.1 Tunability Using Injection Locking

The injection locking technique discussed in the previous section does not only enable a low-jitter clock signal. It also provides some frequency tunability of the resonant clocking driver. Using the 130-nm test chip the injected signal frequency is moved away from the natural oscillation frequency of 1.56 GHz (Core-A) until the oscillator goes out of lock. The tuning range utilizing the injection locking signal is found to be 400 MHz from a lower frequency of 1.3 GHz to the highest frequency of 1.7 GHz.

6.5.2 Capacitive Tuning on a Oscillator Test Chip

A common technique to tune oscillators, when used for instance as VCOs in phase-locked loops, is to use switchable capacitors or varactors to change the capacitance in the LC-tank resonant circuit [1] - [3]. This technique can also be incorporated into the resonant clock driver to provide additional tuning capability. The first implemented resonant test chip did not incorporate any capacitive tuning elements, so in order to test the tunability using digitally

controlled capacitive elements a simple test circuit was fabricated in a 7-metal layer 90-nm CMOS process. Figure 6.17 shows the chip photo of the oscillator test chip, and Figure 6.18 shows the implemented tunable oscillator used in the tunability test. The structure is the same as the one utilized in the previously presented 130-nm resonant clocking test chip. The two capacitors added at each output are switchable through transmission-gates, which has a size of 50 µm for both the NMOS and PMOS transistor. The capacitance value C_0 is equal to 1.053 pF, while C_1 is 2X larger. The digital tuning thus enables four discrete capacitive load settings.

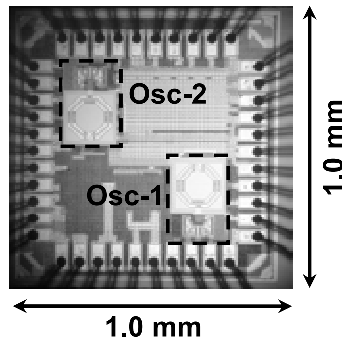


Figure 6.17: Chip photo of the reconfigurable oscillator test chip in 90-nm CMOS.

6.5.3 Switchable Inductance

As a part of the test circuit the inductance in the LC-oscillator is implemented with two NMOS-switches, which enables a deactivation of the inductance. This could opportunistically be utilized in order to drive the oscillator with a low-frequency signal whenever the clocked system is tested or goes into a low-speed mode. The NMOS switches are implemented using 1500-µm wide transistors to reduce the additional resistance added in series with the inductance. As a reference design the same oscillator as shown in Figure 6.18 is implemented but with the two NMOS-switches on the inductor path bypassed with wide metal interconnects. This enables a fair comparison of the impact on the power efficiency, between using a switchable inductance and non-switchable inductance for the LC-tank oscillator.

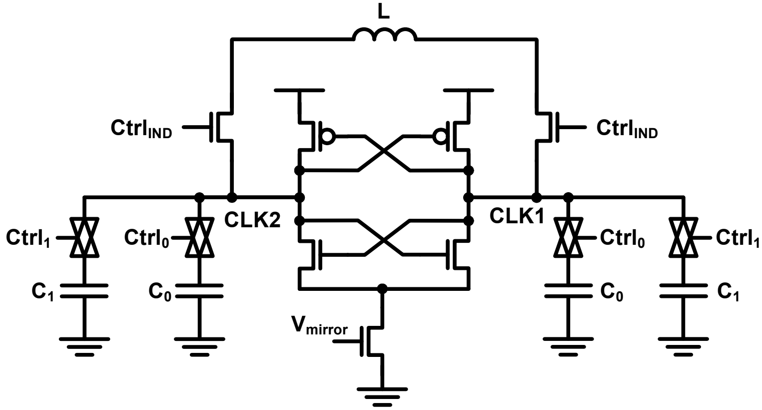


Figure 6.18: Oscillator with digitally tunable capacitance and switchable inductance.

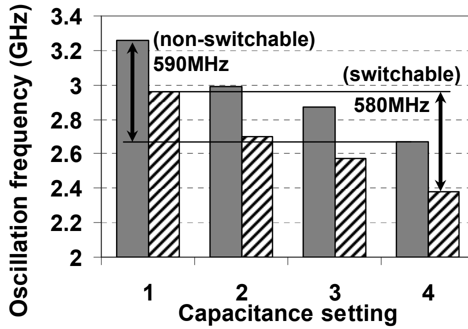


Figure 6.19: Measured oscillation frequency using different capacitance settings and using switchable inductor.

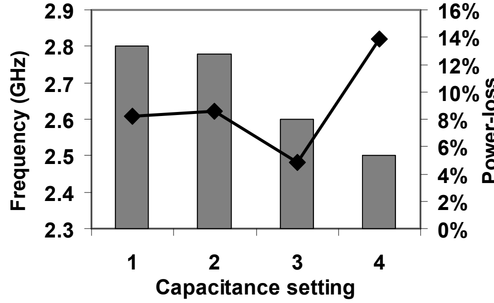


Figure 6.20: Forced oscillation frequency and power loss due to switched inductor.

6.5.4 Chip Measurement Results

Figure 6.19 shows the measured free-running oscillation frequency for the two oscillator implementations. The frequency tuning-range for the two oscillators is around 22% or almost 600 MHz. As can be seen the oscillator implemented with the switchable inductance results in a lower free-running oscillation frequency, which is expected due to the increased series resistance in the inductive path reducing the oscillation frequency according to

$$\omega_0 = \sqrt{\frac{1}{LC} - \frac{R_L^2}{L^2}}\,. \tag{6.2}$$

Obviously, as the resistances added in the inductive paths are causing more parasitic resistance in the tank, the switches will result in a power penalty for the resonant clock driver. Figure 6.20 shows measurements of the power loss from using the implemented switchable inductor compared to the bypassed implementation. Here the oscillators are injection locked to be able to compare both implementations at the same frequency. As can be seen the switches results in a power loss between 5% and 14%. In order to reduce the power loss, wider switch transistors or large transmission-gates should be utilized. However, the switches will add additional capacitance, which need to be considered.

6.6 Bibliography

[1]. R.L. Bunch and S. Raman, "Large-Signal Analysis of MOS Varactors in CMOS $-G_m$ LC VCOs," in *IEEE Journal of Solid-State Circuits*, vol. 38, no. 8, pp. 1325-1332, 2003.

[2]. A. Hajimiri and T.H. Lee, "Design Issues in CMOS Differential LC Oscillators," in *IEEE Journal of Solid-State Circuits*, vol. 34, no. 5, pp. 717-724, 1999.

[3]. D. Ham and A. Hajimiri, "Concepts and Methods in Optimization of Integrated LC VCOs," in *IEEE Journal of Solid-State Circuits*, vol. 36, no. 6, pp. 896-909, 2001.

[4]. H. R. Rategh and T. H. Lee, "Superharmonic injection-locked frequency dividers," in *IEEE Journal of Solid-State Circuits*, vol. 34, pp. 813–821, June 1999.

[5]. B. Mesgarzadeh and A. Alvandpour, "A study of injection locking in ring oscillators," in *Proceedings of the IEEE International Symposium on Circuits and Systems*, vol. 2, pp. 5465–5468, May 2005.

[6]. B. Mesgarzadeh, M. Hansson, and A. Alvandpour, "Jitter Characteristic in Charge Recovery Resonant Clock Distribution," in *IEEE Journal of Solid-State Circuits*, vol. 42, no. 7, pp. 1618-1625, 2007.

[7]. B. Razavi, "A study of injection locking and pulling in oscillators," in *IEEE Journal of Solid-State Circuits*, vol. 39, pp. 1415–1424, Sept. 2004.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

7.1.1 Low-Power Resonant Clocking

A low-power resonant clocking technique is presented, which is aimed at high-performance, multi-GHz VLSI systems [1] - [4]. Compared to conventional clocking solution based on rigid buffer driven networks, which offers small possibilities for reduced power at high-frequency operation, the resonant clocking technique enables substantial power reduction. The proposed clocking technique relies on energy recovering in the clock network enabled by using an LC-tank resonant circuit, where the energy is oscillated between the clock capacitance in the clock network, and the inductance in the oscillator. Power analysis and comparison between the conventional buffer driven clock distribution and the proposed LC-tank-based methodology is presented.

Furthermore, a successful chip implementation of the proposed resonant clocking approach in a 130-nm CMOS technology is presented, where a fully integrated differential LC-tank oscillator directly drives 2x896 flip-flops without intermediate buffers. The experimental chip enables proof-of-concept and accurate power comparisons. The resonant clocking technique is also shown to be fully functional at multi-GHz clock frequencies using fully integrated on-chip spiral inductors enabling 57% clock power reduction at a clock frequency of 1.56 GHz. By utilizing experimental off-chip bondwire-based inductors,

operational frequencies up to 1.76 GHz and up to 73% clock power reduction are presented.

An injection-locking-based circuit technique is proposed in order to mitigate some of the data dependent jitter coming from the resonant oscillator when driving the clock load without intermediate buffers. The proposed technique results in 50% reduction of the data dependent jitter from a worst-case peak-to-peak jitter of 28.4 ps down to 14.5 ps. The implemented jitter suppression technique also enables 400 MHz of frequency tuning range around the natural oscillation frequency of 1.56 GHz. A second tuning technique based on switchable capacitive loads is also studied on an experimental oscillator chip in 90-nm CMOS. The capacitive tuning enables 22% tuning range or 600 MHz around a center frequency of 2.9 GHz. In order to incorporate low-frequency testability in future resonant clocking systems, a switchable inductance is briefly studied, which shows that the energy loss in the switch transistors requires careful design considerations.

7.1.2 Flip-Flop Behavior in Resonant Clocking Systems

A performance and power study, based on power-delay simulations comparing five flip-flops, is presented [5]. The flip-flops suffer a general delay penalty when a sinusoidal GHz-rate clock signal is used compared to a conventional clock signal. As a consequence of the lower edge rate on the sinusoidal clock, hold times are reduced for all studied flip-flop topologies, which degrades the race immunities of the flip-flops. However, conventional TG-MSFF remain robust against hold time violations even with the edge rates degraded by 10X, displaying a 20-30% delay penalty at iso-power compared to a conventionally clocked version of the same flip-flop.

Moreover, the generation of any additional clock phase locally in the flip-flops consumes considerable power due to short-circuit currents in the local clock inverter. Given the large amount of flip-flops in modern digital systems a more power-efficient alternative is to distribute both clock phases from the differential LC-tank oscillator [1] - [4], and in that way remove the power in the local clock buffer. Simulations show that the performance penalty, for a conventional TG-MSFF, is just marginally degraded using two-phase sinusoidal clocks compared to a one-phase sinusoidal clock, while the power dissipation is reduced considerably.

To mitigate the problem with requiring multiple clock phases, single-phase flip-flops such as SAFFs and TSPC-FFs are analyzed. High-performance pulse-latched flip-flops are also analyzed. SAFFs are shown to enable robust behavior even when clocked with a GHz-range sinusoidal clock, thus providing an attractive alternative for flip-flops in resonant clocking systems. High-

performance flip-flops such as the pulsed HLFF and dynamic TSPC-FF suffer from degraded hold times and race margins when clocked with slow-edge clocks. Hence, they require extensive timing verification due to the degraded edge rate and are not suitable flip-flops in a resonant clocking system.

7.2 Future Work

7.2.1 Low-Power Resonant Clocking

The proposed resonant clocking technique shows substantial clock power reduction capabilities, making it a very promising technique to reduce the clock power. Nevertheless, a number of issues need to be address in order to make the proposed technique feasible, and taking it from research to industrial production in high-performance VLSI designs.

The presented 130-nm test chip included a small data path region using very simple dummy logic. An interesting continuation to the resonant clocking technique would be to implement more advanced computing blocks, which are clocked by the resonant clock driver. Moreover, the question on how large clock regions that can be driven by a single oscillator still require further analysis. One solution could be to divide the synchronous circuit into a number of clock sub-regions, and incorporate several oscillators, which are driving one sub-region each. This requires methods to tune the individual oscillators to the same frequencies, and techniques to be able to frequency-lock the oscillators together. Mesochronous communication interfaces could then be used for the data transfer between the different sub-regions.

A related field that requires further research is how to integrate clock gating, when using the proposed resonant clocking technique. The gate transistors incorporated in the 90-nm test chip provide a solution, where the oscillator can be turned off, or driven by a conventional clock driver in order to enable low-frequency operation using a more conventional clocking approach. However, it requires large switch transistors in order to minimize the resistive power losses in the gates. This also touches on another issue that requires further study, and that is how to incorporate low-frequency testing.

Robust and repeatable design methodologies are also needed in order to integrate the resonant clocking technique in a standard VLSI design flows. This includes better techniques for modeling of the required inductors. Furthermore, accurate extraction models for the clock capacitance are needed in order to enable pre-silicon design and tuning of the passives in order to reach desired frequency targets.

7.2.2 Flip-Flops for Resonant Clocking Systems

Another important issue that requires further research is more specialized flip-flops and latches that ensure robust functionality using sinusoidal clocks over a wide range of clock frequencies. The study presented in this thesis concludes that conventional TG-MSFF and SAFF will be feasible alternatives in a resonant clocking system, if special care is taken to optimize and characterize the flip-flops for the targeted clocking approach. Nevertheless, the performance penalty in both these flip-flops will limit their usage in high-speed applications. Therefore, further research is needed on specialized high-speed flip-flops and latches.

7.3 Bibliography

[1]. M. Hansson, B. Mesgarzadeh, and A. Alvandpour, "1.56 GHz On-chip Resonant Clocking in 130nm CMOS," in *Proceedings of the IEEE Custom Integrated Circuit Conference*, pp. 241-244, 2006.

[2]. B. Mesgarzadeh, M. Hansson, and A. Alvandpour, "Jitter Characteristic in Resonant Clock Distribution," in *Proceedings of the 32nd European Solid-State Circuit Conference*, pp. 464-467, 2006.

[3]. B. Mesgarzadeh, M. Hansson, and A. Alvandpour, "Jitter Characteristic in Charge Recovery Resonant Clock Distribution," in *IEEE Journal of Solid-State Circuits*, vol. 42, no. 7, pp. 1618-1625, 2007.

[4]. B. Mesgarzadeh, M. Hansson, and A. Alvandpour, "Low-Power Bufferless Resonant Clock Distribution Networks," in *50th IEEE International Midwest Symposium on Circuits and Systems*, pp. 960-963, 2007.

[5]. M. Hansson and A. Alvandpour, "Power-Performance Analysis of Sinusoidally Clocked Flip-Flops," in *Proceedings of 23rd IEEE NORCHIP Conference,* pp. 153-156, 2005.

Part III

Leakage Tolerant Circuit Design

Chapter 8

Background

8.1 Introduction

With device sizes deep in the nanometer range, leakage currents have continued to increase, and have become one of the primary design constraints in VLSI systems. All from battery powered mobile processors to high-performance server processors, where the cooling cost is limiting the power envelop, requires power constrained design [1], [2]. With the trend being that large scale VLSI systems incorporates more and more complex functionality, a large number of transistors are added, which will further exacerbate the leakage problem [2], [3]. Furthermore, increasing leakage does not only increase the power dissipation, but also impacts circuit robustness severely, because of diminishing I_{ON}/I_{OFF}-ratios. This is leading to performance and area penalties in order to compensate for the increasing leakage and retain circuit robustness [1]. This is especially problematic for dynamic circuit styles, which rely on temporary charge storage on parasitic capacitors. Hence, leakage seriously threatens to limit the general usage of dynamic circuit techniques in nanoscale CMOS [4]. As a common example, dynamic wide-OR domino circuits are utilized in high-speed performance-critical memory components such as register files in microprocessors. With the continuing scaling the escalating leakage has forced the amount of pull-down paths (width) of the OR-gates to reduce, and the keeper circuitry to be upscaled in order to maintain the robustness [5]. This has lead to a reduction in the performance.

This chapter briefly describes some commonly used leakage reduction techniques. Moreover, a dynamic flip-flop topology is used as an example, explaining the concept of floating nodes, and the impact that the growing leakage have on dynamic circuits. Some common circuit techniques, used to compensate for leakage and noise on dynamic flip-flops and latches, are also discussed.

8.2 Leakage Reduction Techniques

The impact of leakage on the power dissipation of CMOS VLSI systems was introduced in Chapter 2, where the main leakage sources were discussed. Among these, subthreshold current still remains the largest cause of the total leakage power, which is a growing contributor to the total power dissipation in high-performance digital systems. Therefore, there is a continuing need for leakage reduction techniques [7].

8.2.1 Power Gating and Multiple-V_{th} Techniques

For a long time process manufacturers have provided the designer with a choice between transistor versions with different threshold voltages. This enables a trade-off between low-V_{th} transistors with high-performance and high-leakage, or slower transistors with higher threshold voltage and much lower leakage. High-V_{th} transistors can be incorporated in a circuit design as sleep transistors, turning off the active part by gating the power supply during standby [2], [8]. The high-V_{th} transistors used as the power-gate devices are leaking less, thus reducing the static power dissipation during standby. However, in order to impact the performance of the circuit as little as possible during the active phase, these devices need to be made large. Therefore, buffer stages are needed to drive the power-gates, which leads to switching power dissipation and might limit the reduction of the total power [2], [8].

Another way to integrate high-V_{th} transistors is to selectively replace non-critical transistors in the design with low-leakage, high-V_{th} devices. For instance, for a critical data path in the design, the transistors with the lowest threshold voltage are utilized to maximize the performance, while non-critical paths utilize high-V_{th} transistors to reduce the overall leakage. As an example, the keeper transistors in a static flip-flop are not in the critical data path, and could thereby be replaced with low-leakage transistors [9], [10].

8.2.2 Selective Long-Channel Insertion

Increasing the channel length is an efficient method to reduce the leakage currents in the transistor, because the threshold voltage can be modulated by the channel length for small dimension transistors as discussed in section 2.2. This in turn leads to an exponential reduction of the subthreshold current [1], [11]. Selective long-channel insertion is commonly used in memory cells for non-critical transistors. Because memory cells are mostly in their hold phase, reducing the static currents will be effective to reduce the total power [11]. However, the increased channel length has a direct relationship on the delay through a digital gate, thus leading to a performance penalty [1], [11]. Furthermore, the increased gate capacitance could also lead to an increase of the dynamic switching power. Nevertheless, selective insertion of long-channel transistors is still an attractive method to achieve multiple-V_{th} design without the penalty of adding extra masks in the design process for different channel doping profiles of different devices [11].

8.2.3 Threshold Voltage Modulation

The threshold voltages of the transistor depend on the voltage potential between the source and body terminals, as discussed in section 2.2. Therefore, if the body voltage can be altered, the threshold voltage of the device can be modulated both up and down. This is the basic principle of a number of body bias techniques, which have been proposed in the literature. If a circuit is in standby mode, a higher body potential would increase the threshold voltage according to equation (2.1) in section 2.2. This increase in threshold voltage reduces the weak conduction current, causing the total leakage for the circuit to decrease. On the other hand, during high-frequency operation the circuits need to have low delay, therefore a low threshold voltage is required. A lower body potential can then be applied, reducing the threshold voltage and reducing the delay of the circuits [10], [12]. A drawback with this technique is that it requires access to the body terminals for individual transistors, or at least blocks of transistors. Usually, all NMOS transistors are manufactured in the same substrate, causing all of their body contact to be common. This is making variable threshold voltage technique impractical for large complex designs. The solution to this is to use so called triple-well technologies, where the body terminal to both the NMOS and PMOS transistors can be accessed and changed individually. This requires a few more mask layers in the manufacturing process, which need to be taken into account [10], [12]. Furthermore, additional circuit overhead is required in order to dynamically tune the bias voltages during operation.

8.3 Dynamic Circuits

Dynamic circuits have been successfully utilized in high-performance digital designs because of their attractive high-speed properties and their low area compared to conventional static logic [5], [13]. As an example, wide dynamic OR-gates are commonly used as high-performance register files in performance-critical parts of multi-GHz VLSI designs [5].

The increasing clock power [7] and the increasing amount of flip-flops in modern microprocessors emphasize the need for low clock load topologies. Dynamic flip-flops represent a category of high-speed timing circuits, which due to their low number of transistors also leads to low area and low clock power. However, in contrast to static logic, where all nodes in the circuits are tied to either ground or the power supply at steady state, dynamic circuits relies on the ability to temporarily store charge on the parasitic capacitance of the storage node. Consequently, this makes dynamic circuits highly susceptible to noise and leakage, which can lead to functional failures [14].

8.3.1 Low-Power Dynamic Flip-Flop

Figure 8.1 shows a fully dynamic TG-MSFF. This implementation of an edge-triggered register is very appealing from an area perspective, because it only requires eight transistors. This is also an attractive feature for both high-performance and low-power systems [15]. The low number of transistors in the implementation of the flip-flop also results in a low clock load compared to a static TG-MSFF using clocked keepers (i.e. Figure 3.12). Moreover, without the keeper circuits, the capacitive load on the data path and the contention currents are removed. These properties lead to a reduced delay through the flip-flop, and therefore higher performance compared to a conventional static TG-MSFF. However, as there are no keepers to resupply the storage nodes when they are floating, the memory in dynamic latches and flip-flops relies solely on charge storage on the storage node capacitance. This means that dynamic flip-flops are sensitive to leakage and noise of all sorts, and the storage time will be limited. A dynamic flip-flop or latch therefore relies on a constant refreshing of the stored charge. Nonetheless, flip-flops and latches are used as temporary memory cells in registers and pipeline paths, which are clocked with a continuous clock signal under normal operation. This relaxes the robustness constraint, as data is either updated or refreshed every clock cycle [15]. This is true especially for high-performance custom designed data paths clocked at multi-GHz clock frequencies, which is common for high-performance microprocessors.

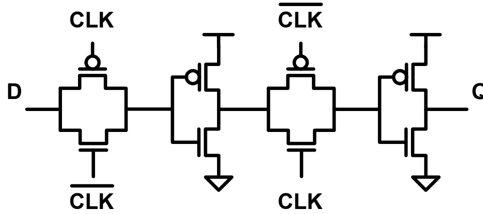


Figure 8.1: Dynamic transmission-gate master-slave flip-flop.

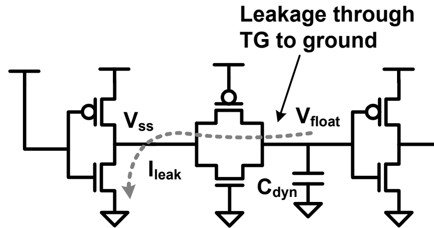


Figure 8.2: Floating node susceptible to leakage currents.

8.3.2 Leakage Robustness Issues

With escalating leakage, floating nodes will lose their states faster and the time that the node could be left floating reduces dramatically [14]. Figure 8.2 shows a floating node in a dynamic TG-MSFF with the transmission-gate in its non-conducting state. V_{float} is here charged to the power supply voltage (V_{dd}) before the transmission-gate closes. The input side of the transmission-gate has the opposite voltage causing the maximum drain-source potential difference. Thus, DIBL, GIDL, and short-channel effects are maximizing the leakage current, and create the worst-case leakage situation. If the total leakage from the floating node is equal to I_{leak}, and the parasitic capacitance on the floating node is C_{dyn}. Then the time it takes for the leakage to discharge the floating node in Figure 8.2 from V_{float} to V_{ss} (ground) is described by equation (8.1).

$$t_{discharge} = \frac{C_{dyn} V_{float}}{I_{leak}} \tag{8.1}$$

Obviously, the node in Figure 8.2 would have lost its state if it were to discharge all the way to ground. If a voltage drop of ΔV could be accepted for the floating node, a maximum time can be defined for which the node can be left floating before it needs to be refreshed. This time is defined as the so called survival time of the dynamic circuit, here denoted $t_{survival}$. The survival time for the floating node can be expressed as equation (8.2), which gives a minimum refresh frequency described by the relationship in equation (8.3).

$$t_{survival} = \frac{C_{dyn} \Delta V}{I_{leak}} \tag{8.2}$$

$$f_{refresh,min} = 1/t_{survival} \tag{8.3}$$

8.3.3 Static Weak-Keeper Flip-Flops

As low clock load is an extremely important property, this benefit of dynamic flip-flops is hard to ignore. However, their poor leakage and noise robustness require constant refreshing, which poses a considerable problem. Especially as clock gating remains an attractive method to reduce clock power [17] - [19]. This will effectively stop the refreshing of the floating nodes, which will eventually lead to an erroneous discharging/charging of the state in the dynamic flip-flops and latches. To mitigate this, a common technique to improve the robustness, but keep the low number of clocked devices, is to use an uninterrupted keeper as shown in Figure 8.3 [15], and make the dynamic flip-flop fully static again. This static keeper technique is commonly refereed to as ratioed or weak keeper. The keeper is directly connected to the storage node of each latch. During a state change the previous value on the output of each latch will drive the keeper to the opposite supply compared to the input. Hence, the driver (*invD*) of the latch has to fight with and overpower the keeper (*invF*) in order to change the state. This results in a significant contention with the input data, which increases the delay [15]. This contention will also, during the switching event, lead to a direct path from power supply to ground between the keeper and the input driver as shown in Figure 8.3. This leads to additional power dissipation, and therefore, it is essential to minimize the size of the keeper in order to reduce the power dissipation and the delay.

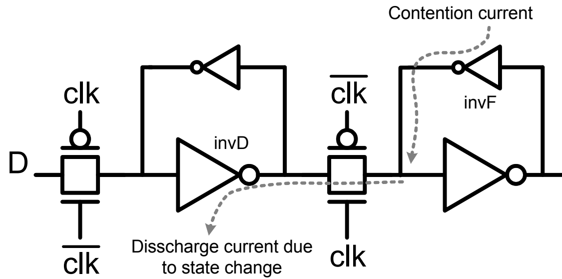


Figure 8.3: TG-MSFF with uninterrupted weak keeper.

Theoretically, as weaker the feedback inverter in the keeper becomes, the faster the flip-flop will be. However, there is not much room to make the feedback inverter weaker. Apart from sizing the transistor width to minimum size, the gate length could be upsized or a stacked keeper could be incorporated in order to make the keeper strength even weaker. Both of these techniques will increase the output load of the driving inverter (*invD* in Figure 8.3), and once again degrade the performance of the flip-flop. Another option is to size up the rest of the flip-flop with respect to the feedback inverter, which instead results in larger clock load and power consumption. Moreover, there is a minimum keeper strength that is set by the robustness conditions. Therefore, sizing the keepers will be a careful tradeoff between power and performance on one hand, and static robustness on the other. Nevertheless, for non-critical data-paths where the data activity is low, a weak-keeper flip-flop is a feasible alternative in order to reduce the clock load of the entire chip.

8.4 Bibliography

[1]. B. Chattarjee, M. Sachdev, R. Krishnamurthy, "Leakage Control Techniques for Designing Robust, Low Power Wide-OR Domino Logic for Sub-130nm CMOS Technologies," in *Proceeding of the 5th International Symposium on Quality Electronics*, pp. 415-420, 2004.

[2]. J.W. Tschanz, S.G. Narendra, Y. Ye, B.A. Bloechel, S. Borkar, and V. De, "Dynamic Sleep Transistor and Body Bias for Active Leakage Power Control of Microprocessors," in *IEEE Journal of Solid-State Circuits*, vol. 38, no. 11, pp. 1838-1845, 2003.

[3]. S. Heo, K. Barr, M. Hampton, and K. Asanović, "Dynamic Fine-Grain Leakage Reduction Using Leakage-Biased Bitlines," in *Proceeding of the 29th Annual Symposium on Computer Architechture*, pp. 137-147, 2002.

[4]. R.K. Krishnamurthy, A. Alvandpour, S. Mathew; M. Anders, V. De, S. Borkar, "High–performance, low–power, and leakage–tolerance challenges for sub–70nm microprocessor circuits," in *Proceeding of the European Solid-State Circuits Conference*, pp. 315-321, 2002.

[5]. R.K. Krishnamurthy, A. Alvandpour, G. Balamurugan, N.R. Shanbhag, K. Soumyanath, and S.Y. Borkar, "A 130-nm 6-GHz 256 x 32 bit Leakage-Tolerant Register File," in *IEEE Journal of Solid-State Circuits*, vol. 37, no. 5, pp. 624-632, 2002.

[6]. K. Mistry, C. Allen, C. Auth, B. Beattie, D. Bergstrom, M. Bosi, M. Brazier, M. Buehler, A. Cappellani, R. Chau, C.-H. Choi, G. Ding, K. Fisher, T. Ghani, R. Grover, W. Han, D. Hanken, M. Hattendorf, J. He, J. Hicks, R. Huessner, D. Ingerly, R. Jain, R: James, L. Jong, S. Joshi, C. Kenyon, K. Kuhn, K. Lee, H. Lin, J. Maiz, B. McIntyre, P. Moon, J. Neirynck, S. Pae, C. Parker, D, Parsons, C. Prasad, L. Pipes, M. Prince, R. Ranade, T. Reynolds, J. Sandford, L. Shifren, J. Sebastian, J. Seiple, D. Simon, S. Sivakumar, P. Smith, C. Thomas, T. Troeger, P. Vandervoorn, S. Williams, K. Zawadzki, "A 45nm Logic Technology with High-k+Metal Gate Transistors, Strained Silicon, 9 Cu Interconnect Layer, 193nm Dry Patterning, and 100% Pb-free Packaging," in *IEEE International Electron Device Meeting*, pp. 247-250, 2007.

[7]. S. Naffziger, B. Stackhouse, T. Grutkowski, "The Implementation of a 2-core Multi-Threaded Itanium®-Family Processor," in *Digest of Technical Papers IEEE International Solid-State Circuit Conference,* pp. 182-183, 2005.

[8]. D. Duarte, Y.-F. Tsai, N. Vijaykrishnan, and M.J. Irwin, "Evaluating Run-Time Techniques for Leakage Power Reduction," in *Procceding of the 15th International Conference on VLSI Design*, pp. 31-38, 2002.

[9]. V. De, Y. Ye, A. Keshavarzi, S. Nerendra, J. Kao, D. Somasekhar, R. Nair, S. Borkar, "Techniques for Leakage Power Reduction," in A.Chandrakasan, W.J. Bowhill, F. Fox, *Design of High-Performance Microprocessor Circuits*, IEEE Press, 2001, ISBN: 0-7803-6001-X.

[10]. K. Roy, S. Mukhopadhyay, and H. Mahmoodi-Meimand, "Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicrometer CMOS Circuits," in *Proceeding of the IEEE*, vol. 91, no. 2, pp. 305-327, 2003.

[11]. F. Fallah and M. Pedram, "Standby and Active Leakage Current Control and Minimization ion CMOS VLSI Circuits," in *IEICE Transaction on Electronics*, vol. E88-C, no. 4, 2005.

[12]. J. Kao, S. Narendra, A. Chandrakasan, "Subthreshold Leakage Modeling and Reduction Techniques," in *International Conference on Computer Aided Design*, pp. 141-148, 2002.

[13]. Y. Lih, N. Tzartzanis, W.W. Walker, "A Leakage Current Replica Keeper for Dynamic Circuits," in *Digest of Technical Papers IEEE International Solid-State Circuit Conference*, pp. 442-443, 2006.

[14]. P. Gronowski, "Issues in Dynamic Logic Design," in A. Chandrakasan, W.J. Bowhill, F. Fox, *Design of High-Performance Microprocessor Circuits*, IEEE Press, 2001, ISBN: 0-7803-6001-X.

[15]. J.M. Rabaey, A. Chandrakasan, B. Nikolic, *Digital Integrated Circuits – A Design Perspective*, Prentice-Hall, 2003, ISBN: 0-13-597444-5.

[16]. S. Chou, "Integration and Innovation in the Nonoelectronics Era," in *Digest of Technical Papers IEEE International Solid-State Circuit Conference,* pp. 36-41, 2005.

[17]. Q. Wu, M. Pedram, and X. Wu, "Clock-Gating and Its Application to Low Power Design of Sequential Circuits," in *IEEE Transaction on Circuits and Systems-I: Fundamental Theory and Applications*, vol. 47, no. 103, pp. 415-420, 2000.

[18]. J. Oh and M. Pedram, "Gated Clock Routing for Low-Power Microprocessor Design," in *IEEE Transaction on Computer-Aided Design of Integrated Circuits and Systems*, vol. 20, no. 6, pp. 715-722, 2001.

[19]. H. Jacobson, P. Bose, Z. Hu, A. Buyuktosunoglo, V. Zyuban, R. Eickemeyer, L. Eisen, J. Griswell, D. Logan, B. Sinharoy, J. Tendel, "Stretching the Limits of Clock-Gating Efficiency in Server-Class Processors," in *Proceedings of the 11th International Symposium on High-Performance Computer Architectures*, pp. 238-242, 2005.

Chapter 9

Leakage Compensation Keeper

9.1 Introduction

As discussed in the previous section, dynamic implementations of flip-flops can be used to reduce the clock load in the design, and thereby reduce the clock and total power dissipation. However, the noise and leakage robustness for dynamic circuits is in general poor, which creates a need to use some compensation technique in order to retain the state in a high-leakage environment. Weak keeper techniques are commonly used to retain a static behavior without adding more transistors to the clock load, but with the price of increased delay and power dissipation in the flip-flop.

In this chapter an alternative keeper technique, aimed for dynamic latches and flip-flops, is presented. The proposed keeper technique is implemented on a reconfigurable flip-flop. During normal operation mode the flip-flop is clocked with multi-GHz operating frequency, and thereby the storage nodes are refreshed every clock cycle. The proposed flip-flop then acts as a dynamic topology, using a simple and scalable leakage compensation technique to avoid incorrect charging/discharging of the floating nodes due to leakage. To enable robust functionality during any low-frequency operation or test, the proposed keeper technique can be reconfigured to a static weak-keeper feedback, thereby retaining full static functionality. Analysis and measurement are presented on the proposed leakage current compensation technique. The main goal is to retain

the traditional operation of dynamic flip-flops and latches even in the presence of the large leakage currents in nanoscale CMOS technologies.

9.2 Reconfigurable Leakage Compensation Keeper

As subthreshold currents have been increasing 3-5X for every new technology generation [1] up until now, obtaining correct functionality for uncompensated dynamic latch will be hard for future nanoscale technologies, even at very high clock frequencies. Hence, it will become extremely difficult to take advantage of the great benefits of dynamic flip-flops, like low clock load and low contention currents, in the future without new keeper techniques.

9.2.1 Principle of Operation

A keeper is proposed that compensates for the leakage in a dynamic latch, while imposing minimal contention with the input driver. The proposed leakage compensation keeper improves the leakage tolerance, therefore increasing the usability of low-power dynamic latches in future CMOS technologies. Figure 9.1 shows a transmission-gate latch implemented with the proposed leakage compensation keeper. The gate terminals of the keeper transistors (M_{n2} and M_{p2}) are connected to the power supply and ground, respectively. The keeper is therefore in OFF-mode, and the storage node is exposed with minimal contention current during the switching of states. Furthermore, the keeper adds

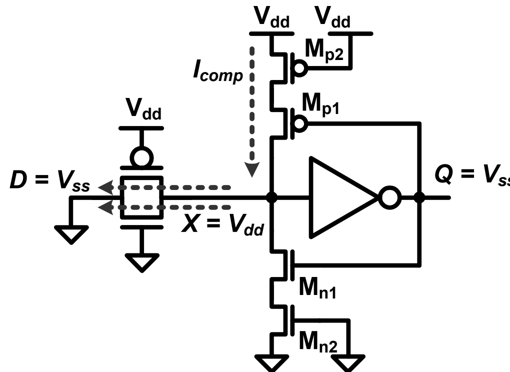


Figure 9.1: Transmission-gate latch with leakage compensation keeper at worst-case leakage condition.

no additional transistors to the clock, which reduce the total clock load compared to a flip-flop with interrupted keepers (e.g. Figure 3.12).

$$I_{subth} \propto W \cdot e^{1/mv_T(V_{GS}+\eta V_D)} \tag{9.1}$$

$$Sizing\ Ratio = \frac{W_{Mp2}}{W_{TGn}+W_{TGp}} = \frac{\alpha \cdot W_{Mn2}}{W_{TGn}+W_{TGp}} \tag{9.2}$$

Assuming that a voltage of V_{dd} is stored at node X, the worst-case leakage condition is when the input is asserted a low voltage (V_{ss}). The subthreshold leakage currents through the transmission gate transistors are then maximized due to DIBL and other short-channel effects according to equation (9.1) [2]. The output of the inverter is also low (V_{ss}), which turns transistor M_{p1} on and transistor M_{n1} off. Both M_{p1} and M_{n1} are realized with minimum size transistors to minimize the load on the storage node. The destructive subthreshold leakage through the transmission-gate and additional destructive leakage such as any gate leakage will start to discharge node X. When the voltage on X decreases the drain-source voltage of transistor M_{p2} will increase, which is causing the subthreshold current through M_{p2} to increase. This additional leakage current will compensate the leakage current through the transmission gates. The subthreshold leakage current is linearly proportional to the transistor width according to equation (9.1), which enables the magnitude of the compensation current (I_{comp}) to be adjusted by sizing of the transistor M_{p2}. This sizing has a relatively weak impact on the output load of the driving inverter M_{p3}/M_{n3} and will have minor impact on the contention of the latch. A sizing ratio for the PMOS-keeper transistor is defined according to equation (9.2), where α is a constant that compensates for the relatively lower subthreshold leakage in the PMOS device compared to the NMOS transistor. For a certain sizing ratio the compensation leakage will compensate the leakage currents through the transmission-gate (destructive leakage), eventually leading to a steady-state voltage on the storage node. Provided that this steady-state voltage is larger[1] than the trip point of the output inverter, the latch will be held stable even though subject to leakage. Figure 9.2 shows a DC simulation for the proposed leakage compensation keeper when the sizing ratio is increased. The simulation is run on a standard 130-nm dual-V_{th} 1.2-V CMOS technology, at 125 °C, and at worst-case leakage corner. The sizing of the PMOS transistors is here increased by a factor of three to compensate for the relatively higher leakage in the NMOS

[1] In the case presented in Figure 9.1.

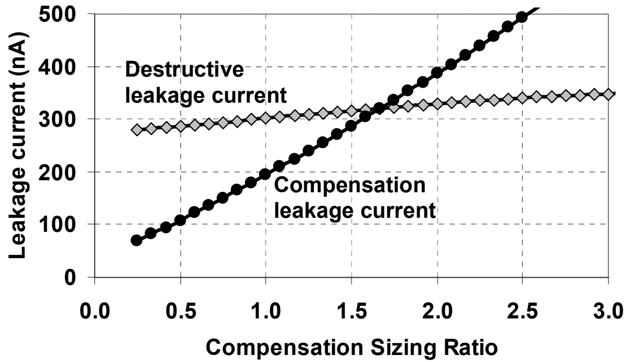


Figure 9.2: Simulated destructive and compensation DC-leakage currents.

transistors ($\alpha = 3$). Additionally, utilizing low-V_{th} devices for M_{n2} and M_{p2} would increase the compensating current in a multi-V_{th} CMOS processes. The destructive leakage is here defined as the sum of all leakage currents erroneously discharging or charging the floating node *X*. As the Figure 9.2 shows, by sizing M_{p2}, the destructive leakage can be compensated with the compensation leakage. The case where the storage node is holding a low (V_{ss}) voltage can be described analog to the high case by sizing M_{n2} according to equation (9.2). An obvious advantage of this natural compensation is that the technique is insensitive to temperature variation as any change in temperature will change both the compensating and destructive leakage in the same way.

Figure 9.3 shows the transient behavior of the floating node *X* in Figure 9.1. The transient simulation is based on a relatively slow (pessimistic) clock frequency of 16 MHz. The data input node changes to its opposite value right after the transmission-gate becomes opaque, leading to maximum subthreshold leakage in the transmission-gate. The voltage of the floating node *X* shows that the proposed compensation technique in Figure 9.1 increases the leakage robustness of the latch. The transient behavior of an uncompensated dynamic latch is included to illustrate the leakage tolerance improvement. In Figure 9.3 a sizing ratio of two is used, which leads to a slight overcompensation according to Figure 9.2. This overcompensation can opportunistically be used to increase the noise robustness of the floating node *X* in Figure 9.1.

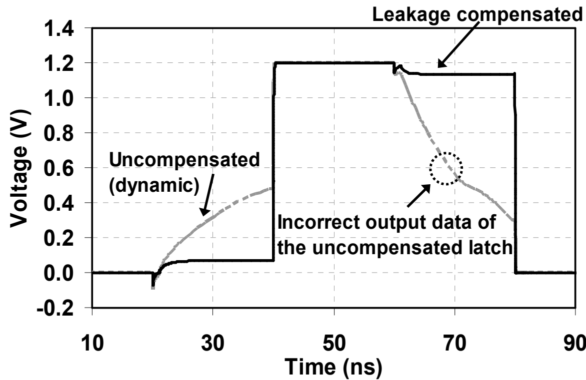


Figure 9.3: Storage node voltage of an uncompensated dynamic latch and a leakage compensated latch (f_{clk} = 16 MHz).

9.2.2 Reconfigurable Dynamic Flip-Flop

In order to reduce the clock power for digital synchronous designs, different clock-gating approaches are extensively utilized. The principle is to turn off the clock for blocks that are idle [4] - [6]. Synchronous memory elements should by definition hold their states even when the clock is gated. Moreover, the reliability of digital designs are usually verified and tested under extreme conditions, so called burn-in tests [7], [8]. Elevated temperature and power-supplies are used to accelerate the time to failure due to manufacturing defects. During the test the functionality of the chip is controlled and should be maintained, even at the low clock frequencies used during testing [7], [8]. Both described scenarios sets extreme noise robustness constrains on uncompensated dynamic latches and flip-flops in nanometer-scale CMOS technologies. Flip-flops or latches, implemented with the proposed leakage compensation technique in Figure 9.1, are when it comes to conventional DC-noise robustness metrics still dynamic and must be treated as such. Due to this fact, low-frequency operation like during testing or clock gating will eventually lead to that the flip-flop or latch will lose its stored data due to noise. To counteract this, a conditional leakage compensation keeper is proposed. An implementation of a TG-MSFF utilizing the proposed reconfigurable keeper is shown in Figure 9.4. The proposed reconfigurable flip-flop has two operation modes, a dynamic mode, which is used during normal operation (Mode B), and a low-

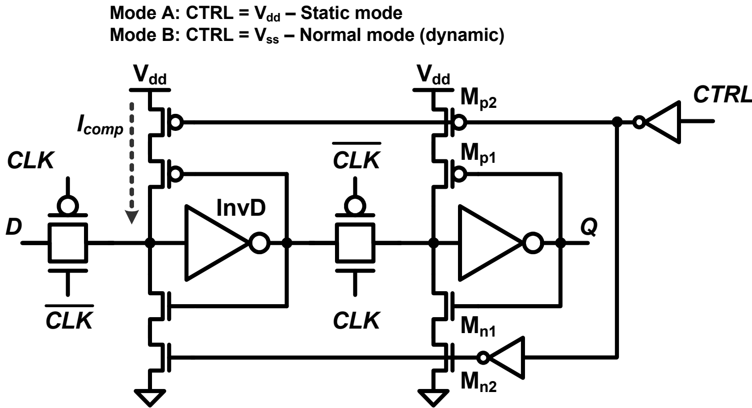


Figure 9.4: Proposed conditional transmission-gate flip-flop.

frequency/stand-by static mode (Mode A). During normal operation the flip-flop is aimed for multi-GHz synchronous designs, where the high-frequency is assumed to enable sufficient refreshing of the semi-floating nodes. The proposed leakage compensation keeper is here used to increase the leakage tolerance, while interfering as little as possible with the data path through contention. At any clock gating scenario or at burn-in tests, the flip-flop is configured into the stand-by mode, where the keeper is configured as a weak static keeper. Hence, the functional robustness is retained also at low-frequency operation, where the performance penalty due to the weak keeper contention can be omitted.

9.3 Simulation Results

9.3.1 Robustness using Leakage Compensation

As discussed in section 8.3.2, robustness of a dynamic circuit can be expressed as a maximum acceptable deviation from the desired voltage on a floating node. This gives a minimum required refresh frequency in order to keep the stored data on the node. For the leakage compensated reconfigurable flip-flop described above, this frequency is defined as minimum required clock frequency leading to maximum 10% degradation from the desired value on the outputs. Figure 9.5 shows the minimum required clock frequency versus the sizing ratio

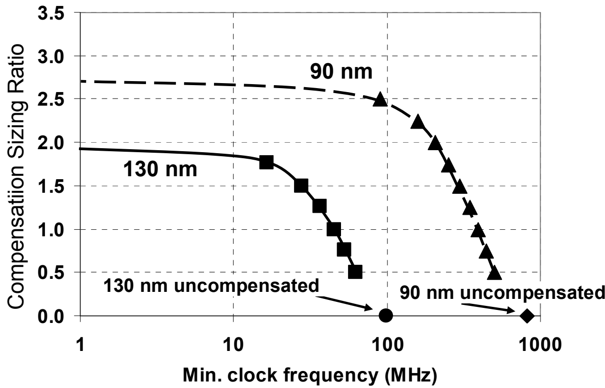


Figure 9.5: Simulated minimum required frequency for correct functionality.

in equation (9.2) for both uncompensated and leakage compensated flip-flops[2], in 130-nm and 90-nm CMOS technologies, respectively. Simulations are run for a worst-case leakage corner at 110 ºC. Figure 9.5 shows that full compensation is achieved for a sizing ratio above 2.0 in 130-nm CMOS, not considering other noise sources than leakage. The leakage compensation requires a slightly higher compensation factor of 2.5 in 90-nm CMOS.

The proposed technique utilizes either an NMOS or a PMOS transistor to compensate for leakage in both NMOS and PMOS transistors, which makes it matching sensitive. A plot showing the leakage compensation technique at worst-case matching conditions simulated with 130-nm process data is shown in Figure 9.6. The curve refereed to as Fast N/Slow P represents the worst-case matching condition when the NMOS in the transmission-gate has a high destructive leakage and the compensating PMOS transistor has low leakage. The curve refereed to as Fast N/P is the worst-case leakage corner for the proposed leakage compensation technique. The results indicate that the leakage robustness can be improved considerably even when considering worst-case matching. By scaling the compensation transistor 3X the robustness can be improved 2.4X compared to the minimum sized compensation transistor. Compared to an uncompensated dynamic flip-flop the robustness is improved 3.8X.

[2] Mode B in Figure 9.4

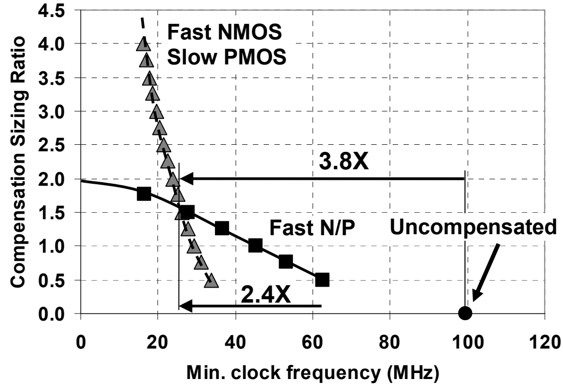


Figure 9.6: Proposed leakage compensation under worst-case matching corner (130-nm data).

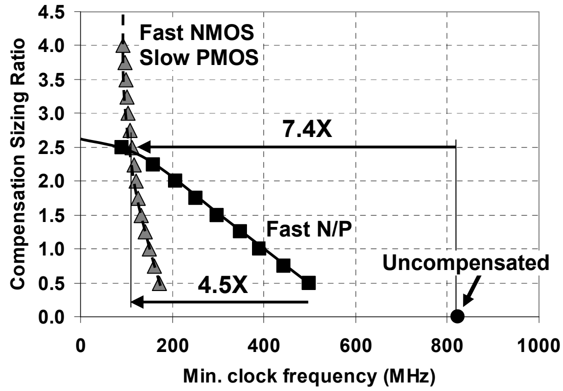


Figure 9.7: Proposed leakage compensation under worst-case matching corner (90-nm data).

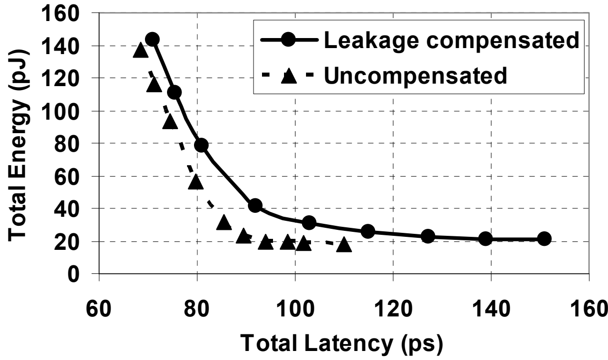


Figure 9.8: Power vs. total flip-flop delay.

Figure 9.7 shows the worst-case matching condition for the conditional flip-flop configured in dynamic mode for a 90-nm CMOS technology. The leakage tolerance compared to the uncompensated dynamic flip-flop is increased by more than 7.4X resulting in a reduction of the minimum required operational frequency from 830 MHz down to less than 110 MHz. Furthermore, a sizing-ratio range of 0.5-to-2.5 gives a leakage tolerance increase of more than 4.5X for the proposed flip-flop. Comparing the results in Figure 9.6 and Figure 9.7 shows that a sizing ratio of 3.0 increases worst-case leakage tolerance 5X and 8X in 130 nm and 90 nm, respectively. This shows that the proposed leakage technique enable good scaling properties into sub-90-nm technologies.

9.3.2 Performance Impact of Leakage Compensation Keeper

The proposed keeper technique increases the leakage robustness for dynamic flip-flop dramatically. However, due to the introduced keeper circuitry the load on the storage nodes of the flip-flop in Figure 9.4 is increased compared to uncompensated dynamic flip-flops (e.g. Figure 8.1). This leads to a performance penalty. Figure 9.8 shows the energy-delay plot of the proposed reconfigurable flip-flop, using a constant sizing ration of two, and a dynamic uncompensated flip-flop. Total energy is calculated from the average currents into the data and clock inputs and the current from the power supply. The complementary clock phase is generated internally in all flip-flops. The clock drivers are sized for equal edge rates using a clock frequency of 3 GHz. However, the sizing of M_{p2} and M_{n2} in Figure 9.4 has a relatively weak impact on the output load of the driving inverter *InvD*. This is due to the isolation provided by the inverter

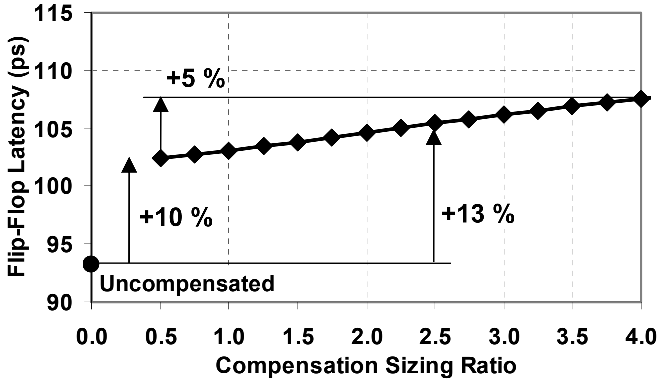


Figure 9.9: Performance impact due to leakage-compensation circuitry.

transistors M_{p1}/M_{n1}. The impact of the keeper load is shown in Figure 9.9. For a sizing-ratio from 0.5 to 4.0 the latency of the proposed flip-flop in dynamic mode is only increased by 5%, while the leakage tolerance is increased more than 4.5X. Compared to the uncompensated flip-flop the latency increases 13% for a sizing ratio of 2.5 where the dominating contributor to the performance impact is due to the increased load (10%). The 3% additional delay penalty with the increasing sizing-ratio can be accounted to increasing compensating leakage currents, which gives a small contention current.

9.3.3 Leakage Compensation Keeper for Low Clock Power

As a consequence of the non-clocked keeper the proposed feedback technique will reduce the clock power consumption compared to using interrupted keepers, while preserving the functionality of the dynamic operation in the presence of large leakage currents. Power and performance comparisons in a 130-nm, 1.2-V standard CMOS technology at typical process corners, 110 °C is shown in Figure 9.10. Clock energy is here calculated as the total energy at zero data activity. A constant sizing ratio of two is used for the leakage compensation keeper in the proposed flip-flop, where the transistors M_{n2} and M_{p2} are sized according to equation (9.2). The proposed flip-flop achieves more than 30% lower clock power dissipation compared to the conventional static flip-flop with interrupted keeper.

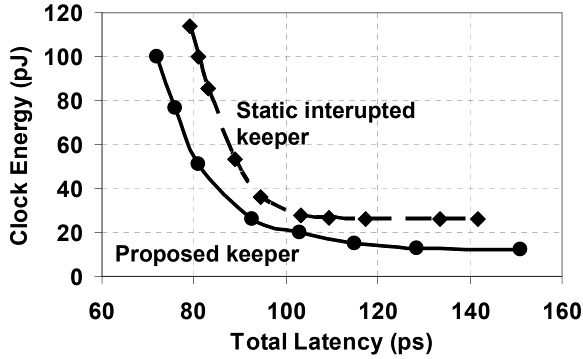


Figure 9.10: Clock load reduction by using proposed leakage compensating keeper.

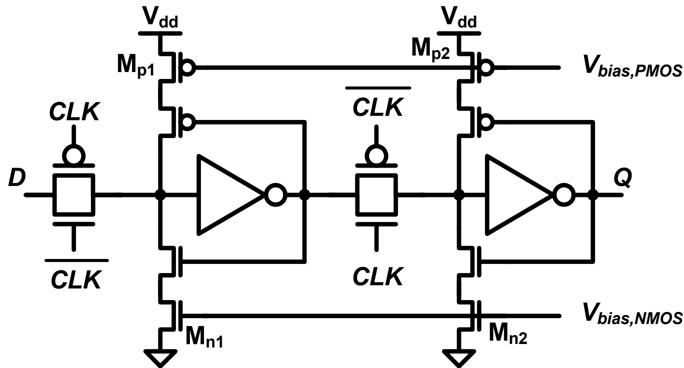


Figure 9.11: Stand-alone test circuit for reconfigurable flip-flop implemented with leakage compensation keeper

9.4 Experimental Chip Results

9.4.1 Chip Implementation

The proposed leakage compensation technique is implemented on a test chip in 0.18-µm CMOS technology. Identifying the impact of the gate voltage of the subthreshold conduction relation in equation (9.1), it is clear that changing the gate bias of the keeper transistors will increase the compensating current. This can be utilized to emulate the proposed sizing method, at least to a fist order. The leakage compensation keeper technique is verified by measurements using a stand-alone flip-flop, which enables external adjustment of the gate voltage on the keeper transistors, M_{n1}, M_{n2}, M_{p1}, and M_{p2} as shown in Figure 9.11.

To demonstrate the global power savings using the proposed technique, a conventional pipelined 32-bit ripple-carry adder (RCA) is implemented. The circuit is designed in a standard 0.18-µm, 1.8-V CMOS process. Figure 9.12 shows the block-level schematic for the adder test circuit. The 32-bit test vectors for A and B are generated either by using 4x32-bit pre-set vectors in a shift register, or by using 32-bit PRBS data generated by linear feedback shift registers. The 32-bit adder is implemented in a 4-stage pipeline, resulting in the critical path equivalent to an 8-bit ripple-carry adder. If the input to the adder is chosen from the fixed input vectors the resulting 32-bit sum and the carry bit are compared to the correct output in a comparator tree. The output of the comparator is used as the test signal, which can be monitored off chip. Two

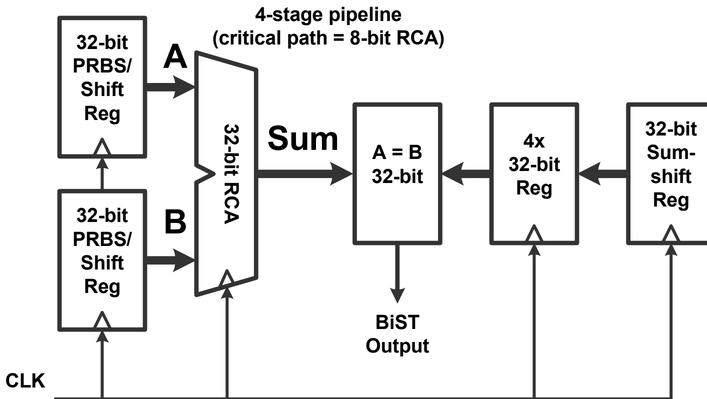
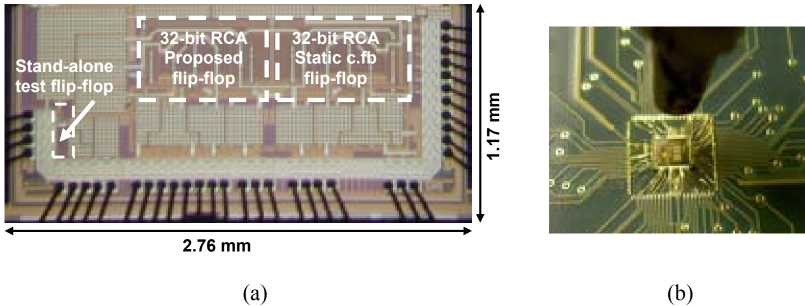


Figure 9.12: Block level schematic of power-performance test circuit.



(a) (b)

Figure 9.13: (a) Photograph of test-chip in 0.18-µm CMOS. (b) Test chip directly bonded on the test PCB.

comparable designs are implemented using the conventional static TG-MSFF with interrupted keepers (e.g. Figure 3.12) and a TG-MSFF using the proposed keeper technique in Figure 9.4. Each design includes over 400 flip-flops in the adder and registers, where 160 of the flip-flops are used in the adder. The two separate test blocks use separate clock distribution networks. The clock distribution network to each evaluation circuit is designed for comparable edge rates of the internal clock signal. Due to the reduced clock load in the flip-flop using the proposed keeper, the clock driver to the adder implemented using the reconfigurable flip-flop is down-sized to achieve equal edge rate. The chip micrograph of the two evaluation circuits and the stand-alone flip-flop is shown in Figure 9.13(a).

9.4.2 Measurement Results

The feasibility of the keeper technique in Figure 9.1 is verified with measurements on the experimental test chip. The measurements are conducted at elevated temperatures in order to increase the leakage currents for the 0.18-µm CMOS process using the test circuit in Figure 9.11. Figure 9.13(b) shows the test chip directly bonded on the PCB. The measured minimum operational frequency is shown in Figure 9.14 at chip temperatures of 110 °C and 130 °C. The measured results displays good agreement with the simulated data presented in section 9.3, hence proves the concept. As a comparison a gate underdrive-voltage is used. In this case the leakage compensation is firmly turned off due to the negative gate-source voltage.

Figure 9.15 shows the comparison between the total and clock power of the leakage compensated flip-flop and the conventional static flip-flop with interrupted keeper. The measurements are done at a clock frequency of

830 MHz and at room temperature using PRBS data on the inputs. The results show that the leakage compensating keeper flip-flop results in a chip clock power reduction of 25% and a total power reduction of 19% compared to the identical adder circuit using flip-flops with interrupted keepers.

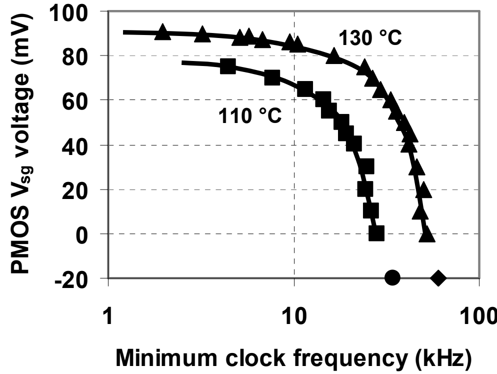


Figure 9.14: Minimum required clock frequency for different gate-bias voltages and temperatures.

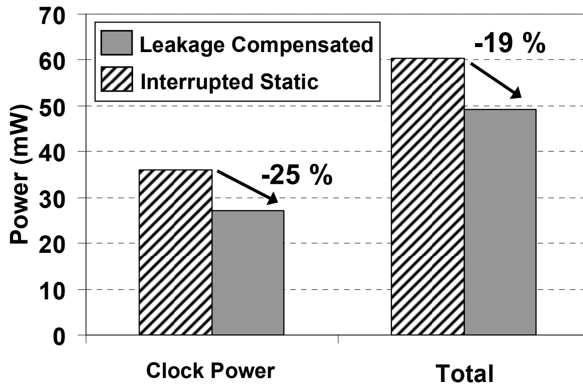


Figure 9.15: Power comparison between RCA using flip-flops with either conventional interrupted keepers or leakage compensated keepers.

9.5 Bibliography

[1]. B. Chattarjee, M. Sachdev, R. Krishnamurthy, "Leakage Control Techniques for Designing Robust, Low Power Wide-OR Domino Logic for Sub-130nm CMOS Technologies," in *Proceeding of the 5th International Symposium on Quality Electronics*, pp. 415-420, 2004.

[2]. K. Roy, S. Mukhopadhyay, H. Mahmoodi-Meimand, "Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicron CMOS Circuits," in *Proceeding of the IEEE*, vol. 91, no. 2, pp. 305-327, 2003.

[3]. R.K. Krishnamurthy, A. Alvandpour, G. Balamurugan, N.R. Shanbhag, K. Soumyanath, and S.Y. Borkar, "A 130-nm 6-GHz 256 x 32 bit Leakage-Tolerant Register File," in *IEEE Journal of Solid-State Circuits*, vol. 37, no. 5, pp. 624-632, 2002.

[4]. Q. Wu, M. Pedram, and X. Wu, "Clock-Gating and Its Application to Low Power Design of Sequential Circuits," in *IEEE Transaction on Circuits and Systems-I: Fundamental Theory and Applications*, vol. 47, no. 103, pp. 415-420, 2000.

[5]. J. Oh and M. Pedram, "Gated Clock Routing for Low-Power Microprocessor Design," in *IEEE Transaction on Computer-Aided Design of Integrated Circuits and Systems*, vol. 20, no. 6, pp. 715-722, 2001.

[6]. H. Jacobson, P. Bose, Z. Hu, A. Buyuktosunoglo, V. Zyuban, R. Eickemeyer, L. Eisen, J. Griswell, D. Logan, B. Sinharoy, J. Tendel, "Stretching the Limits of Clock-Gating Efficiency in Server-Class Processors," in *The 11th International Symposium on High-Performance Computer Architecture*, pp. 238-242, 2005

[7]. A. Alvandpour, R. Krishnamurthy, K. Soumyanath, S. Borkar, "A Sub-130-nm Conditional Keeper Technique," in *IEEE Journal of Solid-State Circuits*, vol. 37, no. 5, pp. 633-638, 2002.

[8]. A. Alvandpour, R. Krishnamurthy, S. Borkar, A. Rahman, C. Webb, "A Burn-In Tolerant Dynamic Circuit Technique," in *Proceedings of the IEEE Custom Integrated Circuits Conference*, pp. 81-84, 2002

[9]. Y. Lih, N. Tzartzanis, W. W. Walker, "A Leakage Current Replica Keeper for Dynamic Circuits," in *Digest of Technical Papers IEEE International Solid-State Circuit Conference*, pp. 442-443, 2006.

[10]. R.K. Krishnamurthy, A. Alvandpour, S. Mathew; M. Anders, V. De, S. Borkar, "High-performance, low-power, and leakage-tolerance challenges for sub-70nm microprocessor circuits," in *Proceedings of the European Solid-State Circuits Conference*, pp. 315-321, 2002.

[11]. N. Vasseghi, K. Yeager, E. Sarto, and M. Seddighnezhad, "200-MHz Superscalar RISC Microprocessor," in *IEEE Journal of Solid-State Circuits*, vol. 31, no. 11, pp. 1675-1686, 1996.

[12]. S. Naffziger, B. Stackhouse, T. Grutkowski, "The Implementation of a 2-core Multi-Threaded Itanium®-Family Processor," in *Digest of Technical Papers IEEE International Solid-State Circuit Conference*, pp. 182-183, 2005.

Chapter 10

Low-Leakage Microcode ROM

10.1 Introduction

Read-only memories (ROMs) are essential components in high-performance microprocessors, DSPs, and special purpose accelerators. They are extensively used to stored fixed information, such as microcode instructions for complex micro operations in processors [1], [2], or data coefficients for digital filters [3]. Reduction of active and leakage power in ROMs has become important because of power constrained high-performance designs, and low-power portable systems, such as cellular phones and notebooks, requiring longer battery life [3] - [5]. As previously discussed, leakage is exponentially increasing as technology is scaled [6], which requires more aggressive low leakage power schemes [7]. The leakage power component can constitute a large part of the total power in conventional ROM circuits. An example is shown in Figure 10.1(a) for a high-performance microcode ROM, where 36% of the overall ROM array power is from leakage. This part becomes an even larger component as leakage current increases for sub-65-nm designs.

High-performance ROM structures are commonly implemented as NOR-type arrays, since they are faster than other topologies [3]. In a NOR-type ROM the bitcell transistors are connected in parallel on a bitline. The bitline is precharged high through a precharge device that is usually clocked. After the precharge phase, the appropriate wordline is asserted and the bitline is discharged, which indicates a one stored in the ROM. On the other hand, a zero stored in the ROM

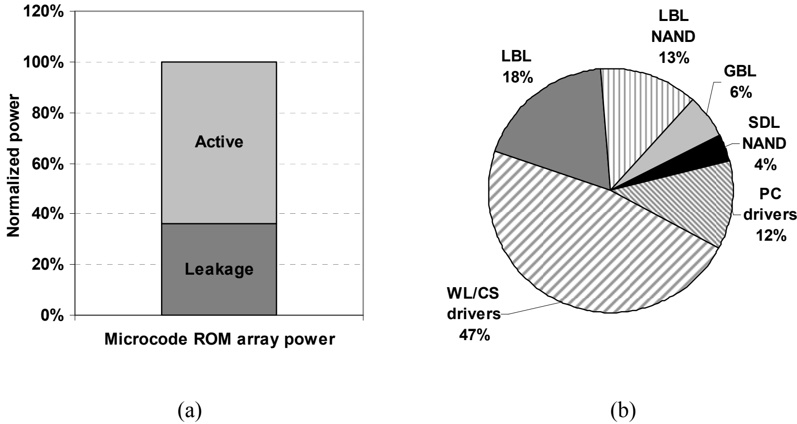


Figure 10.1: (a) ROM array power breakup, (b) ROM array leakage power components.

is implemented by removing the connection of the NOR-pulldown transistor on the bitline for that bit [8]. This means that the bitline will keep its precharged state when the corresponding bitline is asserted. Hence, the main cause of power dissipation in a NOR-based ROM is from ones stored in the memory.

A number of techniques have been proposed to reduce the power dissipation in ROMs. Instead of precharging all bitlines in the design only the bitlines that will be addressed can be precharged. This leads to a reduction of the active power dissipation [3]. Furthermore, as the ones stored in the memory are causing the discharge of the bitlines, a technique to reduce the power dissipation is to minimize the number of ones stored in the ROM. There are several techniques that do just that. If the dominating number of the bits in the memory are one, the entire ROM could be inverted, which results in that a mostly zeros are stored in the ROM. The output value can then simply be inverted to retain the correct result [3]. Other techniques involves more elaborate schemes where either rows or columns in the ROM are inverted, which requires control circuit overhead in order to conditionally invert the output bits [3], [4]. However, conventional ROM design utilizes a worst-case design approach where drivers, pre-charge devices, and column-select multiplexers are designed for a worst-case fully populated ROM array[1]. Typically in a fabricated ROM implementation, the memory array is only sparsely populated, thereby making the drivers oversized. Figure 10.1(b) shows that the majority of the array leakage

[1] Fully populated is here referred to logic value '1' on all bit-entries.

power in a conventional ROM design is dissipated in the drivers for wordline, column-select, and pre-charge signals.

In this chapter a technique is presented, which results in a reduction of both the leakage and active power for a high-performance 25.6-Kb microcode ROM. A programmable logic technique is proposed, which utilizes the data heuristics of the microcode to allow optimal programming of local and global merge circuitry, pre-charge devices, column-select multiplexers, and wordline driver strengths.

10.2 ROM Organization

Figure 10.2 shows the organization of the 25.6-Kb high-performance microcode ROM, consisting of 320 entries x 80 bits. Each bitslice is implemented in a three-level hierarchy, where the bitcells are connected to the bitcell-line that contains eight bitcells and individual pre-charge transistors, as shown in Figure 10.3. By using a 4-to-1 column-select multiplexer four bitcell lines are merged together to form a 32-bit local bitline (LBL). Two LBLs are merged via a static

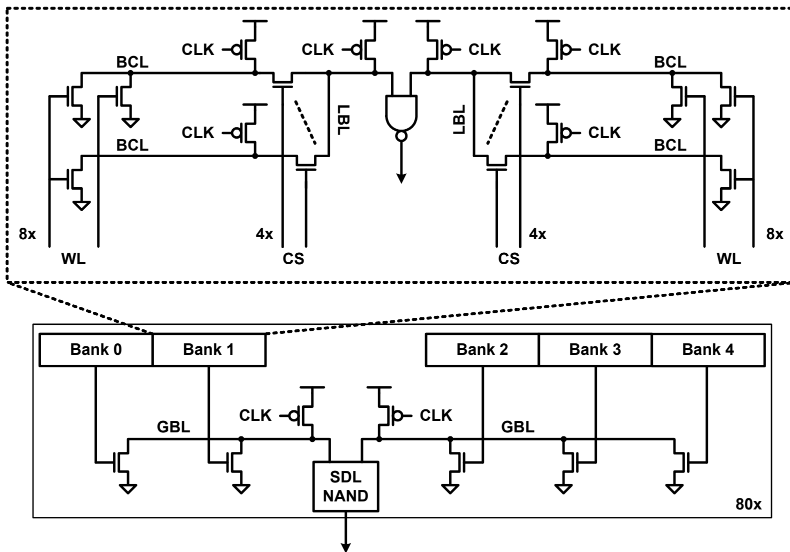


Figure 10.2: Micro-code ROM organization.

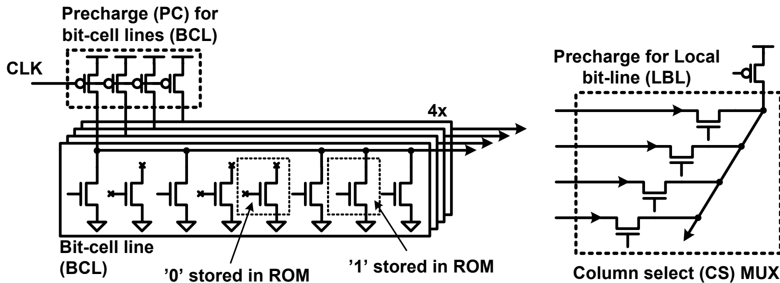


Figure 10.3: LBL including CS-MUX and PC devices.

NAND gate that drives a single global bitline (GBL) pull-down transistor. Finally, a 2-way and a 3-way GBL are merged with a SDL-NAND gate, which forms a 5-way GBL producing a static output.

A gate connection to the wordline and a drain connection to the bitcell line represent a bitcell layout storing a logic value '1'. Removing the direct connections of the gate and drain from the wordline and bitline, respectively, represents a bitcell layout storing a logic value '0'. Prior to fabrication, the ROM contains a fully populated array and is ready for programming. During microcode programming the '0'-valued bitcell layouts replace the appropriate '1'-valued bitcell layouts, which finalizes the layout before fabrication.

A complete read operation for the studied ROM is performed in two cycles. A 2-phase 50% duty-cycle clocking plan allows seamless time borrowing at the phase boundaries. In the first cycle, a partially decoded 9-bit address delivers 2x100 control signals, out of which 2x80 lines drive the wordlines directly, and 2x20 lines drive column-select MUXs. In the next cycle, wordline buffers drive across two 40-bit arrays and bitline evaluation starts.

10.3 Microcode Heuristics

The heuristics of the finalized ROM are shown in Table 10-1. After programming, 80% of the finalized ROM store bitcell layouts that contain logic value '0', thereby removing up to 155 bitcell loads on a single wordline as shown in the histogram in Figure 10.4. Conventional wordline driver strengths allow for full population of 160 bitcell loads. However, the wordline drivers end up oversized based on the actual programmed wordline loads.

Since 48% of the bitcell lines always read a logic value ‘0’ and remain constant at V_{cc}, this results in 24% of the LBLs remaining constant at V_{cc}. Moreover, 14% of the LBL-NAND gate outputs are remaining constant at V_{ss}. These constant voltage bitcell lines, LBL, and LBL-NAND output nodes will never transition due to the programmed data. Alternatively, 2.5% of the bitcell lines always read a logic value ‘1’, which means that these bitcell-lines are always discharged consuming unnecessary power as 0.5% of the LBLs always will be discharge to V_{ss}.

Therefore, many pre-charge transistors, column-select transistors, static NAND gates, and GBL pull-downs are unnecessary and perform no useful operation. These unused devices together with the oversized wordline drivers contribute greatly to the total leakage as shown in Figure 10.1(b).

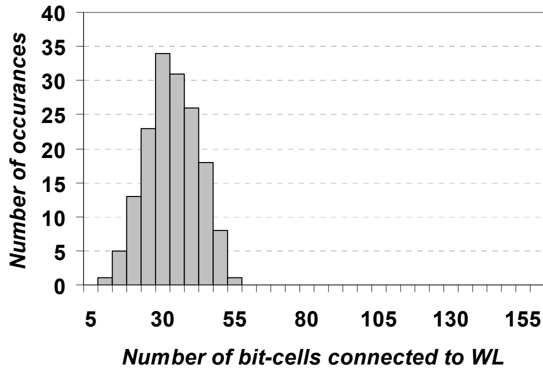


Figure 10.4: Histogram of wordline loading after BC programming.

TABLE 10-1: MICROCODE ROM HEURISTICS.

	Total	**Storing only ‘0’**	**Storing only ‘1’**
Bits	25.6K	20577 (80.4%)	5023 (19.6%)
Bitcell lines	3.2K	1544 (48%)	81 (2.5%)
Local bitlines	800	193 (24%)	4 (0.5%)
NAND / Global bitlines	400	56 (14%)	0 (0%)

10.4 Programmable Logic Technique

Based on the data heuristics of the ROM microcode presented in Table 10-1, pre-fabrication layout programming is not limited to only the bitcell programming. A programmable logic technique is proposed, which allows optimal programming to occur on the unnecessary local/global merge circuitry, precharge devices, column-select multiplexers, and wordline driver layouts. The proposed programmable logic technique is depicted in Figure 10.5 and Figure 10.6.

10.4.1 Removal of Unused Devices

A bitcell-line node that always reads a logic value ‘0’ allows a direct layout connection to V_{cc}, which removes any unnecessary precharge devices and one column-select device as shown in Figure 10.5. Similarly, a LBL node that always reads a logic value ‘0’ allows the removal of the four column-select and precharge devices creating a direct connection to V_{cc}. When a logic value ‘1’ always is read, the bitcell line is implemented with all eight bitcells and precharge devices removed, and a direct connection to V_{ss} is made. Column-select and precharge devices are also removed when a logic value ‘1’ always is read out from the entire LBL, which instead is implemented by tying the LBL output directly to V_{ss}. Bitcell, column-select, and precharge device removal achieves a 22% reduction of the LBL leakage component. Finally, the removal of unnecessary LBL-NAND merge and GBL pull-downs results in a 15% reduction of the respective leakage component.

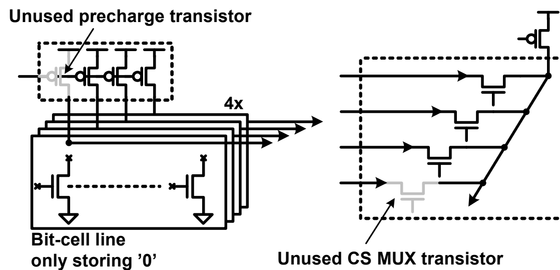


Figure 10.5: Proposed programmable logic technique for bit-cell lines only storing zeros.

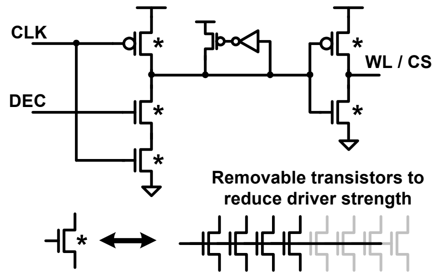


Figure 10.6: Proposed programmable logic technique to optimize word-line driver strength.

10.4.2 Optimization of Driver Strength

After removal of bitcells, column-select, and precharge devices, respective column-select, wordline, and precharge driver loading is reduced because of less gate load. Due to the variation in loading across all wordlines, four differently sized wordline buffer layout designs are proposed, which are designed to drive 40, 80, 120, and 160 bitcells with less than or equivalent the delay of the 160 bitcells worst-case load. Figure 10.7 shows the number of occurrences of the wordline driver strengths, where the majority of wordline drivers only drive up to 40 bitcells. Removing additional legs of the devices in the wordline driver layouts creates the three additional layout cells as shown in Figure 10.6. This enables optimization of the wordline driver strength to the actual pre-programming load. By applying the same approach as for the wordline driver, three additional drivers are designed for the column select and bitcell-line precharge drivers, respectively, with strength optimized for 25%, 50%, and 75% of the fully populated load. The proposed driver optimization technique enables further leakage power reduction in column-select drivers and precharge drivers. Figure 10.8 and Figure 10.9 show the distribution of drivers before and after applying the reprogramming, thus reducing wordline/column-select driver leakage by 52% and precharge driver leakage by 25%.

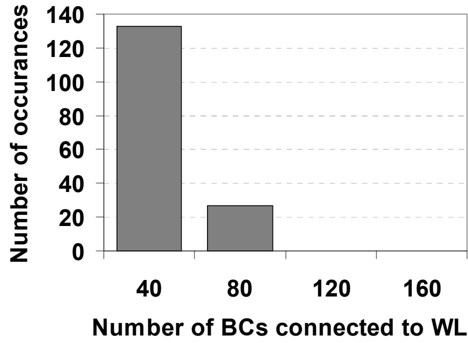


Figure 10.7: Wordline driver histogram for the proposed ROM.

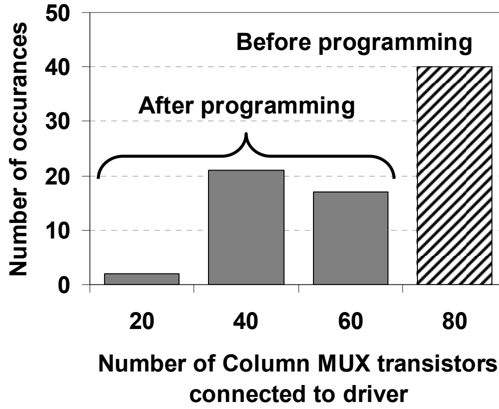


Figure 10.8: CS driver histogram for conventional and proposed ROM.

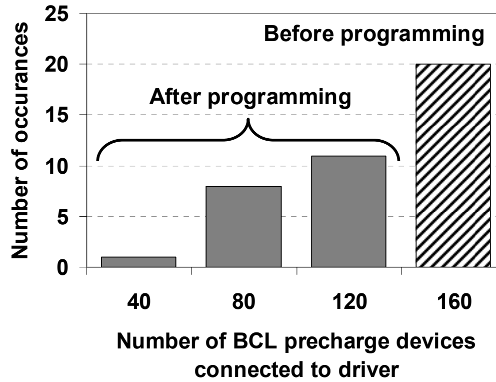


Figure 10.9: PC driver histogram for conventional and proposed ROM.

10.5 Comparison Results and Discussion

The conventional and proposed ROM arrays are implemented in a 1.2-V, 65-nm CMOS technology [9]. The conventional microcode ROM incorporates selective insertion of long-channel transistors in the bitcells and GBL pull-down transistors. By using the proposed programmable logic technique and incorporating the microcode heuristics in Table 10-1, the ROM array leakage is reduced by 32%, from 45.1 mW down to 30.6 mW, and overall array power is reduced 15%, from 118.5 mW down to 101.1 mW for a clock frequency of 9 GHz as shown in Table 10-2. The leakage reduction due to the proposed technique is also shown in the power break-up in Figure 10.10, which clearly shows the large leakage reduction in the wordline/columns-select drivers.

TABLE 10-2: ROM POWER AND FREQUENCY COMPARISON.

<table>
<tr><th colspan="2">Microcode ROM</th><th>Leakage Power (mW)</th><th>Total Power (mW)</th><th>Max. Freq. (GHz)</th></tr>
<tr><td rowspan="2">Inserted long channel length</td><td>Conventional</td><td>45.1</td><td>118.5</td><td rowspan="2">9.0</td></tr>
<tr><td>Proposed</td><td>30.6</td><td>101.1</td></tr>
<tr><td rowspan="2">All nominal channel length</td><td>Conventional</td><td>61.7</td><td>141.4</td><td rowspan="2">9.7</td></tr>
<tr><td>Proposed</td><td>44.8</td><td>121.4</td></tr>
</table>

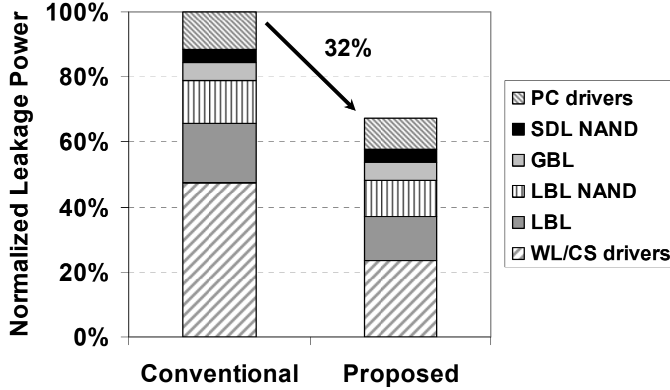


Figure 10.10: ROM leakage power break-up.

In a fully nominal channel length ROM design, maximum frequency is increased by 8% to 9.7 GHz as shown in Table 10-2. The proposed programmable logic technique reduces the leakage power by 27%, from 61.7 mW down to 44.8 mW, in a fully nominal channel length ROM design. Moreover, total ROM array power for the fully nominal channel length ROM is reduced by 14%, from 141.4 mW down to 121.4 mW. This is comparable to the power dissipation in the conventional ROM using inserted long-channel transistors. Hence, 8% performance improvement can be achieved without any leakage or total power penalties.

10.6 Bibliography

[1]. E. Borin, M. Breternitz Jr., Y. Wu, and G. Araujo, "Clustering-Based Microcode Compression," in *International Conference on Computer Design*, pp. 189-196, 2006.

[2]. D. Boggs, A. Baktha, J. Hawkins, D.T. Marr, J.A. Miller, P. Roussel, R. Singhal, Bret Toll, K.S. Venkatraman, "The Microarchitecture of the Intel® Pentium® 4 Processor on 90nm Technology," in *Intel Technology Journal,* Vol. 8, Issue 1, 2004.

[3]. E. de Angel, E.E. Swartzlander Jr., “Survey of Low Power Techniques for ROMs,” in *Proceedings of the International Symposium on Low Power Electronics and Design*, pp. 7-11, 1997.

[4]. K. Inoue, V.G. Moshnyaga, and K. Murakami, “Reducing Power Consumption of Instruction ROMs by Exploiting Instruction Frequency,” in *Asia-Pacific Conference on Circuits and Systems*, vol. 2, pp. 1-6, 2002.

[5]. B.-D. Yang, L.-S. Kim, “A Low-Power ROM using Charge Recycling and Charge Sharing,” in *Digest of Technical Papers IEEE International Solid-State Circuit Conference*, vol. 1, pp. 108-109, February 2002.

[6]. B. Chattarjee, M. Sachdev, R. Krishnamurthy, “Leakage Control Techniques for Designing Robust, Low Power Wide-OR Domino Logic for Sub-130nm CMOS Technologies,” in *Proceeding of the 5th International Symposium on Quality Electronics*, pp. 415-420, 2004.

[7]. T. Ghani, K. Mistry, P. Packan, S. Thomson, M. Stettler, S. Tyagi, M. Bohs, “Scaling Challenges and Device Design Requirements for High Performance Sub-50 nm Gate Length Planar CMOS Transistors,” in *Symposium on VLSI Technology Digest of Technical Papers*, pp. 174-175, 2000.

[8]. J.M. Rabaey, A. Chandrakasan, B. Nikolic, *Digital Integrated Circuits – A Design Perspective*, Prentice-Hall, 2003, ISBN: 0-13-597444-5.

[9]. P. Bai, C. Auth, S. Balakrishnan, M. Bost, R. Brain, V. Chikarmane, R. Heussner, M. Hussein, J. Hwang, D. Ingerly, R. James, J. Jeong, C. Kenyon, E. Lee, S.-H. Lee, N. Lindert, M. Liu, Z. Ma, T. Marieb, A. Murthy, R. Nagisetty, S. Natarajan, J. Neirynck, A. Ott, C. Parker, J. Sebastian, R. Shaheed, S. Sivakumar, J. Steigerwald, S. Tyagi, C. Weber, B. Woolery, A. Yeoh, K. Zhang, M. Bohr, “A 65nm Logic Technology Featuring 35nm Gate Length, Enhanced Channel Strain, 8 Cu Interconnect Layers, Low-k ILD and 0.57 μm^2 SRAM Cell,” *IEEE International Electron Device Meeting Technical Digest*, pp. 657-660, 2004.

Chapter 11

Conclusions and Future Work

11.1 Conclusions

11.1.1 Leakage Compensation Keeper

An analysis of a leakage current compensation technique for dynamic flip-flops is presented [1], [2]. The proposed technique preserves traditional operation of dynamic flip-flops even in the presence of the increased leakage in nano-scale CMOS. The proposed keeper technique enables over 7.4X higher leakage tolerance for a conventional dynamic flip-flop compared to an uncompensated flip-flop for a high-leakage corner in a 90-nm CMOS technology. The compensation can be scaled during the design phase using simple transistor sizing, which is verified by chip measurements at elevated temperatures in a 0.18-μm CMOS process. The proposed technique also enables optional static operation with a weak keeper to provide robust functionality during clock gating and/or at any low frequency testing. As a consequence of the fact that dynamic flip-flops require less clocked transistors the clock load of flip-flops implemented with the proposed leakage compensation technique can be implemented with smaller clock drivers resulting in lower clock power. Power-delay comparisons with conventional flip-flops utilizing interrupted keepers show that the proposed leakage compensation keeper enables more than 30% reduction of the clock load in each individual flip-flop. Utilization of the flip-flop enables 25% reduction of on-chip clock power, due to downsizing of the

clock drivers. This is shown on an experimental test-chip in a 0.18-μm CMOS technology.

11.1.2 Low-Leakage High-Speed ROM

A 9-GHz 25.6-Kb microcode ROM implemented in a 1.2-V, 65-nm CMOS technology [3] is presented [4]. An extended pre-fabrication programmable-logic technique is proposed. The technique utilizes the fact that a sparsely occupied ROM has extensive hardware overhead due to unused precharge and select devices, causing unnecessary extra transistor load on the drivers. The proposed pre-fabrication programmable logic technique removes this hardware overhead based on the heuristics of the ROM contents, and therefore enables driver strengths more optimized to the actual load. The proposed programmable logic technique enables 32% leakage power reduction and 15% total array power reduction without delay or area penalty. The proposed ROM design consumes 101.1 mW total array power with a leakage component of 30.6 mW, showing good sub-65-nm scaling trend. Furthermore, the leakage reduction achieved by the proposed layout programming method enables the utilization of all nominal length transistors, without increasing the leakage or total power dissipation compared to the conventional ROM design. This results in 8% increase in the maximum clock frequency without any power or area penalties.

11.2 Future Work

The proposed leakage compensation technique does not explicitly consider additional noise sources. A comprehensive analysis of the feasibility of the proposed leakage compensation technique under the influence of various AC-noise sources could be an interesting continuation. An implementation of a test chip in a more advanced, and thereby more leaky CMOS process could also be an interesting continuation.

Leakage will remain a large challenge for circuit designers also in future CMOS technologies. Therefore, more leakage energy efficient techniques are needed. Research and studies on leakage reduction and compensation techniques for larger scale digital circuit blocks would be an interesting research path.

11.3 Bibliography

[1]. M. Hansson and A. Alvandpour, "A Low Clock Load Conditional Flip-Flop," in *Proceedings of IEEE International System-on-Chip Conference*, pp. 169-170, 2004.

[2]. M. Hansson and A. Alvandpour, “A Leakage Compensation Technique for Dynamic Latches and Flip-flops in Nano-scale CMOS,” in *Proceedings of IEEE International System-on-Chip Conference*, pp. 83-84, 2006.

[3]. P. Bai, C. Auth, S. Balakrishnan, M. Bost, R. Brain, V. Chikarmane, R. Heussner, M. Hussein, J. Hwang, D. Ingerly, R. James, J. Jeong, C. Kenyon, E. Lee, S-H. Lee, N. Lindert, M. Liu, Z. Ma, T. Marieb, A. Murthy, R. Nagisetty, S. Natarajan, J. Neirynck, A. Ott, C. Parker, J. Sebastian, R. Shaheed, S. Sivakumar, J. Steigerwald, S. Tyagi, C. Weber, B. Woolery, A. Yeoh, K. Zhang, and M. Bohr, “A 65nm Logic Technology Featuring 35nm Gate Lengths, Enhanced Channel Strain, 8 Cu Interconnect Layers, Low-k ILD and 0.57 µm2 SRAM Cell,” in *Technical Digest IEEE International Electron Device Meeting*, pp. 657-660, 2004.

[4]. S.K. Hsu, M. Hansson, A. Agarwal, S. K. Mathew, A. Alvandpour, and R.K. Krishnamurthy, “A 9GHz 320x80bit Low Leakage Microcode Read Only Memory in 65nm CMOS,” in *Proceedings of the 32nd European Solid-State Circuit Conference*, pp.299-302, 2006.

Part IV

Process Variation Aware Design

Chapter 12

Background

12.1 Introduction

CMOS technology scaling is predicted to continue with unaltered pace into the near future [1]. Today the minimum features fabricated in the state-of-the-art device fabrication facilities around the world are approaching the fundamental atomistic limits [2], [3]. As a consequence of the reducing feature sizes and the limited accuracy of several of the tools used in the manufacturing process, the device dimensions and doping levels are becoming harder and harder to control. This results in increasing statistical variation in the device parameters, which is causing a growing spread in the performance and power metrics of the circuits. The increasing variability is considered to be one of the major design problems for the continuing integration in future CMOS processes [4] - [7]. This chapter will briefly discuss the impact on the circuit characteristics due to increase process variation. Moreover, some common techniques that have been proposed to compensate for the process variation will be presented.

12.2 Impact of Process Variation

Parameter variation in transistors and interconnects, due to physical or environmental variations, are important since it causes parametric yield loss. A circuit designed at nominal process corner may fail to satisfy delay and/or

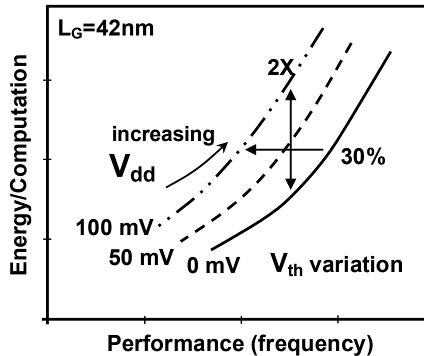


Figure 12.1: Spread in performance and energy per computation due to threshold voltage variation [4].

leakage targets under parameter variations [8]. Statistical variations in device and interconnect parameters, such as channel length and width, metal-wire dimensions, gate oxide thickness, and threshold voltage, produce large spread in the speed, power dissipation, and robustness of integrated circuits.

Among all transistor parameters, threshold voltage constitutes an extremely important parameter both for the driving strength of the transistors (i.e. delay), and the ability to turn off properly (i.e. leakage currents) as was discussed in section 2.2. The standard deviation for the threshold voltage has been shown to be as high as 45 mV for minimum sized transistors in 45-nm CMOS technology [9]. Figure 12.1 illustrates the variation in threshold voltage, which results in 2X variations in the energy per computation and a 30% spread in performance of clock frequency for a high-performance microprocessor [4]. Furthermore, Figure 12.2 shows the distribution of NMOS leakage current in a 150-nm CMOS process, measured at a temperature of 110 ºC. The figure shows a wide range of leakage variations across the process corners with a worst-case spread of 20X from the slowest dies to the fastest most leaky dies [10], [11].

With increasing process variation, fluctuation in circuit performance and behavior becomes increasingly more difficult to handle during circuit design, causing both increased design time and manufacturing effort. Overestimation of the variations during the circuit design leads to longer design times, possibly leading to missed market windows and larger area penalties. On the other hand underestimation of the process variation instead leads to reduced performance, reduced functional yield, and longer time for debugging the manufactured designs [12].

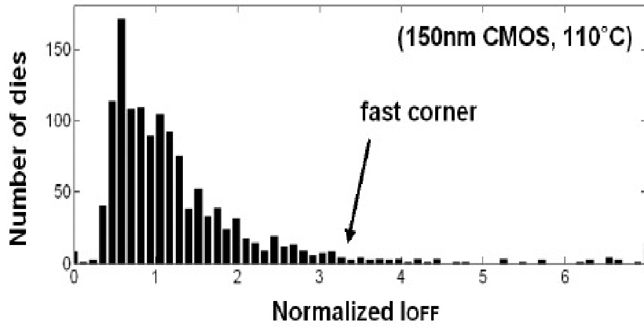


Figure 12.2: Normalized leakage NMOS current distribution [10] [11].

Traditionally during circuit design, worst-case process spreads have been used, where the circuits are designed in order to reach their performance, power, and robustness targets at the worst-case design corners, so called guard-banding [12]. However, because a majority of the measured dies are located in the relatively lower leakage side of the distribution curve (Figure 12.2), guard-banding is an increasingly inefficient method as it results in overall degradation of chip performance, power consumption, and yield.

12.3 Process Variation Compensation Techniques

12.3.1 Power Supply and Body Bias Adjustments

Control of the body bias has been proposed as a compensation technique, where the threshold voltage of the transistors in the circuit is adjusted to compensate for the specific die-to-die process variability. As discussed in section 8.2, body biasing can be used to adjust the leakage in transistors. The same technique has been proposed in order to compensate for process variation [13], [14]. The performance for the slow circuits can be increased by applying forward body bias and thereby lower the threshold voltage. This way static robustness and power is traded against increased performance. Similarly, the fast circuits, which have large leakage currents and low robustness margins, can use a reverse body bias voltage that increases the threshold voltage. That way the higher performance is traded against increased robustness and lower leakage power [8], [13], [14]. Bidirectional adjustments of the body bias have been shown in

measurements using adaptive techniques to reduce both inter-die and intra-die variations [14].

Similarly to adjusting the body bias, power supply voltage have a large effect on both leakage currents and speed of the circuits. Hence, the process variation induced performance and robustness spread could be reduced by using a lower power supply for fast high leakage dies, and a higher power supply for slower low-leakage dies [15], [16].

12.3.2 Reconfigurable Designs

In order to reduce the large spread in performance, which is introduced when robustness needs to be assured over the entire design space, some techniques relying on reconfigurable designs has been proposed. Especially wide dynamic OR-gates for high-performance register files have been proposed using keepers with reconfigurable strengths [10], [11]. The principle is explained by observing that the performance of the circuit is low at the low-leakage side of the graph shown in Figure 12.2, but the robustness is high. This provides a large margin for the specified robustness target. In a reconfigurable design this additional robustness margin can be traded-off against performance, so that slow circuits can be made faster, while losing some of their robustness margin. At the same time the fast circuits, which have the lowest robustness margin, can be reconfigured so that some small performance hit is taken in order to increase their robustness considerably. Consequently, both the spread in robustness (i.e. process and leakage tolerance) and performance are reduced, which increases the total yield of the circuits.

12.3.3 Device Sizing

Another technique to compensate for the increasing process variation during the circuit design is to utilize device sizing. Due to the threshold voltage roll-off in short channel CMOS transistors, a slight increase of the channel length of selective transistors in the design will lead to a higher threshold voltage and a lower leakage current. Moreover, as the threshold voltage variation, due to random dopant fluctuations, is inversely proportional to the gate area the sensitivity to process variations can be reduced by increasing the channel length [5] [17]. This can be a promising method in non-critical data paths where the delay penalty due to the increased channel length can be tolerated. Instead of, or as a complement to sizing the length of the gate, the width of the transistor can be increased, which also leads to a reduction of the process variation due to random dopant fluctuation [17]. By increasing the width of the transistors the driving strength increases, which can compensate for the loss of driving strength due to the increase of the channel length.

12.4 Bibliography

[1]. International Technology Roadmap for Semiconductors (ITRS) – 2007 ed., http://www.itrs.net, accessed: June 2008.

[2]. http://www.intel.com, accessed: June 2008.

[3]. K. Mistry, C. Allen, C. Auth, B. Beattie, D. Bergstrom, M. Bosi, M. Brazier, M. Buehler, A. Cappellani, R. Chau, C.-H. Choi, G. Ding, K. Fisher, T. Ghani, R. Grover, W. Han, D. Hanken, M. Hattendorf, J. He, J. Hicks, R. Huessner, D. Ingerly, R. Jain, R: James, L. Jong, S. Joshi, C. Kenyon, K. Kuhn, K. Lee, H. Lin, J. Maiz, B. McIntyre, P. Moon, J. Neirynck, S. Pae, C. Parker, D, Parsons, C. Prasad, L. Pipes, M. Prince, R. Ranade, T. Reynolds, J. Sandford, L. Shifren, J. Sebastian, J. Seiple, D. Simon, S. Sivakumar, P. Smith, C. Thomas, T. Troeger, P. Vandervoorn, S. Williams, K. Zawadzki, "A 45nm Logic Technology with High-k+Metal Gate Transistors, Strained Silicon, 9 Cu Interconnect Layer, 193nm Dry Patterning, and 100% Pb-free Packaging," in *IEEE International Electron Device Meeting*, pp. 247-250, 2007.

[4]. T.-C. Chen, "Where CMOS is going: trendy hype vs. real technology," in *Digest of Technical Papers IEEE Solid-State Circuits Conference*, pp. 1-8, 2006.

[5]. D. Sylvester, K. Agarwal, S. Shah, "Variability in nanometer CMOS: Impact, analysis, and minimization," in *Integration the VLSI Journal*, 2007, doi:10.1016/j.vlsi.2007.09.001.

[6]. S. Nassif, K. Bernstein, D.J. Frank, A. Gattiker, W. Haensch, B.L. Ji, E. Nowak, D. Pearson, N.J. Rohrer, "High Performance CMOS Variability in the 65nm Regime and Beyond," in *IEEE International Electron Device Meeting*, pp. 569-571, 2007.

[7]. S. Nassif, "Delay Variability: Sources, Impacts and Trends," in *Digest of Technical Papers IEEE International Solid-State Circuit Conference*, pp. 368-369, 2000.

[8]. S. Bhunia, S. Mukhopadhyay, and K. Roy, "Process Variation and Process-Tolerant Design," in *International Conference on VLSI Design*, pp. 699-704, 2007.

[9]. K.J. Kuhn, "Reducing Variation in Advanced Logic Technologies: Approaches to Process and Design for Manufacturability of Nanoscale CMOS," in *IEEE International Electron Device Meeting*, pp. 471-474, 2007.

[10]. C. H. Kim, K. Roy, S. Hsu, A. Alvandpour, R. K. Krishnamurthy, S. Borkar, "A Process Variation Compensation Technique for Sub-90nm Dynamic Circuits," in *Digest of Technical Papers Symposium on VLSI Circuits*, pp. 205-206, 2003.

[11]. A. Agarwal, K. Roy, S. Hsu, R. K. Krishnamurthy, S. Borkar, "A 90nm 6GHz 128x64b 4-Read 4-Write Ported Parameter Variation Tolerant Register File", in *Digest of Technical Papers Symposium on VLSI Circuits*, pp 386-387, 2004.

[12]. D. Boning and S. Nassif, "Models of Process Variations in Device and Interconnects," in A. Chandrakasan, W.J. Bowhill, F. Fox, *Design of High-Performance Microprocessor Circuits*, IEEE Press, 2001, ISBN: 0-7803-6001-X.

[13]. S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, V. De, "Parameter Variation and Impact on Circuits and Microarchitecture," in *Proceedings of Design Automation Conference*, pp. 338-342, 2003.

[14]. J.W. Tschanz, J.T. Kao, S.G. Narendra, R. Nair, D.A. Antoniadis, A.P. Chandrakasan, V. De, "Adaptive Body Bias for Reducing Impacts of Die-to-Die and Within-Die Parameter Variations on Microprocessor Frequency and Leakage," in *IEEE Journal of Solid State Circuits*, vol. 37, no. 11, pp. 1396-1402, 2002.

[15]. M. Elgebaly, M. Sachdev, "Efficient Adaptive Voltage Scaling Systems through Critical Path Emulation," in *ACM/IEEE International Symposium on Low Power Electronic Design*, pp. 375-380, 2004.

[16]. T. Chen and S. Naffziger, "Comparison of Adaptive Body Bias (ABB) and Adaptive Supply Voltage (ASV) for Improving Delay and Leakage under the Presence of Process Variation," in *IEEE Transaction on VLSI Systems*, vol. 11, issue 5, pp 888-898, 2003.

[17]. M.H. Abu-Rahma and M. Anis, "Variability in VLSI Circuits: Sources and Design Considerations," in *IEEE International Symposium on Circuits and Systems*, pp. 3215-3218, 2007.

Chapter 13

Impact of Process Variation on Flip-Flops

13.1 Introduction

As discussed in the Chapter 3, synchronous sequential designs are dominating digital SoCs. Flip-flops and latches are extensively utilized, for instance in the pipelining of the instruction and execution path in microprocessors. With the continuing downscaling of feature sizes, considerable uncertainties in the manufacturing process cause fluctuations in the performance and power dissipation of all circuits [1] - [4]. For extreme high-performance designs, the latency overhead due to the clocked registers, such as flip-flops, consumes a growing part of the total available clock period [5] - [7]. Performance spread, due to increasing process variation for flip-flops, therefore becomes an important roadblock for improved performance with further scaling. Knowledge on how clocked registers behaves under the impact of variation is therefore important. To satisfy the timing constrains for clocked registers, a tight control of timing is required. When this tight control is compromised, by the variation of delays across the die, functionality of the design can be severely degraded.

As for all digital circuits, the performance and functionality of the flip-flops are not only influenced by process parameter variation, but also fluctuations in operating supply voltages and temperature. Some studies on the effect of all these fluctuations on a few flip-flop topologies have been presented in the literature, giving some guideline in the choice of clocking element in various

scenarios [5], [6]. However, these analyses have been based only on deterministic variations of process and environmental parameters.

This chapter treats the statistical impact of process variation on some commonly used flip-flop topologies, and a technique to find the minimum required marginal for a given rate of functionality is discussed. The additional margin is then used to compare the process variation tolerance between different flip-flop topologies.

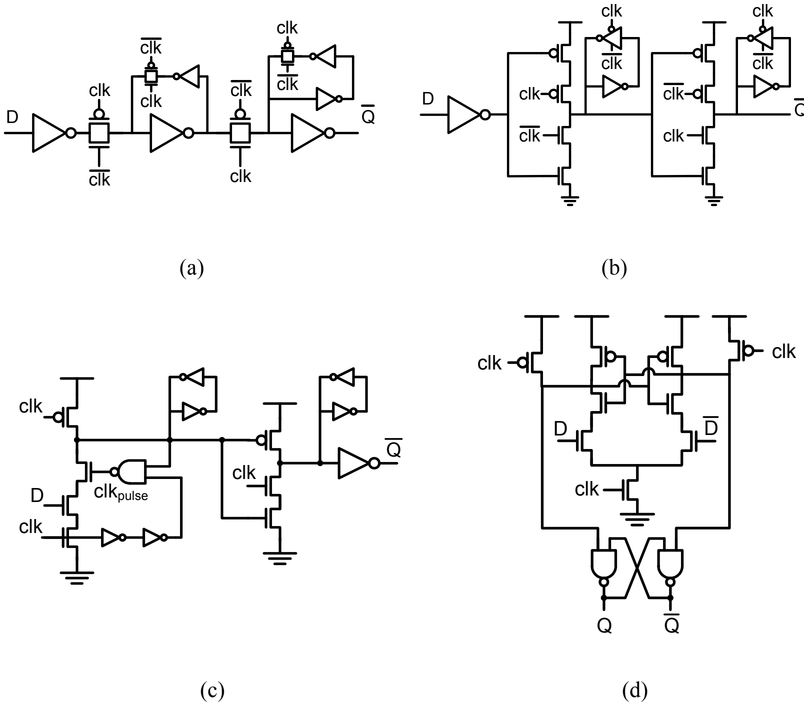


Figure 13.1: (a) Transmission-gate master-slave flip-flop (TG-MSFF), (b) Clocked CMOS master-slave flip-flop (C2MOS-MSFF), (c) Pulsed semi-dynamic flip-flop (SDFF), (d) Sense-amplifier flip-flop (SAFF).

13.2 Flip-Flop Topologies and Optimization

13.2.1 Flip-Flop Topologies

For this analysis four different flip-flops are chosen, which represents some commonly used topologies. Figure 13.1(a) and Figure 13.1(b) shows the static transmission-gate master-slave flip-flop (TG-MSFF) and the clocked CMOS master-slave flip-flop (C^2MOS-MSFF), respectively. Both flip-flops are built up by cascading two complementary clocked latches, which form a robust flip-flop with good hold time behavior. Moreover, these flip-flops are commonly utilized in standard-cell libraries [8], making it important to analyze their process variation tolerance. Figure 13.1(c) shows a high-performance flip-flop usually referred to as semi-dynamic flip-flop (SDFF) [9]. This flip-flop is essentially a pulsed latch, which due to the soft-edge property (transparency period) of the pulse have low total flip-flop latency, and short setup times. Finally, Figure 13.1(d) shows a typical SAFF with a NAND SR-latch acting as a slave latch. This flip-flop is interesting to analyze from a matching perspective due to the differential sense-amplifier in the input latch.

13.2.2 Optimization Approach

In order to design a general energy optimal digital circuit, an optimization of different performance or power targets is done. The studied flip-flops are optimized by iterating the transistor sizes according to the algorithm presented

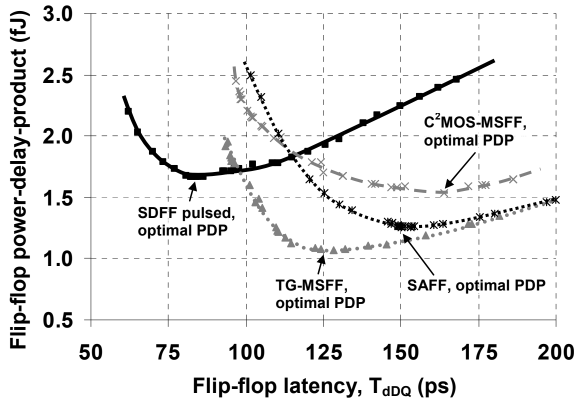


Figure 13.2: Power-delay space with PDP-optimal flip-flop designs.

in section 3.3.3 on page 41. Figure 13.2 shows the PDP versus total flip-flop latency ($t_{d,DQ}$) for the four flip-flops, using process data for a commercially available 90-nm, dual-V_{th} CMOS technology with a power supply of 1.2 V, a temperature of 110 °C, a clock frequency of 1 GHz, and 50% data activity. The optimum PDP point will be refereed to as the optimal point in the remainder of the chapter. The optimal point is chosen as it gives the transistor sizing that result in the lowest energy per switching event. As expected, the SDFF achieves a power-delay optimal point at a rather low latency of 86 ps, while the optimal point for the TG-MSFF is found at a latency of 129 ps. For the SAFF and C^2MOS-MSFF the optimal PDP points are found at a latency of 153 ps and 164 ps, respectively.

13.3 Process Variation Impact on Flip-Flop Timing

The flip-flop delays discussed in section 3.2 could be obtained for any process and environmental variation corner. However, if the flip-flops are optimized at the worst-case corners for speed, the design will be over-designed since only a small portion of the fabricated flip-flops actually will be in that slow corner. On the other hand, if the optimization is done for typical process parameters, there is a need for a larger design margin in order to cover the performance spread. This section introduces a method to find the required delay margins for flip-flops. The method is used in the comparison between the four different flip-flop topologies.

13.3.1 Setup Time Margin

Figure 13.3 shows an example distribution of the setup time (before the clock edge), and the clock-to-output delay for a flip-flop. The typical and slow setup time and clock-to-output delay are indicated in the figure. Obviously, for a slow flip-flop a data transition that arrives at the typical setup time will be sampled incorrectly, which will lead to a timing failure. The slow flip-flops therefore require a longer setup time, which could be defined as a delay margin added to the typical setup time. This delay margin is added in order to guarantee, with a certain probability, that the data will be correctly sampled. Figure 13.4 shows another way to look at the required setup time margin. Here the clock-to-output delay is plotted as a function of the data-to-clock delay at three different process corners. From Figure 13.4 it is clear that the typical setup time results in a metastable flip-flop at the slow corner. Hence, with the added margin a larger part of the process spread can be considered functional. However, the added setup time margin will certainly increase the total latency (data-to-output delay) of the flip-flop. In order to minimize the performance impact the margin need to

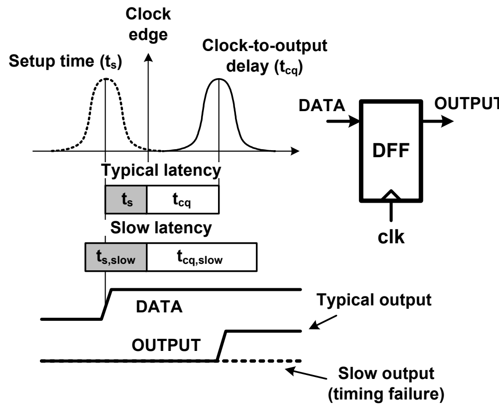


Figure 13.3: Process spread impact on a flip-flop performance.

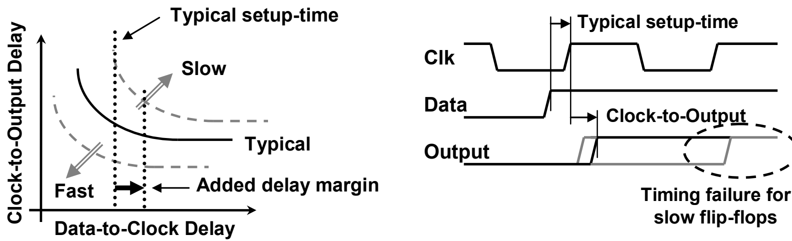


Figure 13.4: Failure to capture data for slow flip-flops.

chosen so that the lowest total latency can be assured, still providing some statistical guarantee of functionality over the design space.

13.3.2 Statistical Simulation Approach

In order to compare process variation impact on different flip-flops a statistical simulation approach is used. Monte-Carlo analysis, including mismatch between transistors, is performed on the flip-flops at the optimal points (Figure 13.2). For the die-to-die variation and systematic within-die variation, the transistor model parameters for the threshold voltage, mobility, drain-source resistance, and junction capacitance are varied uniformly within their respective $\pm 2\sigma$ ranges. For the random local variation the threshold voltage for all transistors are

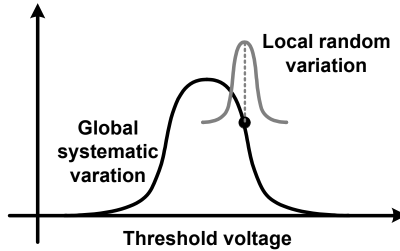


Figure 13.5: Systematic variation with local random mismatch.

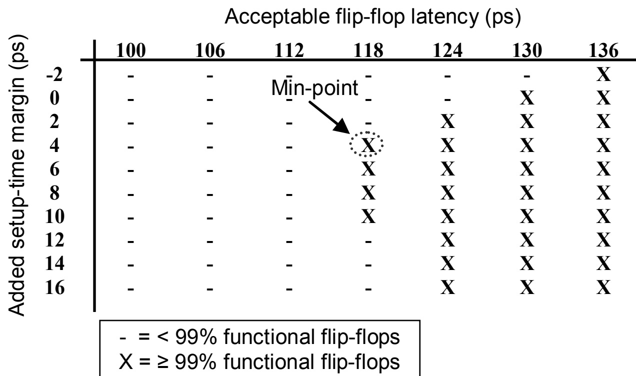


Figure 13.6: Example of an acceptable latency / setup time margin matrix plot.

independently normally distributed around the value given by the systematic variation, as shown in Figure 13.5. This way, systematic within-die variations are modeled at the same time as mismatch effects and local random variations.

In order to make a fair comparison, the total flip-flop latency needs to be compared. To reduce the simulation time, the Monte-Carlo analysis is simulated using a fixed delay between the input data and the clock edge for each Monte-Carlo run. To clarify, the setup time is set to a fixed value, while the process parameters are randomly varied over the design space in a number of Monte-Carlo iterations. The optimum setup times for the flip-flops, found using typical process corners, are used as the starting point for the fixed data-to-clock delay.

However, because of the process variation, this results in that up to 50% of the simulated Monte-Carlo points are slower than the typical corner. The required margin is therefore found by running parametric Monte-Carlo simulations for a range of data-to-clock delays (i.e. different fixed setup times). The ratio of Monte-Carlo simulation points, which results in flip-flops that sample the input data correctly, is evaluated for each fixed setup time margin versus the total latency. The results of the parametric Monte-Carlo simulations are summarized in a matrix-plot like the one shown in Figure 13.6. In order to minimize the performance penalty due to the introduced margin, the point with the lowest acceptable flip-flop latency and minimum setup time margin is chosen from Figure 13.6. This means that in order to get 99% of the pulsed SDFFs to sample the correct data, and thereby be considered functional, a total flip-flop latency of up to 118 ps need to be tolerated, and an additional setup time margin of 4 ps is required. The inserted setup time margin should be seen as a measure of the process tolerance on the setup time for a given flip-flop sizing and topology.

13.4 Proces Variation Simulation Results

All simulations discussed in this chapter are done on a commercially available 90-nm CMOS technology, using a power supply voltage of 1.2 V, and at a temperature of 110 °C. All the flip-flops are clocked with a clock frequency of 1 GHz. The clock signal is fed to the flip-flops through a local buffer chain, which is individually sized for each flip-flop in order to achieve an equivalent FO-4 edge rate on the clock. This way the same driving strength on the clock

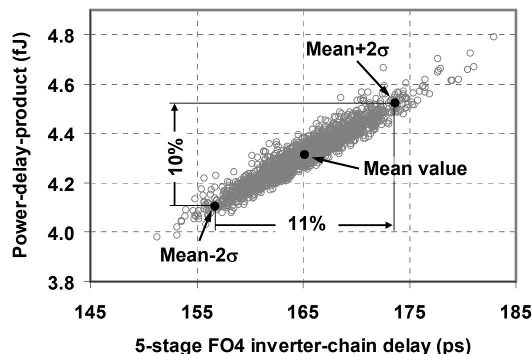


Figure 13.7: Power-performance spread for a 5-stage FO4 inverter-chain.

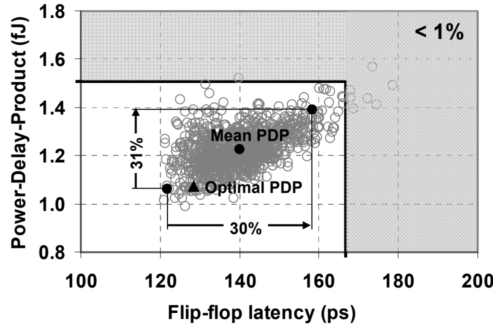


Figure 13.8: Power-performance spread for a TG-MSFF.

driver is achieved for all flip-flops. Moreover, the local clock buffer is also affected by the process variation during the simulations. Finally, the data activity for the flip-flop data input is set to 50% for all simulations.

In order to enable comparison of process variation tolerance between CMOS logic and the flip-flops, a 5-stage FO-4 inverter-chain is also analyzed. The number of stages in the inverter-chain is chosen because the average delay was roughly the same as the average delay through the flip-flops. Figure 13.7 shows the results from a 2500 point Monte-Carlo simulation, which gives the power-performance spread of the inverter-chain due to the process variability. The ratio between the $\pm 2\sigma$ spread around the mean value is used throughout this chapter as a measure of the process variation tolerance of the simulated circuit. Figure 13.7 shows a plot of the power-delay space for the 5-stage FO-4 inverter chain. The plot shows that the $\pm 2\sigma$-delay spread is 11%, and the $\pm 2\sigma$ spread for the PDP is 10%, corresponding to a 3% variation of the power.

Figure 13.8 shows the power-delay space for a TG-MSFF simulated with a 1000-point Monte-Carlo analysis using the setup time margin specified in Table 13-1. The gray area in Figure 13.8 is the defined as the fail region, where the boundary values are determined so that less than 1% of the Monte-Carlo simulation points fall inside the fail region. The triangular point referrers to the PDP point from the initial optimization at the typical process corners. The plot in Figure 13.8 also shows the mean value of the PDP and latency for the flip-flop, as well as the two points located $\pm 2\sigma$ from the mean value. According to the simulation result in Figure 13.8, the TG-MSFF suffers 30% delay variation between the -2σ latency point to $+2\sigma$ latency point, which is 2.7X higher compared to the 5-stage inverter-chain shown in Figure 13.7. This indicates that

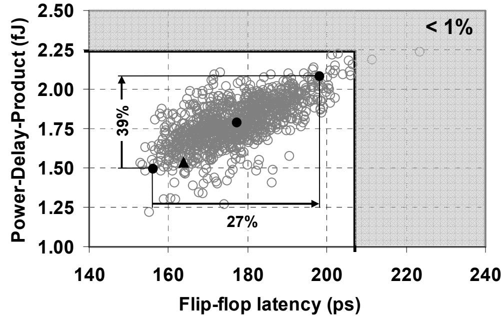


Figure 13.9: Power-performance spread for a C²MOS-MSFF

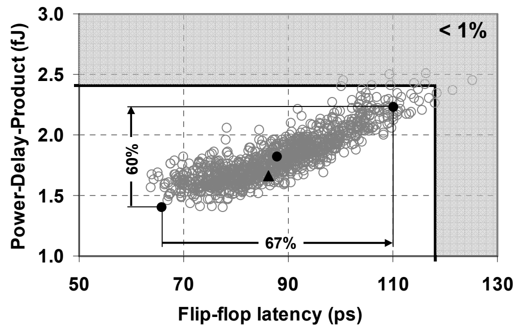


Figure 13.10: Power-performance spread for a SDFF.

in a pipeline stage with both flip-flops and logic, where the logic delay is comparable with the flip-flop delay, the performance spread in the flip-flops will be dominating. Figure 13.9 shows the corresponding Monte-Carlo simulation data for the C²MOS-MSFF, which results in similar process variation behavior as the TG-MSFF. The ±2σ-latency spread is 27%, which is 2.5X higher compared to the 5-stage inverter chain. The process variation induced PDP spread between the ±2σ points are 31% and 39% for the TG-MSFF and C²MOS-MSFF, respectively.

Figure 13.10 shows the result of the Monte-Carlo simulations for the SDFF using the optimal setup time margin of 4 ps. As shown in Figure 13.10, the

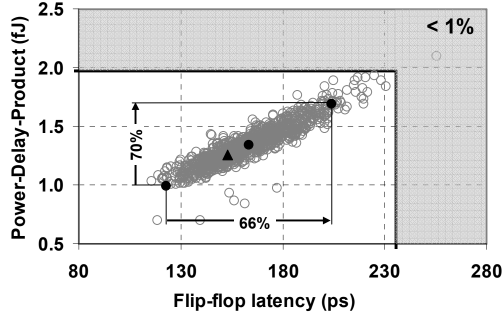


Figure 13.11: Power-performance spread for a SAFF.

difference between the optimal PDP and the average PDP is lower compared to the two MSFFs. This is because the required setup time margin for the pulsed SDFF is considerably lower than for the MSFFs. However, for the total latency spread the SDFF suffers considerably higher process variation impact. The Monte-Carlo simulation results in Figure 13.10 indicate that the SDFF at the optimal PDP point suffers a $\pm 2\sigma$-latency spread of 67%. This is 2.2X higher latency-spread compared to the TG-MSFF, and 6.1X higher compared to the 5-stage inverter chain.

Figure 13.11 shows the results of the power-delay spread for the SAFF. The small distance between the average and the optimal PDP points indicate that the SAFF requires lower additional margin for the setup time. However, just as for the SDFF the SAFF suffer from a $\pm 2\sigma$-latency spread of 66%, which is 2.2X higher than for the TG-MSFF, and 6X higher than for the 5-stage inverter chain.

13.5 Summary and Discussion

Table 13-1 summarizes the simulated results for all flip-flops. The table shows the flip-flop latency, power, and PDP from the optimization at the typical corners (indicated with *opt.*) and the average latency, power, and PDP calculated from the 1000 Monte-Carlo simulation points. Moreover, the relative standard deviation, calculated from the Monte-Carlo simulations is presented together with the required additional setup time margin for each flip-flop.

The introduced setup time margin increases the overall latency of the flip-flops, which is seen in the comparison between optimal $t_{d,DQ}$, and the mean $t_{d,DQ}$ from the statistical simulations. For both the TG-MSFF and the C^2MOS-MSFF

TABLE 13-1: SUMMARY OF FLIP-FLOP RESULTS AT OPTIMAL PDP POINT

	TG-MSFF	C^2MOS-MSFF	SDFF	SAFF
$t_{d,DQ}$ **opt. (ps)**	128.5	163.8	86.3	153.1
$t_{d,DQ}$ **mean (ps)**	140.0	177.2	88.0	163.6
$t_{d,DQ}$, σ **(%)**	6.5	6.0	12.5	12.6
t_{setup} **margin (ps)**	16	18	4	2
Power opt. (μW)	8.34	9.39	19.29	8.22
Power mean (μW)	8.77	10.08	20.72	8.21
Power, σ **(%)**	6.6	6.5	7.4	5.4
PDP opt. (fJ)	1.07	1.54	1.67	1.26
PDP mean (fJ)	1.23	1.79	1.82	1.34
PDP, σ **(%)**	6.8	8.4	11.5	13.0

the required setup time margins were fairly high, 16 ps and 18 ps, respectively. The pulsed SDFF in Figure 13.1(c) requires only 4 ps additional setup time margin. This is due to the soft-edge property of the pulsed latch, which makes the SDFF capable of absorbing some of the clock skew induced by the process variation. The required setup time margin for the SAFF is also low, which indicates that the input stage is fairly robust against variations on the data arrival time. This is not the case in a master-slave flip-flop, which relies on a robust hard edge property. However, as the results in Table 13-1 indicate the master-slave flip-flops have roughly 50% lower delay spread due to the process variations, compared to the SDFF and SAFF.

As discussed before, the design margin is in this analysis defined as the required delay and PDP values in order to cover 99% of the simulated PDP-points. Figure 13.12 shows the flip-flop latency according to the power-delay optimization at the typical corner together with the required total latency margin, which includes also the setup time margin. According to results in Figure 13.12, both master-slave flip-flops require the lowest design margin of the four flip-flops. The additional delay margin for the TG-MSFF and C^2MOS-MSFF is 30% and 26%, respectively. This corresponds well to the low spread shown in Figure 13.8 and Figure 13.9. Although the SDFF suffers from high latency variation, the required delay margin is 37%, which is only slightly higher than for the MSFFs. The largest design margin is needed for the SAFF, which requires an additional 82 ps (54%) to cover 99% of the simulated points. This is consistent with the 66% spread of the delay shown in Figure 13.11. The reason is that the SAFF relies on a symmetric cross-coupled structure, which will be more severely affected by mismatch compared to the other single-ended flip-flops.

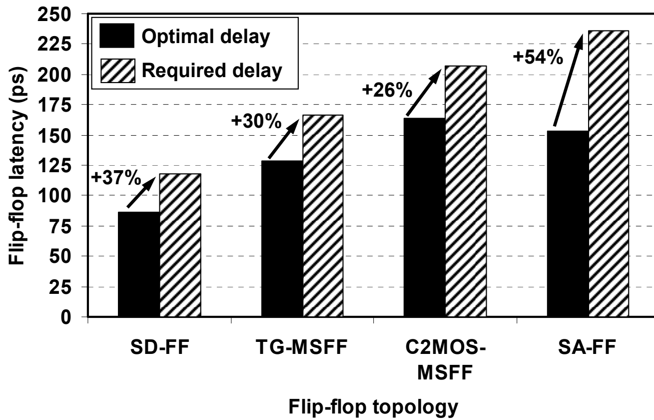


Figure 13.12: Required delay margin for 99% functional flip-flops.

13.6 Bibliography

[1]. S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, V. De, “Parameter Variation and Impact on Circuits and Microarchitecture,” in *Proceedings of Design Automation Conference*, pp. 338-342, 2003.

[2]. S. Nassif, “Delay Variability: Sources, Impacts and Trends,” in *Digest of Technical Papers International Solid-State Circuits Conference*, pp. 368-369, 2000.

[3]. K.A. Bowman, S.G. Duvall, and J.D. Meindl, “Impact on Die-to-Die and Within-Die Parameter Fluctuations on the Maximum Clock Frequency Distribution for Gigascale Integration,” in *IEEE Journal of Solid-State Circuits*, vol. 37, no. 2, pp. 183-190, Feb. 2002.

[4]. H. Masuda, S. Ohkawa, A. Kurokawa, and M. Aoki, “Challenge: Variability Characterization and Modeling for 65- to 90-nm Processes,” in *Proceedings of the Custom Integrated Circuits Conference*, pp. 593-599, 2005.

[5]. P.R. Gada, W.R. Roberts, and D. Velenis, “Effects of Parameter Variations on Timing Characteristics of Clocked Registers,” in

International Conference on Electro Information Technology, pp. 1-4, 2005.

[6]. W.R. Roberts and D. Velenis, “Effects of Process and Environmental Variations on Timing Characteristics of Clocked Registers,” in *Proceedings of the 16th ACM Great Lakes symposium on VLSI*, pp. 165-168, 2006.

[7]. H.Q. Dao, K. Nowka, V.G. Oklobdzija, “Analysis of Clocked Timing Elements for Dynamic Voltage Scaling Effects over Process Parameter Variation,” in *International Symposium on Low Power Electronics and Design*, pp. 56-59, 2001.

[8]. V. Stojanovic and V.G. Oklobdzija, “Comparative Analysis of Master-Slave Latches and Flip-Flops for High-Performance and Low-Power Systems,” in *IEEE Journal of Solid-State Circuits*, vol. 34, no. 4, pp. 536-548, 1999.

[9]. F. Klass, et al., “A New Family of Semidynamic and Dynamic Flip-Flops with Embedded Logic for High-Performance Processors,” in *IEEE Journal of Solid-State Circuits*, vol. 34, no. 5, pp. 712-716, May 1999.

Chapter 14

Process Variation Compensation Keeper

14.1 Introduction

With increasing process variation, digital circuit design is becoming increasingly harder when it comes to reaching desired frequency and robustness targets. The traditional guard-banding approach, where circuits are sized for robustness at worst-case process corners, is becoming increasingly inefficient because it result in an overall degradation of chip performance, power consumption, and/or functional yield. As a solution, reconfigurable techniques have been successfully utilized, in order to squeeze the post fabrication characteristics towards the desired target performance.

In this chapter a process compensation technique for latches and flip-flops is introduced. The internal storage nodes of the clocked registers, such as flip-flops and latches, will experience increasing robustness degradations as the spread of the leakage currents are increased. The proposed circuit technique utilizes a reconfigurable keeper topology, which reduces the delay penalties on low-leakage dies, while maintaining the worst-case leakage robustness. The design flexibility is increased because of the fact that the keeper strength can be optimized for several process corners.

14.2 Reconfigurable Keeper for Latches and Flip-Flops

For static flip-flops the robustness is largely determined by the strength of the keepers, which have large impact on the performance and power dissipation of the flip-flops. Interrupted or uninterrupted keepers are required to hold the actual data in storage nodes of static latches and flip-flops for all design corners during static noise conditions. However, sizing these keepers for worst-case leakage corners will result in a significant delay penalty for a majority of dies.

14.2.1 Circuit Concept

Adjusting the keeper strength for the actual leakage, instead of for the worst-case leakage, would result in the lowest performance penalty for a given robustness constraint. The optimal solution is to have a variable keeper that holds the internal storage node with an optimal driving strength, which is adjusted for the actual leakage currents as shown in Figure 14.1(a). However, a variable keeper with continues range of driving strength imposes a significant design complexity due to the need for routing of the analog control signals across the chip. Therefore, to reduce the complexity, a digital control circuitry is a more feasible solution. With a digitally controlled keeper, as shown in Figure 14.1(b), the driving strength can be adjusted in more than one process corner. This gives the circuit designer more flexibility to optimize the performance, power, and robustness of the latches and flip-flops.

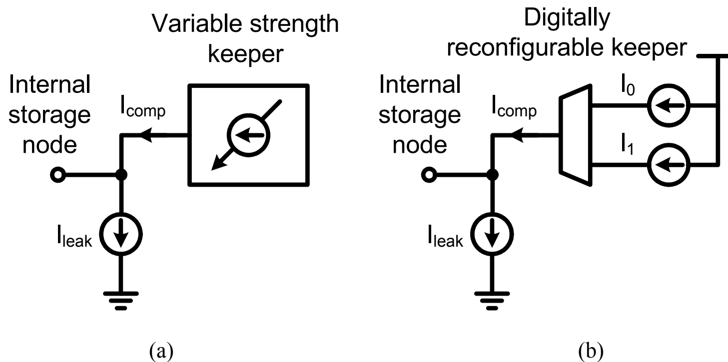


Figure 14.1: Basic concept of a variable strength keeper with (a) continuous adjustment (analog) and (b) digital reconfigurable keeper.

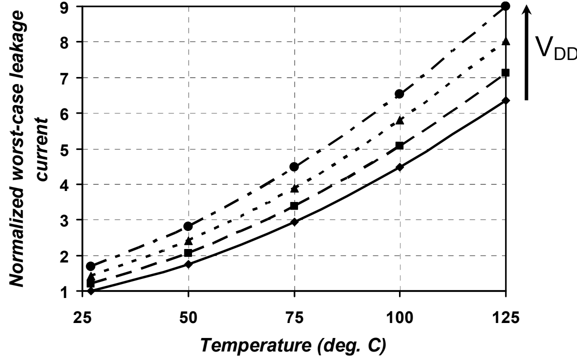


Figure 14.2: Normalized I_{OFF} for a wide MUX-latch at different temperature and supply voltage.

As a consequence of the reconfigurable keeper, the proposed technique can also be used as a burn-in keeper where the digital control signal is utilized as the burn-in enable signal. Burn-in tests are done at elevated temperature and high supply voltage conditions [9], which leads to a large increase in the leakage currents, as shown in Figure 14.2. This increases the need for optional strong keepers in order to obtain circuit functionality during the test, while a weaker keeper can be sufficient during normal operational conditions.

14.2.2 Reconfigurable Keeper for Static MUX-Latches

The proposed process variation tolerant technique is implemented for a wide MUX-latch circuit. Figure 14.3 shows the conventional implementation of an N-bit wide MUX-latch. The worst-case leakage paths, which expose the storage node X, are through the N parallel transmission-gates in the select latches. The conventional N-to-1 MUX-latch in Figure 14.3 is implemented with a weak keeper designed to obtain robustness at the worst-case leakage corner. The weak keeper is used to hold the storage node X when none of the select signals are activated. The keeper is implemented as a stacked pair of transistors in order to decrease the driving strength, and further reduce the contention with the input data.

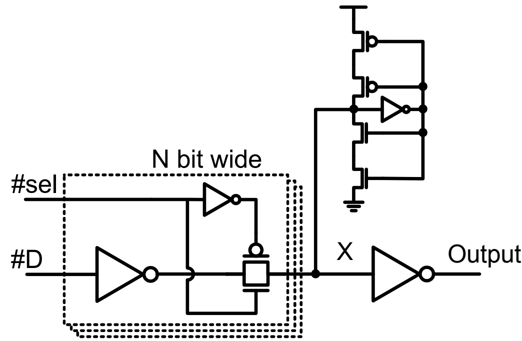


Figure 14.3: Static MUX-latch with conventional weak keeper

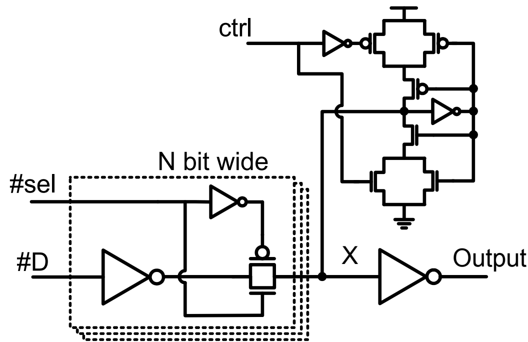


Figure 14.4: Static MUX-latch proposed reconfigurable keeper.

Figure 14.4 shows the N-bit wide MUX-latch utilizing a 1-bit digitally reconfigurable keeper. With the external control signal (*ctrl*), the keeper strength can be digitally adjusted between two separate settings. This leads to an increase of the design flexibility, because the strength can be optimized for two of the process corners instead of only being set by the worst-case corner. With the control signal (*ctrl*) low, the proposed process variation tolerant keeper is configured as a weak keeper, thus providing sufficient feedback strength in order to obtain the desired robustness at the typical leakage condition. For high-leakage conditions, like at fast process corners or during burn-in tests [9], the

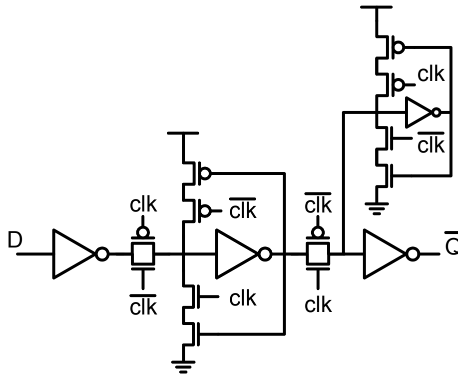


Figure 14.5: Conventional Transmission-Gate Flip-Flop with interrupted keeper.

proposed keeper is reconfigured in the strong mode with *ctrl* high. The amount of keeper strength controllability could be increased by adding more bits and parallel keeper transistor paths. This could be utilized for even wider MUX-latches or as technology scales further.

14.2.3 Reconfigurable Keeper for Static MSFFs

The increase of the design flexibility, which is provided by the reconfigurable keeper technique, could also be used to trade off delay for lower clock power in a static flip-flop. Figure 14.5 shows a conventional TG-MSFF that utilizes two interrupted keepers to hold the data while the latches are opaque. The advantage of using an interrupted keeper is that the delay is relatively insensitive to the keeper strength, which is because the keepers are interrupted when the latches are transparent. Hence, a flip-flop with interrupted keeper is less influenced by the process variation impacts on the keeper strength [1]. However, an interrupted keeper requires that the transistors are clocked, which will result in a higher load on the clock drivers. As previously discussed, a large part of the on-chip power is consumed by the clock driver. Thus any local clock load reduction will also reduce the global power dissipation. An alternative, which was discussed in section 8.3.3, is to use weak keepers to achieve low clock load for static MSFF. However, compared to a flip-flop or latch utilizing an interrupted keeper, the weak keeper approach usually leads to considerable power and performance tradeoffs due to increased contention during data switching in the flip-flop [7]. Therefore, this technique requires careful sizing of the keepers. Moreover, the keeper is constantly on, which makes it sensitive to the increasing

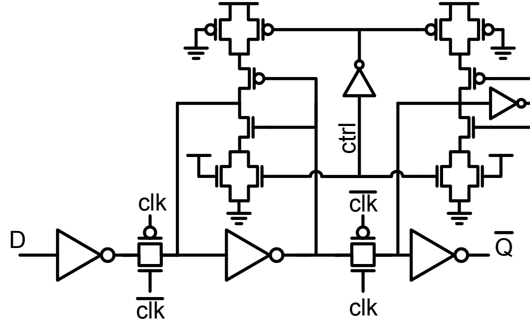


Figure 14.6: TGFF implementation with reconfigurable weak keepers.

process variations. In order to mitigate the problem with higher process variation, the proposed reconfigurable keeper technique is implemented on a weak-keeper TG-MSFF. This enables further opportunities to reduce the keeper overhead, which reduces the performance and power penalty. Therefore, the feasibility of the weak-keeper concept can be increased by providing a better process variation tolerance. Figure 14.6 shows an implementation of a reconfigurable weak-keeper TG-MSFF. The keeper circuit design approach is to reduce the flip-flop DC-robustness at the typical process corner down to the minimum required DC-robustness. Here the minimum robustness is set by the worst-case process corner. At the process corners, where the robustness is higher, the keeper strength is reduced, which lead to lower contention power and reduced latency penalty. Therefore, the majority of the manufactured dies will have better performance and lower power as well as lower clock load, compared to a worst-case design. For those manufactured dies that are leaky the keeper can be reconfigured to strong mode providing sufficient keeper strength to obtain the desired DC-robustness at the worst-case leakage corner.

14.3 Simulation Results

The proposed variation tolerant keeper technique is implemented for a 5-to-1 static MUX-latch and a conventional static TG-MSFF with ratioed keepers. Both designs are verified by simulation in an advanced 65-nm CMOS technology [10]. Power, performance, as well as noise robustness are compared.

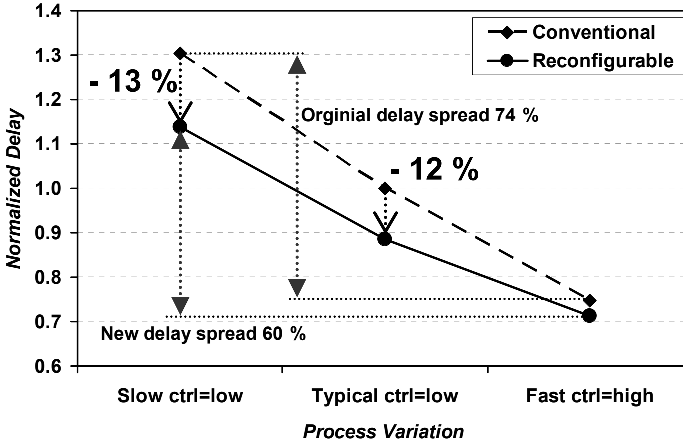


Figure 14.7: Normalized delay for the 5-to-1 MUX-latches.

The robustness is here related to the DC unity-gain noise margin (UGNM), which is defined as the voltage for which the output of the circuit equals the noise on the select or clock node [8].

14.3.1 Reconfigurable Static 5-to-1 MUX-Latch

For the MUX-latch, the delay is defined as the select-to-output delay when data is assumed to be stable on the input[1]. Power for the MUX-latch is the total average power dissipated during a complete cycle when all selects input goes high once. The fast and slow process skew refereed to in this chapter represent the worst-case process variation used for the given CMOS process [10].

Normalized delay simulation results for a 5-to-1 MUX-latch are shown in Figure 14.7. A delay reduction at the typical corner of 12% is observed for the proposed keeper compared to the conventional keeper. At the slow process skew the delay is reduced by 13%. This will result in more than 12% performance improvement for the given circuit on more than 50% of the fabricated dies. Moreover, as the design flexibility is increased more of the high leakage dies can be used without making oversized keepers, which will increase the yield of the processed circuits. As the contention at the typical process corner can be reduced by using the reconfigurable keeper a power reduction is obtained.

[1] This delay is defined the same way as the contamination delay for flip-flops, as discussed in section 3.2.2.

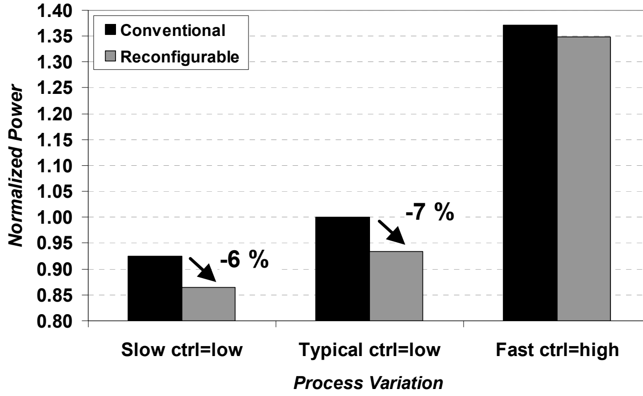


Figure 14.8: Normalized power for the 5-to-1 MUX-latches.

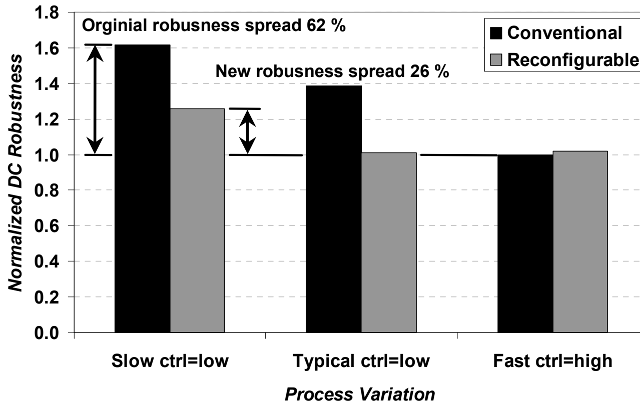


Figure 14.9: Normalized worst-case robustness for 5-to-1 MUX-latches.

Figure 14.8 shows the power simulation of the 5-to-1 MUX-latch circuit. A reduction of the total power of 7% is shown. This power reduction is partly because the load of the internal node is reduced when the reconfigurable keeper is used. This load reduction comes from the fact that the top and bottom transistors in the stacked keepers are downsized compared to the keeper in the

conventional MUX-latch, which also explains the power reduction at the fast corner. However, most of the power reduction is due to the reduced contention current when the keeper is configured in weak mode.

The DC-robustness simulation results for the MUX-latches are shown in Figure 14.9. From the diagram it is clear that the keeper strength is oversized at the typical corner. DC robustness at the typical process corner is close to 40% higher than at the fast corner. The total robustness spread of the conventional MUX-latch is more than 62%. For the circuit with the reconfigurable keeper it is shown that robustness is obtained at the typical skew as well as for the fast process skew. The keeper is thus optimized for acceptable strength at both corners, and robustness spread is decreased by 60%.

14.3.2 Reconfigurable Uninterrupted Keeper for Static Flip-Flops

Simulation results of the clock power for both the MSFF using the proposed reconfigurable keeper and the conventional MSFF using the interrupted keeper is shown in Figure 14.10. The flip-flop using the proposed reconfigurable keeper results in a 9% lower local clock power compared to the conventional MSFF using the interrupted keepers. This power reduction is a direct consequence of the reduced clock load of the proposed flip-flop. Furthermore, the 9% clock power reduction does not include the effect of the possible downsizing of the clock driver, which is enabled due to the reduction of the clock load. This means that an even larger local clock power reduction could be obtained if the clock

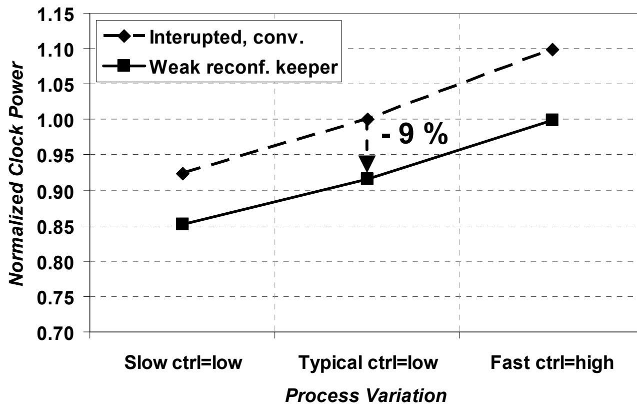


Figure 14.10: Normalized clock power for compared flip-flops.

buffers where sized accordingly. This effect is shown in section 9.4 for an experimental chip. Moreover, by utilizing the proposed reconfigurable keeper the performance penalty due to the uninterrupted keeper can be mitigated at the slow and typical process corners. Table 14-1 shows the normalized delay and DC-robustness for the two compared flip-flops. The results show that the robustness at the typical corner is reduced to match the minimum DC robustness requirements, which are set by the fast process corner. The optimized keeper strength also results in a comparable delay for the two flip-flops. Hence, more controllability of the keeper strength enables lower clock power without any performance penalty.

TABLE 14-1: NORMALIZED FLIP-FLOP ROBUSTNESS AND PERFORMANCE

	Process corner	Conv. TG-MSFF	Prop. TG-MSFF
Delay	Slow	1.26	1.24
	Typical	**1.00**	0.99
	Fast	0.78	0.80
DC-Robustness	Slow	1.35	1.17
	Typical	1.21	1.01
	Fast	**1.00**	1.01

14.4 Bibliography

[1]. S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, V. De, "Parameter Variation and Impact on Circuits and Microarchitecture," in *Proceedings of Design Automation Conference*, pp. 338-342, 2003.

[2]. C. H. Kim, K. Roy, S. Hsu, A. Alvandpour, R. K. Krishnamurthy, S. Borkar, "A Process Variation Compensation Technique for Sub-90nm Dynamic Circuits," in *Digest of Technical Papers. 2003 Symposium on VLSI Circuits*, pp. 205-206, 2003.

[3]. A. Agarwal, K. Roy, S. Hsu, R. K. Krishnamurthy, S. Borkar, "A 90nm 6GHz 128x64b 4-Read 4-Write Ported Parameter Variation Tolerant Register File," in *Digest of Technical Papers. 2004 Symposium on VLSI Circuits*, pp 386-387, 2004.

[4]. J. Tschanz, J. Kao, S. Narendra, R. Nair, D. Antoniadis, A. Chandrakasan, D. Vivek, "Adaptive Body Bias for Reducing Impacts of Die-to-Die and Within-Die Parameter Variations on Microprocessor Frequency and

Leakage," in *Digest of Technical Papers IEEE International Solid-State Circuit Conference*, pp. 422-423, 2002.

[5]. A. Alvandpour, R. Krishnamurthy, S. Borkar, A. Rahman, C. Webb, "A Burn-In Tolerant Dynamic Circuit Technique," in *Proceedings on the Custom Integrated Circuits Conferenc*e, pp. 81-84, 2002.

[6]. T. Sakurai, H. Kawaguchi, and T. Kuroda, "Low-Power CMOS Design through VTH Control and Low-Swing Circuits," in *Proceeding of the International Symposium on Low-Power Electronics and Design*, pp. 1-6, 1997.

[7]. M. Hansson and A. Alvandpour, "A Low Clock Load Conditional Flip-flop", in *Proceedings of the IEEE International SOC Conference*, pp. 169-170, 2004.

[8]. B. Chatterjee, M. Sachdev, R. Krishnamurthy, "Leakage control techniques for designing robust, low power wide-OR domino logic for sub-130nm CMOS technologies," in *Proceedings of the 5th International Symposium on Quality Electronic Design*, pp. 415-420, 2004.

[9]. A. Alvandpour, R.K. Krishnamurthy, K. Soumyanath, and S.K. Borkar, "A Sub-130-nm Conditional Keeper Technique," in *IEEE Journal of Solid-State Circuits*, vol. 37, no. 5, pp. 633-638, 2002.

[10]. P. Bai, C. Auth, S. Balakrishnan, M. Bost, R. Brain, V. Chikarmane, R. Heussner, M. Hussein, J. Hwang, D. Ingerly, R. James, J. Jeong, C. Kenyon, E. Lee, S-H. Lee, N. Lindert, M. Liu, Z. Ma, T. Marieb, A. Murthy, R. Nagisetty, S. Natarajan, J. Neirynck, A. Ott, C. Parker, J. Sebastian, R. Shaheed, S. Sivakumar, J. Steigerwald, S. Tyagi, C. Weber, B. Woolery, A. Yeoh, K. Zhang, and M. Bohr, "A 65nm Logic Technology Featuring 35nm Gate Lengths, Enhanced Channel Strain, 8 Cu Interconnect Layers, Low-k ILD and 0.57 µm2 SRAM Cell," in *Technical Digest IEEE International Electron Device Meeting*, pp. 657-660, 2004.

Chapter 15

Conclusions and Future Work

15.1 Conclusions

15.1.1 Process Variation Impact on Flip-Flop Power-Performance

A comparative analysis of the process variation impact on four common flip-flop topologies is presented [1]. Among the four flip-flops the MSFF based on either transmission-gates or clocked CMOS gates shows the best tolerance for process uncertainties, with up to 30% variation between the $\pm 2\sigma$ points. Compared to the flip-flop performance optimized at typical corners an additional 25-30% latency margin is needed in order to achieve 99% functionality over the process variation in a 90-nm CMOS technology. The SAFF suffers from matching requirements in the sense-amplifier, which is causing 2.2X higher delay spread compared to the MSFF. Hence the SAFF is more sensitive to mismatch and process variation, and requires 54% latency margin to achieve 99% functionality over the process variation design space. Finally, the high-performance pulsed SDFF requires only a small additional setup time margin in order get good functionality. However, still requires 37% latency margin in order to reach 99% functionally.

15.1.2 Reconfigurable Process Variation Tolerant Keeper

A technique to increase the process-variation tolerance in static sequential circuits by using a reconfigurable uninterrupted keeper is presented [2]. The

proposed technique is utilized to reduce the sensitivity to process variations in high-performance static clocked registers. The proposed reconfigurable compensation technique introduces additional flexibility to the designer by making it possible to optimize keeper strengths at two or more process corners. The reconfigurable keeper could also be utilized during any burn-in tests as a strong burn-in keeper.

The proposed reconfigurable keeper technique is implemented on wide transmission-gate MUX-latches. Simulation in a 65-nm CMOS process [3] have shown 12% delay improvement by using the proposed reconfigurable keeper technique compared to conventional weak-keeper designs. Moreover, the reduced keeper strength for the slower dies resulted in 7% power reduction of the studied MUX-latch circuit. Continuing process scaling will result in larger process spreads and higher leakage currents, thus increasing the feasibility of the proposed technique.

An alternative to performance improvement is to trade-off the delay reduction with clock load. The proposed reconfigurable keeper technique is implemented on a TGMS-FF. Results comparing to a conventional interrupted transmission gate flip-flop shows a local clock load reduction of 9% without any performance penalties. This could be utilized to down size the clock drivers to these flip-flops leading to considerable clock power savings on the entire chip.

15.2 Future Work

Because of the fact that process variability is increasing as a consequence of the scaling of the CMOS technologies, further research on circuit techniques to reduce the effect of the parameter variability is needed. An interesting continuation of the proposed reconfigurable technique would be to incorporate the leakage compensation keeper, presented in Chapter 9, as a minimum strength keeper, and then further increase the strength with a multi-bit digital control signal. Furthermore, an implementation of a synchronous system utilizing the proposed reconfigurable keeper on the clocked registers would be interesting in order to verify the implication of the reduced clock load on the global clock power dissipation.

15.3 Bibliography

[1]. M. Hansson and A. Alvandpour, "Comparative Analysis of Process Variation Impact on Flip-Flop Power-Performance," in *IEEE International Symposium on Circuits and Systems,* pp. 3744-3747, 2007.

[2]. M. Hansson, A. Alvandpour, S.K. Hsu, and R.K. Krishnamurthy, "A Process Variation Tolerant Technique for sub-70 nm Latches and Flip-Flops," in *Proceedings of the 23rd IEEE NORCHIP Conference*, pp. 149-152, 2005.

[3]. P. Bai, C. Auth, S. Balakrishnan, M. Bost, R. Brain, V. Chikarmane, R. Heussner, M. Hussein, J. Hwang, D. Ingerly, R. James, J. Jeong, C. Kenyon, E. Lee, S-H. Lee, N. Lindert, M. Liu, Z. Ma, T. Marieb, A. Murthy, R. Nagisetty, S. Natarajan, J. Neirynck, A. Ott, C. Parker, J. Sebastian, R. Shaheed, S. Sivakumar, J. Steigerwald, S. Tyagi, C. Weber, B. Woolery, A. Yeoh, K. Zhang, and M. Bohr, "A 65nm Logic Technology Featuring 35nm Gate Lengths, Enhanced Channel Strain, 8 Cu Interconnect Layers, Low-k ILD and 0.57 µm2 SRAM Cell," in *Technical Digest IEEE International Electron Device Meeting*, pp. 657 – 660, 2004.