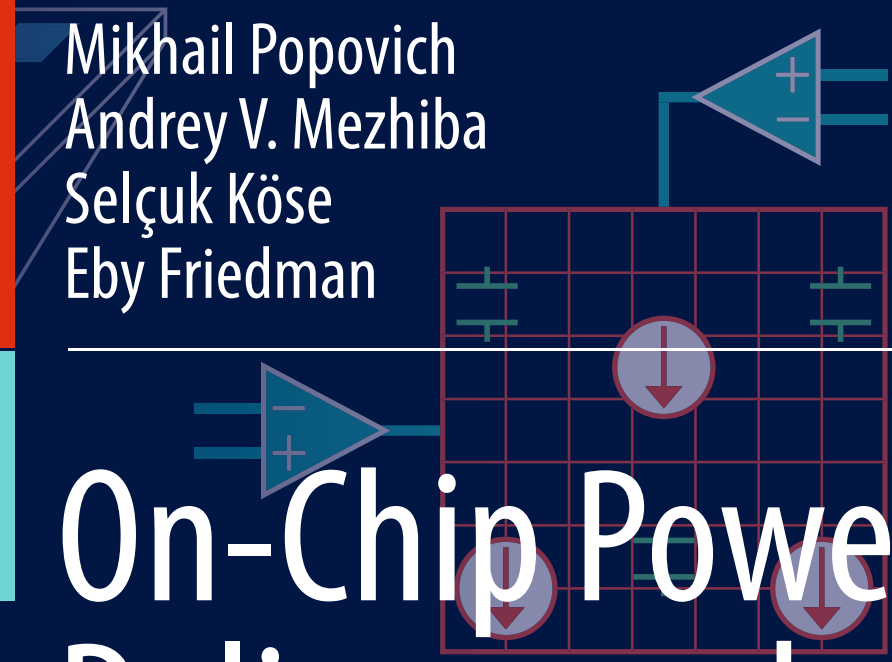
Inna P. Vaisband
Renatas Jakushokas
Mikhail Popovich
Andrey V. Mezhiba
Selçuk Köse
Eby Friedman



On-Chip Power Delivery and Management

Fourth Edition

Springer

On-Chip Power Delivery and Management

Inna P.-Vaisband • Renatas Jakushokas
Mikhail Popovich • Andrey V. Mezhiba
Selçuk Köse • Eby G. Friedman

On-Chip Power Delivery and Management

Fourth Edition

Springer

Inna P.-Vaisband
University of Rochester
Rochester, NY, USA

Renatas Jakushokas
Qualcomm Corporation
San Diego, CA, USA

Mikhail Popovich
Qualcomm Corporation
San Marcos, CA, USA

Andrey V. Mezhiba
Intel Corporation
Hillsboro, OR, USA

Selçuk Köse
University of South Florida
Tampa, NY, USA

Eby G. Friedman
University of Rochester
Rochester, USA

ISBN 978-3-319-29393-6 ISBN 978-3-319-29395-0 (eBook)
DOI 10.1007/978-3-319-29395-0

Library of Congress Control Number: 2016936678

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

To Sasha and Eva
To Victoria and Daniel
To Oksana, Elizabeth, and JulieAnn
To Elizabeth
To the memory of my late father, Nurettin Köse
To Laurie, Joseph, Shlomit, and Samuel

Preface to the Fourth Edition

Novel market segments such as intelligent transportation, revolutionary health care, sophisticated security systems, and smart energy have recently emerged, requiring increasingly diverse functionality such as RF circuits, power control, passive components, sensors/actuators, biochips, optical communication, and microelectromechanical devices. Integration of these non-digital functionalities at the board-level into system platforms such as systems-in-package (SiP), systems-on-chip (SoC), and three-dimensional (3-D) systems is a primary near- and long-term challenge of the semiconductor industry. The delivery and management of high-quality, highly efficient power have become primary design issues in these functionally diverse systems. Integrated in-package and distributed on-chip power delivery is currently under development across a broad spectrum of applications; the power delivery design process, however, is currently dominated by ad hoc approaches.

The lack of methodologies, architectures, and circuits for scalable on-chip power delivery and management is at the forefront of current heterogeneous system design issues. The objective of this book is to describe the many short- and long-term challenges of high-performance power delivery systems, provide insight and intuition into the behavior and design of next-generation power delivery systems, and suggest design solutions while providing a framework for addressing power objectives at the architectural, methodology, and circuit levels.

This book is based on the body of research carried out by the authors of previous editions of this book from 2001 to 2011. The first edition of the book, titled *Power Distribution Networks in High Speed Integrated Circuits*, was published in 2004 by Andrey V. Mezhiba and Eby G. Friedman. This first book focused on on-chip distribution networks, including electrical characteristics, relevant impedance phenomenon, and related design trade-offs. On-chip distributed power delivery, at that time an innovative paradigm shift in power delivery, was also introduced in the book. As the concept of integrated power delivery evolved, the important topic of on-chip decoupling capacitance was added to the book, which was released in 2008 with a new title, *Power Distribution Networks with On-Chip Decoupling Capacitors* by Mikhail Popovich, Andrey V. Mezhiba, and Eby G. Friedman. Later, this book was revised by Renatas Jakushokas, Mikhail Popovich, Andrey V. Mezhiba,

Selçuk Köse, and Eby G. Friedman to address emerging design and analysis challenges in on-chip power networks. This last edition was published with an identical title in 2011. Since the first book was published in 2004, the issue of power delivery has greatly evolved. The concept of on-chip distributed power delivery has been recognized as an important cornerstone to high-performance integrated circuits. A number of ultrasmall on-chip power supplies to support this on-chip focus have also been demonstrated.

While on-chip power integration has become a primary objective for system integration, research has remained focused on developing compact and efficient power supplies, lacking a methodology to effectively integrate and manage in-package and on-chip power delivery systems. The challenge has become greater as the diversity of modern systems increases, and dynamic voltage scaling (DVS) and dynamic voltage and frequency scaling (DVFS) become a part of the power management process. Hundreds of on-chip power domains with tens of different voltage levels have recently been reported, and thousand-core ICs are being considered. Scalable power delivery systems and the granularity of power management in DVS/DVFS multicore systems are limited by existing ad hoc approaches. To cope with this increasing design complexity and the quality and system-wide efficiency challenges of next-generation power delivery systems, enhanced methodologies to design and analyze scalable, hierarchical power management and delivery systems with fine granularity of dynamically controllable voltage levels are necessary. Updating the vision of on-chip power delivery networks, traditionally viewed as a passive network, is the primary purpose for publishing a new (fourth) edition of this book. Emphasis is placed on complex and scalable power delivery systems, system-wide efficiency, quality of power, and intelligent, real-time, fine-grain local power management. A framework that addresses various power objectives at the architectural, methodology, and circuit levels is described, providing a general solution for existing and emerging power delivery challenges and techniques. This book, titled *On-Chip Power Delivery and Management*, is authored by Inna P.-Vaisband, Renatas Jakushokas, Mikhail Popovich, Andrey V. Mezhiba, Selçuk Köse, and Eby G. Friedman as the fourth edition of this series of books.

The chapters of the book are now separated into eight parts. Power networks, inductive properties, electromigration, and decoupling capacitance within integrated circuits are described in Part I (Chaps. 1, 2, 3, 4, 5, and 6). In Part II (Chaps. 7, 8, 9, and 10), the design of on-chip power distribution networks and power supplies is discussed. Circuits for on-chip power delivery and management and integrated power delivery systems are described in Part IV (Chaps. 17, 18, 19, and 20). Closed-form expressions for power grid analysis, modeling and optimization of power networks, and the codesign of power supplies are presented in Part V (Chaps. 21, 22, 23, 24, 25, 26, and 27). Since noise within the power grid is a primary design constraint, this issue is reviewed in Part VI (Chaps. 28, 29, 30, 31, 32, 33, and 34). Multilayer power distribution networks are the focus of Part VII (Chaps. 35, 36, 37, 38, and 39). In Part III (Chaps. 12, 13, 14, and 15), the issue of placing on-chip decoupling capacitors is discussed. In Part VIII (Chaps. 40, 41, 42, and 43), multiple power supply systems are described. The focus of this part is on those integrated

circuits where multiple on-chip power supplies are required. In Part IX, some concluding comments, the appendices, and additional information are provided.

This revised and updated material is based on recent research by Inna P.-Vaisband developed between 2009 and 2015 at the University of Rochester during her doctoral studies under the supervision of Prof. Eby G. Friedman. The new chapters focus on design complexity, system scalability, and system-wide optimization of power delivery and management systems. The concept of intelligent power delivery is introduced, and a framework for on-chip power delivery and management is described that provides local power control and real-time management for sharing energy resources.

The book covers a wide spectrum of issues related to on-chip power networks and systems. The authors believe that this revised edition provides the latest information on a dynamic and highly significant topic of primary importance to both the industrial and academic research and development communities.

Acknowledgments

The authors would like to thank Chuck Glaser for his sincere encouragement and enthusiastic support of the publication of this book. The authors would also like to thank Burt Price and Jeff Fischer from Qualcomm and Avinoam Kolodny from Technion – Israel Institute of Technology for their collaboration and support.

The research described in this book has been supported in part by the Binational Science Foundation under grant no. 2012139; the National Science Foundation under grant nos. CCF-1329374, CCF-1526466, and CNS-1548078; the IARPA under grant no. W911NF-14-C-0089 and by grants from Qualcomm, Cisco Systems, and Intel.

Rochester, USA — Inna P.-Vaisband
San Diego, USA — Renatas Jakushokas
San Diego, USA — Mikhail Popovich
Hillsboro, USA — Andrey V. Mezhiba
Tampa, USA — Selçuk Köse
Rochester, USA — Eby G. Friedman
December 2015

Preface to the Third Edition

The first planar circuit was fabricated by Fairchild Semiconductor Company in 1959. Since then, the evolution of the integrated circuit has progressed, now providing billions of transistors on a single monolithic substrate. These integrated circuits are an integral and nearly essential part of our modern life. The power consumed by a typical $20 \times 20\,\text{mm}^2$ microprocessor is in the range of several hundreds of watts, making integrated circuits one of the highest power consumers per unit area. With such a high rate of power consumption, the problem of delivering power on-chip has become a fundamental issue. The focus of this book is on distributing power within high-performance integrated circuits.

In 2004, the book titled *Power Distribution Networks in High Speed Integrated Circuits* by A. V. Mezhiba and E. G. Friedman was published to describe, for the first time in book form, the design and analysis of power distribution networks within integrated circuits. The book described different aspects of on-chip power distribution networks, starting with a general introduction and ending with a discussion of various design trade-offs in on-chip power distribution networks. Later, the important and highly relevant topic of decoupling capacitance was added to this book. Due to the significant change in size and focus, the book was released in 2008 as a new first edition with a new title, *Power Distribution Networks with On-Chip Decoupling Capacitors* by M. Popovich, A. V. Mezhiba, and E. G. Friedman. Since this revised book was published, new design and analysis challenges in on-chip power networks have emerged.

The rapidly evolving field of integrated circuits has required an innovative perspective on on-chip power generation and distribution, shifting the authors' research focus to these new challenges. Updating knowledge on chip-based power distribution networks is the primary purpose for publishing a second edition of *Power Distribution Networks with On-Chip Decoupling Capacitors*. Focus is placed on complexity issues related to power distribution networks, developing novel design methodologies and providing solutions for specific design and analysis issues. In this second edition, the authors have revised and updated previously

published chapters and added four new chapters to the book. This second edition has also been partitioned into subareas (called parts) to provide a more intuitive flow to the reader.

The organization of the book is now separated into seven parts. A general background, introducing power networks, inductive properties, electromigration, and decoupling capacitance within integrated circuits, is provided in Part I (Chaps. 1, 2, 3, 4, 5, 6, and 7). In Part II (Chaps. 8, 9, 10, 11, and 12), the design of on-chip power distribution networks is discussed. Since noise within the power grid is a primary design constraint, this issue is reviewed in Part III (Chaps. 13, 14, 15, 16, 17, 18, and 19). In Part IV (Chaps. 20, 21, 22, and 23), the primary issue of placing on-chip decoupling capacitors is discussed. Multilayer power distribution networks are the focus of Part V (Chaps. 24, 25, and 26). In Part VI (Chaps. 27, 28, 29, and 30), multiple power supply systems are described. The focus of this part is on those integrated circuits where several on-chip power supplies are required. In Part VII, some concluding comments, the appendices, and additional information are provided.

This revised and updated material is based on recent research by Renatas Jakushokas and Selçuk Köse developed between 2005 and 2010 at the University of Rochester during their doctoral studies under the supervision of Prof. Eby G. Friedman. The emphasis of these newly added chapters is on the complexity of power distribution networks. Models for commonly used meshed and interdigitated interconnect structures are described. These models can be used to accurately and efficiently estimate the resistance and inductance of complex power distribution networks. With these models, on-chip power networks can be efficiently analyzed and designed, greatly enhancing the performance of the overall integrated circuit.

Acknowledgments

The authors would like to thank Charles Glaser from Springer for making this book a reality. The authors are also grateful to Dr. Sankar Basu of the National Science Foundation for his support over many years. We are sincerely thankful to Dr. Emre Salman for endless conversations and discussions, leading to novel research ideas and solutions.

This research has been supported in part by the National Science Foundation under grant nos. CCF-0541206, CCF-0811317, and CCF-0829915; grants from the New York State Office of Science, Technology and Academic Research to the Center for Advanced Technology in Electronic Imaging Systems; and grants from Intel Corporation, Eastman Kodak Company, and Freescale Semiconductor Corporation.

Rochester, USA — Renatas Jakushokas
San Diego, USA — Mikhail Popovich
Hillsboro, USA — Andrey V. Mezhiba
Rochester, USA — Selçuk Köse
Rochester, USA — Eby G. Friedman
September 2010

Preface to the Second Edition

The purpose of this book is to provide insight and intuition into the behavior and design of power distribution systems with decoupling capacitors for application to high-speed integrated circuits. The primary objectives are threefold. First is to describe the impedance characteristics of the overall power distribution system, from the voltage regulator through the printed circuit board and package onto the integrated circuit to the power terminals of the on-chip circuitry. The second objective of this book is to discuss the inductive characteristics of on-chip power distribution grids and the related circuit behavior of these structures. Finally, the third primary objective is to present design methodologies for efficiently placing on-chip decoupling capacitors in nanoscale integrated circuits.

Technology scaling has been the primary driver behind the amazing performance improvement of integrated circuits over the past several decades. The speed and integration density of integrated circuits have dramatically improved. These performance gains, however, have made distributing power to the on-chip circuitry a difficult task. Highly dense circuitry operating at high clock speeds has increased the distributed current to many tens of amperes, while the noise margin of the power supply has shrunk consistent with decreasing power supply levels. These trends have elevated the problems of power distribution and allocation of the on-chip decoupling capacitors to the forefront of several challenges in developing high-performance integrated circuits.

This book is based on the body of research carried out by Mikhail Popovich from 2001 to 2007 and Andrey V. Mezhiba from 1998 to 2003 at the University of Rochester during their doctoral studies under the supervision of Professor Eby G. Friedman. It is apparent to the authors that although various aspects of the power distribution problem have been addressed in numerous research publications, no text exists that provides a unified focus on power distribution systems and related design problems. Furthermore, the placement of on-chip decoupling capacitors has traditionally been treated as an algorithmic oriented problem. A more electrical perspective, both circuit models and design techniques, has been used in this

book for presenting how to efficiently allocate on-chip decoupling capacitors. The fundamental objective of this book is to provide a broad and cohesive treatment of these subjects.

Another consequence of higher speed and greater integration density has been the emergence of inductance as a significant factor in the behavior of on-chip global interconnect structures. Once clock frequencies exceeded several hundred megahertz, incorporating on-chip inductance into the circuit analysis process became necessary to accurately describe signal delays and waveform characteristics. Although on-chip decoupling capacitors attenuate high-frequency signals in power distribution networks, the inductance of the on-chip power interconnect is expected to become a significant factor in multi-gigahertz digital circuits. An important objective of this book, therefore, is to clarify the effects of inductance on the impedance characteristics of on-chip power distribution grids and to provide an understanding of related circuit behavior.

The organization of the book is consistent with these primary goals. The first eight chapters provide a general description of distributing power in integrated circuits with decoupling capacitors. The challenges of power distribution are introduced and the principles of designing power distribution systems are described. A general background to decoupling capacitors is presented followed by a discussion of the use of a hierarchy of capacitors to improve the impedance characteristics of the power network. An overview of related phenomena, such as inductance and electromigration, is also presented in a tutorial style. The following seven chapters are dedicated to the impedance characteristics of on-chip power distribution networks. The effect of the interconnect inductance on the impedance characteristics of on-chip power distribution networks is evaluated. The implications of these impedance characteristics on circuit behavior are also discussed. On-chip power distribution grids are described, exploiting multiple power supply voltages and multiple grounds. Techniques and algorithms for the computer-aided design and analysis of power distribution networks are also described; however, the emphasis of the book is on developing circuit intuition and understanding the electrical principles that govern the design and operation of power distribution systems. The remaining five chapters focus on the design of a system of on-chip decoupling capacitors. Methodologies for designing power distribution grids with on-chip decoupling capacitors are also presented. These techniques provide a solution for determining the location and magnitude of the on-chip decoupling capacitance to mitigate on-chip voltage fluctuations.

Acknowledgments

The authors would like to thank Alex Greene and Katelyn Stanne from Springer for their support and assistance. We are particularly thankful to Bill Joyner and Dale Edwards from the Semiconductor Research Corporation and Marie Burnham, Olin Hartin, and Radu Secareanu from Freescale Semiconductor Corporation for

their continued support of the research project that culminated in this book. The authors would also like to thank Emre Salman for his corrections and suggestions on improving the quality of the book. Finally, we are grateful to Michael Sotman and Avinoam Kolodny from Technion – Israel Institute of Technology for their collaboration and support.

The original research work presented in this book was made possible in part by the Semiconductor Research Corporation under contract nos. 99–TJ–687 and 2004–TJ–1207; the DARPA/ITO under AFRL contract F29601–00–K–0182; the National Science Foundation under contract nos. CCR–0304574 and CCF–0541206; grants from the New York State Office of Science, Technology and Academic Research to the Center for Advanced Technology in Electronic Imaging Systems; and by grants from Xerox Corporation, IBM Corporation, Lucent Technologies Corporation, Intel Corporation, Eastman Kodak Company, Intrinsix Corporation, Manhattan Routing, and Freescale Semiconductor Corporation.

Rochester, USA — Mikhail Popovich
Rochester, USA — Eby G. Friedman
Hillsboro, USA — Andrey V. Mezhiba
June 2007

Preface to the First Edition

The primary purpose of this book is to provide insight and intuition into the behavior and design of power distribution systems for high-speed integrated circuits. The objective is twofold. First is to describe the impedance characteristics of the overall power distribution system, from the voltage regulator through the printed circuit board and package onto the integrated circuit to the power terminals of the on-chip circuitry. The second objective of this book is to discuss the inductive characteristics of on-chip power distribution grids and the related circuit behavior of these structures.

Technology scaling has been the primary driver behind improving the performance characteristics of integrated circuits over the past several decades. The speed and integration density of integrated circuits have dramatically improved. These performance gains, however, have made distributing power to the on-chip circuitry a difficult task. Highly dense circuitry operating at high clock speeds has increased the distributed current to tens of amperes, while the noise margin of the power supply has been shrunk consistent with decreasing power supply levels. These trends have elevated the problem of power distribution to the forefront of challenges in developing high-performance integrated circuits.

This monograph is based on the body of research carried out by Andrey V. Mezhiba from 1998 to 2003 at the University of Rochester during his doctoral study under the supervision of Professor Eby G. Friedman. It has become apparent to the authors during this period that although various aspects of the power distribution problem have been addressed in numerous research publications, no text provides a unified description of power distribution systems and related design problems. The primary objective of this book is therefore to provide a broad and cohesive, albeit not comprehensive, treatment of this subject.

Another consequence of higher speed and greater integration density has been the emergence of inductance as a significant factor in the behavior of on-chip global interconnect structures. Once clock frequencies exceeded several hundred megahertz, incorporating on-chip line inductance into the circuit analysis process became necessary to accurately describe signal delays and rise times. Although on-chip decoupling capacitors attenuate high-frequency signals in power distribution

networks, the inductance of the on-chip power interconnect is expected to become a significant factor in multi-gigahertz digital circuits. Another objective of this book, therefore, is to describe the effects of inductance on the impedance characteristics of on-chip power distribution grids and to develop an understanding of related circuit behavior.

The organization of the book is consistent with these primary goals. The first eight chapters provide a general description of distributing power in integrated circuits. The challenges of power distribution are introduced and the principles of designing power distribution systems are described. A hierarchy of decoupling capacitors used to improve the impedance characteristics is reviewed. An overview of related phenomena, such as inductance and electromigration, is also presented in a tutorial style. The following six chapters are dedicated to the impedance characteristics of on-chip power distribution networks. The effect of the interconnect inductance on the impedance characteristics of on-chip power distribution networks is evaluated. The implications of these impedance characteristics for the circuit behavior are also discussed. Techniques and algorithms for the computer-aided design and analysis of power distribution networks are also described; however, the emphasis of the book is on developing circuit intuition and understanding the principles that govern the design and operation of power distribution systems.

Acknowledgments

The authors would like to thank Michael Hackett from Kluwer Academic Publishers for his support and assistance. We are particularly thankful to Bill Joyner from the Semiconductor Research Corporation for his continuing support of the research project that culminated in this book. Finally, we are sincerely grateful to Bilyana Boyadjieva for designing the cover of the book.

The original research work presented in this monograph was made possible in part by the Semiconductor Research Corporation under contract no. 99-TJ-687; the DARPA/ITO under AFRL contract F29601-00-K-0182; grants from the New York State Office of Science, Technology and Academic Research to the Center for Advanced Technology-Electronic Imaging Systems and the Microelectronics Design Center; and grants from Xerox Corporation, IBM Corporation, Intel Corporation, Lucent Technologies Corporation, and Eastman Kodak Company.

Rochester, USA

Andrey V. Mezhiba
Eby G. Friedman

Contents

About the Authors

Inna Vaisband received the Bachelor of Science degree in computer engineering and the Master of Science degree in electrical engineering from the Technion-Israel Institute of Technology, Haifa, Israel in, respectively, 2006 and 2009, and the Ph.D. degree in electrical engineering from the University of Rochester, Rochester, New York in 2015.

She is currently a post-doctoral researcher with the Department of Electrical Engineering, University of Rochester, Rochester, New York. Between 2003 and 2009, she held a variety of software and hardware R&D positions at Tower Semiconductor Ltd., G-Connect Ltd., and IBM Ltd., all in Israel. In summer 2012, Inna was a Visiting Researcher at Stanford University. Her current research interests include the analysis and design of high performance integrated circuits, analog circuits, and on-chip power delivery and management. Dr. Vaisband is an Associate Editor of the *Microelectronics Journal*.

Renatas Jakushokas was born in Kaunas, Lithuania. He received the B.Sc. degree in electrical engineering from Ort-Braude College, Karmiel, Israel in 2005, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Rochester, Rochester, New York in, respectively, 2007 and 2011.

He was previously an intern at Intrinsix Corporation, Fairport, New York, in 2006, working on Sigma Delta ADCs. In the summer of 2007, he interned with Eastman Kodak Company, Rochester, New York, where he designed a high speed and precision comparator for high performance ADCs. During the summer of 2008, he was with Freescale Semiconductor Corporation, Tempe, Arizona where he worked on developing a noise coupling estimation calculator, supporting the efficient evaluation of diverse substrate isolation techniques. In 2011, Renatas joined Qualcomm Inc., where he works on custom high speed circuit design, power and signal integrity, power distribution networks, development/optimization/placement of on-die decoupling capacitors, and power estimation/correlation/optimization.

He currently holds a US patent and is the author of additional disclosed patents. He has authored a book and published over ten journal and conference papers. Dr. Jakushokas participates in conference committees and is currently serving as an editor for the *Microelectronics Journal*. His research interests are in the areas of power distribution, noise evaluation, signal and power integrity, substrate modeling/analysis, and optimization techniques for high performance integrated circuit design.

Mikhail Popovich was born in Izhevsk, Russia in 1975. He received the B.S. degree in electrical engineering from Izhevsk State Technical University, Izhevsk, Russia in 1998, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Rochester, Rochester, New York in, respectively, 2002 and 2007.

He was an intern at Freescale Semiconductor Corporation, Tempe, Arizona in the summer of 2005, where he worked on signal integrity in RF and mixed-signal ICs and developed design techniques and methodologies

for placing distributed on-chip decoupling capacitors. His professional experience also includes characterization of substrate and interconnect crosstalk noise in CMOS imaging circuits for Eastman Kodak Company, Rochester, New York. He has authored several conference and journal papers in the areas of power distribution networks in CMOS VLSI circuits, placement of on-chip decoupling capacitors, and the inductive properties of on-chip interconnect. He holds several US patents. In 2007, Mikhail joined Qualcomm Corporation, where he works on power distribution networks, power and signal integrity, low power techniques, and interconnect design including on-chip inductive effects, noise coupling, and placement of on-chip decoupling capacitors.

Mr. Popovich received the Best Student Paper Award at the ACM Great Lakes Symposium on VLSI in 2005, and the GRC Inventor Recognition Award from the Semiconductor Research Corporation in 2007.

Andrey V. Mezhiba graduated from the Moscow Institute of Physics and Technology in 1996 with a Diploma in Physics. He continued his studies at the University of Rochester where he received the Ph.D. degree in electrical and computer engineering in 2004. Andrey authored several conference and journal papers in the areas of power distribution networks, on-chip inductance, circuit coupling, and signal integrity; he holds several patents. Andrey is currently with Intel Corporation working on phase-locked loops and other mixed-signal circuits in advanced CMOS technologies.

Selçuk Köse received the B.S. degree in electrical and electronics engineering from Bilkent University, Ankara, Turkey in 2006, and the M.S. and Ph.D. degrees in electrical engineering from the University of Rochester, Rochester, NY, respectively, in 2008 and 2012.

He is currently an Assistant Professor at the Department of Electrical Engineering, University of South Florida, Tampa, Florida. He was a part-time engineer at the Scientific and Technological Research Council (TÜBİTAK), Ankara, Turkey in 2006. He was with the Central Technology and Special Cir-

cuits Team in the enterprise microprocessor division of Intel Corporation, Santa Clara, California in 2007 and 2008. He was with the RF, Analog, and Sensor Group, Freescale Semiconductor, Tempe, Arizona in 2010. His current research interests include the analysis and design of high performance integrated circuits, on-chip DC-DC voltage converters, and interconnect related issues with specific emphasis on the design, analysis, and management of on-chip power delivery networks, 3-D integration, and hardware security.

Dr. Köse received the National Science Foundation CAREER Award in 2014, University of South Florida College of Engineering Outstanding Junior Research Achievement Award in 2014, and Cisco Research Award in 2015. He is currently serving on the editorial board of the *Journal of Circuits, Systems, and Computers* and the *Microelectronics Journal*. He is a member of the technical program committee of a number of conferences.

Eby G. Friedman received the B.S. degree from Lafayette College in 1979, and the M.S. and Ph.D. degrees from the University of California, Irvine, in 1981 and 1989, respectively, all in electrical engineering.

From 1979 to 1991, he was with Hughes Aircraft Company, rising to the position of manager of the Signal Processing Design and Test Department, responsible for the design and test of high performance digital and analog ICs. He has been with the Department of Electrical and Computer Engineering at the University of Rochester since 1991, where he is a Distinguished Professor and the Director of the High Performance VLSI/IC Design and Analysis Laboratory. He is also a Visiting Professor at the Technion—Israel Institute of Technology. His current research and teaching interests are in high performance synchronous digital and mixed-signal microelectronic design and analysis with application to high speed portable processors and low power wireless communications.

He is the author of almost 500 papers and book chapters, 13 patents, and the author or editor of 16 books in the fields of high speed and low power CMOS design techniques, 3-D integration, high speed interconnect, and the theory and application of synchronous clock and power distribution networks. Dr. Friedman is the Editor-in-Chief of the *Microelectronics Journal*, a Member of the editorial boards of the *Analog Integrated Circuits and Signal Processing*, *Journal of Low Power Electronics*, and *Journal of Low Power Electronics and Applications*, and

a Member of the technical program committee of numerous conferences. He previously was the Editor-in-Chief and Chair of the steering committee of the *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, the Regional Editor of the *Journal of Circuits, Systems and Computers*, a Member of the editorial board of the *Proceedings of the IEEE*, *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, and *Journal of Signal Processing Systems*, a Member of the Circuits and Systems (CAS) Society Board of Governors, Program and Technical chair of several IEEE conferences, and a recipient of the IEEE Circuits and Systems 2013 Charles A. Desoer Technical Achievement Award, a University of Rochester Graduate Teaching Award, a College of Engineering Teaching Excellence Award, and is a member of the University of California, Irvine Engineering Hall of Fame. Dr. Friedman is a Senior Fulbright Fellow and an IEEE Fellow.

Part I
General Background

A general background of on-chip power distribution networks is described in Part I. These chapters familiarize the reader with topics relevant to power supply networks. Different aspects of inductance and inductive loops are also reviewed in this part. These chapters provide sufficient background to enable the reader to follow the remainder of the book. Greater detail describing each chapter in this part is provided below.

An introduction to the evolution of integrated circuits and problems related to power distribution are presented in Chap. 1. Technology trends describing microprocessor transistor count, clock frequency, and power are summarized in this chapter. The important issue of noise within power distributions networks is also discussed.

The inductive properties of interconnect are described in Chap. 2. Different methods of characterizing the inductance of complex interconnect systems as well as limitations of these methods are also discussed. The concept of a partial inductance is reviewed. This concept is helpful in describing the inductive properties of complex structures. The distinction between the absolute inductance and the inductive behavior is emphasized and the relationship between these concepts is discussed.

The inductive properties of interconnect structures where current flows in long loops are described in Chap. 3. The variation of the partial inductance with line length is compared to the loop inductance. The inductance of a long current loop increases linearly with loop length. Similarly, the effective inductance of several long loops connected in parallel decreases inversely linearly with the number of loops. Exploiting these properties to enhance the efficiency of the circuit analysis process is discussed.

The phenomenon of electromigration and implications on related circuit reliability are the subject of Chap. 4. With increasing current density in on-chip interconnect lines, the transport of metal atoms under an electric driving force, known as electromigration, becomes more significant. Metal depletion and accumulation occur at the sites of electromigration atomic flux divergence. Voids and protrusions are formed, respectively, at the sites of metal depletion and accumulation, causing,

respectively, open circuit and short-circuit faults in interconnect structures. The mechanical characteristics of the interconnect structures are critical in determining electromigration reliability. Power and ground lines are particularly susceptible to electromigration damage as these lines carry a significant amount of unidirectional current.

Scaling trends of on-chip power distribution noise are discussed in Chap. 5. A model for scaling power distribution noise is described. Two scenarios of interconnect scaling are analyzed. The effects of scaling trends on the design of next generation complementary metal-oxide semiconductor (CMOS) circuits are also discussed.

Chapter 1
Introduction

In July 1958, Jack Kilby of Texas Instruments suggested building all of the components of a circuit completely in silicon [1]. By September 12, 1958, Kilby had built a working model of the first "solid circuit," the size of a pencil point. A couple of months later in January 1959, Robert Noyce of Fairchild Semiconductor developed a better way to connect the different components of a circuit [2, 3]. Later, in the spring of 1959, Fairchild Semiconductor demonstrated the first planar circuit—a "unitary circuit." The first monolithic integrated circuit (IC) was born, where multiple transistors coexisted with passive components on the same physical substrate [4]. Microphotographs of the first IC (Texas Instruments, 1958), the first monolithic IC (Fairchild Semiconductor, 1959), and the high performance i7-6700K Skylake Quad-Core microprocessor with up to 4.2 GHz clock frequency (Intel Corporation, 2015) are depicted in Fig. 1.1.

In 1960, Jean Hoerni invented the planar process [5]. Later, in 1960, Dawon Kahng and Martin Atalla demonstrated the first silicon based metal oxide semiconductor field effect transistor (MOSFET) [6], followed in 1967 by the first silicon gate MOSFET [7]. These seminal inventions resulted in the explosive growth of today's multi-billion dollar microelectronics industry. The fundamental cause of this growth in the microelectronics industry has been made possible by technology scaling, particularly in CMOS technology.

The goal of this chapter is to provide a brief perspective on the development of ICs, introduce power delivery and management in the context of this development, motivate the use of on-chip voltage regulators and decoupling capacitors, and provide guidance and perspective to the rest of this book. The evolution of integrated circuit technology from the first ICs to highly scaled CMOS technology is described in Sect. 1.1. As manufacturing technologies supported higher integration densities and switching speeds, the primary constraints and challenges in the design of integrated circuits have also shifted, as discussed in Sect. 1.2. The basic nature of the problem of distributing power and ground in integrated circuits is described

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0_1

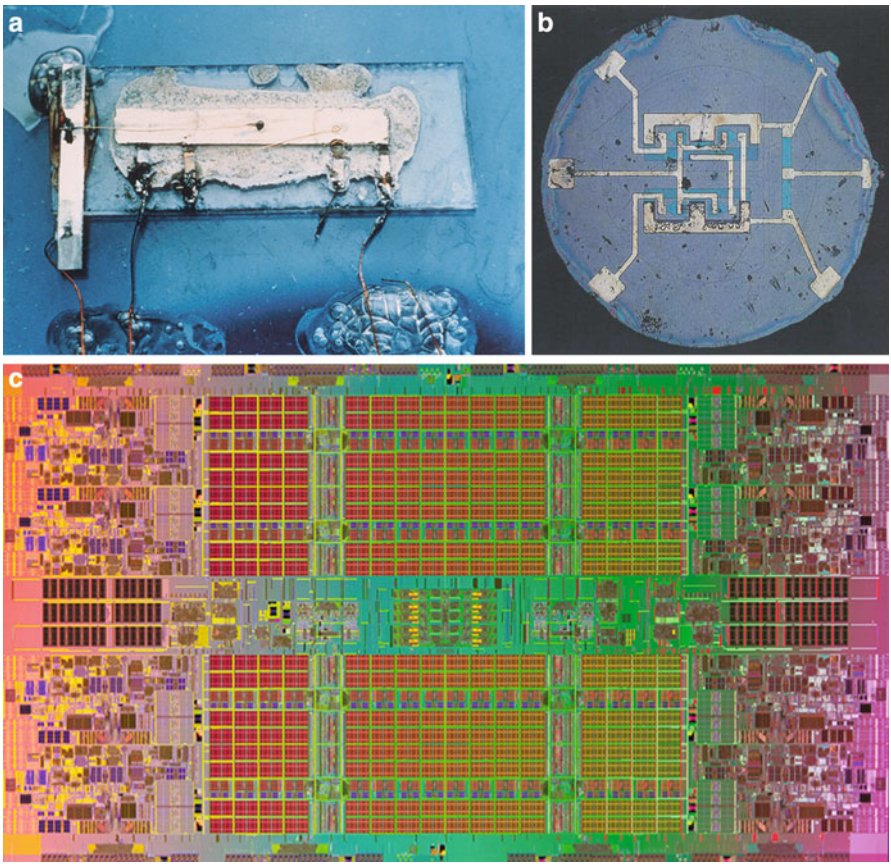


Fig. 1.1 Microphotographs of early and recent integrated circuits (IC) (the die size is not to scale); (**a**) the first IC (Texas Instruments, 1958), (**b**) the first monolithic IC (Fairchild Semiconductor, 1959), (**c**) the high performance i7-6700K Skylake Quad-Core microprocessor (Intel Corporation, 2015)

in Sect. 1.3. The adverse effects of variations in the power supply voltage on the operation of a digital integrated circuit are discussed in Sect. 1.4. Finally, the chapter is summarized in Sect. 1.5.

1.1 Evolution of Integrated Circuit Technology

The pace of IC technology over the past three decades is well characterized by Moore's law. As noted in 1965 by Gordon Moore, the integration density of the first commercial integrated circuits has doubled approximately every year [8]. A prediction was made that the economically effective integration density, i.e.,

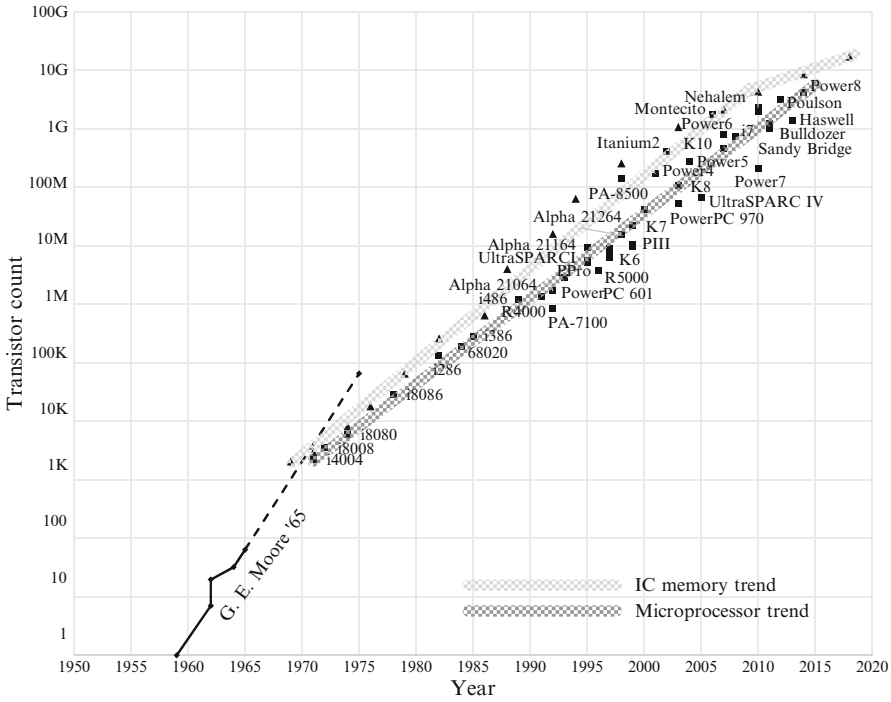


Fig. 1.2 Evolution of transistor count of CPU/microprocessor and memory ICs. In the lower left corner, the original Moore's data [8] is displayed followed by the extrapolated prediction (*dashed line*). The *wide lines* are linearized trends for both IC memory and microprocessors

the number of transistors on an integrated circuit leading to the minimum cost per integrated component, will continue to double every year for another decade. This prediction has held true through the early 1970s. In 1975, the prediction was revised to suggest a new, slower rate of growth–doubling of the IC transistor count every two years [9]. This trend of exponential growth of IC complexity is commonly referred to as "Moore's law." Since the start of commercial production of integrated circuits in the early 1960s, circuit complexity has risen from a few transistors to several billions of transistors functioning together on a single monolithic substrate. This trend is expected to continue at a comparable pace for another decade [10].

The evolution of the integration density of microprocessor and memory ICs is shown in Fig. 1.2 along with the original prediction described in [8]. As seen from the data illustrated in Fig. 1.2, DRAM IC complexity has grown at an even higher rate, quadrupling roughly every three years. The progress of microprocessor clock frequencies is shown in Fig. 1.3. Associated with increasing IC complexity and clock speed is an exponential increase in microprocessor performance (doubling every 18 to 24 month). This performance trend is also referred to as Moore's law.

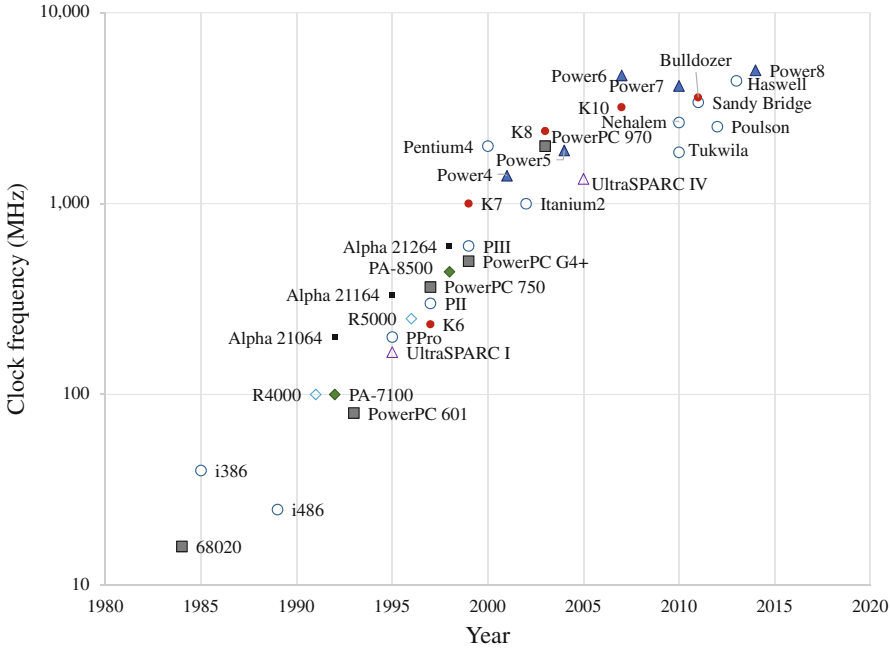


Fig. 1.3 Evolution of microprocessor clock frequency. Several lines of microprocessors are shown in *different colors and shapes*

The principal driving force behind this spectacular improvement in circuit complexity and performance has been the steady decrease in the feature size of semiconductor devices. Advances in optical lithography have allowed manufacturing of on-chip structures with increasingly higher resolution. The area, power, and speed characteristics of transistors with a planar structure, such as MOS devices, improve with the decrease (i.e., scaling) of the lateral dimensions of the devices. These technologies are therefore referred to as *scalable*. The maturing of scalable planar circuit technologies, first PMOS and later NMOS, has allowed the potential of technology scaling to be fully exploited as lithography has improved. The development of planar MOS technology culminated in CMOS circuits. The low power characteristics of CMOS technology deferred the effects of thermal limitations on integration complexity and permitted technology scaling to continue unabated through the 1980s, 1990s, 2000s, and 2010s making CMOS the digital circuit technology of choice.

From a historical perspective, the development of scalable ICs was simultaneously circuitous and serendipitous, as described by Murphy, Haggan, and Troutman [11]. Although the ideas and motivation behind scalable ICs seem straightforward from today's vantage point, the emergence of scalable commercial ICs was neither inevitable nor a result of a well guided and planned pursuit. Rather, the original motivation for the development of integrated circuits was circuit

miniaturization for military and space applications. Although the active devices of the time, discrete transistors, offered smaller size (and also lower power dissipation with higher reliability) as compared to vacuum tubes, much of this advantage was lost at the circuit level, as the size and weight of electronic circuits were dominated by passive components, such as resistors, capacitors, and diodes. Thus, the original objective was to reduce the size of the passive elements through integration of these elements onto the same die as the transistors. The cost effectiveness and commercial success of high complexity ICs were highly controversial for several years after the integrated circuit was invented. Successful integration of a large number of transistors on the same die seemed infeasible, considering the yield of discrete devices at the time [11].

Many obstacles precluded early ICs from scaling. The bulk collector bipolar transistors used in these early ICs suffered from performance degradation due to high collector resistance and, more importantly, the collectors of all of the on-chip transistors were connected through the bulk substrate. The speed of a bipolar transistor does not, in general, scale with the lateral dimensions (i.e., vertical NPN and PNP doping structures typically determine the performance). In addition, early device isolation approaches were not amenable to scaling and consumed significant die area. On-chip resistors and diodes also made inefficient use of die area. Scalable schemes for device isolation and interconnection were therefore essential to truly scale ICs. It was not until these problems were solved and the structure of the bipolar transistor was improved that device miniaturization led to dramatic improvements in IC complexity.

The concept of scalable ICs received further development with the maturation of the MOS technology. Although the MOS transistor is a contemporary of the first ICs, the rapid progress in bipolar devices delayed the development of MOS ICs at the beginning of the IC era. The MOS transistor lagged in performance as compared with existing bipolar devices and suffered from reproducibility and stability problems. The low current drive capability of MOS transistors was also a serious disadvantage at low integration densities. Early use of the MOS transistor was limited to those applications that exploited the excellent switch-like characteristics of the MOS devices. Nevertheless, the circuit advantages and scaling potential of MOS technology were soon realized, permitting MOS circuits to gain increasingly wider acceptance. Gate insulation and the enhancement mode of operation made MOS technology ideal for direct-coupled logic [12]. Furthermore, MOS did not suffer from punch-through effects and could be fabricated with higher yield. The compactness of MOS circuits and the higher yield eventually resulted in a fourfold density advantage in devices per IC as compared to bipolar ICs. Ironically, it was the refinement of bipolar technology that paved the path to these larger scales of integration, permitting the efficient exploitation of MOS technology. With advances in lithographic resolution, the MOS disadvantage in switching speed as compared to bipolar devices gradually diminished. The complexity of bipolar ICs had become primarily constrained by power dissipation. As a result, MOS emerged as the dominant digital integrated circuit technology.

1.2 Evolution of Design Objectives

Advances in fabrication technology and the emergence of new applications have induced several shifts in the principal objectives in the design of integrated circuits over the past 50 years. The evolution of the IC design paradigm is illustrated in Fig. 1.4.

In the 1960s and 1970s, yield concerns served as the primary limitation to IC integration density and, as a consequence, circuit compactness and die area were the primary criteria in the IC design process. Due to limited integration densities, a typical system at the time would contain dozens to thousands of small ICs. As a result, chip-to-chip communications traversing board-level interconnect limited overall system performance. As compared to intra-chip interconnect, board level interconnect have high latency and dissipate large amounts of power, limiting the speed and power of a system. Placing as much functionality as possible into a yield limited silicon die supported the realization of electronic systems with fewer ICs. Fewer board level contacts and interconnections in systems comprised of fewer ICs improved system reliability and lowered system cost, increased system speed (due to lower communication latencies), reduced system power consumption, and decreased the size and weight of the overall system. Producing higher functionality per silicon area with the ensuing reduction in the number of individual ICs typically achieved an improved cost/performance tradeoff at the system level. A landmark example of that era is the first Intel microprocessor, the 4004, commercialized at the end of 1971 [13]. Despite the limitation to 4-bit word processing and initially operating at a mere 108 kHz, the 4004 microprocessor was a complete processor core built on a monolithic die containing approximately 2300 transistors. A microphotograph of the 4004 microprocessor is shown in Fig. 1.5.

The impact of off-chip communications on overall system speed decreased as the integration density increased with advances in fabrication technology, lowering the

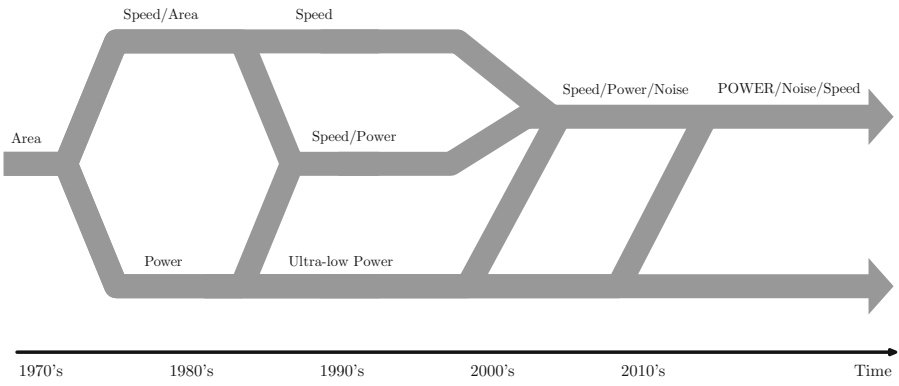


Fig. 1.4 Evolution of design criteria in CMOS integrated circuits

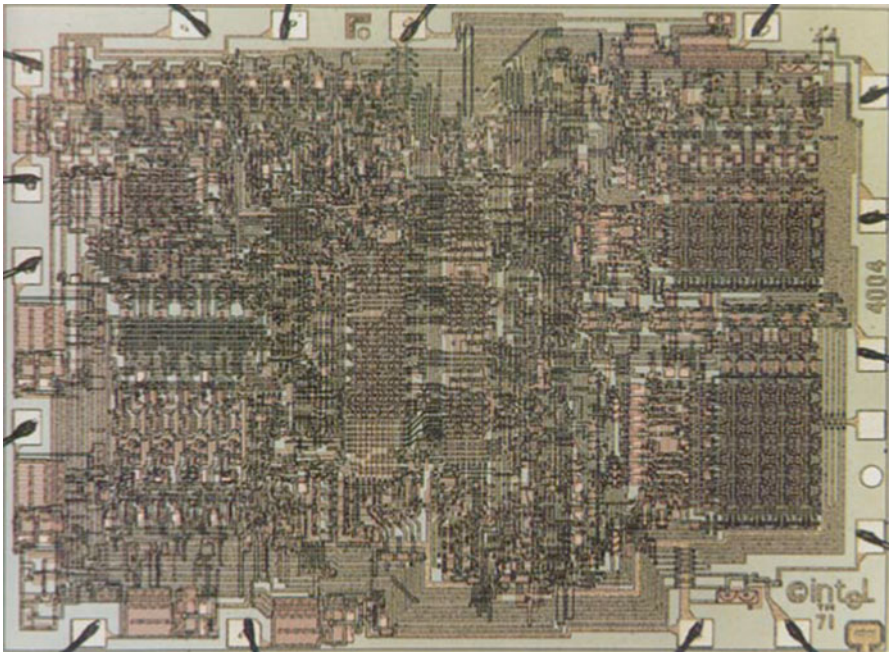


Fig. 1.5 Microphotograph of the 4004—the first microprocessor manufactured on a monolithic die

number of ICs comprising a system. System speed became increasingly dependent on the speed of the component ICs (and less dependent on the speed of the board-level communications). By the 1980s, circuit speed had become the design criterion of greatest significance. Concurrently, a new class of applications emerged, principally restricted by the amount of power consumed. These applications included digital wrist watches, handheld calculators, pacemakers, and satellite electronics. These applications established a new design concept—design for ultra-low power, i.e., power dissipation being the primary design criterion, as illustrated by the lowest path shown in Fig. 1.4.

As device scaling progressed and a greater number of components were integrated onto a single die, on-chip power dissipation began to produce significant economic and technical difficulties. While the market for high performance circuits could support the additional cost, the design process in the 1990s had focused on optimizing both speed and power, borrowing a number of design approaches previously developed for ultra-low power products. The proliferation of portable electronic devices further increased the demand for power efficient and ultra-low power ICs, as shown in Fig. 1.4.

A continuing increase in power dissipation exacerbated system price and reliability concerns, making power a primary design metric across an entire range of applications. The evolution of power consumed by several lines of commercial

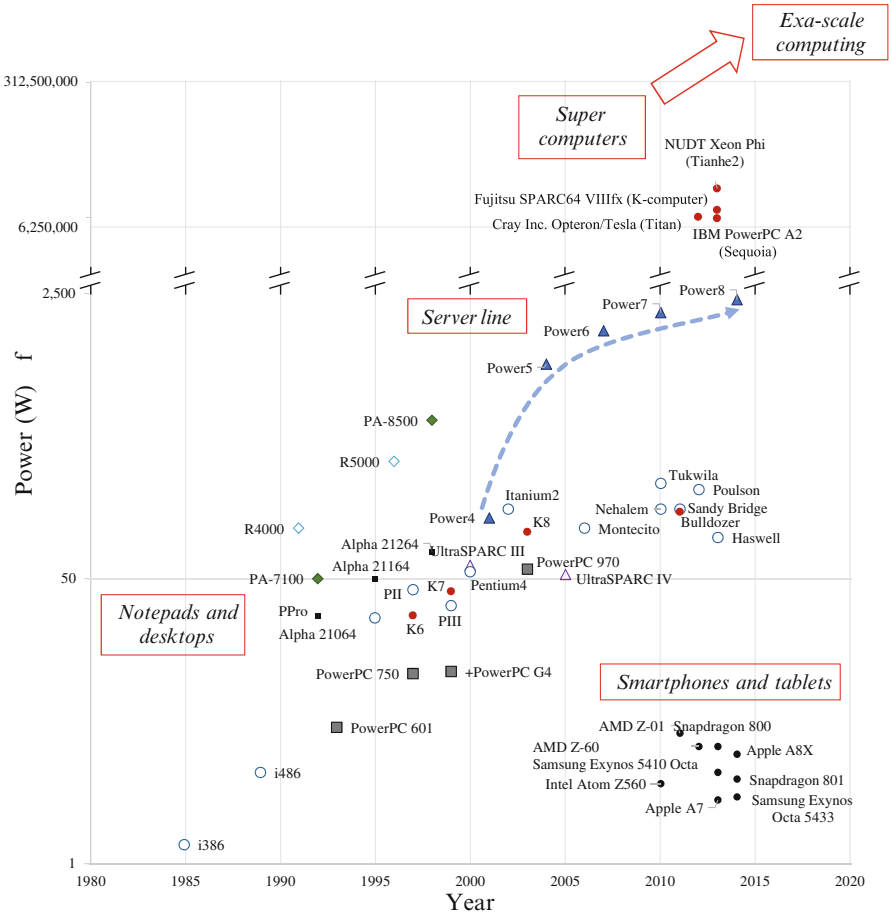


Fig. 1.6 Evolution of microprocessor power consumption. Several lines of microprocessors are shown in *different colors and shapes*

microprocessors is shown in Fig. 1.6. Furthermore, aggressive device scaling and increasing circuit complexity have caused severe noise (or signal integrity) issues in high complexity, high speed integrated circuits. Although digital circuits have traditionally been considered immune to noise due to the inherently high noise margins, circuit coupling through the parasitic impedances of the on-chip interconnect has significantly increased with technology scaling. Ignoring the effects of on-chip noise is no longer possible in the design of high speed digital ICs. These changes are reflected in the convergence of "speed" and "speed/power" design criteria to "speed/power/noise," as depicted in Fig. 1.4.

As device scaling continued in the twenty first century, more than seven billions transistors have successfully been integrated onto a single die [14], keeping up with Moore's law. As a result, the overall power dissipation increased accordingly,

exceeding the maximum capability of conventional cooling technologies. Any further increase in on-chip power dissipation would require either expensive and challenging technology solutions, such as liquid cooling, significantly increasing the overall cost of a system, or innovations in system architecture that exploit massive integration levels or local functional characteristics. Moreover, an explosive growth of portable and handheld devices, such as cell phones and personal device assistants (PDAs), resulted in a shift of design focus towards low power. As an architectural solution for low power in high performance ICs, multi-core systems emerged [15–18], trading off silicon area with on-chip power dissipation. Since the emphasis on ultra-low power design continues in the second decade of the twenty first century, major design effort is focused on reducing system-level power dissipation.

1.3 The Issue of Power Delivery and Management

The issue of power delivery is illustrated in Fig. 1.7, where a simple power delivery system is shown. The system consists of a power supply, a power load, and interconnect lines connecting the supply to the load. The power supply is assumed to behave as an ideal voltage source providing nominal power and ground voltages, V_{dd} and V_{gnd}. The power load is modeled as a variable current source $I(t)$. The interconnect lines connecting the supply and the load are not ideal; the power and ground lines have, respectively, a finite parasitic resistance R_p and R_g, and inductance L_p and L_g. Resistive voltage drops $\Delta V_R = IR$ and inductive voltage drops $\Delta V_L = L\, dI/dt$ develop across the parasitic interconnect impedances, as the load draws current $I(t)$ from the power delivery system. The voltage levels across the load terminals, therefore, change from the nominal level provided by the supply, dropping to $V_{\text{dd}} - IR_p - L_p\, dI/dt$ at the power terminal and rising to $V_{\text{gnd}} + IR_g + L_g\, dI/dt$ at the ground terminal, as shown in Fig. 1.7.

This uncertainty in the supply voltages is referred to as power supply noise. Power supply noise adversely affects circuit operation through several mechanisms,

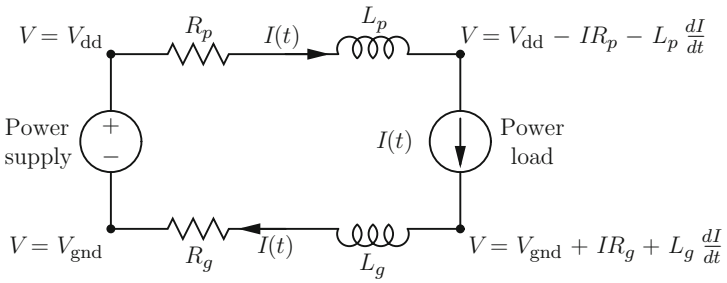


Fig. 1.7 Power delivery system consisting of the power supply, power load, and non-ideal interconnect lines

as described in Sect. 1.4. Proper design of the load circuit ensures correct operation under the assumption that the supply levels are maintained within a certain range near the nominal voltage levels. This range is called the power noise margin. The primary objective in the design of the power delivery system is to supply sufficient current to each transistor on an integrated circuit while ensuring that the power noise does not exceed target noise margins.

The on-going miniaturization of integrated circuit feature size has placed significant requirements on the on-chip power and ground distribution networks. Circuit integration densities rise with each nanometer technology generation due to smaller devices and larger dies; the current density and total current increase accordingly. Simultaneously, the higher speed switching of smaller transistors produces faster current transients within the power distribution network. Both the average current and the transient current are rising exponentially with technology scaling. The evolution of the average current of high performance microprocessors is illustrated in Fig. 1.8.

With thermal design power (TDP) of over 130 W (e.g., the TDP of the Intel Sandy Bridge, Poulson, and Tukwila microprocessors is, respectively, 130, 170, and 185 W [19]) and power supply voltage as low as 0.8 V [20], the current in contemporary microprocessors is approaching 200 A and will further increase with technology scaling. Forecasted demands in the power current of high performance

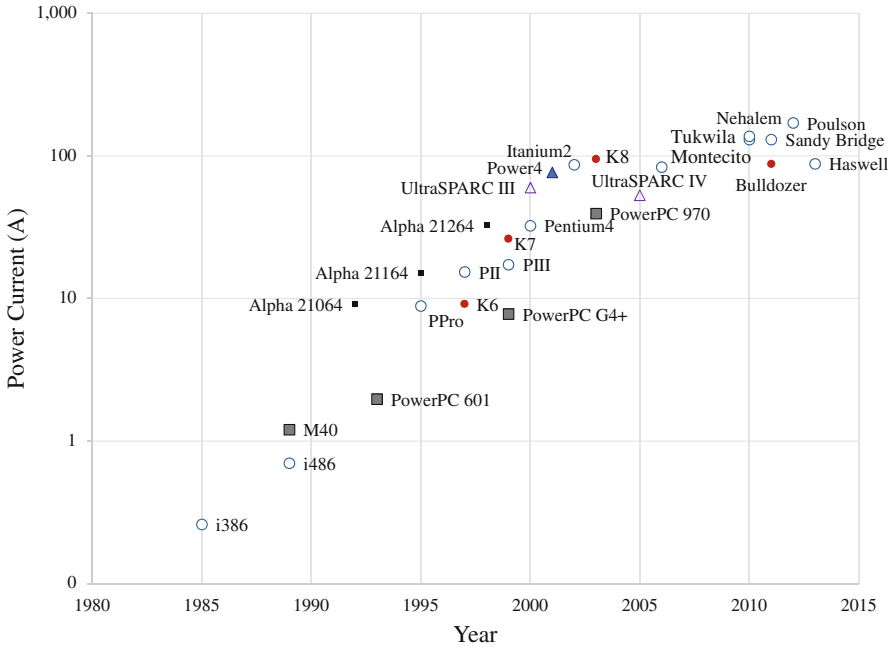


Fig. 1.8 Evolution of the average current in high performance microprocessors. Several lines of microprocessors are shown in *different colors and shapes*

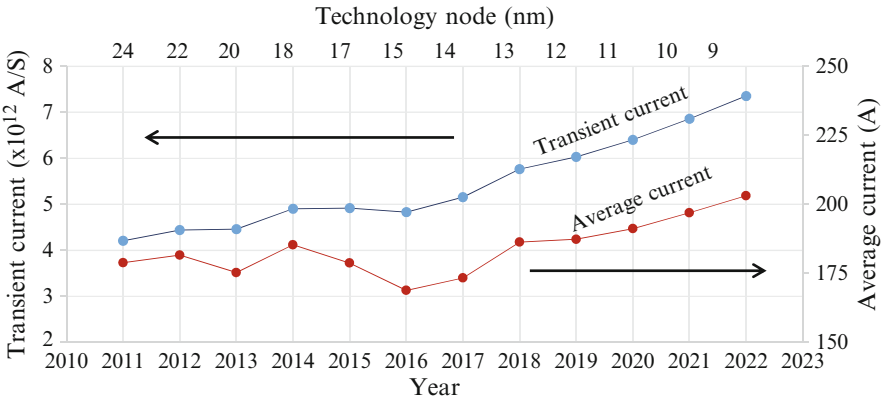


Fig. 1.9 Increasing power current requirements of high performance microprocessors with technology scaling, according to the ITRS roadmap [10]. The average current is the ratio of the circuit power to the supply voltage. The transient current is the product of the average current and the on-chip clock rate, $2\pi f_{clk}$

microprocessors are illustrated in Fig. 1.9. The rate of increase in the transient current is expected to more than double the rate of increase in the average current, as indicated by the slope of the trend lines depicted in Fig. 1.9.

The faster rate of increase in the transient current as compared to the average current is due to increasing on-chip clock frequencies. The transient current in modern high performance microprocessors is approximately one teraampere per second (10^{12} A/s) and is expected to rise, exceeding seven teraamperes per second by 2022. A transient current of this high magnitude is due to switching hundreds of amperes within tens to hundreds of picoseconds. Fortunately, the rate of increase in the transient current has slowed with the introduction of lower speed multi-core microprocessors. In a multi-core microprocessor, similar performance is achieved at a lower frequency at the expense of increased circuit area.

Insuring adequate signal integrity of the power supply under these high current requirements has become a primary design issue in high performance, high complexity integrated circuits. The high average currents produce large ohmic IR voltage drops [21], and the fast transient currents cause large inductive $L\ dI/dt$ voltage drops [22] (ΔI noise) within power distribution networks [23]. Power distribution networks are designed to minimize these voltage drops, maintaining the local supply voltage within specified noise margins. If the power supply voltage sags too low, the performance and functionality of the circuit is severely compromised. Alternatively, excessive overshoot of the supply voltage can affect circuit reliability. Further exacerbating these issues is the reduced noise margins of the power supply as the supply voltage is reduced with each new generation of nanometer process technology, as shown in Fig. 1.10.

To maintain the local supply voltage within specified design margins, the output impedance of a power delivery system should be low as seen at the power

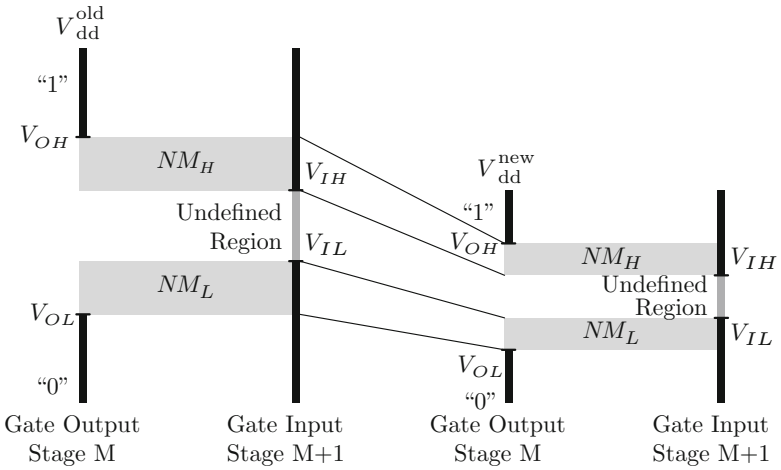


Fig. 1.10 Reduction in noise margins of CMOS circuits with technology scaling. NM_H and NM_L are the noise margins, respectively, in the high and low logic state

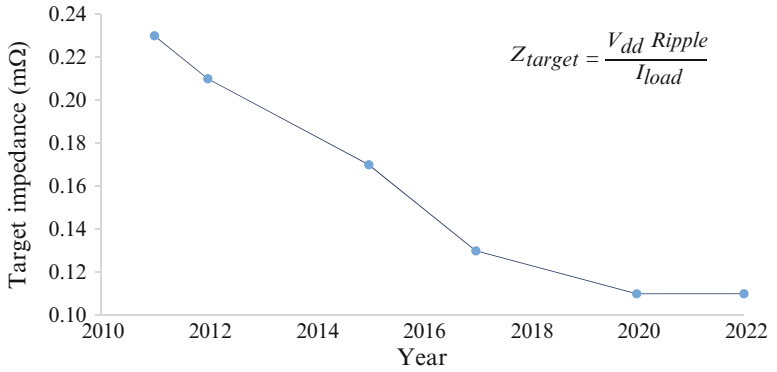


Fig. 1.11 Projections of the target impedance of a power delivery system. The target impedance will continue to drop for future technology generations at an aggressive rate of 1.25 X per technology node [24]

terminals of the circuit elements. IC technologies are expected to scale for another decade [10]. As a result, the average and transient currents drawn from the power delivery network will continue to rise. Simultaneously scaling the power supply voltage, however, has become limited due to threshold variations. The target output impedance of a power delivery system in high speed, high complexity ICs such as microprocessors will therefore continue to drop, reaching an inconceivable level of 150 μΩ by the year 2022 [24], as depicted in Fig. 1.11.

With transistor switching times as short as a few picoseconds, on-chip signals typically contain harmonic frequencies as high as ∼100 GHz. For on-chip wires, the inductive reactance ωL dominates the overall wire impedance beyond ∼10 GHz.

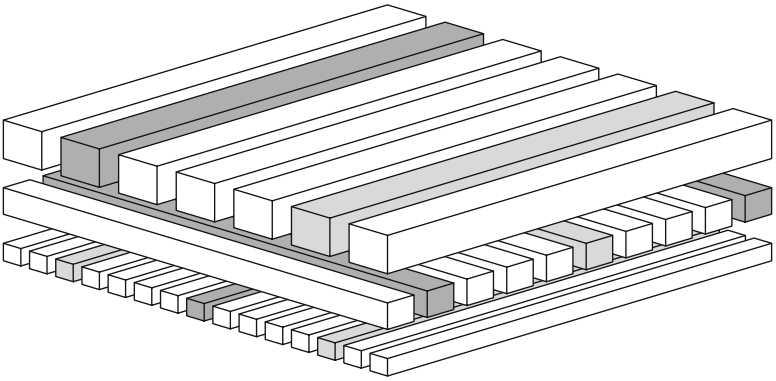


Fig. 1.12 A grid structured power distribution network. The ground lines are *light gray*, the power lines are *dark gray*, and the signal lines are *white*

The on-chip inductance affects the integrity of the power supply through two phenomena. First, the magnitude of the ΔI noise is directly proportional to the power network inductance as seen at the current sink. Second, the network resistance, inductance, and decoupling capacitance form an *RLC* tank circuit with multiple resonances. The peak impedance of this *RLC* circuit must be lowered to guarantee that target power supply noise margins are satisfied. Thus, information characterizing the inductance is needed to correctly design and analyze power delivery systems.

Power distribution networks in high performance digital ICs are commonly structured as a multilayer grid. In such a grid, straight power/ground (P/G) lines in each metalization layer span the entire die (or a large functional unit) and are orthogonal to the lines in the adjacent layers. The power and ground lines typically alternate in each layer. Vias connect a power (ground) line to another power (ground) line at the overlap sites. This power grid organization is illustrated in Fig. 1.12, where three layers of interconnect are depicted with the power lines shown in *dark gray* and the ground lines shown in *light gray*. The power/ground lines are surrounded by signal lines.

A significant fraction of the on-chip resources is committed to insure the integrity of the power supply voltage levels. The global on-chip power delivery system is typically determined at early stages of the design process, when little is known about the local current demands at specific locations on an IC. Additional metal resources for the global power delivery system are typically allocated at later stages of the design process to improve the local electrical characteristics of the power network. A complete redesign of the surrounding global signals can be prohibitively expensive and time consuming. For these reasons, power delivery systems tend to be conservatively designed [25], sometimes using more than a third of the on-chip metal resources [26, 27]. Overengineering the power delivery system is, therefore, costly in modern interconnect limited, high complexity digital integrated circuits.

Performance objectives in power delivery systems, such as low impedance (low inductance and resistance) to satisfy noise specifications under high current loads, small physical area, and low current densities (for improved reliability) are typically in conflict. Widening the lines to increase the conductance and improve the electromigration reliability also increases the grid area. Replacing wide metal lines with narrow interdigitated P/G lines increases the line resistance if the grid area is maintained constant or increases the physical area if the net cross section of the lines is maintained constant. It is therefore important to make a balanced choice under these conditions. A quantitative model of the inductance/area/resistance tradeoff in high performance power distribution networks is therefore needed to achieve an efficient power delivery system. Another important objective is to provide quantitative tradeoff guidelines and intuition in the design of high performance power delivery systems.

Decoupling capacitors are often used to reduce the impedance of a power distribution system and provide the required charge to the switching circuits, lowering the power supply noise [28]. At high frequencies, however, the on-chip decoupling capacitors can be effective due to the high parasitic impedance of the power network connecting a decoupling capacitor to the current load [29]. On-chip decoupling capacitors, however, reduce the self-resonant frequency of a power delivery system, resulting in high amplitude power supply voltage fluctuations at the resonant frequencies. A hierarchical system of on-chip decoupling capacitors should therefore be carefully designed to provide a low impedance, resonant-free power delivery system over the entire range of operating frequencies, while delivering sufficient charge to the switching circuits to maintain the local power supply voltages within target noise margins [30].

In earlier technology generations, high quality DC voltages and currents were delivered from off-chip voltage converters to on-chip load circuitry within carefully designed electrical power grids, producing a power system which was passive in nature. To maintain sufficient quality of power under increasing current densities and parasitic impedances, the power needs to be locally regulated with distributed on-chip voltage converters close to the load. This concept of distributed power delivery poses new power design challenges in modern ICs, requiring circuit level techniques to convert and regulate power at points-of-load (POL), methodological solutions for distributing on-chip power supplies, and automated design techniques to co-design distributed power supplies and decoupling capacitors.

While the quality of power can be addressed with a POL approach, the emerging trends of heterogeneity, on-chip integration, and dynamic control require fundamental changes in traditional power delivery approaches—power delivery systems should not be viewed as a passive power distribution network but rather as systems that need to be efficiently and proactively managed. The regulation of DC voltages close to the load, distributed on-chip current delivery, and local intelligence are all required to efficiently manage power resources in high performance ICs. To address these novel challenges, traditional power delivery and management systems need to be conceptually reorganized. Specialized power delivery circuits,

locally intelligent power routers, microcontrollers, and power managing policies have become basic building blocks for delivering and managing power in modern heterogeneous systems.

1.4 Deleterious Effects of Power Distribution Noise

Power noise adversely affects the operation of an integrated circuit through several mechanisms. These mechanisms are discussed in this section. Power supply noise produces uncertainty in the delay of the clock and data signals, as described in Sect. 1.4.1. Power supply noise also increases the uncertainty of the timing reference signals generated on-chip (clock jitter), lowering the clock frequency of the circuit, as discussed in Sect. 1.4.2. The reduction of noise margins is discussed in Sect. 1.4.3. Power supply variations diminish the maximum supply voltage, degrading the speed of operation, as described in Sect. 1.4.4.

1.4.1 Signal Delay Uncertainty

The propagation delay of on-chip signals depends on the power supply voltage during a signal transition. The source of the PMOS transistors in pull-up networks within logic gates is connected to the highest supply voltage directly or through other PMOS transistors. Similarly, the source of the NMOS transistors within a pull-down networks is connected to the lowest supply voltage (directly or through other NMOS transistors). The drain current of an MOS transistor increases with the voltage difference between the transistor gate and source. When the rail-to-rail power voltage is reduced due to power supply variations, the gate-to-source voltage of the NMOS and PMOS transistors is less, lowering the output current of the transistors. The signal delay increases accordingly as compared to the delay under a nominal power supply voltage. Conversely, a higher power voltage and a lower ground voltage shortens the propagation delay. The effect of the power noise on the propagation delay of the clock and data signals is, therefore, an increase in both delay uncertainty and the delay of the data paths [31, 32]. Consequently, power supply noise limits the maximum operating frequency of an integrated circuit [33–35].

1.4.2 On-Chip Clock Jitter

A phase-locked loop (PLL) is often used to generate the on-chip clock signal. An on-chip PLL generates an on-chip clock signal by multiplying the system clock signal. Certain changes in the electrical environment of a PLL, power supply voltage

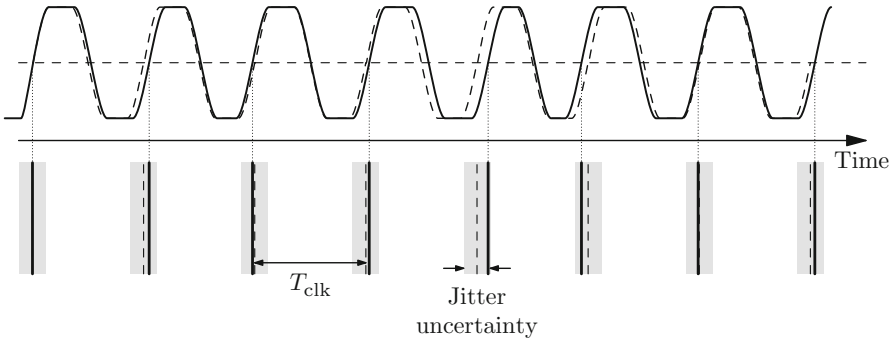


Fig. 1.13 Cycle-to-cycle jitter of a clock signal. The phase of the clock signal (*solid line*) randomly deviates from the phase of an ideal clock signal (*dashed line*)

variations in particular, affect the phase of the on-chip clock signal. A feedback loop within the PLL controls the phase of the PLL output and aligns the output signal phase with the phase of the system clock. Ideally, the edges of the on-chip clock signal are at precisely equidistant time intervals determined by the system clock signal. The closed loop response time of modern PLL is typically hundreds of nanoseconds (e.g., 300 ns in [36]). Disturbances of shorter duration than the PLL response time result in deviations of the on-chip clock phase from ideal timing objectives. These deviations are referred to as clock jitter [37, 38]. The clock jitter is classified into two types: cycle-to-cycle jitter and peak-to-peak jitter.

Cycle-to-cycle jitter refers to *random* deviations of the clock phase from the ideal timing, as illustrated in Fig. 1.13 [39]. Deviation from the ideal phase at one edge of a clock signal is independent of the deviations at other edges. That is, the cycle-to-cycle jitter characterizes the variation of the time interval between two adjacent clock edges. The average cycle-to-cycle jitter asymptotically approaches zero with an increasing number of samples. This type of jitter is therefore characterized by a mean square deviation. This type of phase variation is produced by disturbances of duration shorter or comparable to the clock period. Active device noise and high frequency power supply noise (i.e., of a frequency higher or comparable to the clock frequency) contribute to the cycle-to-cycle jitter. Due to the stochastic nature of phase variations, the cycle-to-cycle jitter directly contributes to the uncertainty of the time reference signals across an integrated circuit. Increased uncertainty of an on-chip timing reference results in a reduced operating frequency [39].

The second type of jitter, peak-to-peak jitter, refers to *systematic* variations of on-chip clock phase *as compared to the system clock*. Consider a situation where several consecutive edges of an on-chip clock signal have a positive cycle-to-cycle variation, i.e., several consecutive clock cycles are longer than the ideal clock period, as illustrated in Fig. 1.14 (due to, for example, prolonged power supply variations from the nominal voltage). The timing requirements of the on-chip circuits are not violated provided that the cycle-to-cycle jitter is sufficiently small. The phase difference between the system clock and the on-chip clock,

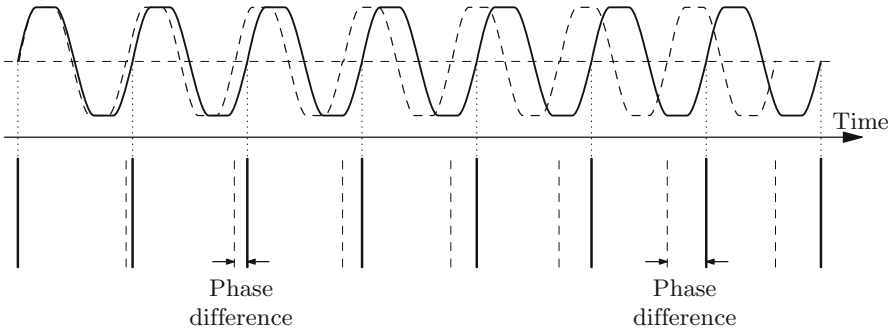


Fig. 1.14 Peak-to-peak jitter of a clock signal. The period of the clock signal (the *solid line*) systematically deviates from the period of the reference clock (the *dashed line*), leading to accumulation of the phase difference

however, accumulates with time. Provided the disturbance persists, the phase difference between the system and the on-chip clocks can accumulate for tens or hundreds of clock cycles, until the PLL feedback adjustment becomes effective. This phase difference degrades the synchronization among different clock domains (i.e., between one portion of an integrated circuit and other system components controlled by different clock signals). Synchronizing the clock domains is critical for reliable communication across these domains. The maximum phase difference between two clock domains is characterized by the peak-to-peak jitter.

The feedback response time is highly sensitive to the power supply voltage [40]. For example, the PLL designed for the 400 MHz IBM S/390 microprocessor exhibits a response time of approximately 50 clock cycles when operating at a 2.5 V power supply and disturbed by a 100 mV drop in supply voltage. The recovery time from the same disturbance increases manyfold when the supply voltage is reduced to 2.3 V and below [40].

1.4.3 Noise Margin Degradation

In digital logic styles with single-ended signaling, the power and ground delivery system also serves as a voltage reference for the on-chip signals. If a transmitter communicates a low voltage state, the output of the transmitter is connected to the ground distribution network. Alternatively, the output is connected to the power distribution network to communicate the high voltage state. At the receiver end of the communication line, the output voltage of the transmitter is compared to the power or ground voltage *local to the receiver*. Spatial variations in the power supply voltage create a discrepancy between the power and ground voltage levels at the transmitter and receiver ends of the communication line. The power noise induced uncertainty in these reference voltages degrades the noise margins of the

on-chip signals. As the operating speed of integrated circuits has risen, crosstalk noise among on-chip signals has also increased. Providing sufficient noise margins of the on-chip signals is therefore a design issue of paramount importance.

1.4.4 Degradation of Gate Oxide Reliability

The performance characteristics of an MOS transistor depend on the thickness of the gate oxide. The current drive of the transistor increases as the gate oxide thickness is reduced, improving the speed and power characteristics. Reduction of the gate oxide thickness in process scaling has therefore been instrumental in improving transistor performance. A thin oxide layer, however, poses the problems of electron tunneling and oxide layer reliability [41]. As the thickness of the gate silicon oxide has reached several molecular layers (tens of angstroms) in contemporary digital CMOS processes, the power supply voltage is limited by the maximum electric field across the gate oxide layer [35]. Variations in the power supply voltage can increase the voltage applied across the ultra-thin gate oxide layer above the nominal power supply, degrading the long term reliability of the semiconducting devices [42]. Overshoots of the power and ground voltages should be limited to avoid significant degradation in the transistor reliability characteristics.

1.5 Summary

A historical background, general motivation, and relevant aspects related to integrated circuits in general and on-chip power networks in particular are presented in this introductory chapter. This chapter is summarized as follows.

- The development of integrated circuits has rapidly progressed after the first planar circuit—a “unitary circuit”
- Current microprocessors integrate many billions of transistors on a single monolithic substrate
- The clock frequency of modern microprocessors is in the range of several gigahertz
- The power consumption of mobile, notepad/desktop, and supercomputing microprocessor-based server farms, respectively, are in the range of a few watts, several hundreds of watts, and millions of watts.
- Different design criteria for integrated circuits have evolved over the past several decades with changing technology and application characteristics
- The issue of effective power delivery is fundamental to the successful operation of high complexity ICs. As current demand requirements have increased, voltage margins have been reduced, constraining the impedance of the power delivery system

- Voltage fluctuations within the power delivery system are causing a variety of problems, such as signal delay uncertainty, clock jitter, smaller noise margins, and reliability concerns due to degradation of the gate oxide
- Point-of-load power delivery is fundamental to maintain high quality of power as current densities and parasitic impedances have increased
- To support heterogeneous dynamically on-chip controlled systems, power resources should be intelligently managed in real-time

Chapter 2
Inductive Properties of Electric Circuits

Characterizing the inductive properties of the power and ground interconnect is essential in determining the impedance characteristics of a power distribution system. Several of the following chapters are dedicated to the inductive properties of on-chip power distribution networks. The objective of this chapter is to introduce the concepts used in these chapters to describe the inductive characteristics of complex interconnect structures.

The magnetic properties of circuits are typically described using circuits with inductive coils. The inductive characteristics of such circuits are dominated by the self- and mutual inductances of these coils. The inductance of a coil is well described by the classical definition of inductance based on the magnetic flux through a current loop. The situation is more complex in circuits with no coils where no part of the circuit is inductively dominant and the circuit elements are strongly inductively coupled. The magnetic properties in this case are determined by the physical structure of the entire circuit, resulting in complex inductive behavior. The loop inductance formulation is inconvenient to represent the inductive characteristics of these circuits. The objective of this chapter is to describe various ways to represent a circuit inductance, highlighting specific assumptions. Intuitive interpretations are offered to develop a deeper understanding of the limitations and interrelations of these approaches. The variation of inductance with frequency and the relationship between the absolute inductance and the inductive behavior are also discussed in this chapter.

These topics are discussed in the following order. Several formulations of the circuit inductive characteristics as well as advantages and limitations of these formulations are described in Sect. 2.1. Mechanisms underlying the variation of inductance with frequency are examined in Sect. 2.2. The relationship between the absolute inductance and the inductive behavior of circuits is discussed in Sect. 2.3. The inductive properties of on-chip interconnect structures are analyzed in Sect. 2.4. The chapter is summarized in Sect. 2.5.

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0_2

2.1 Definitions of Inductance

There are several ways to represent the magnetic characteristics of a circuit. Understanding the advantages and limitations of these approaches presents the opportunity to choose the approach most suitable for a particular task. Several representations of the inductive properties of a circuit are presented in this section. The field energy formulation of inductive characteristics is described in Sect. 2.1.1. The loop flux definition of inductance is discussed in Sect. 2.1.2. The concept of a partial inductance is described in Sect. 2.1.3. The net inductance formulation is described in Sect. 2.1.4.

2.1.1 Field Energy Definition

Inductance represents the capability of a circuit to store energy in the form of a magnetic field. Specifically, the inductance relates the electrical current to the magnetic flux and magnetic field energy. The magnetic field is interrelated with the electric field and current, as determined by Maxwell's equations and constitutive relations,[1]

$$\nabla \boldsymbol{D} = \rho, \tag{2.1}$$

$$\nabla \boldsymbol{B} = 0, \tag{2.2}$$

$$\nabla \times \boldsymbol{H} = \boldsymbol{J} + \frac{\partial \boldsymbol{D}}{\partial t}, \tag{2.3}$$

$$\nabla \times \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t}, \tag{2.4}$$

$$\boldsymbol{D} = \epsilon \boldsymbol{E}, \tag{2.5}$$

$$\boldsymbol{B} = \mu \boldsymbol{H}, \tag{2.6}$$

$$\boldsymbol{J} = \sigma \boldsymbol{E}, \tag{2.7}$$

assuming a linear media. The domain of circuit analysis is typically confined to those operational conditions where the electromagnetic radiation phenomena are negligible. The direct effect of the displacement current $\frac{\partial \boldsymbol{D}}{\partial t}$ on the magnetic field, as expressed by (2.3), can be neglected under these conditions (although the displacement current can be essential to determine the current density $\boldsymbol{J}$). The magnetic field is therefore determined only by the circuit currents. The local current density determines the local behavior of the magnetic field, as expressed by Ampere's law in the differential form,

[1] Vector quantities are denoted with bold italics, such as $\boldsymbol{H}$.

$$\nabla \times \boldsymbol{H} = \boldsymbol{J}. \tag{2.8}$$

Equivalently, the elemental contribution to the magnetic field $d\boldsymbol{H}$ is expressed in terms of an elemental current $d\boldsymbol{J}$, according to the Biot-Savart law,

$$d\boldsymbol{H} = \frac{d\boldsymbol{J} \times \boldsymbol{r}}{4\pi r^3}, \tag{2.9}$$

where $\boldsymbol{r}$ is the distance vector from the point of interest to the current element $d\boldsymbol{J}$ and $r = |\boldsymbol{r}|$.

It can be demonstrated that the magnetic energy in a linear media can be expressed as [43]

$$W_{\mathrm{m}} = \frac{1}{2} \int \boldsymbol{J} \cdot \boldsymbol{A}\, d\boldsymbol{r}, \tag{2.10}$$

where $\boldsymbol{A}$ is the magnetic vector potential of the system, determined as

$$\boldsymbol{A}(\boldsymbol{r}) = \frac{\mu}{4\pi} \int \frac{\boldsymbol{J}(\boldsymbol{r}')\, d\boldsymbol{r}'}{|\boldsymbol{r} - \boldsymbol{r}'|}. \tag{2.11}$$

Substituting (2.11) into (2.10) yields the expression of the magnetic energy in terms of the current distribution in a system,

$$W_{\mathrm{m}} = \frac{\mu}{8\pi} \iint \frac{\boldsymbol{J}(\boldsymbol{r}) \cdot \boldsymbol{J}(\boldsymbol{r}')}{|\boldsymbol{r} - \boldsymbol{r}'|}\, d\boldsymbol{r}\, d\boldsymbol{r}'. \tag{2.12}$$

If the system is divided into several parts, each contained in a volume V_i, the magnetic energy expression (2.12) can be rewritten as

$$W_{\mathrm{m}} = \frac{\mu}{8\pi} \sum_i \sum_j \int_{V_i} \int_{V_j} \frac{\boldsymbol{J}(\boldsymbol{r}) \cdot \boldsymbol{J}(\boldsymbol{r}')}{|\boldsymbol{r} - \boldsymbol{r}'|}\, d\boldsymbol{r}\, d\boldsymbol{r}'. \tag{2.13}$$

Assuming that the relative distribution of the current in each volume V_i is independent of the current magnitude, the current density distribution $\boldsymbol{J}$ can be expressed in terms of the overall current magnitude I and current distribution function $\boldsymbol{u}(\boldsymbol{r})$, so that $\boldsymbol{J}(\boldsymbol{r}) = I\boldsymbol{u}(\boldsymbol{r})$. The magnetic field energy can be expressed in terms of the overall current magnitudes I_i,

$$W_{\mathrm{m}} = \frac{1}{2} \sum_i \sum_j L_{ij} I_i I_j, \tag{2.14}$$

where

$$L_{ij} \equiv \frac{\mu}{4\pi} \int_{V_i} \int_{V_j} \frac{\boldsymbol{u}(\boldsymbol{r}) \cdot \boldsymbol{u}(\boldsymbol{r}')}{|\boldsymbol{r} - \boldsymbol{r}'|}\, d\boldsymbol{r}\, d\boldsymbol{r}' \tag{2.15}$$

is a mutual inductance between the system parts i and j. In a matrix formulation, the magnetic energy of a system consisting of N parts can be expressed as a positively defined binary form[2] $\mathbf{L}$ of a current vector $\boldsymbol{I} = \{I_1, \ldots, I_N\}$,

$$W_{\mathrm{m}} = \frac{1}{2}\boldsymbol{I}^T\mathbf{L}\boldsymbol{I} = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} L_{ij}I_iI_j. \tag{2.16}$$

Each diagonal element L_{ii} of the binary form $\mathbf{L}$ is a self-inductance of the corresponding current I_i and each non-diagonal element L_{ij} is a mutual inductance between currents I_i and I_j. Note that according to the definition of (2.15), the inductance matrix is symmetric, i.e., $L_{ij} = L_{ji}$.

While the field approach is general and has no limitations, determining the circuit inductance through this approach is a laborious process, requiring numerical field analysis except for the simplest structures. The goal of circuit analysis is to provide an efficient yet accurate description of the system in those cases where the detail and accuracy of a full field analysis are unnecessary. Resorting to a field analysis to determine specific circuit characteristics greatly diminishes the efficiency of the circuit analysis formulation.

2.1.2 Magnetic Flux Definition

The concept of inductance is commonly described as a constant L relating a magnetic flux Φ through a circuit loop to a current I' in another loop,

$$\Phi = LI'. \tag{2.17}$$

In the special case where the two circuit loops are the same, the coefficient is referred to as a loop self-inductance; otherwise, the coefficient is referred to as a mutual loop inductance.

For example, consider two isolated complete current loops ℓ and ℓ', as shown in Fig. 2.1. The mutual inductance M between these two loops is a coefficient relating a magnetic flux Φ through a loop ℓ due to a current I' in loop ℓ',

$$\Phi = \iint_S \boldsymbol{B}' \cdot \boldsymbol{n}\, ds, \tag{2.18}$$

where S is a smooth surface bounded by the loop ℓ, $\boldsymbol{B}'$ is the magnetic flux created by the current in the loop ℓ', and $\boldsymbol{n}$ is a unit vector normal to the surface element ds. Substituting $\boldsymbol{B}' = \nabla \times \boldsymbol{A}'$ and using Stokes's theorem, the loop flux is expressed as

[2] Matrix entities are denoted with bold roman symbols, such as $\mathbf{L}$.

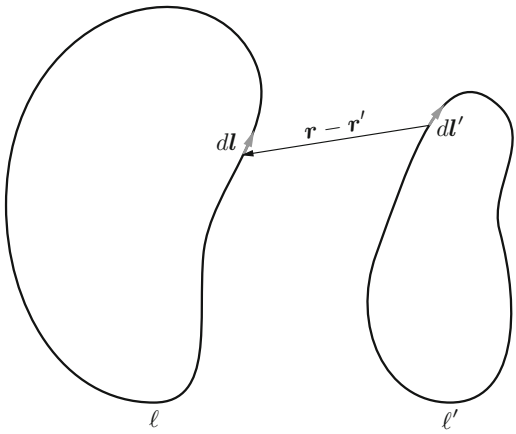


Fig. 2.1 Two complete current loops. The relative position of two differential loop elements $d\boldsymbol{l}$ and $d\boldsymbol{l}'$ is determined by the vector $\boldsymbol{r}-\boldsymbol{r}'$

$$\Phi = \iint_S (\nabla \times \boldsymbol{A}') \cdot \boldsymbol{n}\, ds = \oint_\ell \boldsymbol{A}'\, d\boldsymbol{l}, \tag{2.19}$$

where $\boldsymbol{A}'$ is the vector potential created by the current $\boldsymbol{I}'$ in the loop ℓ'. The magnetic vector potential of the loop ℓ' $\boldsymbol{A}'$ is

$$\boldsymbol{A}'(\boldsymbol{r}) = \frac{\mu}{4\pi} \int_V \frac{\boldsymbol{J}'(\boldsymbol{r}')\, d\boldsymbol{r}'}{|\boldsymbol{r}-\boldsymbol{r}'|} = I' \frac{\mu}{4\pi} \oint_{\ell'} \frac{d\boldsymbol{l}'}{|\boldsymbol{r}-\boldsymbol{r}'|}, \tag{2.20}$$

where $|\boldsymbol{r}-\boldsymbol{r}'|$ is the distance between the loop element $d\boldsymbol{l}'$ and the point of interest $\boldsymbol{r}$. Substituting (2.20) into (2.19) yields

$$\Phi = I' \frac{\mu}{4\pi} \oint_\ell \oint_{\ell'} \frac{d\boldsymbol{l}\, d\boldsymbol{l}'}{|\boldsymbol{r}-\boldsymbol{r}'|} = MI', \tag{2.21}$$

where

$$M \equiv \frac{\mu}{4\pi} \oint_\ell \oint_{\ell'} \frac{d\boldsymbol{l}\, d\boldsymbol{l}'}{|\boldsymbol{r}-\boldsymbol{r}'|} \tag{2.22}$$

is a mutual inductance between the loops ℓ and ℓ'. As follows from the derivation, the integration in (2.20), (2.21), and (2.22) is performed in the direction of the current flow. The mutual inductance (2.22) and associated magnetic flux (2.21) can therefore be either positive or negative, depending on the relative direction of the current flow in the two loops.

Note that the finite cross-sectional dimensions of the loop conductors are neglected in the transition between the general volume integration to a more constrained but simpler contour integration in (2.20). An entire loop current is therefore confined to an infinitely thin filament.

The thin filament approximation of a mutual inductance is acceptable where the cross-sectional dimensions of the conductors are much smaller than the distance $|\boldsymbol{r} - \boldsymbol{r}'|$ between any two points on loop ℓ and loop ℓ'. This approximation becomes increasingly inaccurate as the two loops are placed closer together. More importantly, the thin filament approach cannot be used to determine a self-inductance by assuming ℓ to be identical to ℓ', as the integral (2.22) diverges at the points where $\boldsymbol{r} = \boldsymbol{r}'$.

To account for the finite cross-sectional dimensions of the conductors, both (2.19) and (2.20) are amended to include an explicit integration over the conductor cross-sectional area a,

$$\Phi = \frac{1}{I} \oint_{\ell} \int_{a} \boldsymbol{A}' J d\boldsymbol{l}\, da, \tag{2.23}$$

and

$$\boldsymbol{A}' = \frac{\mu}{4\pi} \oint_{\ell'} \int_{a'} \frac{J'\, d\boldsymbol{l}'\, da'}{|\boldsymbol{r} - \boldsymbol{r}'|}, \tag{2.24}$$

where a and a' are the cross sections of the elemental loop segments $d\boldsymbol{l}$ and $d\boldsymbol{l}'$, da and da' are the differential elements of the respective cross sections, $|\boldsymbol{r} - \boldsymbol{r}'|$ is the distance between da and da', and J is a current density distribution over the wire cross section a, $d\boldsymbol{J} = J\, d\boldsymbol{l}\, da$, and $I = \int_a J\, da$. These expressions are more general than (2.19) and (2.20); the only constraint on the current flow imposed by formulations (2.23) and (2.24) is that the current flow has the same direction across the cross-sectional areas a and a'. This condition is satisfied in relatively thin conductors without sharp turns. Formulas (2.23) and (2.24) can be significantly simplified assuming a uniform current distribution (i.e., $J = \text{const}$ and $I = aJ$) and a constant cross-sectional area a,

$$\Phi = \frac{1}{a} \oint_{\ell} \int_{a} \boldsymbol{A}'\, d\boldsymbol{l}\, da, \tag{2.25}$$

and

$$\boldsymbol{A}' = \frac{\mu}{4\pi} \frac{I'}{a'} \oint_{\ell'} \int_{a'} \frac{d\boldsymbol{l}'\, da'}{|\boldsymbol{r} - \boldsymbol{r}'|}. \tag{2.26}$$

The magnetic flux through loop ℓ is transformed into

$$\Phi = \frac{\mu}{4\pi}\frac{I'}{a\,a'}\oint_{\ell}\oint_{\ell'}\int_{a}\int_{a'}\frac{da\,da'\,d\boldsymbol{l}\,d\boldsymbol{l}'}{|\boldsymbol{r}-\boldsymbol{r}'|} = MI'. \tag{2.27}$$

The mutual loop inductance is therefore defined as

$$M_{\ell\ell'} \equiv \frac{\mu}{4\pi}\frac{1}{a\,a'}\oint_{\ell}\oint_{\ell'}\int_{a}\int_{a'}\frac{da\,da'\,d\boldsymbol{l}\,d\boldsymbol{l}'}{|\boldsymbol{r}-\boldsymbol{r}'|}\,. \tag{2.28}$$

The loop self-inductance L_ℓ is a special case of the mutual loop inductance where the loop ℓ is the same as loop ℓ',

$$L_\ell \equiv M_{\ell\ell} = \frac{\mu}{4\pi}\frac{1}{a^2}\oint_{\ell}\oint_{\ell}\int_{a}\int_{a}\frac{da\,da'\,d\boldsymbol{l}\,d\boldsymbol{l}'}{|\boldsymbol{r}-\boldsymbol{r}'|}\,. \tag{2.29}$$

While straightforward and intuitive, the definition of the loop inductance as expressed by (2.17) cannot be applied to most practical circuits. Only the simplest circuits consist of a single current loop. In practical circuits with branch points, the current is not constant along the circumference of the conductor loops, as shown in Fig. 2.2. This difficulty can be circumvented by employing Kirchhoff's voltage law and including an inductive voltage drop within each loop equation. For example, two independent current loops carrying circular currents I_A and I_B can be identified in the circuit shown in Fig. 2.2. The inductive voltage drops V_A and V_B in loops A and B are

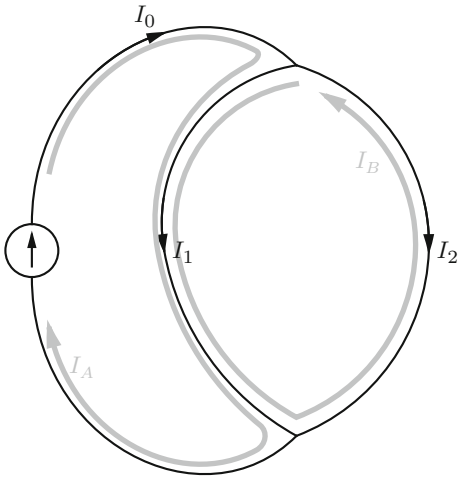


Fig. 2.2 A circuit with branch points. The current in each loop is not uniform along the circumference of the loop

$$\begin{bmatrix} V_A \\ V_B \end{bmatrix} = \begin{bmatrix} L_{AA} & L_{AB} \\ L_{AB} & L_{BB} \end{bmatrix} \begin{bmatrix} I_A \\ I_B \end{bmatrix}. \tag{2.30}$$

The magnetic energy of the system is, analogous to (2.16),

$$W_{\mathrm{m}} = \frac{1}{2} \boldsymbol{I}^T \mathbf{L} \boldsymbol{I} = \frac{1}{2} \begin{bmatrix} I_A & I_B \end{bmatrix} \begin{bmatrix} L_{AA} & L_{AB} \\ L_{AB} & L_{BB} \end{bmatrix} \begin{bmatrix} I_A \\ I_B \end{bmatrix}. \tag{2.31}$$

Note that in a circuit with branch points, two current loops can share common parts, as illustrated in Fig. 2.2. The inductance between these two loops is therefore a hybrid between the self- and mutual loop inductance, as defined by (2.28) and (2.29).

The flux formulation of the inductive characteristics, as expressed by (2.29) and (2.31), is a special case of the field formulation, as expressed by (2.15) and (2.16). The magnetic field expressions (2.16) and (2.31) are the same, while the definition of the loop inductance as expressed by (2.29) is obtained from (2.15) by assuming that the current flows in well formed loops; the thin filament definition of the mutual inductance (2.22) is the result of further simplification of (2.15). The magnetic energy and field flux derivations of the inductance are equivalent; both (2.15) and (2.29) can be obtained from either the energy formulation expressed by (2.31) or the flux formulation expressed by (2.22).

The loop inductance approach provides a more convenient description of the magnetic properties of the circuit with little loss of accuracy and generality, as compared to the field formulation as expressed by (2.16). Nevertheless, significant disadvantages remain. In the magnetic flux formulation of the circuit inductance, the basic inductive element is a closed loop. This aspect presents certain difficulties for a traditional circuit analysis approach. In circuit analysis, the impedance characteristics are described in terms of the circuit elements connecting two circuit nodes. Circuit analysis tools also use a circuit representation based on two-terminal elements. Few circuit elements are manufactured in a loop form. In a strict sense, a physical inductor is also a two terminal element. The current flowing through a coil does not form a complete loop, therefore, the definition of the loop inductance does not apply. The loop formulation does not provide a direct link between the impedance characteristics of the circuit and the impedance of the comprising two terminal circuit elements.

It is therefore of practical interest to examine how the inductive characteristics can be described by a network of two terminal elements with self- and mutual impedances, without resorting to a multiple loop representation. This topic is the subject of the next section.

2.1.3 Partial Inductance

The loop inductance, as defined by (2.28), can be deconstructed into more basic elements if the two loops are broken into segments, as shown in Fig. 2.3. The loop ℓ is broken into N segments $S_1, \ldots, S_N$ and loop ℓ' is broken into N' segments $S'_1, \ldots, S'_{N'}$. The definition of the loop inductance (2.28) can be rewritten as

$$M_{\ell\ell'} = \sum_{i=1}^{N} \sum_{j=1}^{N'} \frac{\mu}{4\pi} \frac{1}{a_i\, a'_j} \oint_{S_i} \oint_{S'_j} \int_{a_i} \int_{a'_j} \frac{da_i\, da'_j\, d\boldsymbol{l}\, d\boldsymbol{l}'}{|\boldsymbol{r} - \boldsymbol{r}'|} = \sum_{i=1}^{N} \sum_{j=1}^{N'} L_{ij}, \tag{2.32}$$

where

$$L_{ij} \equiv \frac{\mu}{4\pi} \frac{1}{a_i\, a'_j} \oint_{S_i} \oint_{S'_j} \int_{a_i} \int_{a'_j} \frac{da_i\, da'_j\, d\boldsymbol{l}\, d\boldsymbol{l}'}{|\boldsymbol{r} - \boldsymbol{r}'|}\,. \tag{2.33}$$

The integration along segments S_i and S'_j in (2.32) and (2.33) is performed in the direction of the current flow.

Equation (2.33) defines the mutual partial inductance between two arbitrary segments S_i and S'_j. Similar to the loop inductance, the mutual partial inductance can be either positive or negative, depending on the direction of the current flow in the two segments. In the special case where S_i is identical to S'_j, (2.33) defines the partial self-inductance of S_i. The partial self-inductance is always positive.

The partial inductance formulation, as defined by (2.33), is more suitable for circuit analysis as the basic inductive element is a two terminal segment of interconnect. Any circuit can be decomposed into a set of interconnected two terminal elements. For example, the circuit shown in Fig. 2.2 can be decomposed

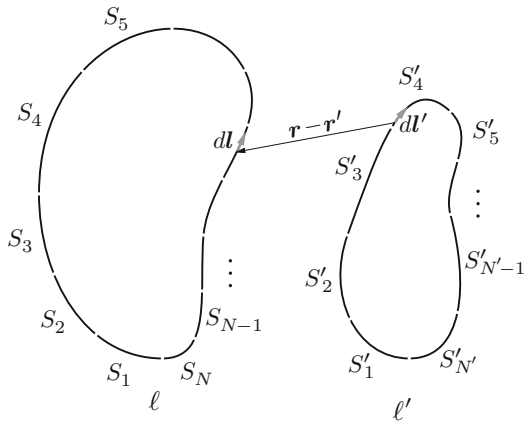


Fig. 2.3 Two complete current loops broken into segments

into three linear segments instead of two loops as in the case of a loop analysis. The magnetic properties of the circuit are described by a partial inductance matrix $\mathbf{L} = \{L_{ij}\}$. Assigning to each element S_i a corresponding current I_i, the vector of magnetic electromotive forces $\boldsymbol{V}$ across each segment is

$$\boldsymbol{V} = \mathbf{L}\frac{d\boldsymbol{I}}{dt}. \tag{2.34}$$

The magnetic energy of the circuit in terms of the partial inductance is determined, analogously to the loop inductance formulation (2.31), as

$$W_{\mathrm{m}} = \frac{1}{2}\boldsymbol{I}^T\mathbf{L}\boldsymbol{I} = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} L_{ij}I_iI_j. \tag{2.35}$$

The partial inductance has another practical advantage. If the self- and mutual partial inductance of a number of basic segment shapes is determined as a function of the segment dimensions and orientations, the partial inductance matrix of any circuit composed of these basic shapes can be readily constructed according to the segment connectivity, permitting the efficient analysis of the magnetic properties of the circuit. In this regard, the partial inductance approach is more flexible than the loop inductance approach, as loops exhibit a greater variety of shapes and are difficult to precharacterize in an efficient manner.

For the purposes of circuit characterization, it is convenient to separate the sign and the absolute magnitude of the inductance. During precharacterization, the absolute magnitude of the mutual partial inductance L_{ij}^{abs} between basic conductor shapes (such as straight segments) is determined. During the process of analyzing a specific circuit structure, the absolute magnitude is multiplied by a sign function s_{ij}, resulting in the partial inductance as defined by (2.33), $L_{ij} = s_{ij}L_{ij}^{\mathrm{abs}}$. The sign function equals either 1 or -1, depending upon the sign of the scalar product of the segment currents: $s_{ij} = \mathrm{sign}\,(\boldsymbol{I}_i \cdot \boldsymbol{I}_j')$.

The case of a straight wire is of particular practical importance. A conductor of any shape can be approximated by a number of short straight segments. The partial self-inductance of a straight round wire is [44]

$$L_{\mathrm{line}} = \frac{\mu l}{2\pi}\left(\ln\frac{2l}{r} - \frac{3}{4}\right), \tag{2.36}$$

where l is the length of the wire and r is the radius of the wire cross section, as shown in Fig. 2.4. The precise analytic expressions for the partial inductance are generally not available for straight conductors with a radially asymmetric cross section. The partial inductance of a straight line with a square cross section can be evaluated with good accuracy using approximate analytic expressions augmented with tables of correction coefficients [44], or expressions suitable for efficient numerical evaluation [45].

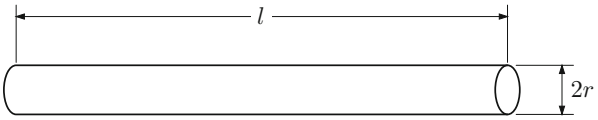


Fig. 2.4 A straight round wire

The partial self-inductance, as expressed by (2.33), depends only on the shape of the conductor segment. It is therefore possible to assign a certain partial self-inductance to an individual segment of the conductor. It should be stressed, however, that the partial self-inductance of the comprising conductors by itself provides no information on the inductive properties of the circuit. For example, a loop of wire can have a loop inductance that is much greater than the sum of the partial self-inductance of the comprising segments (where the wire is coiled) or much smaller than the sum of the comprising partial self-inductances (where the wire forms a narrow long loop). The inductive properties of a circuit are described by *all* partial inductances in the circuit, necessarily including all mutual partial inductances between all pairs of elements, as expressed in (2.32) for the specific case of a current loop.

Unlike the loop inductance, the partial inductance cannot be measured experimentally. The partial inductance is, essentially, a convenient mathematical construct used to describe the inductive properties of a circuit. This point is further corroborated by the fact that the partial inductance is not uniquely defined. An electromagnetic field is described by an infinite number of vector potentials. If a specific field is described by a vector potential $\boldsymbol{A}$, any vector potential $\boldsymbol{A}'$ differing from $\boldsymbol{A}$ by a gradient of an arbitrary scalar function Ψ, i.e., $\boldsymbol{A}' = \boldsymbol{A} + \boldsymbol{\nabla}\Psi$, also describes the field.[3] The magnetic field is determined through the curl operation of the vector potential and is not affected by the $\boldsymbol{\nabla}\Psi$ term, $\nabla \times \boldsymbol{A} = \nabla \times \boldsymbol{A'}$ as $\boldsymbol{\nabla} \times \boldsymbol{\nabla}\Psi = 0$. The choice of a specific vector potential is inconsequential. The vector potential definition (2.11) is therefore not unique. The choice of a specific vector potential is also immaterial in determining the loop inductance as expressed by (2.28), as the integration of a gradient of any function over a closed contour yields a null value. The choice of the vector potential, however, affects the value of the partial inductance, where the integration is performed over a conductor segment. Equation (2.33) therefore defines only one of many possible partial inductance matrices. This ambiguity does not present a problem as long as all of the partial inductances in the circuit are consistently determined using the same vector potential. The contributions of the function gradient to the partial inductance cancel out, where the partial inductances are combined to describe the loop currents.

In the case of straight line segments, the partial inductance definition expressed by (2.33) has an intuitive interpretation. For a straight line segment, the partial

[3]This property of the electromagnetic field is referred to in electrodynamics as gauge invariance.

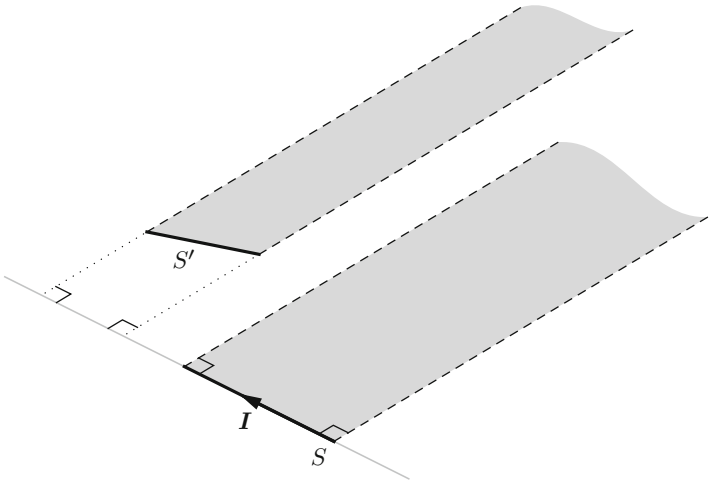


Fig. 2.5 Self- and mutual partial inductance of a straight segment of wire. The partial self-inductance of a segment S, as described by Rosa [46], is determined using the magnetic flux created by current $\boldsymbol{I}$ in segment S through an infinite contour formed by wire segment S (the *bold arrow*) and two rays perpendicular to the segment (the *dashed lines*). Similarly, the partial mutual inductance with another wire segment S' is determined using the flux created by current $\boldsymbol{I}$ through the contour formed by the segment S' and straight lines originating from the ends of the segment S' and perpendicular to segment S

self-inductance is a coefficient of proportionality between the segment current and the magnetic flux through the infinite loop formed by a line segment S and two rays perpendicular to the segment, as illustrated in Fig. 2.5.

This flux is henceforth referred to as a partial flux. This statement can be proved as follows. The flux through the aforementioned infinite loop is determined by integrating the vector potential $\boldsymbol{A}$ along the loop contour, according to (2.25). The magnetic vector potential $\boldsymbol{A}$ of a straight segment, as determined by (2.11), is parallel to the segment. The integration of the vector potential along the rays perpendicular to the segment is zero. The integration of the vector potential along the segment completing the loop at infinity is also zero as the vector potential decreases inversely proportionally with distance. Similarly, the mutual partial inductance between segments S and S' can be interpreted in terms of the magnetic flux through the infinite loop formed by segment S' and two rays perpendicular to the segment S, as illustrated in Fig. 2.5.

This interpretation of the partial inductance in terms of the partial flux is in fact the basis for the original introduction of the partial inductance by Rosa in 1908 in application to linear conductors [46]. Attempts to determine the inductance of a straight wire segment using the total magnetic flux were ultimately unsuccessful as the total flux of a segment is infinite. Rosa made an intuitive argument that only the partial magnetic flux, as illustrated in Fig. 2.5, should be associated with the

self-inductance of the segment. The concept of partial inductance proved useful and was utilized in the inductance calculation formulæ and tables developed by Rosa and Cohen [47], Rosa and Grover [48], and Grover [44]. A rigorous theoretical treatment of the subject was first developed by Ruehli in [45], where a general definition of the partial inductance of an arbitrarily shaped conductor (2.33) is derived. Ruehli also coined the term "partial inductance."

Connections between the loop and partial inductance can also be established in terms of the magnetic flux. The magnetic flux through a specific loop is a sum of all of the partial fluxes of the comprising segments. The contribution of a magnetic field created by a specific loop segment to the loop flux is also the sum of all of the partial inductances of this segment with respect to all segments of the loop. This relationship is illustrated in Fig. 2.6.

2.1.4 Net Inductance

The inductance of a circuit without branch points (i.e., where the current flowing in all conductor segments is the same) can also be expressed in a form with no explicit mutual inductances. Consider a current loop consisting of N segments. As discussed in the previous section, the loop inductance L_{loop} can be described in terms of the partial inductances L_{ij} of the segments,

$$L_{\text{loop}} = \sum_{i=1}^{N} \sum_{j=1}^{N} L_{ij}. \tag{2.37}$$

This sum can be rearranged as

$$L_{\text{loop}} = \sum_{i=1}^{N} L_i^{\text{eff}}, \tag{2.38}$$

where

$$L_i^{\text{eff}} \equiv \sum_{i,j=1}^{N} L_{ij}. \tag{2.39}$$

The inductance L_i^{eff}, as defined by (2.39), is often referred to as the *net* inductance [49–51]. The net inductance also has an intuitive interpretation in terms of the magnetic flux. As illustrated in Fig. 2.6, a net inductance (i.e., the partial self-inductance plus the partial mutual inductances with all other segments) of the segment determines the contribution of the segment current to the overall magnetic flux through the circuit.

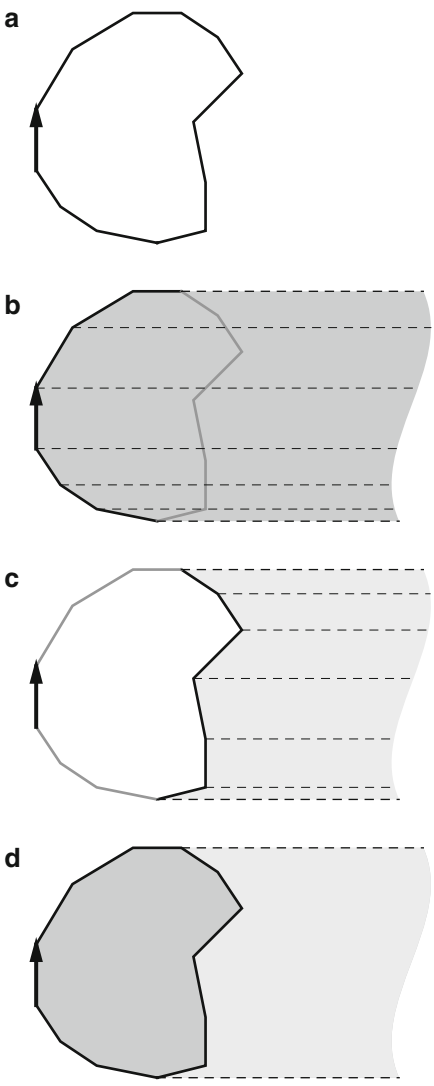


Fig. 2.6 The contribution of a current in a specific loop segment (shown with a *bold arrow*) to the total flux of the current loop is composed of the partial flux of this segment with all other segments of the loop; (**a**) a piecewise linear loop, (**b**) partial flux of the segment with all other segments carrying current in the same direction (i.e., the scalar product of the two segment vectors is positive)—this flux is positive, (**c**) the partial flux of the segment with all other segments carrying current in the opposite direction (i.e., the scalar product of the two segment vectors is negative)—this flux is negative, (**d**) the sum of the positive and negative fluxes, shown in (**b**) and (**c**) (i.e., the geometric difference between the contours (**b**) and (**c**)), is the overall contribution of the segment to the magnetic flux of the loop—this contribution is expressed as the net inductance of the segment

The net inductance describes the behavior of coupled circuits without using explicit mutual inductance terms, simplifying the circuit analysis process. For example, consider a current loop consisting of a signal current path with inductance L_{sig} and return current path with inductance L_{ret}, as shown in Fig. 2.7. The mutual inductance between the two paths is M. The net inductance of the two paths is $L_{\text{sig}}^{\text{eff}} = L_{\text{sig}} - M$ and $L_{\text{ret}}^{\text{eff}} = L_{\text{ret}} - M$. The loop inductance in terms of the net inductance is $L_{\text{loop}} = L_{\text{sig}}^{\text{eff}} + L_{\text{ret}}^{\text{eff}}$. The inductive voltage drop along the return current path is $V_{\text{ret}} = L_{\text{ret}}^{\text{eff}} \frac{dI}{dt}$.

The net inductance has another desirable property. Unlike the partial inductance, the net inductance is independent of the choice of the magnetic vector potential,

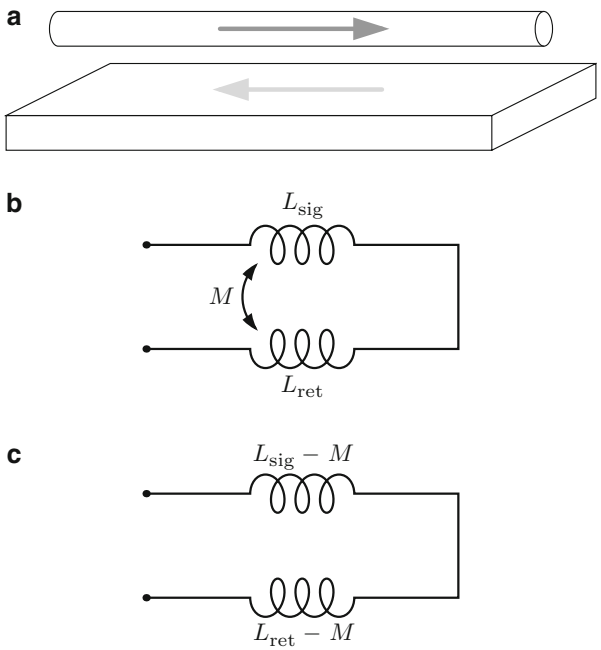


Fig. 2.7 The signal and return current paths. (**a**) The physical structure of the current loop. (**b**) The equivalent partial inductance model. (**c**) The equivalent net inductance model

because, similar to the loop inductance, the integration of the vector potential is performed along a complete loop, as implicitly expressed by (2.39). The net inductance is therefore uniquely determined.

Note that the net inductance of a conductor depends on the structure of the overall circuit as indicated by the mutual partial inductance terms in (2.39). Modifying the shape of a single segment in a circuit changes the net inductance of *all* of the segments. The net inductance is, in effect, a specialized form of the partial inductance and should only be used in the specific circuit where the net inductance terms are determined according to (2.39).

2.2 Variation of Inductance with Frequency

A circuit inductance, either loop or partial, depends upon the current distribution across the cross section of the conductors, as expressed by (2.23) and (2.24). The current density is assumed constant across the conductor cross sections in the inductance formulas described in Sect. 2.1, as is commonly assumed in practice. This assumption is valid where the magnetic field does not appreciably change the path of the current flow. The conditions where this assumption is accurate

are discussed in Sect. 2.2.1. Where the effect of the magnetic field on the current path is significant, the current density becomes non-uniform and the magnetic properties of the circuit vary significantly with frequency. The mechanisms causing the inductance to vary with frequency are described in Sect. 2.2.2. A circuit analysis of the variation of inductance with frequency is performed in Sect. 2.2.3 based on a simple circuit model. The section concludes with a discussion of the relative significance of the different inductance variation mechanisms.

2.2.1 Uniform Current Density Approximation

The effect of the magnetic field on the current distribution can be neglected in two general cases. First, the current density is uniform where the magnetic impedance $L\,dI/dt$ is much smaller than the resistive impedance R of the interconnect structure. Under this condition, however, the magnetic properties of the circuit do not significantly affect the circuit behavior and are typically of little practical interest. The second case is of greater practical importance, where the magnetic impedance to the current flow, although greater than R, is uniform across the cross section of a conductor. This condition is generally satisfied where the separation between conductors is significantly greater that the cross-sectional dimensions. It can be shown by inspecting (2.11) that at a distance d much greater than the conductor cross-sectional dimension a, a non-uniform current distribution within the conductor contributes only a second order correction to the magnetic vector potential $\boldsymbol{A}$. The significance of this correction as compared to the primary term decreases with distance as a/d.

Where the separation of two conductors is comparable to the cross-sectional dimensions, the magnetic field significantly affects the current distribution within the conductors. The current density distribution across the cross section becomes non-uniform and varies with the signal frequency. In this case, the magnetic properties of an interconnect structure cannot be accurately represented by a constant value. Alternatively stated, the inductance varies with the signal frequency.

The frequency variation of the current density distribution and, consequently, of the conductor inductance can be explained from a circuit analysis point of view if the impedance characteristics of different paths *within the same conductor* are considered, as described in Sect. 2.2.2. The resistive properties of alternative parallel paths within the same conductors are identical, provided the conductivity of the conductor material is uniform. The magnetic properties of the paths however can be significantly different. At low frequencies, the impedance of the current paths is dominated by the resistance. The current density is uniform across the cross section, minimizing the overall impedance of the conductor. At sufficiently high frequencies, the impedance of the current paths is dominated by the inductive reactance. As the resistive impedance becomes less significant (as compared to the inductive impedance) at higher frequencies, the distribution of the current density

asymptotically approaches the density profile that yields the minimum overall inductance of the interconnect structure. The inductance of the on-chip interconnect structures can therefore decrease significantly with signal frequency.

2.2.2 *Inductance Variation Mechanisms*

As discussed, the variation of inductance is the result of the variation of the current density distribution. The variation of the current distribution with frequency can be loosely classified into several categories.

Skin Effect

With the onset of the skin effect, the current becomes increasingly concentrated near the line surface, causing a decrease in the magnetic field within the line core, as illustrated in Fig. 2.8. The magnetic field outside the conductor is affected relatively little. It is therefore convenient to divide the circuit inductance into "internal" and "external" parts, $L = L_{\text{internal}} + L_{\text{external}}$, where L_{external} is the inductance associated with the magnetic field outside the conductors and L_{internal} is the inductance associated with the magnetic field inside the conductors. In these terms, a well developed skin effect produces a significant decrease in the internal inductance L_{internal}. For a round wire at low frequency (where the current distribution is uniform across the line cross section), the internal inductance is $0.05\,\frac{\text{nH}}{\text{mm}}$, independent of the radius (see the derivation in [52]). The external inductance of the round wire is unaffected by the skin effect.

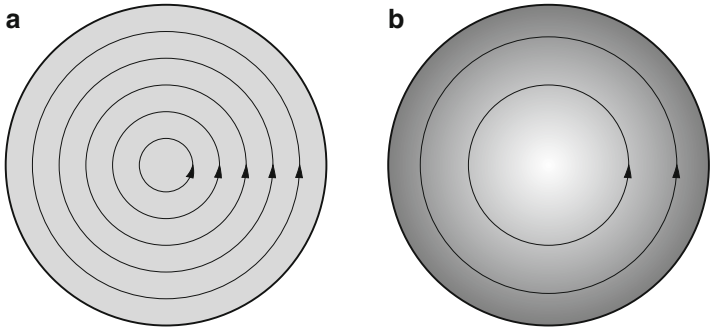


Fig. 2.8 Internal magnetic flux of a round conductor; (**a**) at low frequencies, the current density, as shown by the *shades of gray*, is uniform, resulting in the maximum magnetic flux inside the conductor, as shown by the *circular arrows*, and the associated internal inductance, (**b**) at high frequencies, the current flow is redistributed to the surface of the conductor, reducing the magnetic flux inside the conductor

Fig. 2.9 Proximity effect in two closely spaced lines. Current density distribution in the cross section of two closely spaced lines at high frequencies is shown in *shades of gray*. *Darker shades of gray* indicate higher current densities. In lines carrying current in the same direction (parallel currents), the current concentration is shifted away from the parallel current. In lines carrying current in opposite directions (antiparallel currents), the current concentrates toward the antiparallel current, minimizing the circuit inductance

Proximity Effect

The current distribution also varies with frequency due to the proximity effect. At high frequencies, the current in the line concentrates along the side of the line facing an adjacent current return path, thereby reducing the effective area of the current loop and thus the loop inductance, as illustrated in Fig. 2.9.

The skin and proximity effects are closely related. These effects represent basically the same phenomenon—the tendency of the current to move closer to the current return path in order to minimize the interconnect inductance at high frequencies. When a conductor is surrounded by several alternative current return paths, leading to a relatively symmetric current distribution at high frequency, the effect is typically referred to as the skin effect. The classical example of such an interconnect structure is a coaxial cable, where the shield provides equivalent current return paths along all sides of the core conductor. In the case where the current distribution is significantly asymmetric due to the close proximity of a dominant return path, the effect is referred to as the proximity effect.

Multi-path Current Redistribution

The concept of current density redistribution within a conductor can be extended to redistribution of the current among several separate parallel conductors. This mechanism is henceforth referred to as *multi-path current redistribution*. For example, in standard single-ended digital logic, the forward current path is typically composed of a single line. No redistribution of the forward current occurs. The current return path, though, is not explicitly specified (although local shielding for particularly sensitive nets is becoming more common [53, 54]). Adjacent signal lines, power lines, and the substrate provide several alternative current return paths. A significant redistribution of the return current among these return paths can occur as signal frequencies increase. At low frequencies, the line impedance $Z(\omega) = R(\omega) + j\omega L(\omega)$ is dominated by the interconnect resistance R. In this case, the distribution of the return current over the available return paths is determined by the path resistance, as shown in Fig. 2.10a. The return current spreads out far

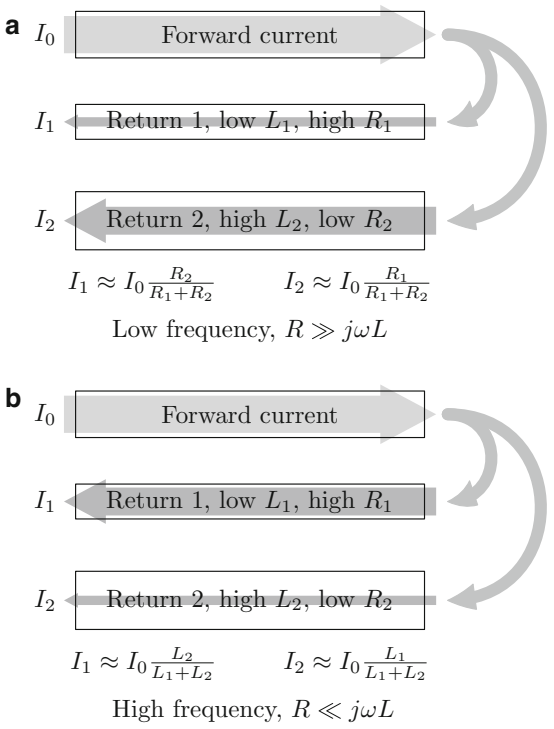


Fig. 2.10 Current loop with two alternative current return paths. The forward current I_0 returns both through return path one with resistance R_1 and inductance L_1, and return path two with resistance R_2 and inductance L_2. In this structure, $L_1 < L_2$ and $R_1 > R_2$. At low frequencies (**a**), the path impedance is dominated by the line resistance and the return current is distributed between two return paths according to the resistance of the lines. Thus, at low frequencies, most of the return current flows through the return path of lower resistance, path two. At high frequencies (**b**), however, the path impedance is dominated by the line inductance and the return current is distributed between two return paths according to the inductance of the lines. Most of the return current flows through the path of lower inductance, path one, minimizing the overall inductance of the circuit

from the signal line to reduce the resistance of the return path and, consequently, the impedance of the current loop. At high frequencies, the line impedance $Z(\omega) = R(\omega) + j\omega L(\omega)$ is dominated by the reactive component $j\omega L(\omega)$. The minimum impedance path is primarily determined by the inductance $L(\omega)$, as shown in Fig. 2.10b. Multi-path current redistribution, as described in Fig. 2.10, is essentially a proximity effect extended to several separate lines connected in parallel. In power grids, both the forward and return currents undergo multi-path redistribution as both the forward and return paths consist of multiple conductors connected in parallel.

The general phenomenon underlying these three mechanisms is, as viewed from a circuit perspective, the same. Where several parallel paths with significantly different electrical properties are available for current flow, the current is distributed

among the paths so as to minimize the total impedance. As the frequency increases, the circuit inductance changes from the low frequency limit, determined by the ratio of the resistances of the parallel current paths, to the high frequency value, determined by the inductance ratios of the current paths. At high signal frequencies, the inductive reactance dominates the interconnect impedance; therefore, the path of minimum inductance carries the largest share of the current, minimizing the overall impedance (see Fig. 2.10). Note that parallel current paths can be formed either by several physically distinct lines, as in multi-path current redistribution, or by different paths within the same line, as in skin and proximity effects, as shown in Fig. 2.11. The difference is merely in the physical structure, the electrical behavior is fully analogous. A thick line can be thought of as being composed of multiple thin lines bundled together in parallel. The skin and proximity effects in such a thick line can be considered as a special case of current redistribution among multiple thin lines forming a thick line.

2.2.3 Simple Circuit Model

A simple model of current redistribution provides deeper insight into the process of inductance variation. This approach can be used to estimate the relative significance of the different current distribution mechanisms in various interconnect structures as well as the frequency characteristics of the inductance. Consider a simple case of two current paths with different inductive properties (for example, as shown in Fig. 2.11). The impedance characteristics are represented by the circuit diagram shown in Fig. 2.12, where the inductive coupling between the two paths is neglected for simplicity. Assume that $L_1 < L_2$ and $R_1 > R_2$.

For the purpose of evaluating the variation of inductance with frequency, the electrical properties of the interconnect are characterized by the inductive time

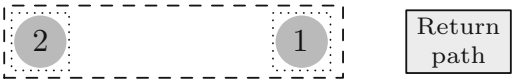
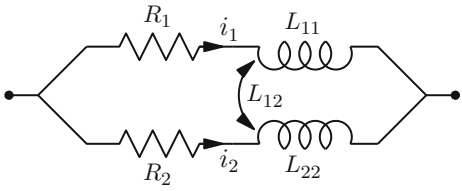


Fig. 2.11 A cross-sectional view of two parallel current paths (*gray circles*) sharing the same current return path (*gray rectangle*). The path closest to the return path, path 1, has a lower inductance than the other path, path 2. The parallel paths can be either two physically distinct lines, as shown by the *dotted line*, or two different paths within the same line, as shown by the *dashed line*

Fig. 2.12 A circuit model of two current paths with different inductive properties



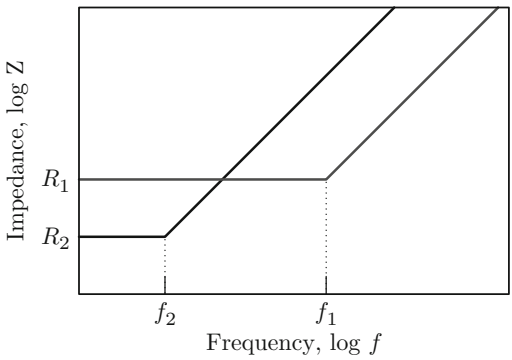


Fig. 2.13 Impedance magnitude versus frequency for two paths with dissimilar impedance characteristics

constant $\tau = L/R$. The impedance magnitude of these two paths is schematically shown in Fig. 2.13. The impedance of the first path is dominated by the inductive reactance above the frequency $f_1 = \frac{1}{2\pi}\frac{R_1}{L_1} = \frac{1}{2\pi\tau_1}$. The impedance of the second path is predominantly inductive above the frequency $f_2 = \frac{1}{2\pi}\frac{R_2}{L_2} = \frac{1}{2\pi\tau_2}$, such that $f_2 < f_1$. At low frequencies, i.e., from DC to the frequency f_1, the ratio of the two impedances is constant. The effective inductance at low frequencies is therefore also constant, determining the low frequency inductance limit. At high frequencies, i.e., frequencies exceeding f_2, the ratio of the impedances is also constant, determining the high frequency inductance limit, $\frac{L_1L_2}{L_1+L_2}$. At intermediate frequencies from f_1 to f_2, the impedance ratio changes, resulting in a variation of the overall inductance from the low frequency limit to the high frequency limit. The frequency range of inductance variation is therefore determined by the two time constants, τ_1 and τ_2. The magnitude of the inductance variation depends upon both the difference between the time constants τ_1 and τ_2 and on the inductance ratio L_1/L_2. Analogously, in the case of multiple parallel current paths, the frequency range and the magnitude of the variation in inductance is determined by the minimum and maximum time constants as well as the difference in inductance among the current paths.

The decrease in inductance begins when the inductive reactance $j\omega L$ of the path with the lowest R/L ratio becomes comparable to the path resistance R, $R \sim j\omega L$. The inductance, therefore, begins to decrease at a lower frequency if the minimum R/L ratio of the current paths is lower.

Due to this behavior, the proximity effect becomes significant at higher frequencies than the frequencies at which multi-path current redistribution becomes significant. Significant proximity effects occur in conductors containing current paths with significantly different inductive characteristics. That is, the inductive coupling of one edge of the line to the "return" current (i.e., the current in the opposite direction) is substantially different from the inductive coupling of the other edge of the line to the same "return" current. In geometric terms, this characteristic means that the line width is larger than or comparable to the distance between the line and the return current. Consequently, the line with significant proximity effects

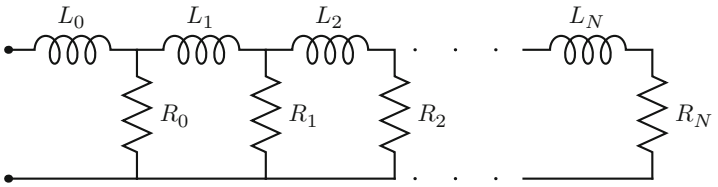


Fig. 2.14 An *RL* ladder circuit describing the variation of inductance with frequency

is typically the immediate neighbor of the current return line. A narrower current loop is therefore formed with the current return path as compared to the other lines participating in the multi-path current redistribution. A smaller loop inductance L results in a higher R/L ratio. Referring to Fig. 2.10, current redistribution between paths one and two develops at frequencies lower than the onset frequency of the proximity effect in path one.

Efficient and accurate lumped element models are necessary to incorporate skin and proximity effects into traditional circuit simulation tools. Developing such models is an area of ongoing research [55–61]. The resistance and internal inductance of conductors are typically modeled with *RL* ladder circuits [55], as shown in Fig. 2.14. The sections of the *RL* ladder represent the resistance and inductance of the conductor parts at different distances from the current return path. Different methods for determining the value of the R and L elements have been developed [56–58]. Analogously, *RL* ladders can also be extended to describe multi-path current redistribution [59, 60]. Techniques for reducing the order of a transfer function of an interconnect structure have also been described [61].

2.3 Inductive Behavior of Circuits

The strict meaning of the term "inductance" is the *absolute inductance*, as defined in Sect. 2.1. The absolute inductance is measured in henrys. Sometimes, however, the same term "inductance" is loosely used to denote the *inductive behavior* of a circuit; namely, overshoots, ringing, signal reflections, etc. The inductive behavior of a circuit is characterized by such quantities as a damping factor and the magnitude of the overshoot and reflections of the signals. While any circuit structure carrying an electrical current has a finite absolute inductance, as defined in Sect. 2.1, not every circuit exhibits inductive behavior. Generally, a circuit exhibits inductive behavior if the absolute inductance of the circuit is sufficiently high. The relationship between the inductive behavior and the absolute inductance is, however, circuit specific and no general metrics for the onset of inductive behavior have been developed.

Specific metrics have been developed to evaluate the onset of inductive behavior in high speed digital circuits [62–64]. A digital signal that is propagating in an underdriven uniform lossy transmission line exhibits significant inductive effects if the line length l satisfies the following condition [63],

$$\frac{t_\mathrm{r}}{2\sqrt{LC}} < l < \frac{2}{R}\sqrt{\frac{L}{C}}, \tag{2.40}$$

where R, L, and C are the resistance, inductance, and capacitance per line length, respectively, and t_r is the rise time of the signal waveform.

The two inequalities comprising condition (2.40) have an intuitive circuit interpretation. The velocity of the electromagnetic signal propagation along a line is $v_\mathrm{c} = \frac{1}{\sqrt{LC}}$. The left inequality of (2.40) therefore transforms into

$$t_\mathrm{r} < \frac{2l}{v_\mathrm{c}}, \tag{2.41}$$

i.e., the signal rise time should be smaller than the round trip time of flight. Alternatively stated, the line length l should be a significant fraction of the shortest wavelength of significant signal frequencies λ_r. The spectral content of the signal with rise time t_r rolls off at -20 dB/decade above the frequency $f_\mathrm{r} = 1/\pi t_\mathrm{r}$. The shortest effective wavelength of the signal is therefore $\lambda_\mathrm{r} = v_c/f_\mathrm{r} = \pi v_c t_\mathrm{r}$. The condition (2.41) can be rewritten as

$$\frac{l}{\lambda_\mathrm{r}} > \frac{1}{2\pi}. \tag{2.42}$$

The dimensionless ratio of the physical size of a circuit to the signal wavelength, l/λ, is referred to as the *electrical size* in high speed interconnect design [51, 65]. Circuits with an electrical size much smaller than unity are commonly called electrically small (or short), otherwise circuits are called electrically large (or long) [51, 65]. Electrically small circuits belong to the realm of classical circuit analysis and are well described by lumped circuits. Electrically large circuits require distributed circuit models and belong to the domain of high speed interconnect analysis techniques. The left inequality of condition (2.40) therefore restricts significant inductive effects to electrically long lines.

With the notion that the damping factor of the transmission line is $\zeta = \frac{R_0}{2}\sqrt{\frac{C_0}{L_0}}$, where $R_0 = Rl$, $L_0 = Ll$, and $C_0 = Cl$ are the total resistance, inductance, and capacitance of the line, respectively, the right inequality in condition (2.40) transforms into

$$\zeta < 1, \tag{2.43}$$

constraining the damping factor to be sufficiently small. Given a line with a specific R, L, and C, the inductive behavior is confined to a certain range of line length, as shown in Fig. 2.15. The upper bound of this range is determined by the damping factor of the line, while the lower bound is determined by the electrical size of the line.

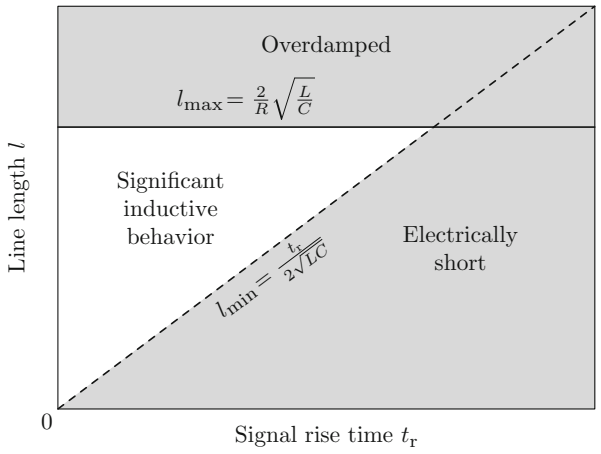


Fig. 2.15 The range of transmission line length where the signal propagation exhibits significant inductive behavior. The area of inductive behavior (the *unshaded area*) is bounded by the conditions of large electrical size (the *dashed line*) and insufficient damping (the *solid line*), as determined by (2.40). In the region where either of these conditions is not satisfied (the *shaded area*), the inductive effects are insignificant

Alternatively, condition (2.40) can be interpreted as a bound on the overall line inductance $L_0 = Ll$. The signal transmission exhibits inductive characteristics if the overall line inductance satisfies both of the following conditions,

$$L_0 > \frac{t_r^2}{4C_0} \tag{2.44}$$

and

$$L_0 > \frac{1}{4}R_0^2 C_0 \,. \tag{2.45}$$

Conditions (2.44) and (2.45) thereby quantify the term "inductance sufficiently large to cause inductive behavior" as applied to transmission lines. The design space for a line inductance with the region of inductive behavior, as determined by (2.44) and (2.45), is illustrated in Fig. 2.16.

2.4 Inductive Properties of On-Chip Interconnect

The distinctive feature of on-chip interconnect structures is the small cross-sectional dimensions and, consequently, a relatively high line resistance. For example, the resistance of a copper line with a $1 \times 3\,\mu$m cross section is approximately $7\,\Omega/$mm.

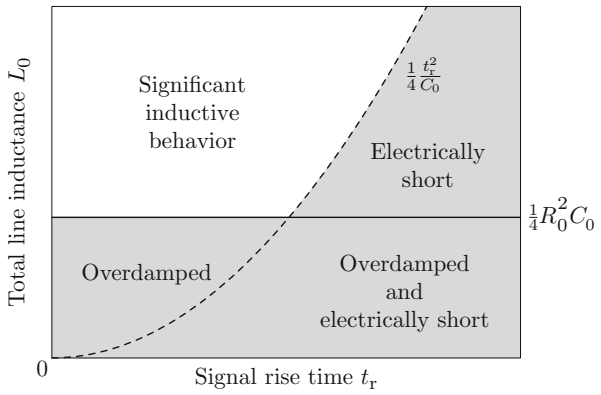


Fig. 2.16 The design space characterizing the overall transmission line inductance is divided into a region of inductive behavior and a region where inductive effects are insignificant. The region of inductive behavior (the *unshaded area*) is bounded by the conditions of large electrical size (the *dashed line*) and low damping (the *solid line*), as determined by (2.44) and (2.45). In the region where either of these conditions is not satisfied (the *shaded area*), the inductive effects are insignificant

The loop inductance of on-chip lines is typically between 0.4 nH/mm and 1 nH/mm. At frequencies lower than several gigahertz, the magnetic characteristics do not significantly affect the behavior of on-chip circuits.

As the switching speed of digital integrated circuits increases with technology scaling, the magnetic properties have become essential for accurately describing on-chip circuit operation. The density and complexity of the on-chip interconnect structures preclude exploiting commonly assumed circuit simplifications, rendering the accurate analysis of inductive properties particularly challenging. Large integrated circuits contain many tens of millions of interconnect segments while the segment spacing is typically either equal to or less than the cross-sectional dimensions. Accurate treatment of magnetic coupling in these conditions is especially important. Neither the loop nor the partial inductance formulation can be directly applied to an entire circuit as the size of the resulting inductance matrices makes the process of circuit analysis computationally infeasible. Simplifying the inductive properties of a circuit is also difficult. Simply omitting relatively small partial inductance terms can significantly change the circuit behavior, possibly causing instability in an originally passive circuit. Techniques to simplify the magnetic characteristics so as to allow an accurate analysis of separate circuit parts is currently an area of focused research [66–69].

The problem is further complicated by the significant variation of inductance with frequency. As discussed in Sect. 2.2, the inductance variation can be described in terms of the skin effect, proximity effect, and multi-path current redistribution. For a line with a rectangular cross section, the internal inductance is similar to the internal inductance of a round line, i.e., 0.05 nH/mm, decreasing with the aspect ratio of the cross section. Over the frequency range of interest, up to 100 GHz, the

skin effect reduces the internal inductance by only a small fraction. The reduction in the internal inductance due to the skin effect is, therefore, relatively insignificant, as compared to the overall inductance. Due to the relatively high resistance of on-chip interconnect, the proximity effect is significant only in immediately adjacent wide lines that carry high frequency current. Where several parallel lines are available for current flow, redistribution of the current among the lines is typically the primary cause in integrated circuits of the decrease in inductance with frequency. The proximity effect and multi-path current redistribution are therefore two mechanisms that can produce a significant change in the on-chip interconnect inductance with signal frequency.

Note that the statement "sufficiently high inductance causes inductive behavior" does not necessarily mean "any change in the interconnect physical structure that increases the line inductance increases the inductive behavior of the line." In fact, the opposite is often the case in an integrated circuit environment, where varying a single physical interconnect characteristic typically affects many electrical characteristics. The relationship between the physical structure of interconnect and the inductive behavior of a circuit is highly complex.

Consider a 3 mm long copper line with a $1 \times 1\,\mu$m cross section. The resistance, inductance, and capacitance per length of the current loop (including both the line itself and the current return path) are, respectively, $R = 25\,\Omega/\text{mm}$, $L = 0.8\,\text{nH/mm}$, and $C = 100\,\text{fF/mm}$. The velocity of the electromagnetic wave propagation along the line is 0.11 mm/ps. This velocity is somewhat smaller than the speed of light, 0.15 mm/ps, in the media with an assumed dielectric constant of 4 and is due to the additional capacitive load of the orthogonal lines in the lower layer. For a signal with a 30 ps rise time, the line is electrically long. The line damping factor, however, is $\zeta = \frac{Rl}{2}\sqrt{\frac{C}{L}} = 1.33 > 1$. The line is therefore overdamped and, according to the metrics expressed by (2.44) and (2.45), does not exhibit inductive behavior, as shown in Fig. 2.17a.

Assume now that the line width is $4\,\mu$m and the resistance, inductance, and capacitance of the line change, respectively, to $R = 10\,\Omega/\text{mm}$, $L = 0.65\,\text{nH/mm}$, and $C = 220\,\text{fF/mm}$. The decrease in the loop resistance and inductance are primarily due to the smaller resistance and partial self-inductance of the line. The increase in the line capacitance is primarily due to the greater parallel plate capacitance between the signal line and the perpendicular lines in the lower layer. This capacitive load becomes more significant, as compared to the capacitance between the line and the return path, further slowing the velocity of the electromagnetic wave propagation to 0.084 mm/ps. For the same signal with a 30 ps transition time, the signal line becomes underdamped, $\zeta = 0.87 < 1$, and exhibits significant inductive behavior, as shown in Fig. 2.17b.

The *inductive behavior* has become significant while the *absolute inductance* has *decreased* from $3\,\text{mm} \times 0.8\frac{\text{nH}}{\text{mm}} = 2.40$ to 1.95 nH. The reason for this seeming contradiction is that the inductance is a weak function of the cross-sectional dimensions, as compared to the resistance and capacitance. In integrated circuits, the signal lines that exhibit inductive behavior are the lowest resistance lines, i.e., the

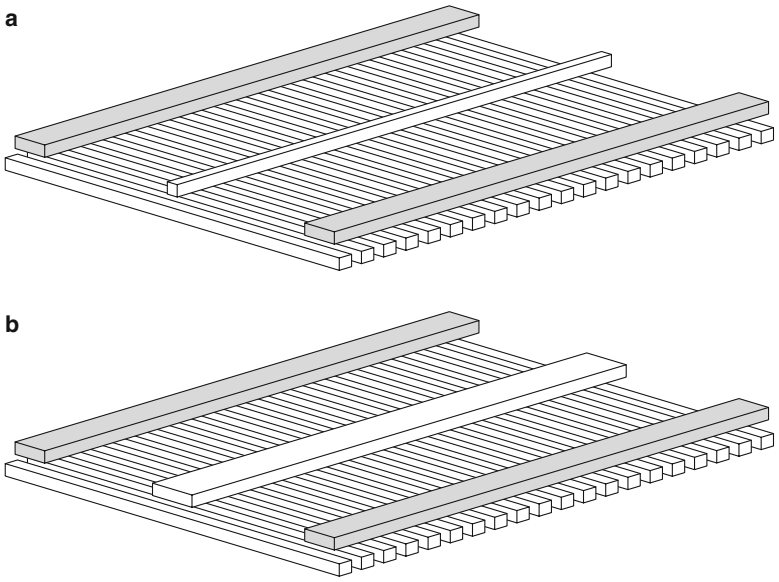


Fig. 2.17 A signal line within an integrated circuit. The power and ground lines (*shaded gray*) parallel to the signal line serve as a current return path. The lines in the lower metal layer increase the capacitive load of the line. The inductive behavior of a wide line, as shown in (**b**), is more significant as compared to a narrow line, as shown in (**a**)

wide lines in the thick upper metalization layers. These lines typically have a lower absolute inductance than other signal lines. It would therefore be misleading to state that the inductive behavior of on-chip interconnect has become important due to the increased inductance. This trend is due to the shorter signal transition times and longer line lengths, while maintaining approximately constant the resistive properties of the upper metal layers.

2.5 Summary

The preceding discussion of the inductive characteristics of electric circuits and different ways to represent these characteristics can be summarized as follows.

- The thin filament approximation is valid only for determining the mutual inductance of relatively thin conductors
- The partial inductance formulation is better suited to describe the inductive properties of circuits with branch points
- The partial inductance is a mathematical construct, not a physically observable property, and should only be used as part of a complete description of the circuit inductance

- The circuit inductance varies with frequency due to current redistribution within the circuit conductors. The current redistribution mechanisms can be classified as the skin effect, proximity effect, and multi-path current redistribution
- Signal propagation along a transmission line exhibits inductive behavior if the line is both electrically long and underdamped
- Characterizing on-chip inductance in both an efficient and accurate manner is difficult due to the density and complexity of on-chip interconnect structures
- The relationship between the physical structure of on-chip interconnect and the inductive behavior of a circuit is complex, as many electrical properties can be affected by changing a specific physical characteristic of an interconnect line

Chapter 3
Properties of On-Chip Inductive Current Loops

The inductive characteristics of electric circuits are described in Chap. 2. Both accurate and efficient characterization of on-chip interconnect inductance is difficult, as discussed in the previous chapter. The objective of this chapter is to demonstrate that the task of inductance characterization, however, is considerably facilitated in certain interconnect structures. These results will be used in Chap. 28 to provide insight into the inductive properties of power distribution grids.

This chapter is organized as follows. A brief overview of the problem is presented in Sect. 3.1. The dependence of inductance on line length is discussed in Sect. 3.2. The inductive coupling of parallel conductors is described in Sect. 3.3. Application of these results to the circuit analysis process is discussed in Sect. 3.4. The conclusions are summarized in Sect. 3.5.

3.1 Introduction

IC performance has become increasingly constrained by on-chip interconnect impedances. Determining the electrical characteristics of the interconnect at early stages of the design process has therefore become necessary. The layout process is driven now by interconnect performance where the electrical characteristics of the interconnect are initially estimated and later refined. Impedance estimation is repeated throughout the circuit design and layout process and should, therefore, be computationally efficient.

The inductance of on-chip interconnect has become an important issue due to the increasing switching speeds of digital integrated circuits [64]. The inductive properties must, therefore, be effectively incorporated at all levels of the design, extraction, and simulation phases in the development of high speed ICs.

Operating an inductance extractor within a layout generator loop is computationally expensive. The efficiency of inductance estimation can be improved by

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0_3

extrapolating the inductive properties of a circuit from the properties of a precharacterized set of structures. Extrapolating inductance, however, is not straightforward as the inductance, in general, varies nonlinearly with the circuit dimensions. The partial inductance of a straight line, for example, is a nonlinear function of the line length. The inductive coupling of two lines decreases slowly with increasing line-to-line separation. The partial inductance representation of a circuit consists of strongly coupled elements with a nonlinear dependence on the geometric dimensions. The inductive properties of a circuit, therefore, do not, in general, vary linearly with the circuit geometric dimensions.

The objective of this chapter is to explore the dependence of inductance on the circuit dimensions, to provide insight from a circuit analysis perspective, and to determine the specific conditions under which the inductance properties scale linearly with the circuit dimensions with the objective of enhancing the layout extraction process. A comparison of the properties of the partial inductance with the properties of the loop inductance is shown to be effective for this purpose. The results of this investigation will be exploited in Chap. 28 to analyze the inductive properties of the power distribution grids.

The analysis is analogous to the procedure described in Sect. 28.3. The inductance extraction program FastHenry [70] is used to explore the inductive properties of interconnect structures. A conductivity of $58\,\mathrm{S/\mu m} \simeq (1.72\,\mu\Omega\cdot\mathrm{cm})^{-1}$ is assumed for the interconnect material.

3.2 Dependence of Inductance on Line Length

A non-intuitive property of the partial inductance is the characteristic that the partial inductance of a line is a nonlinear function of the line length. The partial inductance of a straight line is a superlinear function of length. The partial self-inductance of a rectangular line at low frequency can be described by [44]

$$L_{\mathrm{part}} = 0.2l\left(\ln\frac{2l}{T+W} + \frac{1}{2} - \ln\gamma\right)\ \mu\mathrm{H}, \tag{3.1}$$

where T and W are the thickness and width of the line, and l is the length of the line in meters. The $\ln\gamma$ term is a function of only the T/W ratio, is small as compared to the other terms (varying from 0 to 0.0025), and has a negligible effect on the dependence of the inductance with length. This expression is an approximation, valid for $l \gg T + W$; a precise formula for round conductors can be found in [46].

From a circuit analysis point of view, this nonlinearity is caused by the significant inductive coupling among the different segments of the same line. Consider a straight line; the corresponding circuit representation of the self-inductance of the line is a single inductor, as shown in Fig. 3.1a. Consider also the same line as two shorter lines connected in series. The corresponding circuit representation of the inductance of these two lines is two coupled inductors connected in series, as shown in Fig. 3.1b.

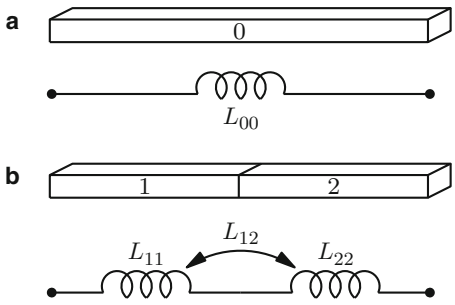


Fig. 3.1 Two representations of a straight line inductance; (**a**) the line can be considered as a single element, with a corresponding circuit representation as a single inductor, (**b**) the same line can also be considered as two lines connected in series, with a corresponding circuit representation as two coupled inductors connected in series

The inductance of this circuit is

$$L_{1+2} = L_{11} + L_{22} + 2L_{12}. \tag{3.2}$$

If the partial inductance is a linear function of length, the inductance of the circuit is the sum of the inductance of its elements, i.e.,

$$L_{linear} = L_{11} + L_{22}. \tag{3.3}$$

The difference between the nonlinear dependence [see (3.2)] and the linear dependence [see (3.3)] is the presence of the cross coupling term $2L_{12}$. This term increases the inductance beyond the linear value, i.e., the sum $L_{11} + L_{22}$. The cross coupling term increases with line length and does not become small as compared to the self-inductance of the lines L_{11} as the line length increases.

However, the loop inductance of a complete current loop formed by two parallel straight rectangular conductors, shown in Fig. 3.2, is given by [44]

$$L_{\text{loop}} = 0.4l\left(\ln\frac{P}{H+W} + \frac{3}{2} - \ln\gamma + \ln k\right)\ \mu\text{H}, \tag{3.4}$$

where P is the distance between the center of the lines (the pitch) and $\ln k$ is a tabulated function of the H/W ratio. This expression is accurate for long lines (i.e., for $l \gg P$). The expression is a linear function of the line length l.

To compare the dependence of inductance on length for both a single line and a complete loop and to assess the accuracy of the long line approximation assumed in (3.4), the inductance extractor FastHenry is used to evaluate the partial inductance of a single line and the loop inductance of two identical parallel lines forming forward and return current paths. The cross section of the lines is $1 \times 1\,\mu$m. The spacing between the lines in the complete loop is $4\,\mu$m. The length is varied from $10\,\mu$m to 10 mm.

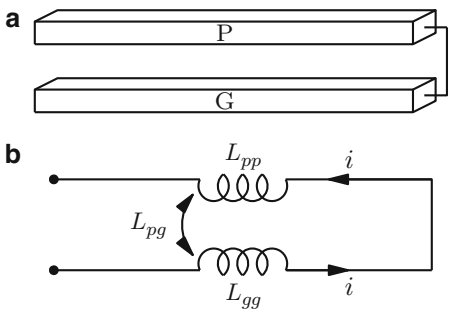


Fig. 3.2 A complete current loop formed by two straight parallel lines; (**a**) physical structure, (**b**) circuit diagram of the partial inductance

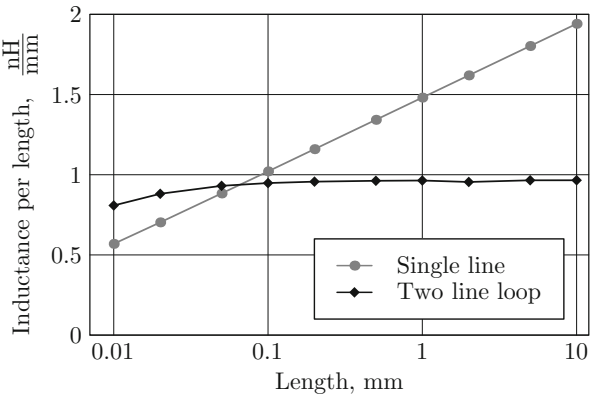


Fig. 3.3 Inductance per length versus line length for a single line and for a loop formed by two identical parallel lines. The line cross section is $1 \times 1\,\mu\text{m}$

The linearity of a function is difficult to visualize when the function argument spans three orders of magnitude (particularly when plotted in a semi-logarithmic coordinate system). The inductance per length, alternatively, is a convenient measure of the linearity of the inductance. The inductance per length (the inductance of the structure divided by the length of the structure) is independent of the length if the inductance is linear with length, and varies with length otherwise.

The inductance per length is therefore determined for a single line and a two-line loop of various length. The results are shown in Fig. 3.3. The linearity of the data rather than the absolute magnitude of the data is of primary interest here. To facilitate the assessment of the inductance per length in relative terms, the data shown in Fig. 3.3 are recalculated as a per cent deviation from the reference value and are shown in Fig. 3.4. The inductance per length at a length of 1 mm is chosen as a reference. Thus, the inductance per length versus line length is plotted in Fig. 3.4 as a per cent deviation from the magnitude of the inductance per length at a line length of 1 mm. As shown in Fig. 3.4, the inductance per length of a single line changes significantly with length. The inductance per length of a complete loop is practically constant for a wide range of lengths [71] (varying approximately 4 %

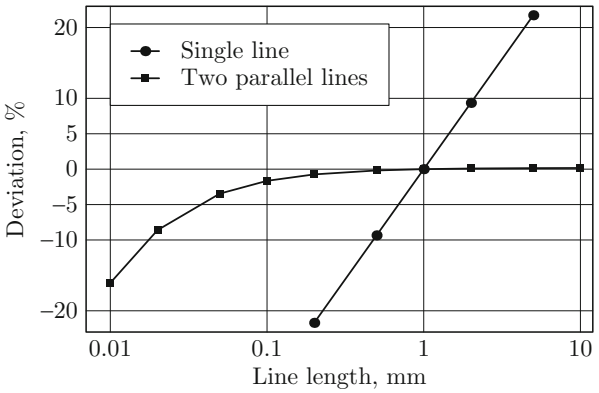


Fig. 3.4 Inductance per length versus line length in terms of the per cent difference from the inductance per length at a 1 mm length. The data is the same as shown in Fig. 3.3 but normalized to the value of inductance per length at 1 mm (0 % deviation). The (loop) inductance per length of a two line loop is virtually constant over a wide range of length, while the (partial) inductance per length of a single line varies linearly with length

over the range from 50 to 10,000 μm). The inductance of a complete loop can, therefore, be considered linear when the length of a loop exceeds the loop width by approximately a factor of ten. Note that in the case of a simple structure, such as the two line loop shown in Fig. 3.2, it is not necessary to use FastHenry to produce the data shown in Fig. 3.4. The formulæ for inductance in [44] can be applied to derive the data shown in Fig. 3.4 with sufficient accuracy.

To gain further insight into why the inductance of a complete loop increases linearly with length while the inductance of a single line increases nonlinearly, consider a complete loop as two loop segments connected in series, as shown in Fig. 3.5. The inductance of the left side loop segment (formed by line segments one and two) is

$$L_{1+2} = L_{11} + L_{22} - 2L_{12}, \tag{3.5}$$

while the inductance of the right side segment of the loop (formed by line segments three and four) is, analogously,

$$L_{3+4} = L_{33} + L_{44} - 2L_{34}. \tag{3.6}$$

The inductance of the entire loop is

$$\begin{aligned} L_{\text{loop}} &= \sum_{i,j=1}^{4} L_{ij} \\ &= L_{11} + L_{22} + L_{33} + L_{44} - 2L_{12} - 2L_{34} \\ &\quad +2L_{13} - 2L_{14} + 2L_{24} - 2L_{23}. \end{aligned} \tag{3.7}$$

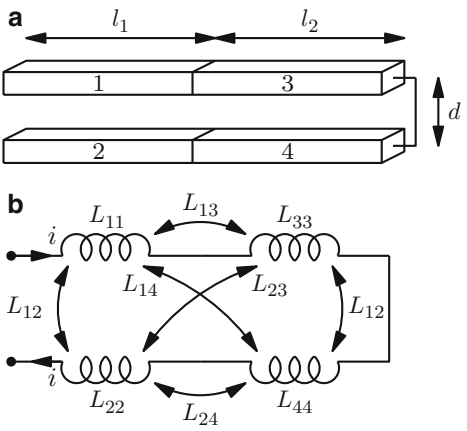


Fig. 3.5 A complete current loop formed by two straight parallel lines consists of two loop segments in series; (**a**) physical structure, (**b**) circuit diagram of the partial inductance

Considering that $L_{13} = L_{24}$ and $L_{14} = L_{23}$ due to the symmetry of the structure and substituting (3.5) and (3.6) into (3.7), this expression reduces to

$$L_{\text{loop}} = L_{1+2} + L_{3+4} + 2M, \tag{3.8}$$

where $M = L_{13} - L_{14} + L_{24} - L_{23} = 2(L_{13} - L_{14})$ is the coupling between the two loop segments. This expression for a complete loop is completely analogous to (3.2) for a single line. Similar to (3.2), the nonlinear term within the parenthesis augments the inductance beyond a linear value of $L_{1+2} + L_{3+4}$. An important difference, however, is that the nonlinear part is a difference of two terms close in value, because $L_{13} \approx L_{14}$ and $L_{24} \approx L_{23}$.

This behavior can be intuitively explained as follows. Segments three and four are physically (and inductively) much closer to each other than these segments are to segment one. The effective distance (the distance to move one segment so as to completely overlap the other segment) between line segment three and line segment four is small as compared to the effective distance between segment three and segment one. The inductive coupling is a smooth function of distance. The magnitude of the inductive coupling of segment one to segment three is, therefore, quite close to the magnitude of the coupling of segment one to segment four (but of opposite sign due to the opposite direction of the current flow).

A mathematical treatment of this phenomenon confirms this intuitive insight. The mutual partial inductance of two parallel line segments, for example, segments one and four shown in Fig. 3.5b, is

$$L_M = 0.1 \left(l_1 \ln \frac{l_1 + l_2}{l_1} + l_2 \ln \frac{l_1 + l_2}{l_2} - d \right) \ \mu\text{H}, \tag{3.9}$$

where l_1 and l_2 are the segment lengths, and d is the distance between the center axes of the segments, as shown in Fig. 3.5a. Consider, for example, the case where

the line segments are of equal length, $l_1 = l_2 = l/2$. The mutual inductance as a function of the axis distance d is

$$L_M(d) = (l \ln 2 - d) \ \mu\text{H}, \tag{3.10}$$

where $L_M(d)$ is a weak function of d if $d \ll l$. The mutual inductance of two segments forming a straight line, i.e., L_{12} in Fig. 3.1 and L_{13} and L_{24} in Fig. 3.2, is $L_M(0)$. The mutual inductance L_{14} and L_{23} is $L_M(d)$. The effective inductive coupling of two loop segments, therefore, simplifies to the following expression,

$$4(L_{13} - L_{14}) = 4(L_M(0) - L_M(d)) = 0.4d \ \mu\text{H}\,. \tag{3.11}$$

Note that this coupling is much smaller than the coupling of two straight segments, $L_M(0)$, and is independent of the loop segment length l. As the loop length l exceeds several loop widths d (as $l \gg d$), the coupling becomes negligible as compared to the self-inductance of the loop segments. As compared to the coupling between two single line segments, the effective coupling is reduced by a factor of

$$\frac{M_{\text{line}}}{M_{\text{loop}}} = \frac{L_M(0)}{2(L_M(0) - L_M(d))} = \frac{\ln 2}{2}\frac{l}{d}\,. \tag{3.12}$$

In general, it can be stated that at distances much larger than the effective separation between the forward and return currents, the inductive coupling is dramatically reduced as the coupling of the forward current and return current is mutually cancelled. Hence, the inductance of a long loop ($l \gg d$) depends linearly upon the length of the loop.

3.3 Inductive Coupling Between Two Parallel Loop Segments

As shown in the previous section, the relative proximity of the forward and return current paths is the reason for the cancellation of the inductive coupling between two loop segments connected in series (i.e., different sections of the same current loop).

The same argument can be applied to show that the effective inductive coupling between two sections of parallel current loops is also reduced [71]. As in the case of the collinear loop segments considered above, this behavior is due to cancellation of the coupling if the distance between the forward and return current paths (the loop width) is much smaller than the distance between the parallel loop segments. The physical structure and circuit diagram of two parallel loop segments are shown in Fig. 3.6.

The mutual loop inductance of the two loop segments is

$$M_{\text{loop}} = L_{13} - L_{14} + L_{24} - L_{23}. \tag{3.13}$$

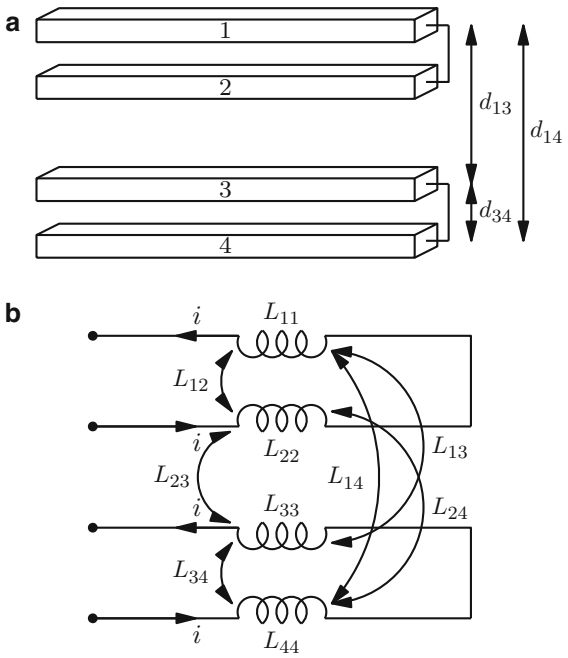


Fig. 3.6 Two parallel loop segments where each loop segment is formed by two straight parallel lines; (**a**) physical structure, (**b**) circuit diagram of the partial inductance

The mutual inductance between two parallel straight lines of equal length is [46]

$$M_{\text{loop}} = 0.2l\left(\ln\frac{2l}{d} - 1 + \frac{d}{l} - \ln\gamma + \ln k\right)\ \mu\text{H}, \tag{3.14}$$

where l is the line length, and d is the distance between the line centers. This expression is an approximation for the case where $l \gg d$. The mutual inductance of two straight lines is a weak function of the distance between the lines. Therefore, if the distance between lines one and three d_{13} is much greater than the distance between lines three and four d_{34}, such that $d_{13} \approx d_{14}$, then $L_{13} \approx L_{14}$. Analogously, $L_{23} \approx L_{24}$. The coupling of the loops M_{loop} is a difference of two similar values, which is small as compared to the self-inductance of the loop segments. The loop segments can be considered weakly coupled in this case. This effective decoupling means that the reluctance of the loop segments wired in parallel is the sum of the reluctances of the individual loop segments. Alternatively, the inductance of the parallel connection is the inductance of two parallel uncoupled inductors, $L_{||} = \frac{L_{11}L_{22}}{L_{11}+L_{22}}$, similar to the resistance of parallel elements. The circuit inductance, therefore, varies linearly with the circuit "width": as more identical circuit elements (i.e., loop segments) are connected in parallel, the inductance of the circuit decreases inversely linearly with the number of parallel elements. This linear property of inductance is demonstrated in [72, 73] with specific application to high performance power grids.

3.4 Application to Circuit Analysis

Although rectangular lines with a unity height to width aspect ratio are used in the case studies described above, the conclusions are quite general and hold for different wire shapes and aspect ratios. At low frequencies, where the current density is uniform throughout a wire cross section, the self-inductance of a wire is determined by the *geometric mean distance* of the wire cross section and the mutual inductance of two wires is determined by the *geometric mean distance* between two cross sections, as first described by Maxwell [74]. Rosa and Grover systematized and tabulated the geometric mean distance data for several practically important cases [44, 48]. Both the self- and mutual inductance of a conductor is moderately dependent on the perimeter length of the inductor cross section and is virtually independent of the conductor cross-sectional shape. For example, the self-inductance of rectangular conductors with aspect ratios of 1 and 10, but with the same perimeter length-to-line length ratio of 40, differ by only 0.012 %, according to (3.1).

Skin effects reduce the current density as well as the magnetic field within the core of the conductor. This effect slightly reduces the self-inductance of a wire and has virtually no effect on the mutual inductance. Proximity effects in two neighboring wires carrying current in the opposite directions (as in a loop formed by two parallel wires) can only reduce the effective distance between the two currents, making the loop effectively "longer." Therefore, a uniform current density distribution is a conservative assumption regarding the linearity of inductance with loop length.

The particular on-chip current return path is rarely known before analysis of the circuit. Nevertheless, if an approximate but conservative estimate of the distance d between the signal wire and the current return path can be made which satisfies the long loop condition $l \gg d$, the inductance can be considered to vary linearly with the length of the structure. Similar to the resistance and capacitance, the inductance of such a structure is effectively a local characteristic, independent of the length of the structure. The analysis of a large interconnect structure is therefore not necessary to obtain the local inductive characteristics, greatly simplifying the circuit analysis process. This property is demonstrated in Chap. 28 on an example of regularly structured power distribution grids.

3.5 Summary

The variation of the partial and loop inductance with line length is analyzed in this chapter. The primary conclusions are summarized as follows.

- The nonlinear variation of inductance with length is due to inductive coupling between circuit segments

- In long loops, the effective coupling between loop segments is small as compared to the self-inductance of the segments due to the mutual cancellation of the forward current coupling with the return current coupling
- The inductance of long loops increases virtually linearly with length
- In a similar manner, the effective inductive coupling between two parallel current loops is greatly reduced due to coupling cancellation as compared to the coupling between line segments of the same length
- As a general rule, the inductance of circuits scales *linearly* with the circuit dimensions where the distance between the forward and return currents is much smaller than the dimensions of the circuit
- The linear variation of inductance with the circuit dimensions can be exploited to simplify the inductance extraction process and the related circuit analysis of on-chip interconnect

Chapter 4
Electromigration

The power current requirements of integrated circuits are rapidly rising, as discussed in Chap. 1 and throughout this book. The current density in on-chip power and ground lines can reach several hundred thousands of amperes per square centimeter. At these current densities, electromigration becomes significant. Electromigration is the transport of metal atoms under the force of an electron flux. The depletion and accumulation of the metal material resulting from the atomic flow can lead to the formation of extrusions (or hillocks) and voids in the metal structures. The hillocks and voids can lead to short circuit and open circuit faults [75], respectively, as shown in Fig. 4.1, degrading the reliability of an integrated circuit.

The significance of electromigration was established early in the development of integrated circuits [76, 77]. Electromigration should be considered in the design process of an integrated circuit to ensure reliable operation over the intended lifetime. Electromigration reliability and related design implications are the subject of this chapter. A more detailed discussion of the topic of electromigration can be found in the literature [78–80].

This chapter is organized as follows. The physical mechanism of electromigration is described in Sect. 4.1. The role of mechanical stress in electromigration reliability is discussed in Sect. 4.2. The steady state conditions of electromigration damage in interconnect lines are established in Sect. 4.3. The dependence of electromigration reliability on the dimensions of the interconnect line is discussed in Sect. 4.4. The statistical distribution of the electromigration lifetime is reviewed in Sect. 4.5. Electromigration reliability under an AC current is discussed in Sect. 4.6. A comparison of electromigration effects within aluminum and copper based interconnect technologies is described in Sect. 4.7. Low-k dielectric materials are also briefly discussed in this section. Certain approaches to designing for electromigration reliability are briefly reviewed in Sect. 4.8. The chapter concludes with a summary.

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0_4

Fig. 4.1 Electromigration induced circuit faults; (**a**) line extrusions formed due to metal accumulation can short circuit the adjacent line, (**b**) voids formed due to metal depletion increase the line resistance and can lead to an open circuit

4.1 Physical Mechanism of Electromigration

Electromigration is a microscopic mass transport of metal ions through diffusion under an electrical driving force. An atomic flux under a driving force F is

$$J_{\mathrm{a}} = C_a \mu F, \tag{4.1}$$

where C_a is the atomic concentration and μ is the mobility of the atoms. From the Einstein relationship, the mobility can be expressed in terms of the atomic diffusivity D_{a} and the thermal energy kT,

$$\mu = \frac{D_{\mathrm{a}}}{kT}. \tag{4.2}$$

Two forces act on metal ions: an electric field force and an electron wind force. The electric field force is proportional to the electric field E and acts in the direction of the field. The electric field also accelerates the conduction electrons in the direction opposite to the electric field. The electrons transfer the momentum to the metal ions in the course of scattering, exerting the force in the direction opposite to the field E. This force is commonly referred to as the electron wind force. The electron wind force equals the force exerted by the electric field on the conduction electrons, which is proportional to the electric field intensity E.

In the metals of interest, such as aluminum and copper, the electron wind force dominates and the net force acts in the direction opposite to the electric current. The resulting atomic flux is therefore in the direction opposite to the electric current j, as shown in Fig. 4.2. The net force acting on the metal atoms is proportional to the electric field and is typically expressed as $-Z^*eE$ or $-Z^*e\rho j$, where Z^*e is the effective charge of the metal ions, j is the current density, and ρ is the resistivity of the metal. The electromigration atomic flux is therefore

$$J_{\mathrm{a}} = -C_a \frac{D_{\mathrm{a}} Z^* e \rho j}{kT}, \tag{4.3}$$

where the minus sign indicates the direction opposite to the field E. The diffusivity D_{a} has an Arrhenius dependence on temperature, $D_{\mathrm{a}} = D_0 \exp(-Q/kT)$, where

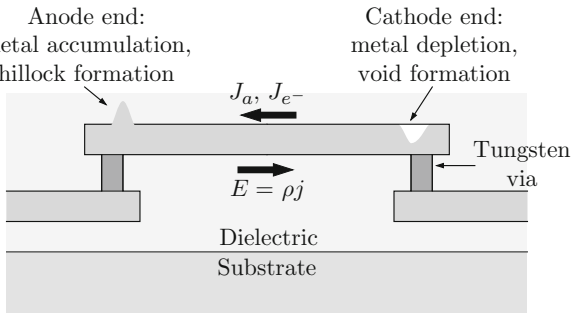


Fig. 4.2 Electromigration mass transport in an interconnect line. An electron flux J_{e^-} flowing in the opposite direction to the electric field $E = \rho j$ induces an atomic flow J_a in the direction of the electron flow. Diffusion barriers, such as tungsten vias, create an atomic flux divergence, leading to electromigration damage in the form of voids and hillocks

Q is the activation energy of diffusion. Substituting this relationship into (4.3), the atomic flux becomes

$$J_a = -C_a \frac{Z^* e \rho j}{kT} D_0 \exp\left(-\frac{Q}{kT}\right). \tag{4.4}$$

The material properties C_a, ρ, and Z^*e are difficult to change. The diffusion coefficient D_0 and the activation energy Q, although also material specific, vary significantly depending upon the processing conditions and the resulting microstructure of the metal film. There are several paths for electromigration in interconnect lines, such as diffusion through the lattice, along the grain boundary, and along the metal surface. Each path is characterized by the individual diffusion coefficient and activation energy. The atomic flux depends exponentially on the activation energy Q and is therefore particularly sensitive to variations in Q. The diffusion path with the lowest activation energy dominates the overall atomic flux.

The on-chip metal films are polycrystalline, consisting of individual crystals of various size. The individual crystals are commonly referred to as grains. The activation energy Q is relatively low for diffusion along a grain boundary, as low as 0.5 eV for aluminum, while the activation energy for diffusion through the lattice is the highest, up to 1.4 eV in aluminum. The activation energy in copper is higher, 1.2 eV for diffusion along a grain boundary, 0.7 eV for the surface diffusion, and 2.1 eV for the lattice diffusion due to a higher melting point, as wells as higher electrical and thermal conductivity as compared to aluminum. The diffusion path with lowest activation energy is a dominant path for electromigration. Diffusion along the grain boundary is therefore the primary path of electromigration in relatively wide aluminum interconnect lines, while surface diffusion is dominant in copper lines.

An atomic flux does not cause damage to on-chip metal structures where the influx of the atoms is balanced by the atom outflow. The depletion and accumulation

of metal (and the associated damage) develop at those sites where the influx and outflux are not equal, i.e., the flux divergence is not zero, $\nabla \cdot J_a \neq 0$. Flux divergence can be caused by several factors, including an interface between materials with dissimilar diffusivity and resistivity, inhomogeneity in the microstructure, and a temperature gradient.

Consider, for example, the current flow in an aluminum metal line segment connected to two tungsten vias at the ends, as shown in Fig. 4.2. At the current densities of interest, tungsten is not susceptible to electromigration and can be considered as an ideal barrier for the diffusion of aluminum atoms. The tungsten-aluminum interface at the anode line end, where the electric current enters the line (i.e., the electron flux exits the line), prevents the outflow of the aluminum atoms from the line. The incoming atomic flux causes an accumulation of aluminum atoms. At the cathode end of the line, the tungsten-aluminum interface blocks the atomic flux from entering the line. The electron flux enters the line at this end and carries away aluminum atoms, leading to metal depletion and, potentially, formation of a void.

In a similar manner, an inhomogeneity in the microstructure can cause flux divergence. As has been discussed, aluminum grain boundaries have significantly lower activation energy, facilitating atomic flow. The electromigration atomic flux is enhanced at those locations with small grains, where the grain boundaries provide numerous paths that facilitate diffusion. At those sites where the grain size changes abruptly, the high atomic flux in the region of the smaller grain size is mismatched with the relatively low atomic flux in the larger grain region. The atomic flux divergence leads to a material depletion or accumulation, depending upon the direction of current flow. These sites are particularly susceptible to electromigration damage [81]. In copper interconnects, the electromigration flux is constrained to the surface (due to the smallest activation energy for diffusion in a surface), exhibiting a bamboo grain structure [82] which leads to lower reliability in smaller size lines [83].

4.2 Electromigration-Induced Mechanical Stress

The depletion and accumulation of material at the sites of atomic flux divergence induce a mechanical stress gradient in the metal structures. At the sites of metal accumulation the stress is compressive, while at the sites of metal depletion the stress is tensile. The resulting stress gradient in turn induces a flux of metal atoms J_a^{stress},

$$J_a^{\text{stress}} = C_a \frac{D_a}{kT} \Omega \frac{\partial \sigma}{\partial l}, \tag{4.5}$$

where σ is the mechanical stress, assumed positive in tension, Ω is the atomic volume, and l is the line length. The atomic flux due to the stress gradient is opposite

in direction to the electromigration atomic flux, counteracting the electromigration mass transport. This flux is therefore often referred to as an atomic backflow. The distribution of the mechanical stress and the net atomic flux are therefore interrelated. Accurate modeling of the mechanical stress is essential in predicting electromigration reliability. Mechanical stress can also have components unrelated to electromigration atomic flux, such as a difference in the thermal expansion rates of the materials.

The on-chip metal structures are encapsulated in a dielectric material, such as silicon dioxide or more advanced low-k material. The stiffness of the dielectric material significantly affects the electromigration reliability. A rigid dielectric limits the variation in metal volume at the sites of the metal depletion and accumulation, resulting in greater electromigration-induced mechanical stress as compared to a metal line in a less rigid environment. The greater mechanical stress induces a greater atomic flux in the direction opposite to the electromigration atomic flux, limiting the net atomic flux and the related structural damage. Rigid encapsulation therefore significantly improves the electromigration reliability of the metal interconnect. Low-k dielectric material as compared with silicon dioxide is less rigid, resulting in lower electromigration-induced mechanical stress and decreasing electromigration reliability.

A rigid dielectric can lead to structural defects due to a mismatch in the thermal expansion rate of the materials. For example, as the silicon wafer is cooled from the temperature of silicon dioxide deposition, the rigid and well adhering silicon dioxide prevents the aluminum lines from contracting according to the thermal expansion rate of aluminum. The resulting tensile stress in the aluminum lines can cause void formation [84]. This effect is exacerbated in narrow lines.

4.3 Steady State Limit of Electromigration Damage

Under certain conditions, the stress induced atomic flux can fully compensate the atomic flux due to electromigration, preventing further damage. In this case, the atomic concentration along a metal line is stationary, $\frac{\partial C_a}{\partial t} = 0$. The net atomic flow is related to the atomic concentration by the continuity equation,

$$\frac{\partial C_a}{\partial t} = -\nabla J_{\mathrm{a}} = \frac{\partial J_{\mathrm{a}}}{\partial l} = 0. \tag{4.6}$$

The atomic flow is uniform along the line length l, i.e., $J_{\mathrm{a}}(l)$ is constant. In a line segment confined by diffusion barriers, the steady state atomic flux is zero. Under this condition, the diffusion due to the electrical driving force is compensated by the diffusion due to the mechanical driving force. The net atomic flux J_{a} is the sum of the electromigration and stress induced fluxes,

$$J_{\mathrm{a}} = J_{\mathrm{a}}^{\mathrm{em}} + J_{\mathrm{a}}^{\mathrm{stress}} = C_a \frac{D_{\mathrm{a}}}{kT} \left(\Omega \frac{\partial \sigma}{\partial l} - Z^* e \rho j \right). \tag{4.7}$$

The steady state condition is established where

$$\Omega \frac{\partial \sigma}{\partial l} = Z^* e \rho j. \tag{4.8}$$

High mechanical stress gradients are required to compensate the electromigration atomic flow at high current densities. The magnitude of the stress gradient depends upon the formation of voids and hillocks.

Consider a line segment of length l_0 between two sites of flux divergence, such as tungsten vias or severe microstructural irregularities. The compressive stress at the anode end should reach a certain yield stress σ_y to develop extrusion damage [85]. Assuming a near zero stress at the cathode end of the segment, the damage critical stress gradient is $\left(\frac{\partial \sigma}{\partial l}\right)_{\max} = \sigma_y / l_0$. Substituting this expression for the stress gradient into (4.8) yields the maximum current density,

$$j_{\max}^{\text{hillock}} = \frac{\Omega \sigma_y}{Z^* e \rho} \frac{1}{l_0}. \tag{4.9}$$

If the current density is lower than the limit determined by (4.9), the steady state condition is reached before formation of the hillocks at the anode, and the interconnect line is highly resistive to electromigration damage. Resistance to electromigration damage below a certain current threshold was first experimentally observed by Blech in 1976 [86].

Void formation also causes mechanical stress that counteracts electromigration atomic flow. As the stress becomes sufficiently high to fully compensate the electromigration flow, the void stops growing and remains at the steady state size [87]. The steady state void size is well described by [88, 89]

$$\Delta l = \frac{Z^* e \rho}{2B\Omega} j l_0^2, \tag{4.10}$$

where B is the elastic modulus relating the line strain to the line stress. Equation (4.10) can be obtained from (4.8) by assuming that a void of size Δl induces a line stress $2B\frac{\Delta l}{l_0}$.

If the steady state void does not lead to a circuit failure, the electromigration lifetime of the line is practically unlimited. A formation of a void in a line does not necessarily lead to circuit failure. Metal lines in modern semiconductor processes are covered with a thin refractory metal film, such as Ta, Ti, TaN, TiN, or $TiAl_3$ (based on the metal interconnect material). These metal films are highly resistant to electromigration damage, providing an alternative current path in parallel to the metal core of the line. A void spanning the line width will therefore increase the resistance of the line, rather than lead to an open circuit fault. The increase in line resistance is proportional to the void size Δl. An increase in the line resistance from 10 % to 20 % is typically considered critical, leading to a circuit failure. A critical increase in the resistance occurs when the void size reaches a certain critical

value Δl_{crit}. The current density resulting in a void of the critical steady state size is determined from (4.10),

$$j_{\max}^{\text{void}} = \frac{2B\Omega}{Z^* e\rho} \frac{\Delta l_{\text{crit}}}{l_0^2}. \tag{4.11}$$

If the current density of the line is lower than $j_{\max}^{\text{void}}$, the void damage saturates at a subcritical size and the lifetime of the line becomes practically unlimited.

The critical current density as determined by (4.9) for hillock formation and (4.11) for void formation increases with shorter line length l_0. For a given current density there exists a certain critical line length l_{crit}. Lines shorter than l_{crit} are highly resistant to electromigration damage. This phenomenon is referred to as the electromigration threshold or Blech effect.

4.4 Dependence of Electromigration Lifetime on the Line Dimensions

While the electromigration reliability depends upon many parameters, as expressed by (4.4), most of these parameters cannot be varied due to material properties and manufacturing process characteristics. The parameters that can be flexibly varied at the circuit design phase are the line width, length, and current. Varying the current is often restricted by circuit performance considerations. The dependence of electromigration reliability on the line width and length is discussed in this section.

The dependence of the electromigration lifetime on the width of an aluminum alloy line is relatively complex [90, 91], as illustrated in Fig. 4.3. Trends of the line width on the lifetime for a copper interconnect due to electromigration are similar to aluminum alloy interconnects, as discussed in [80], however higher in magnitude.

In relatively wide lines, i.e., where the average grain size is much smaller than the line width, the grain boundaries form a continuous diffusion path along the line length, as shown in Fig. 4.4b. Although the grain boundary diffusion path enhances the electromigration atomic flow due to a higher diffusion coefficient along the grain boundary, the probability of abrupt microstructural inhomogeneity along the line length is small, due to a large number of grains spanning the width of the line. The probability of an atomic flux divergence in these polygranular lines is relatively low and the susceptibility of the line to electromigration damage is moderate.

As the line width approaches the average grain size, the polygranular line structure is likely to be interrupted by grains spanning the entire width of the line, disrupting the boundary diffusion path, as shown in Fig. 4.4c. Electromigration transport in the spanning grain can occur only through the lattice or along the surface of the line. The diffusivity of these paths is significantly lower than the diffusivity of the polycrystalline segments. The spanning grains therefore present a barrier to an atomic flux in relatively long polygranular segments of the line. The atomic

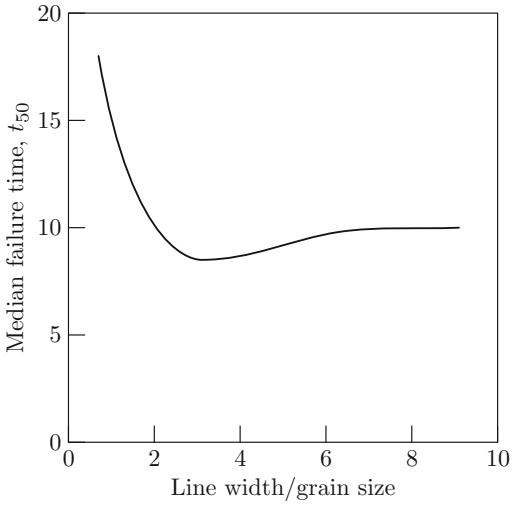


Fig. 4.3 Representative variation of the median electromigration lifetime with line width (Based on data obtained from [91])

flux discontinuity at the boundary of the spanning grains renders these lines more susceptible to electromigration damage, shortening the electromigration lifetime.

As the line width is reduced below the average grain size, the spanning grains dominate the line microstructure, forming the so-called "bamboo" pattern, as shown in Fig. 4.4d. The polygranular segments become sparse and short. The shorter polygranular segments are resistant to electromigration damage, increasing the expected lifetime of the narrow lines.

The electromigration lifetime also varies with line length. Shorter lines have a longer lifetime than longer lines [90]. The longer lines are more likely to contain a significant microstructural discontinuity, such as a spanning grain in wide lines or a long polygranular segment in narrow lines. The longer lines are therefore more susceptible to electromigration damage.

4.5 Statistical Distribution of Electromigration Lifetime

Electromigration failure is a statistical process. Identically designed interconnect structures fail at different times due to variations in the microstructure. Failure times are typically described by a log-normal distribution. A variable distribution is log-normal if the distribution of the logarithm of the variable is normal. The log-normal probability density function $p(t)$ is

$$p(t) = \frac{1}{\sqrt{2\pi}\sigma t} \exp\left(-\frac{(\ln(t/t_{50}))^2}{2\sigma^2}\right). \tag{4.12}$$

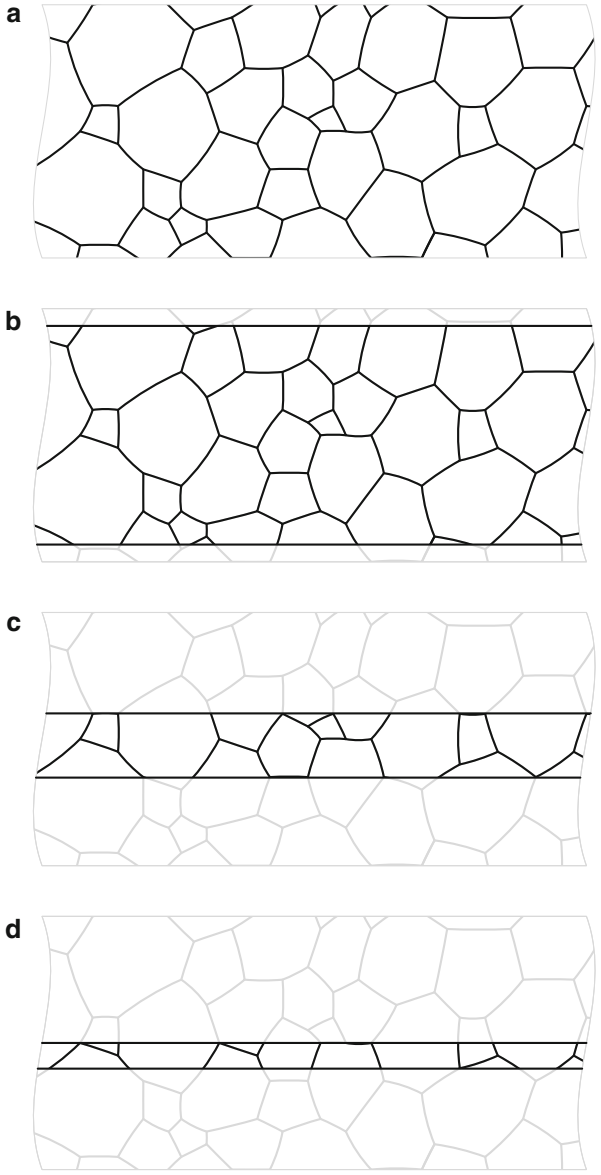


Fig. 4.4 Grain structure of interconnect lines; (**a**) grain structure of an unpatterned thin metalization film, (**b**) structure of the wide lines is polygranular along the entire line length, (**c**) lines with a width close to the average grain size, polygranular segments are interrupted by the grains spanning the entire line width, (**d**) narrow lines, most of the grains span the entire line width, forming a "bamboo" pattern

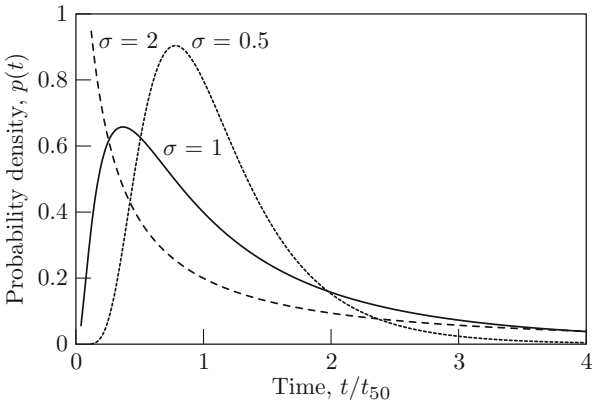


Fig. 4.5 Log-normal distribution of electromigration failures. The distribution is unimodal and is determined by the median time to failure t_{50} and the shape parameter σ

The log-normal distribution is characterized by the median time to failure t_{50} and the shape factor σ. The probability density function for several values of σ are shown in Fig. 4.5.

Accelerated electromigration testing is performed to evaluate the lifetime distribution parameters for a specific manufacturing process. The interconnect structures are subjected to a current and temperature significantly greater than the target specifications to determine the statistical characteristics of the interconnect failures within a limited time period. The following relationship, first proposed by Black in 1967 [77, 92], is commonly used to estimate the median time to failure at different temperatures and current densities,

$$t_{50} = \frac{A}{j^n} \exp\left(\frac{Q}{kT}\right), \tag{4.13}$$

where A and n are empirically determined parameters. In the absence of Joule heating effects, the value of n varies from one to two, depending on the characteristics of the manufacturing process [75]. As demonstrated by models of the electromigration process, $n = 1$ corresponds to the case of void induced failures, and $n = 2$ represents the case of hillock induced failures [93].

4.6 Electromigration Lifetime Under AC Current

On-chip interconnect lines typically carry time-varying AC current. It is necessary to determine the electromigration lifetime under AC conditions based on accelerated testing, which is typically performed with DC current.

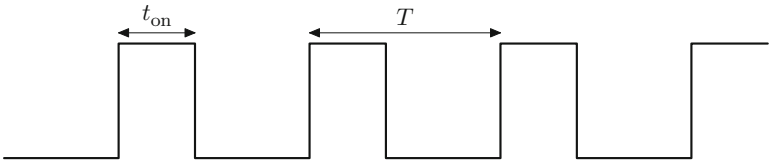


Fig. 4.6 A train of current pulses

Consider a train of square current pulses with a duty ratio $d = t_{on}/T$, as shown in Fig. 4.6, versus a DC current of the same magnitude. Assuming that a linear accumulation of the electromigration damage during the active phase t_{on} of the pulses results in the pulsed current lifetime t_{50}^{pulsed} that is $1/d$ times longer than the DC current lifetime t_{50}^{dc}, i.e., $t_{50}^{dc}/t_{50}^{pulsed} = d$. This estimate is, however, overly conservative, suggesting a certain degree of electromigration damage repair during the quiet phase of the pulses. This "self-healing" effect significantly extends the lifetime of a line. Experimental studies [94–96] have demonstrated that, in the absence of Joule heating effects, the pulsed current lifetime is determined by the average current j_{avg},

$$t_{50}^{pulsed} = \frac{A}{j_{avg}^{n}} \exp\left(\frac{Q}{kT}\right). \tag{4.14}$$

As $j_{avg} = dj_{dc}$, the pulsed current lifetime is related to the DC current lifetime as $t_{50}^{dc}/t_{50}^{pulsed} = d^n$.

Electromigration reliability is greatly enhanced under bidirectional current flow. Accurate characterization of electromigration reliability becomes difficult due to the long lifetimes. Available experimental data are in agreement with the average current model, as expressed by (4.14). According to (4.14), the lifetime becomes infinitely long as the DC component of the current approaches zero. The current density these lines can carry, however, is also limited. As the magnitude of the bidirectional current becomes sufficiently high, Joule heating becomes significant, degrading the self-healing process and, consequently, the electromigration lifetime.

Clock and data lines in integrated circuits are usually connected to a single driver. The average current in these lines is zero and the lines are typically highly resistant to electromigration failure in the absence of significant Joule heating. Power and ground lines carry a high unidirectional current. Power and ground lines are therefore particularly susceptible to electromigration damage.

4.7 A Comparison of Aluminum and Copper Interconnect Technologies

With technology scaling, the resistance of an aluminum interconnect has become increasingly high. Copper interconnect technology has therefore been developed since the resistivity of copper is almost half lower than aluminum, permitting the cross section of an interconnect to be further scaled while satisfying impedance constraints [97]. Copper interconnects also exhibit a higher thermal conductivity and higher melting point, enhancing the reliability of the metal. The metal used for interconnects in high performance integrated circuits is therefore currently copper. The electromigration characteristics of these interconnects are different as compared to aluminum based interconnects, as demonstrated by early studies [98].

The electromigration atomic flux in copper interconnect is significantly lower than in aluminum interconnect mainly due to the higher activation energy of the diffusion paths, as discussed previously in this chapter. The threshold effect has been observed in copper lines [99–101], with threshold current densities several times higher than in aluminum lines. Copper interconnects are able to tolerate stress significantly longer [97]. Experimental results suggest that surface diffusion and diffusion along the interface between the copper and silicon nitride covering the top of the line dominate the electromigration transport process. The silicon nitride film serves as a diffusion barrier at the upper surface of the line. Several features of the copper interconnect structure, however, exacerbate the detrimental effects of electromigration [98].

The sides and bottom of a copper line are covered with a refractory metal film (Ta, Ti, TaN, or TiN) to prevent copper from diffusing into the dielectric. The thickness of this film is much smaller than the thickness of the film covering the aluminum interconnect. The formation of a void in a copper line therefore leads to a higher relative increase in the line resistance. Thin redundant layers can also lead to an abrupt open-circuit failure rather than a gradual increase in the line resistance [102].

The via structure in dual damascene copper interconnect is susceptible to failure due to void formation [103]. The refractory metal film lining the bottom of the via forms a diffusion barrier between the via and the lower metal line, as shown in Fig. 4.7. The resistance of dual damascene interconnect is particularly sensitive to void formation at the bottom of the via. The structural characteristics of the via contact are crucial to interconnect reliability [104].

To further improve the characteristics of metal interconnects, low-k dielectric materials have been developed to encapsulate the metal interconnects. Low-k dielectric materials are commonly used in current technologies for high performance integrated circuits [105]. This dielectric material is typically used with the higher metal layers to reduce the worst-case signal delay of the global interconnects since low-k dielectric materials decrease the distributed capacitance of the metal line (as compared with silicon dioxide for the same dielectric thickness). The dielectric constant for low-k materials ranges from 3.0 and below (as compared to 4.2 for

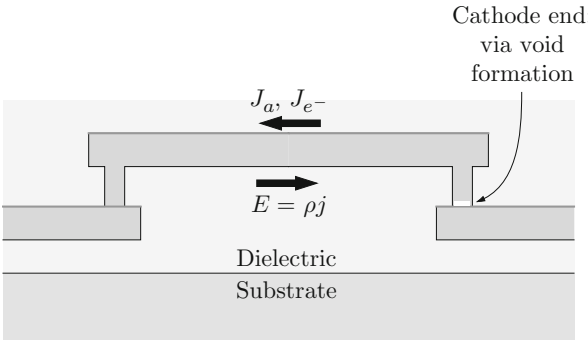


Fig. 4.7 A dual damascene interconnect structure. A diffusion barrier is formed by the refractory metal film lining the side and lower surfaces of the lines and vias. The via bottom at the cathode end is the site of metal depletion and is susceptible to void formation. The resistance of the line is particularly sensitive to void formation at the bottom of the via

silicon dioxide). These materials are a desirable complement to low resistivity copper metallization. Low-*k* dielectrics are also significantly less rigid than silicon dioxide. Metal depletion and accumulation therefore result in less mechanical stress, decreasing the stress-induced backflow and electromigration reliability [106–108]. Low-*k* dielectrics also have a lower thermal conduction coefficient, exacerbating the detrimental effects of Joule heating. A statistical distribution of failure times is more complex in copper interconnects as compared to aluminum and cannot be described by a unimodal probability density distribution [103, 109].

Interconnects based on copper and low-*k* dielectric materials exhibit more desirable performance and reliability characteristics as compared to aluminum and silicon dioxide based interconnects. The reduction in the resistance and capacitance of an interconnect is a key factor in increasing the performance of integrated circuits. Due to technology scaling, electromigration effects require further investigation and novel methodologies are needed to reduce the effects of electromigration within copper and low-*k* interconnect technologies.

4.8 Designing for Electromigration Reliability

Electromigration reliability of integrated circuits has traditionally been ensured by requiring the average current density of each interconnect line to be below a predetermined design rule specified threshold. The cross-sectional dimensions of on-chip interconnect decrease with technology scaling, while the current increases. Simply limiting the line current density by a design rule threshold has become increasingly restrictive under these conditions.

To ensure a target reliability, the current density threshold is selected under the assumption that the current density threshold is reached with a certain number of

interconnect lines. If the number of lines with a critical current density is fewer than the assumed estimate, the design rule is overly conservative. Alternatively, if the number of critical lines is larger than the estimate, the design rule does not guarantee the target reliability characteristics.

This "one size fits all" threshold approach can be replaced with a more flexible statistical electromigration budgeting methodology [110]. If the failure probability of the ith line is estimated as $p_i(t)$ based on the line dimensions and average current, the probability that none of the lines in a circuit fails at time t is $\prod_i (1 - p_i(t))$. The failure probability of the overall system $P(t)$ is therefore

$$P(t) = 1 - \prod_i (1 - p_i(t)) \,. \tag{4.15}$$

A system with a few lines carrying current exceeding the threshold design rule (and therefore exhibiting a relatively high failure probability $p(t)$) can be as reliable as a system with a larger number of lines carrying current at the threshold level.

This statistical approach permits individual budgeting of the line failure probabilities $p_i(t)$, while maintaining the target reliability of the overall system $P(t)$. This flexibility supports more efficient use of interconnect resources, particularly in congested areas, reducing circuit area.

4.9 Summary

The electromigration reliability of integrated circuits has been discussed in this chapter. The primary conclusions of the chapter are the following.

- Electromigration damage develops near the sites of atomic flux divergence, such as vias and microstructural discontinuities
- Electromigration reliability depends upon the mechanical properties of the interconnect structures
- Electromigration reliability of short lines is greatly enhanced due to stress gradient induced backflow, which compensates the electromigration atomic flow
- Power and ground lines are particularly susceptible to electromigration damage as these lines carry a unidirectional current
- Copper based interconnects exhibit enhanced performance and high-er reliability as compared to aluminum based interconnects due to the higher electrical conductivity, melting point, and thermal conductivity

Chapter 5
Scaling Trends of On-Chip Power Noise

A scaling analysis of the voltage drop across the on-chip power distribution networks is performed in this chapter. The design of power distribution networks in high performance integrated circuits has become significantly more challenging with recent advances in process technology. Insuring adequate signal integrity of the power supply has become a primary design issue in high performance, high complexity digital integrated circuits. A significant fraction of the on-chip resources is dedicated to achieve this objective.

State-of-the-art circuits consume higher current, operate at higher speeds, and have lower noise tolerance with the introduction of each new technology generation. CMOS technology scaling is forecasted to continue for at least another 10 years [111]. The scaling trend of noise in high performance power distribution grids is, therefore, of practical interest. In addition to the constraints on the noise magnitude, electromigration reliability considerations limit the maximum current density in on-chip interconnect. The scaling of the peak current density in power distribution grids is also of practical interest. The results of this scaling analysis depend upon various assumptions. Existing scaling analyses of power distribution noise are reviewed and compared along with any relevant assumptions. The scaling of the inductance of an on-chip power distribution network as discussed here extends the existing material presented in the literature. Scaling trends of on-chip power supply noise in ICs packaged in high performance flip-chip packages are the focus of this investigation.

This chapter is organized as follows. Related existing work is reviewed in Sect. 5.1. The interconnect characteristics assumed in the analysis are discussed in Sect. 5.2. The model of the on-chip power distribution noise used in the analysis is described in Sect. 5.3. The scaling of power noise is described in Sect. 5.4. Implications of the scaling analysis are discussed in Sect. 5.5. The chapter concludes with a summary.

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0_5

Table 5.1 Ideal scaling of CMOS circuits [112]

Parameter	Scaling factor
Device dimensions	$1/S$
Doping concentrations	S
Voltage levels	$1/S$
Current per device	$1/S$
Gate load	$1/S$
Gate delay	$1/S$
Device area	$1/S^2$
Device density	S^2
Power per device	$1/S^2$
Power density	1
Total capacitance	SS_C^2
Total power	S_C^2
Total current	SS_C^2

5.1 Scaling Models

Ideal scaling of CMOS transistors was first described by Dennard et al. in 1974 [112]. Assuming a scaling factor S, where $S > 1$, all transistor dimensions uniformly scale as $1/S$, the supply voltage scales as $1/S$, and the doping concentrations scale as S. This "ideal" scaling maintains the electric fields within the device constant and ensures a proportional scaling of the I–V characteristics. Under the ideal scaling paradigm, the transistor current scales as $1/S$, the transistor power decreases as $1/S^2$, and the transistor density increases as S^2. The transistor switching time decreases as $1/S$, the power per circuit area remains constant, while the current per circuit area scales as S. The die dimensions increase by a chip dimension scaling factor S_C. The total capacitance of the on-chip devices and the circuit current both increase by SS_C^2 while the circuit power increases by S_C^2. The scaling of interconnect was first described by Saraswat and Mohammadi [113]. These ideal scaling relationships are summarized in Table 5.1.

Several research results have been published on the impact of technology scaling on the integrity of the IC power supply [114–117]. The published analyses differ in the assumptions concerning the on-chip and package level interconnect characteristics. The analyses can be classified according to several categories: whether resistive IR or inductive $L\ dI/dt$ noise is considered, whether wire bond or flip-chip packaging is assumed, and whether packaging or on-chip interconnect parasitic impedances are assumed dominant. Traditionally, the package-level parasitic inductance (the bond wires, lead frames, and pins) has dominated the total inductance of the power distribution system while the on-chip resistance of the power lines has dominated the total resistance of the power distribution system. The resistive noise has therefore been associated with the resistance of the on-chip interconnect and the inductive noise has been associated with the inductance of the off-chip packaging [117–119].

Scaling behavior of the resistive voltage drop in a wire bonded integrated circuit of constant size has been investigated by Song and Glasser in [116]. Assuming that the interconnect thickness scales as $1/S$, the ratio of the supply voltage to the resistive noise, i.e., the signal-to-noise ratio (SNR) of the power supply voltage, scales as $1/S^3$ under ideal scaling (as compared to $1/S^4$ under constant voltage scaling). Song and Glasser proposed a multilayer interconnect stack to address this problem. Assuming that the top metal layer has a constant thickness, scaling of the power supply signal-to-noise ratio improves by a power of S as compared to standard interconnect scaling.

Bakoglu [114] investigated the scaling of both resistive and inductive noise in wire bonded ICs considering the increase in die size by S_C with each technology generation. Under the assumption of ideal interconnect scaling (i.e., the number of interconnect layers remains constant and the thickness of each layer is reduced as $1/S$), the SNR of the resistive noise decreases as $1/S^4S_C^2$. The SNR of the inductive noise due to the parasitic impedances of the packaging decreases as $1/S^4S_C^3$. These estimates of the SNR are made under the assumption that the number of interconnect levels increases as S. This assumption scales the on-chip capacitive load, average current, and, consequently, the SNR of both the inductive and resistive noise by a factor of S. Bakoglu also considered an improved scaling situation where the number of chip-to-package power connections increases as SS_C^2, effectively assuming flip-chip packaging. In this case, the resistive SNR_R scales as 1 assuming that the thickness of the upper metal levels is inversely scaled as S. The inductive SNR_L scales as $1/S$ under the assumption that the effective inductance per power connection scales as $1/S^2$.

A detailed overview of modeling and mitigation of package-level inductive noise is presented by Larsson [117]. The SNR of the inductive noise is shown to decrease as $1/S^2S_C$ under the assumption that the number of interconnect levels remains constant and the number of chip-to-package power/ground connections increases as SS_C. The results and key assumptions of the power supply noise scaling analyses are summarized in Table 5.2.

The effect of the flip-chip pad density on the resistive drop in power supply grids has been investigated by Arledge and Lynch in [115]. All other conditions being equal, the maximum resistive drop is proportional to the square of the pad pitch. Based on this trend, a pad density of 4000 pads/cm^2 is the minimum density required to assure an acceptable on-chip *IR* drop and input/output (I/O) signal density at the 50 nm technology node [115].

Nassif and Fakhouri describe an analytic expression relating the maximum power distribution noise to the principal design and technology characteristics [120]. The expression is based on a lumped model similar to the model depicted in Fig. 5.3. The noise is shown to increase rapidly with technology scaling based on the ITRS predictions [121]. Assuming constant inductance, a reduction of the power grid resistance and an increase in the decoupling capacitance are predicted to be the most effective approaches to decreasing the power distribution noise.

Table 5.2 Scaling analyses of power distributionDFDF noise

Scaling analysis	Noise type	Noise scaling	SNR scaling	Analysis assumptions	
Glasser and Song [116]	On-chip *IR* noise	S^2	$1/S^3$	Ideal interconnect scaling	Wire bond package, fixed die size
		S	$1/S^2$	Thickness of the top metal remains constant	
Bakoglu [114]	On-chip *IR* noise	$S^3S_C^2$	$1/S^4S_C^2$	Ideal interconnect scaling, wire bond package (the number of power connections is constant)	Current and capacitance scale as $S^2S_C^2$ (due to the scaling of the number of metal levels by S) as compared to SS_C^2
		$1/S$	1	Reverse interconnect scaling ($\propto S$), the number of power connections scale as SS_C^2 (flip-chip)	
	Package $L\frac{dI}{dt}$ noise	$S^3S_C^3$	$1/S^4S_C^3$	Number of power connections remains constant, inductance per connections increases as S_C	
		1	$1/S$	Number of power connections scale as SS_C^2 (flip-chip), inductance per connection scales as $1/S^2$	
Larsson [117]	Package $L\frac{dI}{dt}$ noise	SS_C	$1/S^2S_C$	Wire bond package, number of package connections increases as SS_C	
Mezhiba and Friedman [30]	On-chip *IR* noise	1	$1/S$	Metal thickness remains constant	Area array flip-chip package, pad pitch scales as $1/\sqrt{S}$ (i.e., the number of power connections scales as SS_C^2)
		S	$1/S^2$	Ideal interconnect scaling	
	On-chip $L\frac{dI}{dt}$ noise	S	$1/S^2$	Metal thickness remains constant	
		1	$1/S$	Ideal interconnect scaling	

5.2 Interconnect Characteristics

The power noise scaling trends depend substantially on the interconnect characteristics assumed in the analysis. The interconnect characteristics are described in this section. The assumptions concerning the scaling of the global interconnect are discussed in Sect. 5.2.1. Variation of the grid inductance with interconnect scaling is described in Sect. 5.2.2. Flip-chip packaging characteristics are discussed in Sect. 5.2.3. The impact of the on-chip capacitance on the results of the analysis is discussed in Sect. 5.2.4.

5.2.1 Global Interconnect Characteristics

The scaling of the cross-sectional dimensions of the on-chip global power lines directly affects the power distribution noise. Two scenarios of global interconnect scaling are considered here.

In the first scenario, the thickness of the top interconnect layers (where the conductors of the global power distribution networks are located) is assumed to remain constant. Through several recent technology generations, the thickness of the global interconnect layers has not been scaled in proportion to the minimum local line pitch due to power distribution noise and interconnect delay considerations. This behavior is in agreement with the 1997 edition of the International Technology Roadmap for Semiconductors (ITRS) [122, 123], where the minimum pitch and thickness of the global interconnect are assumed constant.

In the second scenario, the thickness and minimum pitch of the global interconnect layers are scaled in proportion to the minimum pitch of the local interconnect. This assumption is in agreement with the more recent editions of the ITRS [121, 124]. The scaling of the global interconnect in future technologies is therefore expected to evolve in the design envelope delimited by these two scenarios.

The number of metal layers and the fraction of the metal resources dedicated to the power distribution network are also assumed constant. The ratio of the diffusion barrier thickness to the copper interconnect core is assumed to remain constant with scaling. The increase in resistivity of the interconnect due to electron scattering at the interconnect surface interface (significant at line widths below 45 nm [124]) is neglected for relatively thick global power lines.

Under the aforementioned assumptions, in the constant metal thickness scenario, the effective sheet resistance of the global power distribution network remains constant with technology scaling. In the scenario of scaled metal thickness, the grid sheet resistance increases with technology scaling by a factor of S.

5.2.2 Scaling of the Grid Inductance

The inductive properties of power distribution grids are investigated in [72, 73]. It is shown that the inductance of the power grids with alternating power and ground lines behaves analogously to the grid resistance. That is, the grid inductance increases linearly with the grid length and decreases inversely linearly with the number of lines in the grid. This linear behavior is due to the periodic structure of the alternating power and ground grid lines. The long range inductive coupling of a specific (signal or power) line to a power line is cancelled out by the coupling to the ground lines adjacent to the power line, which carry current in the opposite direction [71, 73]. As described in Chap. 28, inductive coupling in periodic grid structures is effectively a short range interaction. Similar to the grid resistance, the grid inductance can be conveniently expressed as a dimension independent grid sheet inductance $L_{\square}$ [73, 125]. The inductance of a specific grid is obtained by multiplying the sheet inductance by the grid length and dividing by the grid width. The grid sheet inductance (for a derivation, see Sect. 28.6.4) can be estimated as

$$L_{\square} = 0.8P\left(\ln\frac{P}{T+W}+\frac{3}{2}\right)\frac{\mu H}{\square}, \tag{5.1}$$

where W, T, and P are, respectively, the width, thickness, and pitch of the grid lines. The sheet inductance is proportional to the line pitch P. The line density is reciprocal to the line pitch. A smaller line pitch means a higher line density and more parallel paths for the current to flow. The sheet inductance is however relatively insensitive to the cross-sectional dimensions of the lines, as the inductance of the individual lines is similarly insensitive to these parameters. Note that while the sheet resistance of the power grid is determined by the metal conductivity and the net cross-sectional area of the lines, the sheet inductance of the grid is determined by the line pitch and the ratio of the pitch to the line width and thickness.

In the constant metal thickness scenario, the sheet inductance of the power grid remains constant since the routing characteristics of the global power grid do not change. In the scaled thickness scenario, the line pitch, width, and thickness are reduced by S, increasing the line density and the number of parallel current paths. The sheet inductance therefore decreases by a factor of S, according to (5.1).

5.2.3 Flip-Chip Packaging Characteristics

In a flip-chip package, the integrated circuit and package are interconnected via an area array of solder bumps mounted onto the on-chip I/O pads [126]. The power supply current enters the on-chip power distribution network from the power/ground pads. A view of the on-chip area array of power/ground pads is shown in Fig. 5.1.

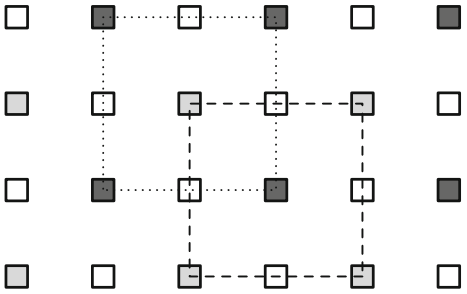


Fig. 5.1 An area array of on-chip power/ground I/O pads. The power pads are colored *dark gray*, the ground pads are colored *light gray*, and the signal pads are *white*. The current distribution area of the power pad (i.e., the power distribution cell) in the center of the figure is delineated by the *dashed line*. The current distribution area of the ground pad in the center of the figure is delineated by the *dotted line*

One of the main goals of this work is to estimate the significance of the *on-chip* inductive voltage drop in comparison to the on-chip resistive voltage drop. Therefore, all of the power/ground pads of a flip-chip packaged IC are assumed to be equipotential, i.e., the variation in the voltage levels among the pads is considered negligible as compared to the noise within the on-chip power distribution network. For the purpose of this scaling analysis, a uniform power consumption per die area is assumed. Under these assumptions, each power (ground) pad supplies power (ground) current only to those circuits located in the area around the pad, as shown in Fig. 5.1. This area is referred to as a power distribution cell (or power cell). The edge dimensions of each power distribution cell are proportional to the pitch of the power/ground pads. The size of the power cell area determines the effective distance of the on-chip distribution of the power current. The power distribution scaling analysis becomes independent of die size.

An important element of this analysis is the scaling of the flip-chip technology. The rate of decrease in the pad pitch and the rate of reduction in the local interconnect half-pitch are compared in Fig. 5.2, based on the ITRS [124]. At the 150 nm line half-pitch technology node, the pad pitch P is 160 μm. At the 32 nm node, the pad pitch is forecasted to be 80 μm. That is, the linear density of the pads doubles for a fourfold reduction in circuit feature size. The pad size and pitch P scale, therefore, as $1/\sqrt{S}$ and the area density ($\propto 1/P^2$) of the pads increases as S with each technology generation. Interestingly, one of the reasons given for this relatively infrequent change in the pad pitch (as compared with the introduction of new CMOS technology generations) is the cost of the test probe head [124]. The maximum density of the flip-chip pads is assumed to be limited by the pad pitch. Although the number of on-chip pads is forecasted to remain constant, some recent research has predicted that the number of on-chip power/ground pads will increase due to electromigration and resistive noise considerations [115, 127].

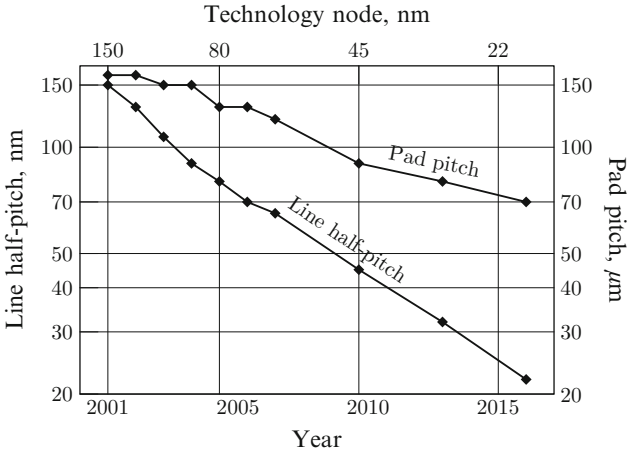


Fig. 5.2 Decrease in flip-chip pad pitch with technology generations as compared to the local interconnect half-pitch

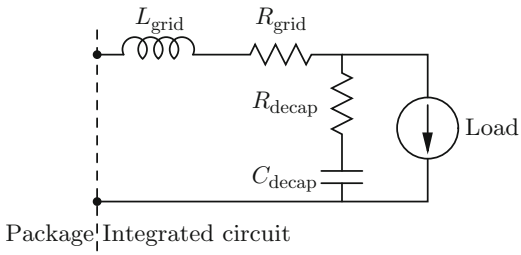


Fig. 5.3 A simplified circuit model of the on-chip power distribution network with a power load and a decoupling capacitance

5.2.4 Impact of On-Chip Capacitance

On-chip capacitors are used to reduce the impedance of the power distribution grid lines as seen from the load terminals. A simple model of an on-chip power distribution grid with a power load and a decoupling capacitor is shown in Fig. 5.3. The on-chip loads are switched within tens of picoseconds in modern semiconductor technologies. The frequency spectrum of the load current therefore extends well beyond 10 GHz. The on-chip decoupling capacitors shunt the load current at the highest frequencies. The bulk of the power current bypasses the on-chip distribution network at these frequencies. At the lower frequencies, however, the capacitor impedance is relatively high and the bulk of the current flows through the on-chip power distribution network. The decoupling capacitors therefore serve as a low pass filter for the power current.

Describing the same effect in the time domain, the capacitors supply the (high frequency) current to the load during a switching transient. To prevent excessive power noise, the charge on the decoupling capacitor should be replenished by the (lower frequency) current flowing through the power distribution network

before the next switching of the load, i.e., typically within a clock period. The effect of the on-chip decoupling capacitors is therefore included in the model by assuming that the current transients within the on-chip power distribution network are characterized by the clock frequency of the circuit, rather than by the switching times of the on-chip load circuits. Estimates of the resistive voltage drop are based on the average power current, which is not affected by the on-chip decoupling capacitors.

5.3 Model of Power Supply Noise

The following simple model is utilized in the scaling analysis of on-chip power distribution noise. A power distribution cell is modeled as a circle of radius r_c with a constant current consumption per area I_a, as described by Arledge and Lynch [115]. The model is depicted in Fig. 5.4. The total current of the cell is $I_{\text{cell}} = I_a \cdot \pi r_c^2$. The power network current is distributed from a circular pad of radius r_p at the center of the cell. The global power distribution network has an effective sheet resistance $\rho_\square$. The incremental voltage drop dV_R across the elemental circular resistance $\rho_\square\, dr/2\pi r$ is due to the current $I_a(\pi r_c^2 - \pi r^2)$ flowing through this resistance toward the periphery of the cell. The voltage drop at the periphery of the power distribution cell is

$$\Delta V_R = \int_{r_p}^{r_c} dV_R = \int_{r_p}^{r_c} I(r) \cdot dR(r)$$

$$= \int_{r_p}^{r_c} \pi(r_c^2 - r^2) I_a \cdot \rho_\square \frac{dr}{2\pi r}$$

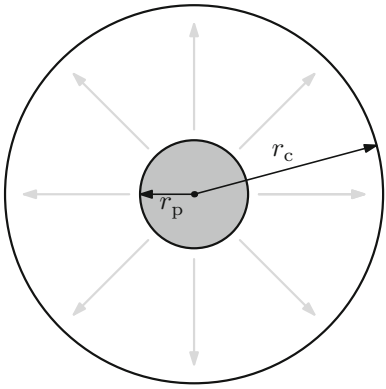


Fig. 5.4 A model of the power distribution cell. Power supply current spreads out from the power pad in the center of the cell to the cell periphery, as shown by the *arrows*

$$= I_a \pi r_c^2 \rho_{\square} \cdot \frac{1}{2\pi} \left(\ln \frac{r_c}{r_p} + \frac{r_p^2}{2r_c^2} - \frac{1}{2} \right)$$

$$= I_{\text{cell}} \rho_{\square} \cdot C \left(\frac{r_c}{r_p} \right). \tag{5.2}$$

The resistive voltage drop is proportional to the product of the total cell current I_{cell} and the effective sheet resistance $\rho_{\square}$ with the coefficient C dependent only on the r_c/r_p ratio. The ratio of the pad pitch to the pad size is assumed to remain constant. The coefficient C, therefore, does not change with technology scaling.

The properties of the grid inductance are analogous to the properties of the grid resistance, as discussed in Sect. 5.2.1. Therefore, analogous to the resistive voltage drop ΔV_R discussed above, the inductive voltage drop ΔV_L is proportional to the product of the sheet inductance $L_{\square}$ of the global power grid and the magnitude of the cell transient current dI_{cell}/dt,

$$\Delta V_L = L_{\square} \frac{dI_{\text{cell}}}{dt} \cdot C \left(\frac{r_c}{r_p} \right). \tag{5.3}$$

5.4 Power Supply Noise Scaling

An analysis of the on-chip power supply noise is presented in this section. The interconnect characteristics assumed in the analysis are described in Sect. 5.2. The power supply noise model is described in Sect. 5.3. Ideal scaling of the power distribution noise in the constant thickness scenario is discussed in Sect. 5.4.1. Ideal scaling of the noise in the scaled thickness scenario is analyzed in Sect. 5.4.2. Scaling of the power distribution noise based on the ITRS projections is discussed in Sect. 5.4.3.

5.4.1 Analysis of Constant Metal Thickness Scenario

The scaling of a power distribution grid over four technology generations according to the constant metal thickness scenario is depicted in Fig. 5.5. The minimum feature size is reduced by $\sqrt{2}$ with each generation. The minimum feature size over four generations is therefore reduced by four, i.e., $\left(\sqrt{2}\right)^4 = 4$, while the size of the power distribution cell (represented by the size of the square grid) is halved ($\sqrt{4} = 2$). As the cross-sectional dimensions of the power lines are maintained constant in this scenario, both the sheet resistance $\rho_{\square}$ and sheet inductance $L_{\square}$ of the power distribution grid remain constant with scaling under these conditions.

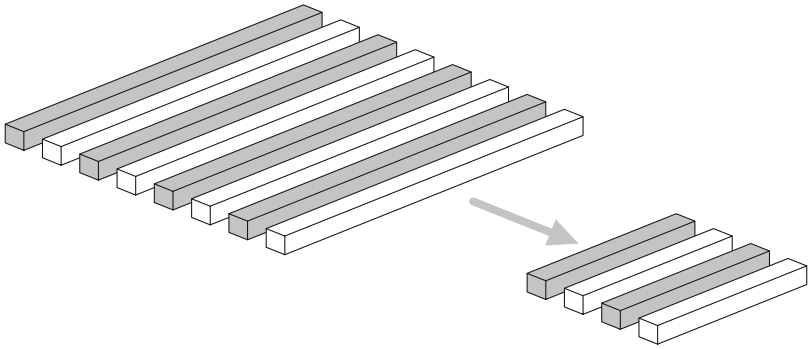


Fig. 5.5 The scaling of a power distribution grid over four technology generations according to the constant metal thickness scenario. The cross-sectional dimensions of the power lines remain constant. The size of the power distribution cell, represented by the size of the *square grid*, is halved

The cell current I_{cell} is the product of the area current density I_a and the cell area πr_c^2. The current per area I_a scales as S; the area of the cell is proportional to P^2 which scales as $1/S$. The cell current I_{cell}, therefore, remains constant (i.e., scales as 1). The resistive drop ΔV_R, therefore, scales as $I_{\text{cell}} \cdot \rho_{\square} \propto 1 \cdot 1 \propto 1$. The resistive SNR_R^{I} of the power supply voltage, consequently, decreases with scaling as

$$\text{SNR}_R^{\text{I}} = \frac{V_{\text{dd}}}{\Delta V_R} \propto \frac{1/S}{1} \propto \frac{1}{S}. \tag{5.4}$$

This scaling trend agrees with the trend described by Bakoglu in the improved scaling regime [114]. Faster scaling of the on-chip current as described by Bakoglu is offset by increasing the interconnect thickness by S which reduces the sheet resistance $\rho_{\square}$ by S. This trend is more favorable as compared to the $1/S^2$ dependence established by Song and Glasser [116]. The improvement is due to the decrease in the power cell area of a flip-chip IC by a factor of S whereas a wire bonded die of constant area is assumed in [116].

The transient current dI_{cell}/dt scales as $I_{\text{cell}}/\tau \propto 1/(1/S) \propto S$, where $\tau \propto 1/S$ is the transistor switching time. The inductive voltage drop ΔV_L, therefore, scales as $L_{\square} \cdot dI_{\text{cell}}/dt \propto 1 \cdot S$. The inductive SNR_L^{I} of the power supply voltage decreases with scaling as

$$\text{SNR}_L^{\text{I}} = \frac{V_{\text{dd}}}{\Delta V_L} \propto \frac{1/S}{S} \propto \frac{1}{S^2}. \tag{5.5}$$

The relative magnitude of the inductive noise therefore increases by a factor of S faster as compared to the resistive noise. Estimates of the inductive and resistive noise described by Bakoglu also differ by a factor of S [114].

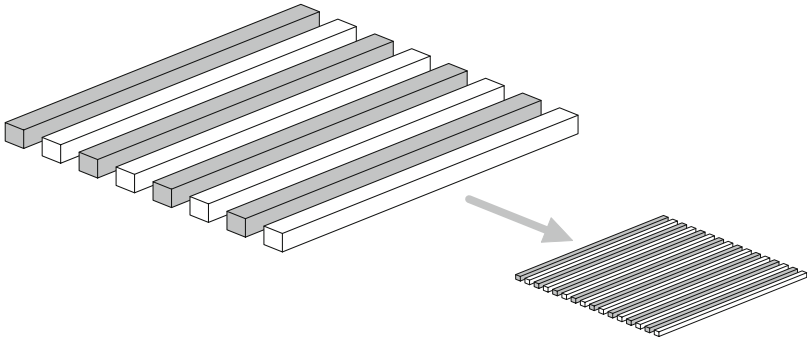


Fig. 5.6 The scaling of a power distribution grid over four technology generations according to the scaled metal thickness scenario. The cross-sectional dimensions of the power lines are reduced in proportion to the minimum feature size by a factor of four. The size of the power distribution cell, represented by the size of the *square grid*, is halved

5.4.2 Analysis of the Scaled Metal Thickness Scenario

The scaling of a power distribution grid over four technology generations according to the scaled metal thickness scenario is depicted in Fig. 5.6. In this scenario, the cross-sectional dimensions of the power lines are reduced in proportion to the minimum feature size by a factor of four, while the size of the power distribution cell is halved. Under these conditions, the sheet resistance $\rho_\square$ of the power distribution grid increases by S, while the sheet inductance $L_\square$ of the power distribution grid decreases by S with technology scaling.

Analogous to the constant metal thickness scenario, the cell current I_{cell} remains constant. The resistive drop ΔV_R, therefore, scales as $I_{\text{cell}} \cdot \rho_\square \propto 1 \cdot S \propto S$. The resistive SNR_R^{II} of the power supply voltage, consequently, decreases with scaling as

$$\text{SNR}_R^{\text{II}} = \frac{V_{\text{dd}}}{\Delta V_R} \propto \frac{1/S}{S} \propto \frac{1}{S^2}. \tag{5.6}$$

As discussed in the previous section, the transient current dI_{cell}/dt scales as $I_{\text{cell}}/\tau \propto S$. The inductive voltage drop ΔV_L, therefore, scales as $L_\square \cdot dI_{\text{cell}}/dt \propto 1/S \cdot S \propto 1$. The inductive SNR_L^{II} of the power supply voltage decreases with scaling as

$$\text{SNR}_L^{\text{II}} = \frac{V_{\text{dd}}}{\Delta V_L} \propto \frac{1/S}{1} \propto \frac{1}{S}. \tag{5.7}$$

The rise of the inductive noise is mitigated if ideal interconnect scaling is assumed and the thickness, width, and pitch of the global power lines are scaled as $1/S$. In this scenario, the density of the global power lines increases as S

and the sheet inductance $L_{\square}$ of the global power distribution grids decreases as $1/S$, mitigating the inductive noise and SNR_L by S. The sheet resistance of the power distribution grid, however, increases as S, exacerbating the resistive noise and SNR_R by a factor of S. Currently, the parasitic resistive impedance dominates the total impedance of on-chip power distribution networks. Ideal scaling of the upper interconnect levels will therefore increase the overall power distribution noise. However, as CMOS technology approaches the nanometer range and the inductive and resistive noise becomes comparable, judicious tradeoffs between the resistance and inductance of the power networks will be necessary to achieve the minimum noise level (see Chap. 28) [128–130].

5.4.3 ITRS Scaling of Power Noise

Although the ideal scaling analysis allows the comparison of the rates of change of both resistive and inductive voltage drops, it is difficult to estimate the *ratio* of these quantities for direct assessment of their relative significance. Furthermore, practical scaling does not accurately follow the concept of ideal scaling due to material and technological limitations. An estimate of the ratio of the inductive to resistive voltage drop is therefore conducted in this section based on the projected 2001 ITRS data [124].

Forecasted demands in the supply current of high performance microprocessors are shown in Fig. 5.7. Both the average current and the transient current are rising exponentially with technology scaling. The rate of increase in the transient current is more than double the rate of increase in the average current as indicated by the

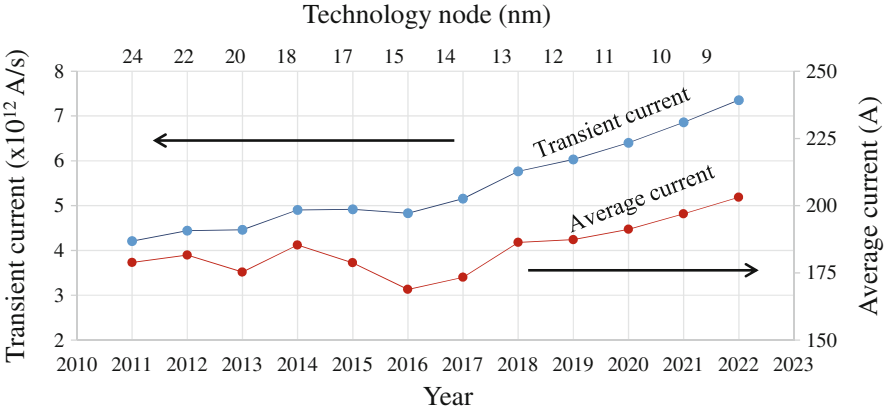


Fig. 5.7 Increase in power current demands of high performance microprocessors with technology scaling, according to the ITRS. The average current is the ratio of the circuit power to the supply voltage. The transient current is the product of the average current and the on-chip clock rate, $2\pi f_{\text{clk}}$

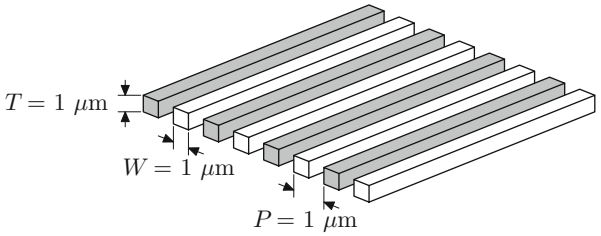


Fig. 5.8 Power distribution grid used to estimate trends in the power supply noise

slope of the trend lines depicted in Fig. 5.7. This behavior is in agreement with ideal scaling trends. The faster rate of increase in the transient current as compared to the average current is due to rising clock frequencies. The transient current of modern high performance processors is approximately one teraampere per second (10^{12} A/s) and is expected to rise, reaching hundreds of teraamperes per second. Such a high magnitude of the transient current is caused by switching hundreds of amperes within a fraction of a nanosecond.

In order to translate the projected current requirements into supply noise voltage trends, a case study interconnect structure is considered. The square grid structure shown in Fig. 5.8 is used here to serve as a model of the on-chip power distribution grid. The square grid consists of interdigitated power and ground lines with a $1 \times 1\,\mu$m cross section and a $1\,\mu$m line spacing. The length and width of the grid are equal to the size of a power distribution cell. The grid sheet inductance is 1.8 picohenrys per square, and the grid sheet resistance is 0.16 ohms per square. The size of the power cell is assumed to be twice the pitch of the flip-chip pads, reflecting that only half of the total number of pads are used for the power and ground distribution as forecasted by the ITRS for high performance ASICs.

The electrical properties of this structure are similar to the properties of the global power distribution grid covering a power distribution cell with the same routing characteristics. Note that the resistance and inductance of the *square* grid are independent of grid dimensions [125] (as long as the dimensions are severalfold greater than the line pitch). The average and transient currents flowing through the grid are, however, scaled from the IC current requirements shown in Fig. 5.7 in proportion to the area of the grid. The current flowing through the square grid is therefore the same as the current distributed through the power grid within the power cell. The power current enters and leaves from the same side of the grid, assuming the power load is connected at the opposite side. The voltage differential across this structure caused by the average and transient currents produces, respectively, on-chip resistive and inductive noise. The square grid has the same inductance to resistance ratio as the global distribution grid with the same line pitch, thickness, and width. Hence, the square grid has the same inductive to resistive noise ratio. The square grid model also produces the same rate of increase in the noise because the current is scaled proportionately to the area of the power cell.

The resulting noise trends under the constant metal thickness scenario are illustrated in Fig. 5.9. As discussed in Sect. 5.2, the area of the grid scales as $1/S$. The

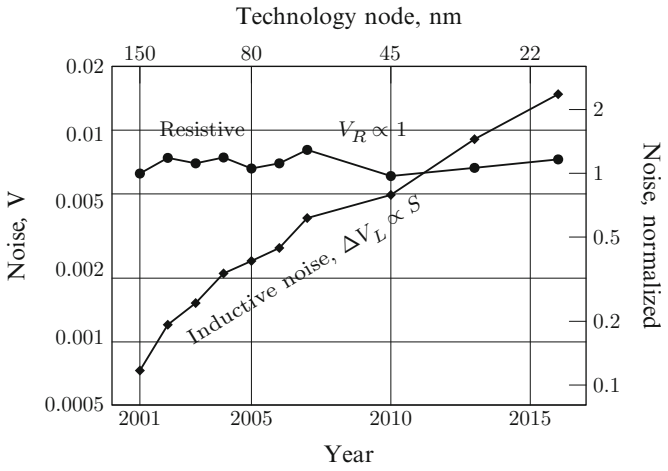


Fig. 5.9 Scaling trends of resistive and inductive power supply noise under the constant metal thickness scenario

current area density increases as S. The total average current of the grid, therefore, remains constant. The resistive noise also remains approximately constant, as shown in Fig. 5.9. The inductive noise, alternatively, rises steadily and becomes comparable to the resistive noise at approximately the 45 nm technology node. These trends are in reasonable agreement with the ideal scaling predictions discussed in Sect. 5.4.1.

The inductive and resistive voltage drops in the scaled metal thickness scenario are shown in Fig. 5.10. The increase in inductive noise with technology scaling is limited, while the resistive noise increases by an order of magnitude. This behavior is similar to the ideal scaling trends for this scenario, as discussed in Sect. 5.4.2.

Note that the structure depicted in Fig. 5.8 has a lower inductance to resistance ratio as compared to typical power distribution grids because the power and ground lines are relatively narrow and placed adjacent to each other, reducing the area of the current loop and increasing the grid resistance [72, 125]. The width of a typical global power line varies from tens to a few hundreds of micrometers, resulting in a significantly higher inductance to resistance ratio. The results shown in Figs. 5.9 and 5.10 can be readily extrapolated to different grid configurations, using the expression for the grid sheet inductance (5.1).

Several factors offset the underestimation of the relative magnitude of the inductive noise due to the relatively low inductance to resistance ratio of the model shown in Fig. 5.8. If the global power distribution grid is composed of several layers of interconnect, the lines in the lower interconnect levels have a smaller pitch and thickness, significantly reducing the inductance to resistance ratio at high frequencies [131]. The transient current is conservatively approximated as the product of the average current I_{avg} and the angular clock frequency $2\pi f_{\text{clk}}$. This estimate, while serving as a useful scaling parameter, tends to overestimate the

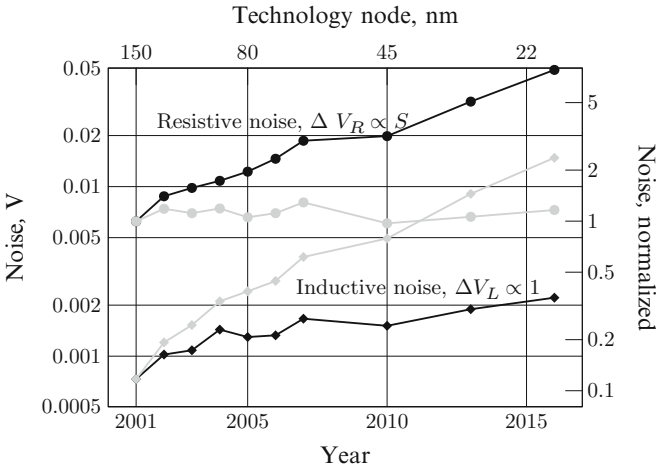


Fig. 5.10 Scaling trends of resistive and inductive power supply noise under the scaled metal thickness interconnect scaling scenario. The trends of the constant metal thickness scenario are also displayed in *light gray* for comparison

absolute magnitude of the current transients, increasing the ratio of the inductive and resistive voltage drops.

5.5 Implications of Noise Scaling

As described in the previous section, the amplitude of both the resistive and inductive noise relative to the power supply voltage increases with technology scaling. A number of techniques have been developed to mitigate the unfavorable scaling of power distribution noise. These techniques are briefly summarized below.

To maintain a constant supply voltage to resistive noise ratio, the effective sheet resistance of the global power distribution grid should be reduced. There are two ways to allocate additional metal resources to the power distribution grid. One option is to increase the number of metalization layers. This approach adversely affects fabrication time and yield and, therefore, increases the cost of manufacturing. The ITRS forecasts only a moderate increase in the number of interconnect levels, from eight levels at the 130 nm line half-pitch node to 11 levels at the 32 nm node [124]. The second option is to increase the fraction of metal area per metal level allocated to the power grid. This strategy decreases the amount of wiring resources available for global signal routing and therefore can also necessitate an increase in the number of interconnect layers.

The sheet inductance of the power distribution grid, similar to the sheet resistance, can be lowered by increasing the number of interconnect levels. Furthermore, wide metal trunks typically used for power distribution at the top levels can be

replaced with narrow interdigitated power/ground lines. Although this configuration substantially lowers the grid inductance, it increases the grid resistance and, consequently, the resistive noise [125].

Alternatively, circuit techniques can be employed to limit the peak transient power current demands of the digital logic. Current steering logic, for example, produces a minimal variation in the current demand between the transient response and the steady state response. In synchronous circuits, the maximum transient currents typically occur during the beginning of a clock period. Immediately after the arrival of a clock signal at the latches, a signal begins to propagate through the blocks of sequential logic. Clock skew scheduling can be exploited to spread in time the periods of peak current demand [39].

The constant metal thickness scaling scenario achieves a lower overall power noise until the technology generation is reached where the inductive and resistive voltage drops become comparable. Beyond this node, a careful tradeoff between the resistance and inductance of the power grid is necessary to minimize the on-chip power supply noise. The increase in the significance of the inductance of the power distribution interconnect is similar to that noted in signal interconnect [64, 132]. The trend, however, is delayed by several technology generations as compared to signal interconnect. As discussed in Sect. 5.2, the high frequency harmonics are filtered out by the on-chip decoupling capacitance and the power grid current has a comparatively lower frequency content as compared to the signal lines.

5.6 Summary

A scaling analysis of power distribution noise in flip-chip packaged integrated circuits is presented in this chapter. Published scaling analyses of power distribution noise are reviewed and various assumptions of these analyses are discussed. The primary conclusions can be summarized as follows.

- Under the constant metal thickness scenario, the relative magnitude (i.e., the reciprocal of the signal-to-noise ratio) of the resistive noise increases by the scaling factor S, while the relative magnitude of the inductive noise increases by S^2
- Under the scaled metal thickness scenario, the scaling trend of the inductive noise improves by a factor of S, but the relative magnitude of the resistive noise increases by S^2
- The importance of on-chip inductive noise increases with technology scaling
- Careful tradeoffs between the resistance and inductance of power distribution networks in nanometer CMOS technologies will be necessary to achieve minimum power supply noise levels

Chapter 6
Conclusions

In Part I, the development of integrated circuits, design objectives, and other general issues are discussed. With increasing current requirements and lower voltage margins, focus is placed on the design of high performance power distribution networks. The quality of the voltage greatly affects the performance of an integrated circuit.

With advancements in technology, additional focus is placed on the inductive properties of the network. The power and ground lines create a closed current loop, where the loop inductance is composed of the self- and mutual inductance of these lines. With increasing number of current loops, accurately estimating the grid inductance has become increasingly complicated since significantly more mutual inductive elements need to be considered.

Electromigration is also reviewed here, since the long term reliability of integrated circuits is a primary concern. High current propagates within power distribution networks, placing stress on the network interconnects. Electromigration is significantly reduced in those lines where the current direction alternates. The current in power supply networks is however primarily unidirectional; therefore, electromigration can significantly degrade the reliability of these networks.

Scaling theory is applied to power distribution networks, providing scaling trends for these networks. On-chip inductive noise has become a primary concern with technology scaling and greatly affects the overall power supply noise in modern high performance systems. The topics reviewed in Part I are intended to provide a general background to the topic of on-chip power delivery, permitting the reader to more easily follow the remainder of the book.

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0_6

Part II
Power Delivery Networks

The focus of Part II is on power distribution systems, specifically on-chip power distribution networks. Different topologies are described and compared. Due to the complex nature of on-chip power distribution networks, computer-aided design processes are required. Different techniques to design and analyze these networks are reviewed in this part. Focusing on a mesh structured power distribution network, efficient analysis methods are presented. A description of each chapter is provided below, finishing with a conclusion chapter summarizing the design and analysis of on-chip power distribution networks.

An overview of hierarchical power distribution networks is presented in Chap. 7. The inductive nature of the board and package interconnect is identified as the primary obstacle towards achieving a low impedance at high frequencies. The effect of decoupling capacitances on the impedance characteristics of power distribution systems is also examined. The use of a hierarchy of decoupling capacitors to reduce the output impedance of power distribution systems is described. Finally, design guidelines for lowering the impedance characteristics of power distribution systems are discussed.

The focus of Chap. 8 is on-chip power and ground distribution networks. Topological variations of the structure of on-chip power distribution networks are described, highlighting the benefits and disadvantages of each network topology. Techniques are reviewed to reduce the impedance of on-chip power distribution networks. A discussion of strategies for allocating the on-chip decoupling capacitance concludes this chapter.

The concept of a power network on-chip (PNoC) is described in Chap. 9 as a systematic methodology for on-chip power delivery and management that provides enhanced power control and real-time locally intelligent management of resource sharing. The PNoC utilizes a modular architecture with intelligent distributed on-chip power routers to address the issues of design complexity and scalability.

Chapter 7
Hierarchical Power Distribution Networks

Supplying power to high performance integrated circuits has become a challenging task. The system supplying power to an integrated circuit greatly affects the performance, size, and cost characteristics of the overall electronic system. This system is comprised of interconnect networks with decoupling capacitors on a printed circuit board, an integrated circuit package, and a circuit die. The entire system is henceforth referred to as the *power distribution system*. The design of power distribution systems is described in this chapter. The focus of the discussion is the overall structure and interaction among the various parts of the system. The impedance characteristics and design of on-chip power distribution networks, the most complex part of the power distribution system, are discussed in greater detail in the following chapters.

This chapter is organized as follows. A typical power distribution system for a high power, high speed integrated circuit is described in Sect. 7.1. A circuit model of a power distribution system is presented in Sect. 7.2. The output impedance characteristics are discussed in Sect. 7.3. The effect of a shunting capacitance on the impedance of a power distribution system is considered in Sect. 7.4. The hierarchical placement of capacitors to satisfy target impedance requirements is described in Sect. 7.5. Design strategies to control the resonant effects in power distribution networks are discussed in Sect. 7.6. A purely resistive impedance characteristic of power distribution systems can be achieved using an impedance compensation technique, as described in Sect. 7.7. A case study of a power distribution system is presented in Sect. 7.8. Design techniques to enhance the impedance characteristics of power distribution networks are discussed in Sect. 7.9. The limitations of the analysis presented herein are discussed in Sect. 7.10. The chapter concludes with a summary.

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0_7

7.1 Physical Structure of a Power Distribution System

A cross-sectional view of a power distribution system for a high performance integrated circuit is shown in Fig. 7.1. The power supply system spans several levels of packaging hierarchy. It consists of a switching voltage regulator module (VRM), the power distribution networks on a printed circuit board (PCB), on an integrated circuit package, and on-chip, plus the decoupling capacitors connected to these networks. The power distribution networks at the board, package, and circuit die levels form a conductive path between the power source and the power load. A switching voltage regulator converts the DC voltage level provided by a system power supply unit to a voltage V_{dd} required for powering an integrated circuit. The regulator serves as a power source, effectively decoupling the power distribution system of an integrated circuit from the system level power supply. The power and ground planes of the printed circuit board connect the switching regulator to the integrated circuit package. The board-level decoupling capacitors are placed across the power and ground planes to provide charge storage for current transients faster than the response time of the regulator. The board power and ground interconnect is connected to the power and ground networks of the package through a ball grid array (BGA) or pin grid array (PGA) contacts. Similar to a printed circuit board, the package power and ground distribution networks are typically comprised of several metal planes. High frequency (i.e., with low parasitic impedance) decoupling capacitors are mounted on the package, electrically close to the circuit die, to ensure the integrity of the power supply during power current surges drawn by the circuit from the package level power distribution network. The package power and ground distribution networks are connected to the on-chip power distribution grid through a flip-chip array of bump contacts or alternative bonding technologies [126]. On-chip decoupling capacitors are placed across the on-chip power and ground networks to maintain a low impedance at signal frequencies comparable to the switching speed of the on-chip devices.

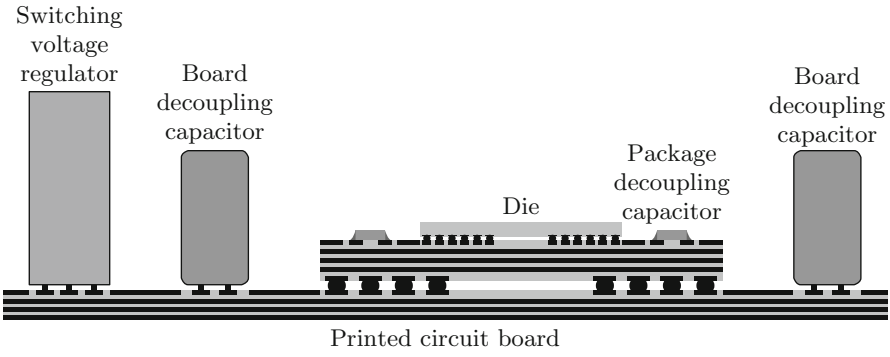


Fig. 7.1 A cross-sectional view of a power distribution system of a high performance integrated circuit

The physical structure of the power distribution system is hierarchical. Each tier of the power distribution system typically corresponds to a tier of packaging hierarchy and consists of a power distribution network and associated decoupling capacitors. The hierarchical structure of the power distribution system permits the desired impedance characteristics to be obtained in a cost effective manner, as described in the following sections.

7.2 Circuit Model of a Power Distribution System

A simplified circuit model of a power supply system[1] is shown in Fig. 7.2. This lumped circuit model is effectively one-dimensional, where each level of the packaging hierarchy is modeled by a pair of power and ground conductors and a decoupling capacitor across these conductors. A one-dimensional model accurately describes the impedance characteristics of a power distribution system over a wide frequency range [133–135]. The conductors are represented by the parasitic resistive and inductive impedances; the decoupling capacitors are represented by a series *RLC* circuit reflecting the parasitic impedances of the capacitors. The italicized superscripts "*p*" and "*g*" refer to the power and ground conductors, respectively; superscript "*C*" refers to the parasitic impedance characteristics of the capacitors. The subscripts "r," "b," "p," and "c" refer to the regulator, board, package, and on-chip conductors. Subscript 1 refers to the conductors upstream of the respective decoupling capacitor, with respect to the flow of energy from the power source to the load. That is, the conductors denoted with subscript 1 are connected to the appropriate decoupling capacitor at the voltage regulator. Subscript 2 refers to the conductors downstream of the appropriate decoupling capacitor. For example, R^g_{b1} refers to

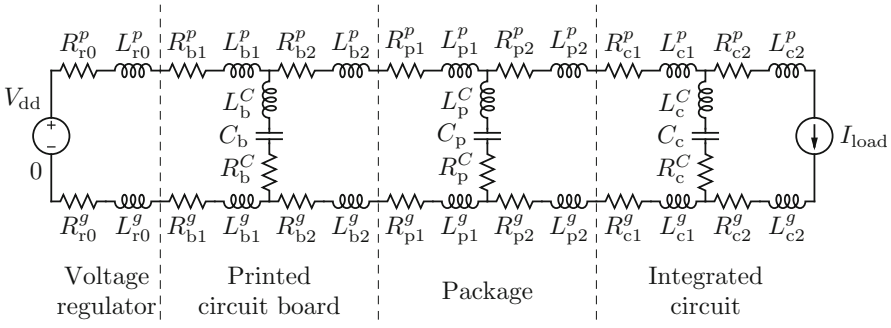


Fig. 7.2 A one-dimensional circuit model of the power supply system shown in Fig. 7.1

[1]The magnetic coupling of the power and ground conductors is omitted in the circuit diagrams of a power distribution system. The inductive elements shown in the diagrams therefore denote a net inductance, as described in Sect. 2.1.4.

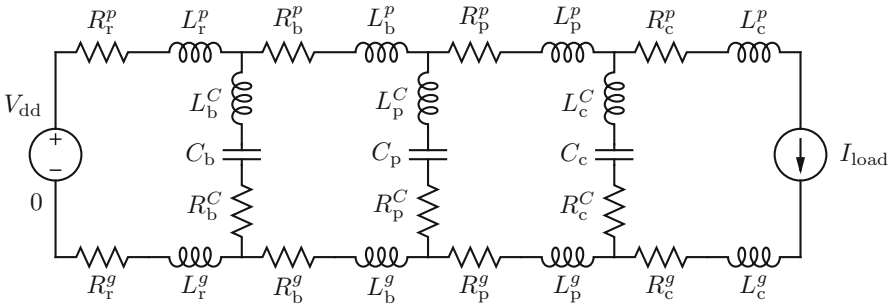


Fig. 7.3 A reduced version of the circuit model characterizing a power supply system

the parasitic resistance of the ground conductors connecting the board capacitors to the voltage regulator, while L_{p2}^p refers to the parasitic inductance of the power conductors connecting the package capacitors to the on-chip I/O contacts. Similarly, L_p^C refers to the parasitic series inductance of the package capacitors.

The model shown in Fig. 7.2 can be reduced by combining the circuit elements connected in series. The reduced circuit model is shown in Fig. 7.3. The circuit characteristics of the circuit shown in Fig. 7.3 are related to the characteristics of the circuit shown in Fig. 7.2 as

$$R_r^p = R_{r0}^p + R_{b1}^p \qquad L_r^p = L_{r0}^p + L_{b1}^p \tag{7.1a}$$

$$R_b^p = R_{b2}^p + R_{p1}^p \qquad L_b^p = L_{b2}^p + L_{p1}^p \tag{7.1b}$$

$$R_p^p = R_{p2}^p + R_{c1}^p \qquad L_p^p = L_{p2}^p + L_{c1}^p \tag{7.1c}$$

$$R_c^p = R_{c2}^p \qquad L_c^p = L_{c2}^p \tag{7.1d}$$

for conductors carrying power current and, analogously, for the ground conductors,

$$R_r^g = R_{r2}^g + R_{b1}^g \qquad L_r^g = L_{r2}^g + L_{b1}^g \tag{7.1e}$$

$$R_b^g = R_{b2}^g + R_{p1}^g \qquad L_b^g = L_{b2}^g + L_{p1}^g \tag{7.1f}$$

$$R_p^g = R_{p2}^g + R_{c1}^g \qquad L_p^g = L_{p2}^g + L_{c1}^g \tag{7.1g}$$

$$R_c^g = R_{c2}^g \qquad L_c^g = L_{c2}^g. \tag{7.1h}$$

As defined, the circuit elements with subscript "r" represent the output impedance of the voltage regulator and the impedance of the on-board current path from the regulator to the board decoupling capacitors. The elements denoted with subscript "b" represent the impedance of the current path from the board capacitors to the package, the impedance of the socket and package pins (or solder bumps in the case of a ball grid array mounting solution), and the impedance of the package lines

and planes. Similarly, the elements with subscript "p" signify the impedance of the current path from the package capacitors to the die mounting site, the impedance of the solder bumps or bonding wires, and, partially, the impedance of the on-chip power distribution network. Finally, the resistors and inductors with subscript "c" represent the impedance of the current path from the on-chip capacitors to the on-chip power load.

The board, package, and on-chip power distribution networks have significantly different electrical characteristics due to the different physical properties of the interconnect at the various tiers of the packaging hierarchy. From the board level to the die level, the cross-sectional dimensions of the interconnect lines decrease while the aspect ratio increases, producing a dramatic increase in interconnect density. The inductance of the board level power distribution network, i.e., L^p_{b1}, L^g_{b1}, L^p_{b2}, and L^g_{b2}, is large as the power and ground planes are typically separated by tens or hundreds of micrometers. The effective output impedance of the voltage switching regulator over a wide range of frequencies can be described as $R + j\omega L$ [133, 136, 137], hence the presence of R^p_{r2} and R^g_{r2}, and L^p_{r2} and L^g_{r2} in the model shown in Fig. 7.2. The inductance of the on-chip power distribution network, L^p_{c1}, L^g_{c1}, L^p_{c2}, and L^g_{c2}, is comparatively low due to the high interconnect density. The inductance L^p_{p1}, L^g_{p1}, L^p_{p2}, and L^g_{p2} of the package level network is of intermediate magnitude, larger than the inductance of the on-chip network but smaller than the inductance of the board level network. The inductive characteristics of the circuit shown in Fig. 7.3 typically exhibit a hierarchical relationship: $L^p_b > L^p_p > L^p_c$ and $L^g_b > L^g_p > L^g_c$. The resistance of the power distribution networks follows the opposite trend. The board level resistances R^p_{b1}, R^g_{b1}, R^p_{b2}, and R^g_{b2} are small due to the large cross-sectional dimensions of the relatively thick board planes, while the on-chip resistances R^p_{c1}, R^g_{c1}, R^p_{c2}, and R^g_{c2} are large due to the small cross-sectional dimensions of the on-chip interconnect. The resistive characteristics of the power distribution system are therefore reciprocal to the inductive characteristics, $R^p_b < R^p_p < R^p_c$ and $R^g_b < R^g_p < R^g_c$. The physical hierarchy is thus reflected in the electrical hierarchy: the progressively finer physical features of the conductors typically result in a higher resistance and a lower inductance.

7.3 Output Impedance of a Power Distribution System

To ensure a small variation in the power supply voltage under a significant current load, the power distribution system should exhibit a small impedance at the terminals of the load within the frequency range of interest, as shown in Fig. 7.4. The impedance of the power distribution system as seen from the terminals of the load circuits is henceforth referred to as the impedance of the power distribution system. In order to ensure correct and reliable operation of an integrated circuit, the impedance of the power distribution system is specified to be lower than a certain upper bound Z_0 in the frequency range from DC to the maximum frequency f_0 [136], as illustrated in Fig. 7.6. Note that the maximum frequency f_0 is determined by the

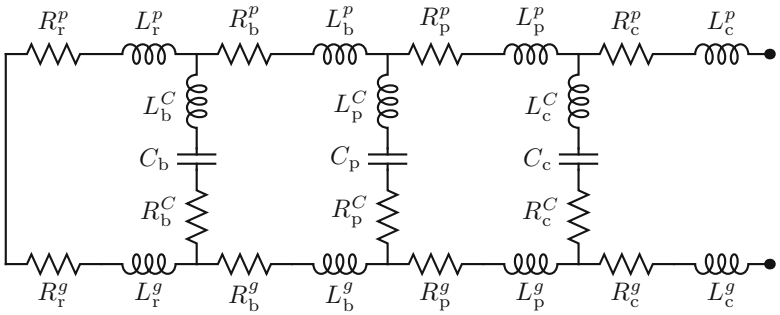


Fig. 7.4 A circuit network representing a power distribution system impedance as seen from the terminals of the power load. The power source at the left side of the network has been replaced with a short circuit. The terminals of the power load are shown at the right side of the network

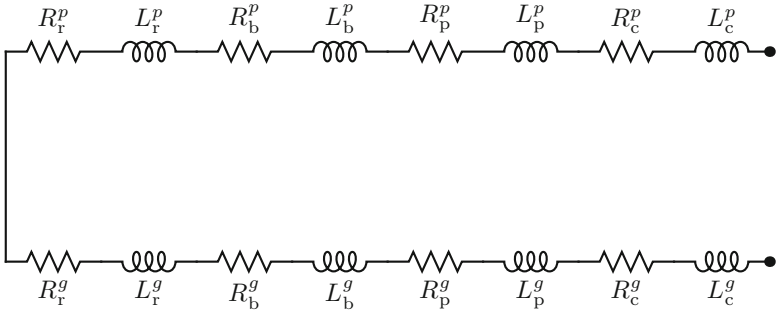


Fig. 7.5 A circuit network representing a power distribution system without decoupling capacitors

switching time of the on-chip signal transients, rather than by the clock frequency f_{clk}. The shortest signal switching time is typically smaller than the clock period by at least an order of magnitude; therefore, the maximum frequency of interest f_0 is considerably higher than the clock frequency f_{clk}.

The objective of designing a power distribution system is to ensure a target output impedance characteristic. It is therefore important to understand how the output impedance of the circuit shown in Fig. 7.4 depends on the impedance of the comprising circuit elements. The impedance characteristics of a power distribution system are analyzed in the following sections. The impedance characteristics of a power distribution system with no capacitors are considered first. The effect of the decoupling capacitors on the impedance characteristics is described in the following sections.

A power distribution system with no decoupling capacitors is shown in Fig. 7.5. The power source and load are connected by interconnect with resistive and inductive parasitic impedances. The magnitude of the impedance of this network is

$$|Z_{\mathrm{tot}}(\omega)| = |R_{\mathrm{tot}} + j\omega L_{\mathrm{tot}}|, \tag{7.2}$$

where R_{tot} and L_{tot} are the total series resistance and inductance of the power distribution system, respectively,

$$R_{\text{tot}} = R^p_{\text{tot}} + R^g_{\text{tot}}, \tag{7.3}$$

$$R^p_{\text{tot}} = R^p_{\text{r}} + R^p_{\text{b}} + R^p_{\text{p}} + R^p_{\text{c}}, \tag{7.4}$$

$$R^g_{\text{tot}} = R^g_{\text{r}} + R^g_{\text{b}} + R^g_{\text{p}} + R^g_{\text{c}}, \tag{7.5}$$

$$L_{\text{tot}} = L^p_{\text{tot}} + L^g_{\text{tot}}, \tag{7.6}$$

$$L^p_{\text{tot}} = L^p_{\text{r}} + L^p_{\text{b}} + L^p_{\text{p}} + L^p_{\text{c}}, \tag{7.7}$$

$$L^g_{\text{tot}} = L^g_{\text{r}} + L^g_{\text{b}} + L^g_{\text{p}} + L^g_{\text{c}}. \tag{7.8}$$

The variation of the impedance with frequency is illustrated in Fig. 7.6. To satisfy the specification at low frequency, the resistance of the power distribution system should be sufficiently low, $R_{\text{tot}} < Z_0$. Above the frequency $f_{L_{\text{tot}}} = \frac{1}{2\pi}\frac{R_{\text{tot}}}{L_{\text{tot}}}$, however, the impedance of the power distribution system is dominated by the inductive reactance $j\omega L_{\text{tot}}$ and increases linearly with frequency, exceeding the target specification at the frequency $f_{\text{max}} = \frac{1}{2\pi}\frac{Z_0}{L_{\text{tot}}}$.

The high frequency impedance should be reduced to satisfy the target specifications. Opportunities for reducing the inductance of the interconnect structures comprising a power system are limited. The feature size of the board and package level interconnect, which largely determines the inductance, depends on the manufacturing technology. Furthermore, the output impedance of the voltage regulator is highly inductive and difficult to reduce.

The high frequency impedance is effectively reduced by placing capacitors across the power and ground conductors. These shunting capacitors terminate the high frequency current loop, permitting the current to bypass the inductive interconnect,

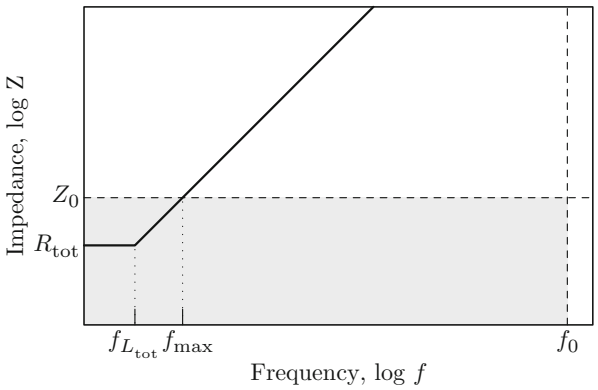


Fig. 7.6 Impedance of the power distribution system with no decoupling capacitors. The *shaded area* denotes the target impedance specifications of the power distribution system

such as the board and package power distribution networks. The high frequency impedance of the system as seen from the load terminals is thereby reduced. The capacitors effectively "decouple" the high impedance parts of the power distribution system from the load at high frequencies. These capacitors are therefore commonly referred to as decoupling capacitors. Several stages of decoupling capacitors are typically used to confine the output impedance within the target specifications.

In the following sections, the impedance characteristics of a power distribution system with several stages of decoupling capacitors are described in several steps. The effect of a single decoupling capacitor on the impedance of a power distribution system is considered in Sect. 7.4. The hierarchical placement of the decoupling capacitors is described in Sect. 7.5. The impedance characteristics near the resonant frequencies are examined in Sect. 7.6.

7.4 A Power Distribution System with a Decoupling Capacitor

The effects of a decoupling capacitor on the output impedance of a power distribution system are the subject of this section. The impedance characteristics of a power distribution system with a decoupling capacitor is described in Sect. 7.4.1. The limitations of using a single stage decoupling scheme are discussed in Sect. 7.4.2.

7.4.1 Impedance Characteristics

A power distribution system with a shunting capacitance is shown in Fig. 7.7. The shunting capacitance can be physically realized with a single capacitor or, alternatively, with a bank of several identical capacitors connected in parallel. Similar to a single capacitor, the impedance of a bank of identical capacitors is

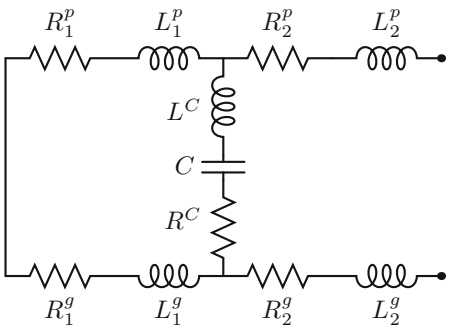


Fig. 7.7 A circuit model of a power distribution network with a decoupling capacitor C

accurately described by a series *RLC* circuit. The location of the capacitors partitions the power network into upstream and downstream sections, with respect to the flow of energy (or power current) from the power source to the load. The upstream section is referred to as stage 1 in Fig. 7.7; the downstream section is referred to as stage 2. Also note that the overall series inductance and resistance of the power distribution network remains the same as determined by (7.3), (7.4), (7.5), (7.6), (7.7), and (7.8): $L_1^p + L_2^p = L_{\text{tot}}^p$, $L_1^g + L_2^g = L_{\text{tot}}^g$, $R_1^p + R_2^p = R_{\text{tot}}^p$, and $R_1^g + R_2^g = R_{\text{tot}}^g$. The impedance of the power distribution network shown in Fig. 7.7 is

$$Z(\omega) = R_2 + j\omega L_2 + (R_1 + j\omega L_1) \left\| \left(R^C + j\left(\omega L^C - \frac{1}{\omega C} \right) \right) \right., \tag{7.9}$$

where $R_1 = R_1^p + R_1^g$ and $L_1 = L_1^p + L_1^g$; analogously, $R_2 = R_2^p + R_2^g$ and $L_2 = L_2^p + L_2^g$.

The impedance characteristics of a power distribution network with a shunting capacitor are illustrated in Fig. 7.8. At low frequencies, the impedance of the capacitor is greater than the impedance of the upstream section. The power current loop extends to the power source, as illustrated by the equivalent circuit shown in Fig. 7.9a. Assuming that the parasitic inductance of the decoupling capacitor is significantly smaller than the upstream inductance, i.e., $L^C < L_1$, the capacitor has a lower impedance than the impedance of the upstream section $R_1 + j\omega L_1$ above the frequency $f_{\text{res}} \approx \frac{1}{2\pi} \frac{1}{\sqrt{L_1 C}}$. Above the frequency f_{res}, the bulk of the power current bypasses the impedance $R_1 + j\omega L_1$ through the capacitor, shrinking the size of the power current loop, as shown in Fig. 7.9b. Note that the decoupling capacitor and the

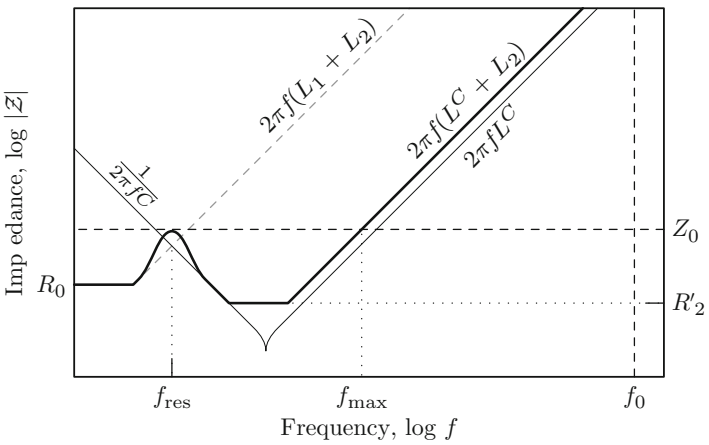


Fig. 7.8 Impedance of the power distribution system with a decoupling capacitor as shown in Fig. 7.7. The impedance of the power distribution system with a decoupling capacitor (the *black line*) exhibits an extended region of low impedance as compared to the impedance of the power distribution system with no decoupling capacitors, shown for comparison with the *dashed gray line*. The impedance of the decoupling capacitor is shown with a *thin solid line*

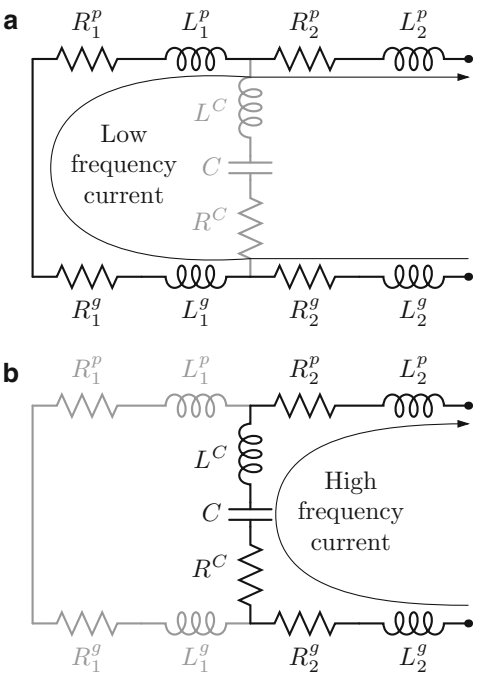


Fig. 7.9 The path of current flow in a power distribution system with a decoupling capacitor; (**a**) at frequencies below $f_{\text{res}} = \frac{1}{2\pi}\frac{1}{\sqrt{L_1 C}}$, the capacitive impedance is relatively high and the current loop extends throughout an entire network, (**b**) at frequencies above f_{res}, the capacitive impedance is lower than the impedance of the upstream section and the bulk of the current bypasses the upstream inductance $L_1^p + L_1^g$ through the decoupling capacitor C

upstream stage form an underdamped LC tank circuit, resulting in a resonant[2] peak in the impedance at frequency f_{res}. The impedance at the resonant frequency f_{res} is *increased* by Q_{tank}, the quality factor of the tank circuit, as compared to the impedance of the power network at the same frequency without the capacitor. Between f_{res} and $f_R = \frac{1}{2\pi}\frac{1}{(R_2+R^C)C}$, the impedance of the power network is dominated by the capacitive reactance $\frac{1}{\omega C}$ and decreases with frequency, reaching $R_2' = R_2 + R^C$. The parasitic resistance of the decoupling capacitor R^C should therefore also be sufficiently low, such that $R_2 + R^C < Z_0$, in order to satisfy the target specification. At frequencies greater than f_R, the network impedance increases as

$$Z(\omega) = R_2 + R^C + j\omega(L_2 + L^C). \tag{7.10}$$

A comparison of (7.10) with (7.2) indicates that placing a decoupling capacitor reduces the high frequency inductance of the power distribution network, as seen by the load, from $L_{\text{tot}} = L_1 + L_2$ to $L_2' = L_2 + L^C$.

The output resistance R_2' and inductance L_2' as seen from the load are henceforth referred to as the *effective* resistance and inductance of the power distribution

[2]The circuit impedance drastically increases near the resonant frequency in parallel (tank) resonant circuits. The parallel resonance is therefore often referred to as the *antiresonance* to distinguish this phenomenon from the series resonance, where the circuit impedance *decreases*. Correspondingly, peaks in the impedance are often referred to as antiresonant.

system to distinguish these quantities from the *overall series* resistance R_{tot} and inductance L_{tot} of the system (which are the effective resistance and inductance at DC). The shunting capacitor "decouples" the upstream portion of the power distribution system from the power current loop at high frequencies, decreasing the maximum impedance of the system. The frequency range where the impedance specification is satisfied is correspondingly increased to $f_{\max} = \frac{1}{2\pi}\frac{Z_0}{L_2'}$.

7.4.2 Limitations of a Single-Tier Decoupling Scheme

It follows from the preceding discussion that the impedance of a power distribution system can be significantly reduced by a single capacitor (or a group of capacitors placed at the same location) over a wide range of frequencies. This solution is henceforth referred to as *single-tier decoupling*. A single-tier decoupling scheme, however, is difficult to realize in practice as this solution imposes stringent requirements on the performance characteristics of the decoupling capacitor, as discussed below.

To confine the network impedance within the specification boundaries at the highest required frequencies, i.e., the $2\pi f_0(L_2 + L^C) < Z_0$, a low inductance path is required between the decoupling capacitance and the load,

$$L_2 + L^C < \frac{1}{2\pi}\frac{Z_0}{f_0}. \tag{7.11}$$

Simultaneously, the capacitance should be sufficiently high to bypass the power current at sufficiently low frequencies, thereby preventing the violation of target specifications above $f_{\max}$ due to a high inductive impedance $j\omega(L_1 + L_2)$,

$$\frac{1}{2\pi f_{\max} C} < 2\pi f_{\max}(L_1 + L_2). \tag{7.12}$$

Condition (7.12) is, however, insufficient. As mentioned above, the decoupling capacitor C and the inductance of the upstream section L_1 form a resonant tank circuit. The impedance reaches the maximum at the resonant frequency $\omega_{\text{res}} \approx \frac{1}{\sqrt{L_1 C}}$, where the inductive impedance of the tank circuit complements the capacitive impedance, $j\omega_{\text{res}} L_1 = -\frac{1}{j\omega_{\max} C}$. Where the quality factor of the tank circuit Q_{tank} is sufficiently larger than unity (i.e., $Q_{\text{tank}} \gtrsim 3$), the impedance of the tank circuit at the resonant frequency ω_{res} is purely resistive and is larger than the characteristic impedance $\sqrt{\frac{L_1}{C}} = \omega_{\text{res}} L_1$ by the quality factor Q_{tank}, $Z_{\text{tank}}(\omega_{\text{res}}) = Q_{\text{tank}}\sqrt{\frac{L_1}{C}}$. The quality factor of the tank circuit is

$$Q_{\text{tank}} = \frac{1}{R_1 + R^C}\sqrt{\frac{L_1}{C}}, \tag{7.13}$$

yielding the magnitude of the peak impedance,

$$Z_{\text{peak}} = \frac{1}{R_1 + R^C} \frac{L_1}{C} . \tag{7.14}$$

To limit the impedance magnitude below the target impedance Z_0, the decoupling capacitance should satisfy

$$C > \frac{L_1}{Z_0(R_1 + R^C)} . \tag{7.15}$$

A larger upstream inductance L_1 (i.e., the inductance that is decoupled by the capacitor) therefore requires a larger decoupling capacitance C to maintain the target network impedance Z_0. The accuracy of the simplifications used in the derivation of (7.14) decreases as the factor Q_{tank} approaches unity. A detailed treatment of the case where $Q \approx 1$ is presented in Sect. 7.6.

These impedance characteristics have an intuitive physical interpretation. From a physical perspective, the decoupling capacitor serves as an intermediate storage of charge and energy. To be effective, such an energy storage device should possess two qualities. First, the device should have a high capacity to store a sufficient amount of energy. This requirement is expressed in terms of the impedance characteristics as the minimum capacitance constraint (7.15). Second, to supply sufficient power at high frequencies, the device should be able to release and accumulate energy at a sufficient rate. This quality is expressed as the maximum inductance constraint (7.11).

Constructing a device with both high energy capacity and high power capability is, however, challenging. The conditions of low inductance (7.11) and high capacitance (7.15) cannot be simultaneously satisfied in a cost effective manner. In practice, these conditions are contradictory. The physical realization of a large decoupling capacitance as determined by (7.15) requires the use of discrete capacitors with a large nominal capacity, which, consequently, have a large form factor. The large physical dimensions of the capacitors have two implications. The parasitic series inductance of a physically large capacitor L^C is relatively high due to the increased area of the current loop within the capacitors, contradicting requirement (7.11). Furthermore, the large physical size of the capacitors prevents placing the capacitors sufficiently close to the power load. Greater physical separation increases the inductance L_2 of the current path from the capacitors to the load, also contradicting requirement (7.11). The available component technology therefore imposes a tradeoff between the high capacity and low parasitic inductance of a capacitor.

Gate switching times of a few tens of picoseconds are common in contemporary integrated circuits, creating high transients in the power load current. Only on-chip decoupling capacitors have a sufficiently low parasitic inductance to maintain a low impedance power distribution system at high frequencies. Placing a sufficiently large on-chip decoupling capacitance, as determined by (7.15), requires a die area

many times greater than the area of the load circuit. Therefore, while technically feasible, the single-tier decoupling solution is prohibitively expensive. A more efficient approach to the problem, a multi-stage hierarchical placement of decoupling capacitors, is described in the following section.

7.5 Hierarchical Placement of Decoupling Capacitance

Low impedance, high frequency power distribution systems are realized in a cost effective way by using a hierarchy of decoupling capacitors. The capacitors are placed in several stages: on the board, package, and circuit die. The impedance characteristics of a power distribution system with several stages of decoupling capacitors are described in this section. The evolution of the system output impedance is described as the decoupling stages are consecutively placed across the network. Arranging the decoupling capacitors in several stages eliminates the need to satisfy both the high capacitance and low inductance requirements, as expressed by (7.11) and (7.15), in the same decoupling capacitor stage.

7.5.1 Board Decoupling Capacitors

Consider a power distribution system with decoupling capacitors placed on the board, as shown in Fig. 7.10. The circuit is analogous to the network shown in Fig. 7.7. Similar to the single-tier decoupling scheme, the board decoupling capacitance should satisfy the following condition in order to meet the target specification at low frequencies,

$$C_{\mathrm{b}} > \frac{L_{\mathrm{r}}}{Z_0(R_{\mathrm{r}} + R_{\mathrm{b}}^C)} \,. \tag{7.16}$$

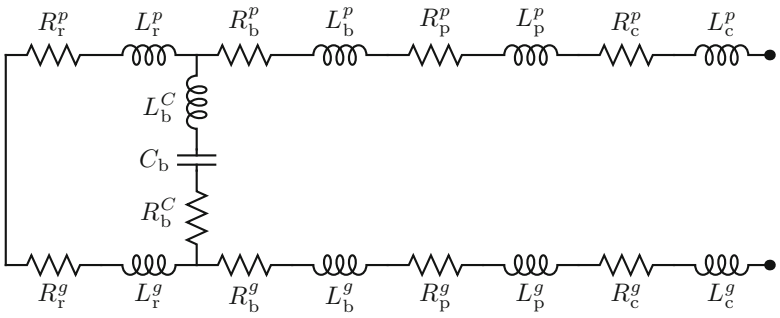


Fig. 7.10 A circuit model of a power distribution system with a board decoupling capacitance

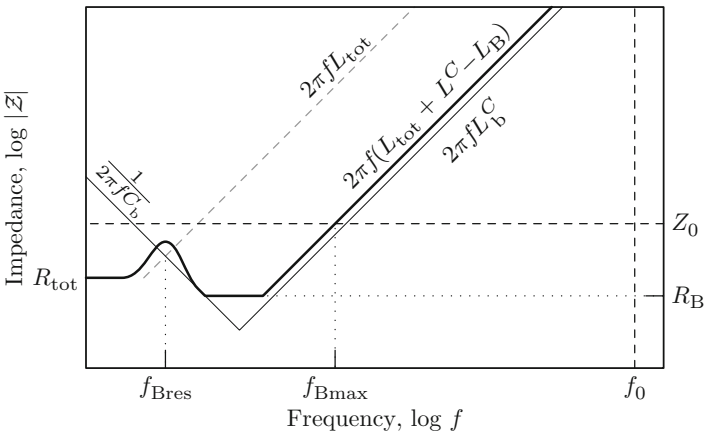


Fig. 7.11 Impedance of the power distribution system with the board decoupling capacitance shown in Fig. 7.7. The impedance characteristic is shown with a *black line*. The impedance of the power distribution system with no decoupling capacitors is shown, for comparison, with a *gray dashed line*. The impedance of the board decoupling capacitance is shown with a *thin solid line*

$L_{\rm r} = L^p_{\rm r} + L^g_{\rm r}$ and $R_{\rm r} = R^p_{\rm r} + R^g_{\rm r}$ are the resistance and inductance, respectively, of the network upstream of the board capacitance, including the output inductance of the voltage regulator and the impedance of the current path between the voltage regulator and the board capacitors, as defined by (7.1). At frequencies greater than $f^{\rm B}_{\rm res}$, the power current flows through the board decoupling capacitors, bypassing the voltage regulator. Condition (7.11), however, is not satisfied due to the relatively high inductance $L_b + L_p + L_c$ resulting from the large separation from the load and the high parasitic series inductance of the board capacitors $L^C_{\rm b}$. Above the frequency $f_{R_{\rm B}} = \frac{1}{2\pi}\frac{1}{R_{\rm B}C_{\rm b}}$, the impedance of the power distribution system is

$$Z_{\rm B} = R_{\rm B} + j\omega L_{\rm B}, \tag{7.17}$$

where $L_{\rm B} = L^C_{\rm b} + L_{\rm b} + L_{\rm p} + L_{\rm c}$ and $R_{\rm B} = R^C_{\rm b} + R_{\rm b} + R_{\rm p} + R_{\rm c}$. Although the high frequency inductance of the network is reduced from $L_{\rm tot}$ to $L_{\rm B}$, the impedance exceeds the target magnitude Z_0 above the frequency $f^{\rm B}_{\rm max} = \frac{1}{2\pi}\frac{Z_0}{L_{\rm B}}$, as shown in Fig. 7.11.

7.5.2 Package Decoupling Capacitors

The excessive high frequency inductance $L_{\rm B}$ of the power distribution system with board decoupling capacitors is further reduced by placing an additional decoupling capacitance physically closer to the power load, i.e., on the integrated circuit package. The circuit model of a power distribution system with board and package decoupling capacitors is shown in Fig. 7.12. The impedance of the network with board and package decoupling capacitors is illustrated in Fig. 7.13. The package

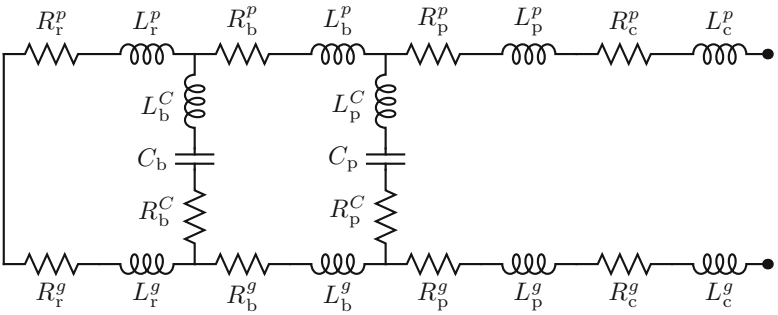


Fig. 7.12 A circuit model of a power distribution system with board and package decoupling capacitances

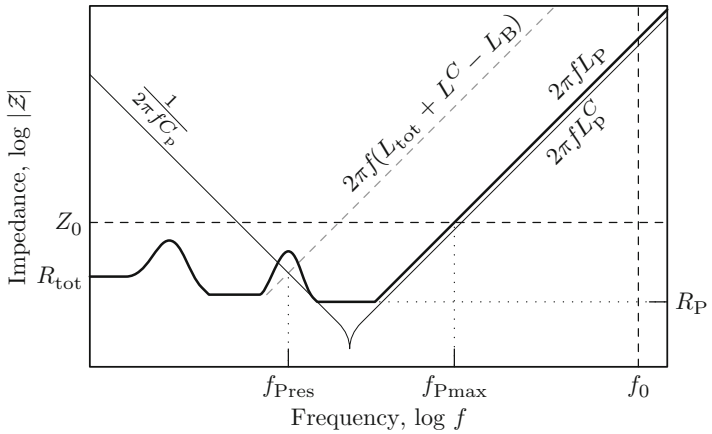


Fig. 7.13 Impedance of the power distribution system with board and package decoupling capacitances as shown in Fig. 7.7. The impedance characteristic is shown with a *black line*. The impedance of the power distribution system with only the board decoupling capacitance is shown with a *dashed gray line* for comparison. The impedance of the package decoupling capacitance is shown with a *thin solid line*

decoupling capacitance decreases the network inductance L_B in a fashion similar to the board decoupling capacitance. A significant difference is that for the package decoupling capacitance the capacity requirement (7.15) is relaxed to

$$C_p > \frac{L'_b}{Z_0(R'_b + R^C_p)}\,, \tag{7.18}$$

where $L'_b = L_b + L^C_b$ and $R'_b = R_b + R^C_b$ are the effective inductance and resistance in parallel with the package decoupling capacitance. The upstream inductance L'_b as seen by the package capacitance at high frequencies (i.e., $f > f_{R_B}$) is significantly lower than the upstream inductance L_r as seen by the board capacitors. The package capacitance requirement (7.18) is therefore significantly less stringent than the board capacitance requirement (7.16).

The lower capacitance requirement is satisfied by using several small form factor capacitors mounted onto a package. As shown in Fig. 7.13, the effect of the package decoupling capacitors on the system impedance is analogous to the effect of the board capacitors, but occurs at a higher frequency. With package decoupling capacitors, the impedance of the power distribution system above a frequency $f_{R_P} = \frac{1}{2\pi}\frac{1}{R_P C_p}$ becomes

$$Z_P = R_P + j\omega L_P, \tag{7.19}$$

where $L_P = L_p^C + L_p + L_c$ and $R_P = R_p^C + R_p + R_c$. Including a package capacitance therefore lowers the high frequency inductance from L_B to L_P. The reduced inductance L_P, however, is not sufficiently low to satisfy (7.11) in high speed circuits. The impedance of the power system exceeds the target magnitude Z_0 at frequencies above $f_{\max}^P = \frac{1}{2\pi}\frac{Z_0}{L_P}$.

7.5.3 On-chip Decoupling Capacitors

On-chip decoupling capacitors are added to the power distribution system in order to extend the frequency range of the low impedance characteristics to f_0. A circuit model characterizing the impedance of a power distribution system with board, package, and on-chip decoupling capacitors is shown in Fig. 7.14. The impedance characteristics of a power distribution system with all three types of decoupling capacitors are illustrated in Fig. 7.15. To ensure that the network impedance does not exceed Z_0 due to the inductance L_P of the current loop terminated by the package capacitance, the on-chip decoupling capacitance should satisfy

$$C_c > \frac{L_p'}{Z_0(R_p' + R_c^C)}, \tag{7.20}$$

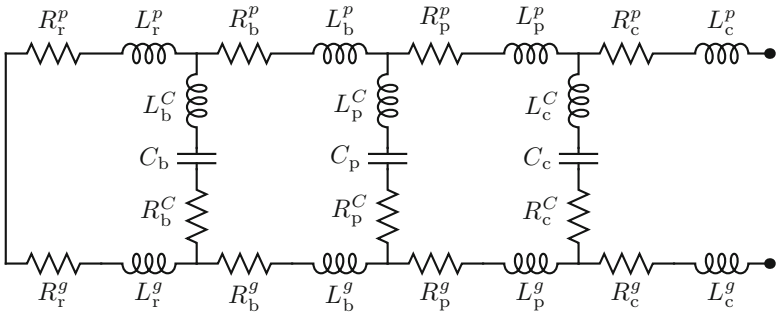


Fig. 7.14 A circuit network characterizing the impedance of a power distribution system with board, package, and on-chip decoupling capacitances

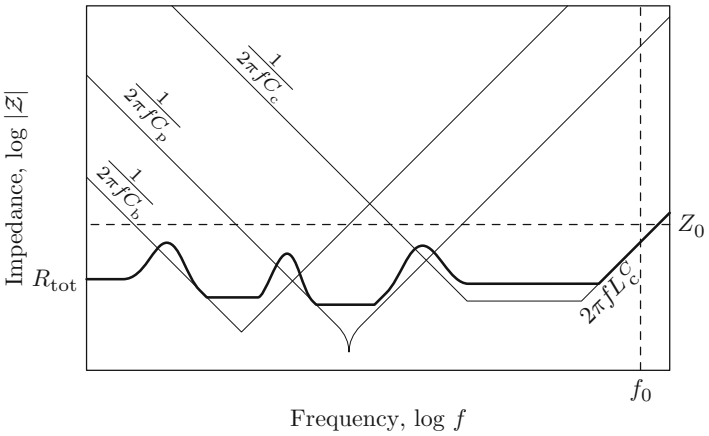


Fig. 7.15 Impedance of a power distribution system with board, package, and on-chip decoupling capacitances, as shown in Fig. 7.7. The impedance is shown with a *black line*. The impedance characteristic is within the target specification outlined by the *dashed lines*. The impedance characteristics of the decoupling capacitances are shown with thin *solid lines*

where $L_p' = L_p + L_p^C$ is the effective inductance in parallel with the package decoupling capacitance. The capacity requirement for the on-chip capacitance is further reduced as compared to the package capacitance due to lower effective inductances, $L_p < L_b$ and $L_p' < L_b'$.

7.5.4 Advantages of Hierarchical Decoupling

Hierarchical placement of decoupling capacitors exploits the tradeoff between the capacity and the parasitic series inductance of a capacitor to achieve an economically effective solution. The total decoupling capacitance of a hierarchical scheme $C_b + C_p + C_c$ is larger than the total decoupling capacitance of a single-tier solution C as determined by (7.15). The advantage of the hierarchical scheme is that the inductance limit (7.11) is imposed only on the final stage of decoupling capacitors which constitute a small fraction of the total decoupling capacitance. The constraints on the physical dimensions and parasitic impedance of the capacitors in the remaining stages are dramatically reduced, permitting the use of cost efficient electrolytic and ceramic capacitors.

The decoupling capacitance at each tier is effective within a limited frequency range, as determined by the capacitance and inductance of the capacitor. The range of effectiveness of the board, package, and on-chip decoupling capacitances overlaps each other, as shown in Fig. 7.15, spanning an entire frequency region of interest from DC to f_0 (the maximum operating frequency).

As described in Sect. 7.4, a decoupling capacitor lowers the high frequency impedance by allowing the power current to bypass the inductive interconnect upstream of the capacitor. In a power distribution system with several stages of

decoupling capacitors, the inductive interconnect is excluded from the power current loop in several steps, as illustrated in Fig. 7.16. At near-DC frequencies, the power current loop extends through an entire power supply system to the power source, as shown in Fig. 7.16a. As frequencies increase, the power current is shunted by the board decoupling capacitors, shortening the current loop as shown in Fig. 7.16b. At higher frequencies, the package capacitors terminate the current loop, further reducing the loop size, as depicted in Fig. 7.16c. Finally, at the highest frequencies, the power current is terminated by the on-chip capacitors, as illustrated in Fig. 7.16d. The higher the frequency, the shorter the distance between the shunting capacitor and the load and the smaller the size of the current loop. At transitional frequencies where the bulk of the current is shifted from one decoupling stage to the next, both decoupling stages carry significant current, giving rise to resonant behavior.

Each stage of decoupling capacitors determines the impedance characteristics over a limited range of frequencies, where the bulk of the power current flows through the stage. Outside of this frequency range, the stage capacitors have an insignificant influence on the impedance characteristics of the system. The lower frequency impedance characteristics are therefore determined by the upstream stages of decoupling capacitors, while the impedance characteristics at the higher frequencies are determined by the capacitors closer to the load. For example, the board capacitors determine the low frequency impedance characteristics but do not affect the high frequency response of the system, which is determined by the on-chip capacitors.

7.6 Resonance in Power Distribution Networks

Using decoupling capacitors is a powerful technique to reduce the impedance of a power distribution system within a significant range of frequencies, as discussed in preceding sections. A decoupling capacitor, however, *increases* the network impedance in the vicinity of the resonant frequency f_{res}. Controlling the impedance behavior near the resonant frequency is therefore essential to effectively use decoupling capacitors. A relatively high quality factor, $Q \gtrsim 3$, is assumed in the expression for the peak impedance of the parallel resonant circuit described in Sect. 7.4. Maintaining low impedance characteristics in power distribution systems necessitates minimizing the quality factor of all of the resonant modes; relatively low values of the quality factor are therefore common in power distribution systems. The impedance characteristics of parallel resonant circuits with a low quality factor, $Q \approx 1$, are the focus of this section.

The purpose of the decoupling capacitance is to exclude the high impedance upstream network from the load current path, as discussed in Sect. 7.4. The decoupling capacitor and the inductive upstream network form a parallel resonant circuit with a significant resistance in both the inductive and capacitive branches. An equivalent circuit diagram is shown in Fig. 7.17.

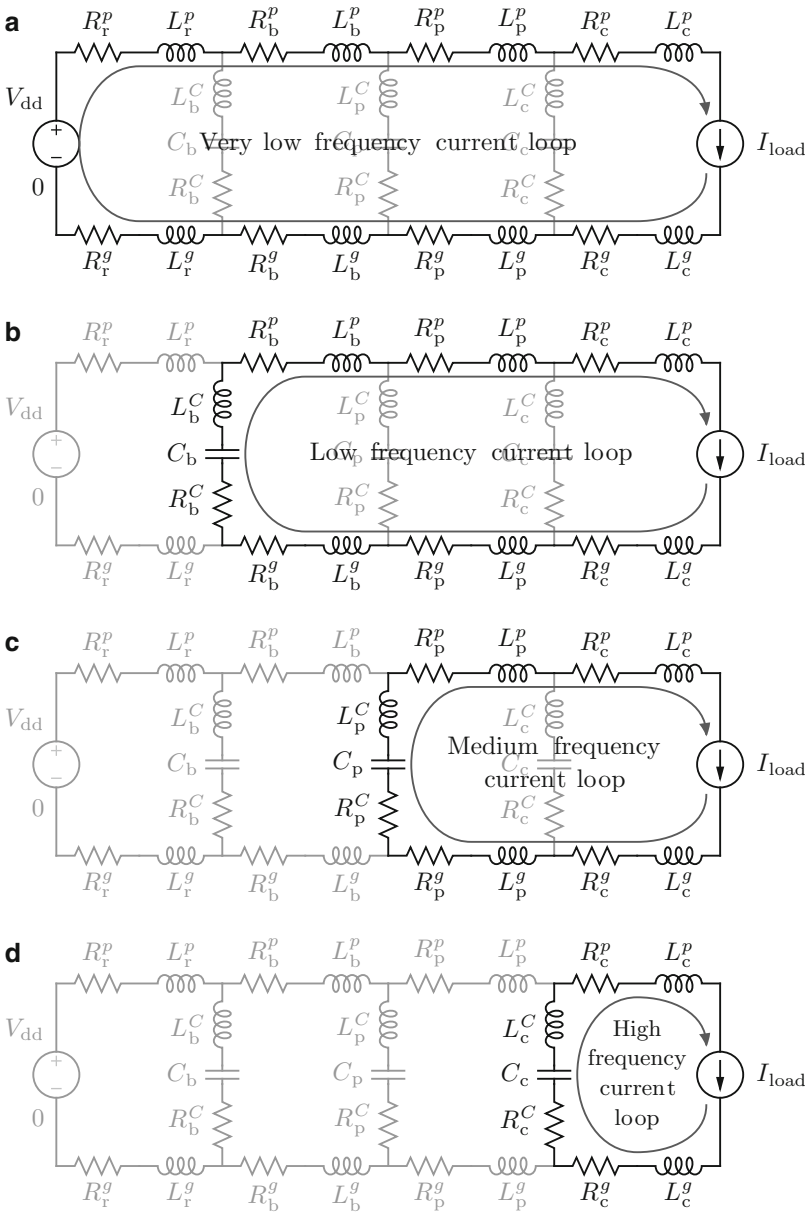


Fig. 7.16 Variation of the power current path with frequency. (**a**) At low frequencies, the current loop extends through an entire system to the power source; as frequencies increase, the current loop is terminated by the board, package, and on-chip decoupling capacitors, as shown in, respectively, (**b**), (**c**), and (**d**)

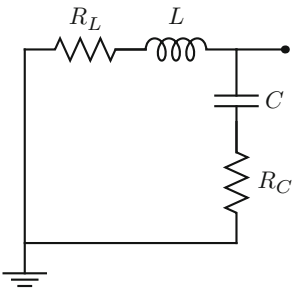


Fig. 7.17 A parallel resonant circuit with a significant parasitic resistance in both branches. L and R_L represent the effective impedance of the upstream network; C and R_C represent the impedance characteristics of the decoupling capacitor

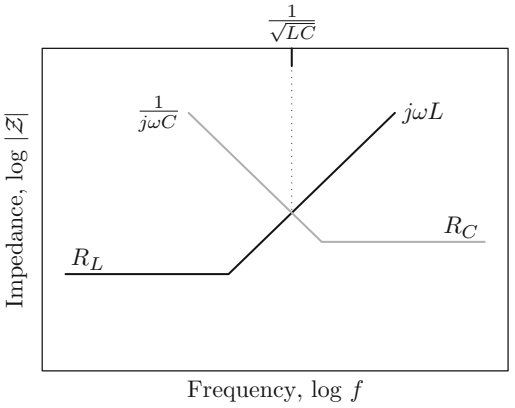


Fig. 7.18 An asymptotic plot of the impedance magnitude of the inductive and capacitive branches forming the tank circuit shown in Fig. 7.17

Near the resonant frequency, the impedance of the upstream network exhibits an inductive-resistive nature, $R_L + j\omega L$. The impedance of the decoupling capacitor is well described near the resonant frequency as $R_C + j\omega C$. The effective series inductance of the capacitor is significantly lower than the upstream network inductance (otherwise the capacitor would be ineffective, as discussed in Sect. 7.4) and can be typically neglected near the resonant frequency. The impedance characteristics of the capacitor and upstream network are schematically illustrated in Fig. 7.18.

The impedance of the tank circuit shown in Fig. 7.17 is

$$\begin{aligned} Z_{\text{tank}} &= (R_L + j\omega L) \parallel \left(R_C + \frac{1}{j\omega C}\right) \\ &= \frac{(R_L + j\omega L)\left(R_C + \frac{1}{j\omega C}\right)}{(R_L + j\omega L) + \left(R_C + \frac{1}{j\omega C}\right)} \\ &= R_C + \frac{s(L - CR_C^2) + (R_L - R_C)}{s^2 LC + sC(R_C + R_L) + 1}\,. \end{aligned} \tag{7.21}$$

The poles of the system described by (7.21) are determined by the characteristic equation,

$$s^2 + 2\zeta\omega_n + \omega_n^2 = 0, \tag{7.22}$$

$$\text{where} \qquad \omega_n = \frac{1}{\sqrt{LC}} \tag{7.23}$$

$$\text{and} \qquad \zeta = \frac{R_L + R_C}{2}\sqrt{\frac{C}{L}}\,. \tag{7.24}$$

Note that the magnitude of the circuit impedance varies from R_L at low frequencies ($\omega \ll \frac{1}{\sqrt{LC}}$) to R_C at high frequencies ($\omega \gg \frac{1}{\sqrt{LC}}$). It is desirable that the impedance variations near the resonant frequency do not significantly increase the maximum impedance of the network, i.e., the variations do not exceed or only marginally exceed $\max(R_L, R_C)$, the maximum resistive impedance. Different constraints can be imposed on the parameters of the tank circuit to ensure this behavior [138]. Several cases of these constraints are described below.

CASE I: The tank circuit has a monotonic (i.e., overshoot- and oscillation-free) response to a step current excitation if the circuit damping is critical or greater: $\zeta \leqslant 1$. Using (7.24), sufficient damping is achieved when

$$R_L + R_C \geqslant 2R_0 = 2\sqrt{\frac{L}{C}}\,, \tag{7.25}$$

where $R_0 = \sqrt{\frac{L}{C}}$ is the characteristic impedance of the tank circuit. Therefore, if the network impedance $R_L + j\omega L$ is to be reduced to the target impedance Z_0 at high frequencies (meaning that $R_C = Z_0$), the required decoupling capacitance is

$$C \geqslant \frac{L}{Z_0^2}\,\frac{4}{(1 + R_L/Z_0)^2}\,. \tag{7.26}$$

CASE II: Alternatively, the monotonicity of the impedance variation with frequency can be ensured. It can be demonstrated [138] that the magnitude of the impedance of the tank circuit varies monotonically from R_L at low frequencies, $\omega \ll \omega_n$, to R_C at high frequencies, $\omega \gg \omega_n$, if

$$R_L R_C \geqslant R_0^2 = \frac{L}{C}\,. \tag{7.27}$$

The required decoupling capacitance in this case is

$$C \geqslant \frac{L}{Z_0^2}\,\frac{1}{R_L/Z_0}. \tag{7.28}$$

Note that condition (7.27) is stronger than condition (7.25), i.e., satisfaction of (7.27) ensures satisfaction of (7.25). Correspondingly, the capacitance requirement described by (7.28) is always greater or equal to the capacitance requirement described by (7.26).

CASE III: The capacitance requirement determined in Case II, as described by (7.28), increases rapidly when either R_L or R_C is small. In this situation, condition (7.28) is overly conservative and restrictive. The capacitive requirement is significantly relaxed if a small peak in the impedance characteristics is allowed. To restrict the magnitude of the impedance peak to approximately 1 % (in terms of the resistive baseline), the following condition must be satisfied [138],

$$(b_2 r^2 + b_1 r + b_0)\, R_{\max}^2 \geqslant R_0^2 = \frac{L}{C}\,, \tag{7.29}$$

where $R_{\max} = \max(R_L, R_C)$, $r = \min\left(\frac{R_L}{R_C}, \frac{R_C}{R_L}\right)$, $b_0 = 0.4831$, $b_1 = 0.4907$, and $b_2 = -0.0139$. The decoupling capacitance requirement is relaxed in this case to

$$C \geqslant \frac{L}{Z_0^2}\,\frac{1}{b_2 r^2 + b_1 r + b_0}\,. \tag{7.30}$$

Case III is the least constrained as compared to Cases I and II.

The design space of the resistances R_L and R_C as determined by (7.25), (7.27), and (7.29) is illustrated in Fig. 7.19 in terms of the circuit characteristic impedance $R_0 = \sqrt{\frac{L}{C}}$. In the unshaded region near the origin, the impedance exhibits significant resonant behavior. The rest of the space is partitioned into three regions. The more constrained regions are shaded in progressively darker tones. In the lightest shaded region, only the Case III condition is satisfied. The Case I and III conditions are satisfied in the adjacent darker region. All three conditions are satisfied in the darkest region.

Alternatively, conditions (7.25), (7.27), and (7.29) can be used to determine the decoupling capacitance requirement as a function of the circuit inductance L and

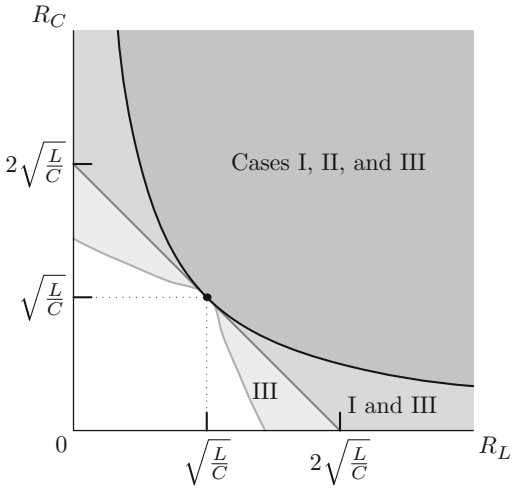


Fig. 7.19 Design space of the resistances R_L and R_C for the tank circuit shown in Fig. 7.17. All three design constraints, as determined by (7.25), (7.27), and (7.29), are satisfied in the *darkest area*. The constraints of Cases I and III are satisfied in the *medium gray area*, while only the Case III constraint is satisfied in the *lightly shaded area*

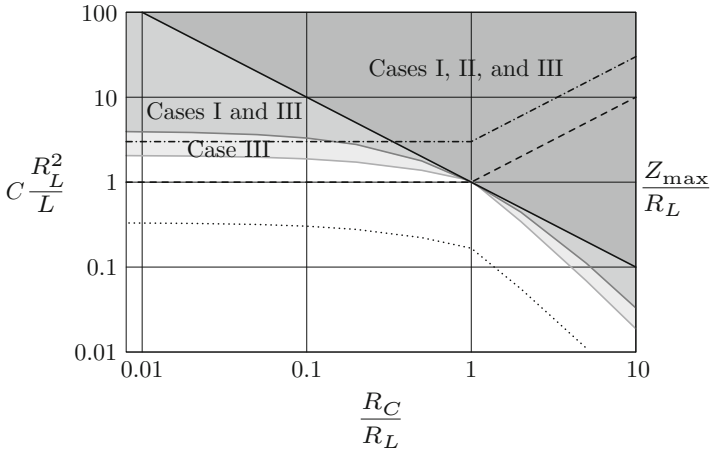


Fig. 7.20 Decoupling capacitance requirements as determined by Cases I, II, and III, respectively, (7.26), (7.28), and (7.30). Case II requires the greatest amount of capacitance while Case III requires the lowest capacitance

resistances R_L and R_C. The normalized capacitance $C\frac{R_L^2}{L}$ versus the normalized resistance R_C/R_L is shown in Fig. 7.20. The shading is analogous to the scheme used in Fig. 7.19. The darkest region boundary is determined by condition (7.28), the boundary of the intermediate region is determined by condition (7.26), and the boundary of the lightest region is determined by condition (7.30). Similar to the resistance design space illustrated in Fig. 7.19, Case II is the most restrictive, while Case III is the least restrictive. The capacitance requirements decrease in all three cases as the normalized resistance $\frac{R_C}{R_L}$ increases.

The normalized maximum impedance of the network $Z_0 = \frac{R_{\max}}{R_L} = \frac{\max(R_L,R_C)}{R_L}$ is also shown in Fig. 7.20 by a dashed line. For $\frac{R_C}{R_L} \leqslant 1$, the maximum network impedance is $Z_0 = R_L \left(\frac{Z_0}{R_L} = 1\right)$. Increasing resistance R_C to R_L therefore reduces the required decoupling capacitance without increasing the maximum impedance of the network. Increasing R_C beyond R_L further reduces the capacitance requirement, but at a cost of increasing the maximum impedance. For $\frac{R_C}{R_L} \geqslant 1$, the maximum impedance becomes $Z_0 = R_C \left(\frac{Z_0}{R_L} = \frac{R_C}{R_L}\right)$.

The requirement for the decoupling capacitance can be further reduced if the resonant impedance is designed to significantly exceed $R_{\max}$. For comparison, the capacitance requirement as determined by (7.15) for the case $Z_0 = 3R_{\max}$ ($Q \approx 3$) is also shown in Fig. 7.20 by the dotted line. This significantly lower requirement, as compared to Cases I, II, and III, however, comes at a cost of a higher maximum impedance, as shown by the dashed line with dots.

7.7 Full Impedance Compensation

A special case of both (7.26) and (7.28) is the condition where

$$R_L = R_C = R_0 = \sqrt{\frac{L}{C}}, \tag{7.31}$$

as shown in Figs. 7.19 and 7.20. Under condition (7.31), the zeros of the tank circuit impedance (7.21) cancel the poles and the impedance becomes purely resistive and independent of frequency,

$$Z_{\text{tank}}(\omega) = R_0. \tag{7.32}$$

This specific case is henceforth referred to as fully compensated impedance. As shown in Fig. 7.20, choosing the capacitive branch resistance R_C in accordance with (7.31) results in the lowest decoupling capacitance requirements for a given maximum circuit impedance.

The impedance of the inductive and capacitive branches of the tank circuit under condition (7.31) is illustrated in Fig. 7.21. The condition of full compensation (7.31) is equivalent to two conditions: $R_L = R_C$, i.e., the impedance at the lower frequencies is matched to the impedance at the higher frequencies, and $\frac{L}{R_L} = R_C C$, i.e., the time constants of the inductor and capacitor currents are matched, as shown in Fig. 7.21.

A constant, purely resistive impedance is achieved across the entire frequency range of interest, as illustrated in Fig. 7.22, if each decoupling stage is fully compensated. The resistance and capacitance of the decoupling capacitors in a fully compensated system are completely determined by the impedance characteristics of the power and ground interconnect and the location of the capacitors. The overall capacitance and resistance of the board decoupling capacitors are, according to (7.31),

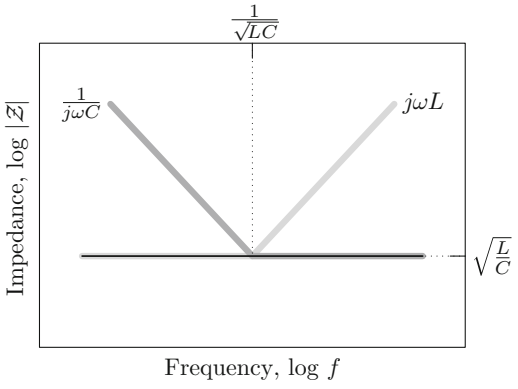


Fig. 7.21 Asymptotic impedance of the inductive (*light gray line*) and capacitive (*dark gray line*) branches of a tank circuit under the condition of full compensation (7.31). The overall impedance of the tank circuit is purely resistive and does not vary with frequency, as shown by the *black line*

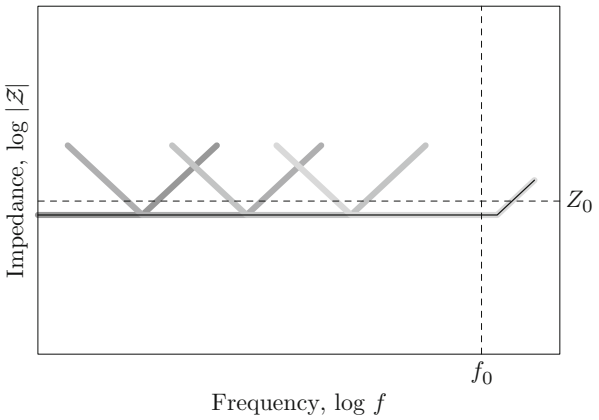


Fig. 7.22 Impedance diagram of a power distribution system with full compensation at each decoupling stage. The impedance of the decoupling stages is shown in *shades of gray*. The resulting system impedance is purely resistive and constant over the frequency range of interest, as shown by the *black line*

$$R_{\mathrm{b}}^{C} = R_{\mathrm{r}}, \tag{7.33}$$

$$C_{\mathrm{b}} = \frac{L_{\mathrm{r}}}{R_{\mathrm{r}}^2}\,. \tag{7.34}$$

The inductance of the upstream part of the power distribution system as seen at the terminals of the package capacitors is $L_{\mathrm{b}}^{C} + L_{\mathrm{b}}$. The capacitance and resistance of the package capacitors are

$$R_{\mathrm{p}}^{C} = R_{\mathrm{b}}^{C} + R_{\mathrm{b}} = R_{\mathrm{r}} + R_{\mathrm{b}}, \tag{7.35}$$

$$C_{\mathrm{p}} = \frac{L_{\mathrm{b}}^{C} + L_{\mathrm{b}}}{(R_{\mathrm{r}} + R_{\mathrm{b}})^2}\,. \tag{7.36}$$

Analogously, the resistance and capacitance of the on-chip decoupling capacitors are

$$R_{\mathrm{c}}^{C} = R_{\mathrm{p}}^{C} + R_{\mathrm{p}} = R_{\mathrm{r}} + R_{\mathrm{b}} + R_{\mathrm{p}}, \tag{7.37}$$

$$C_{\mathrm{c}} = \frac{L_{\mathrm{p}}^{C} + L_{\mathrm{p}}}{(R_{\mathrm{r}} + R_{\mathrm{b}} + R_{\mathrm{p}})^2}\,. \tag{7.38}$$

Note that the effective series resistance of the decoupling capacitors increases toward the load, $R_{\mathrm{b}}^{C} < R_{\mathrm{p}}^{C} < R_{\mathrm{c}}^{C}$, such that the resistance of the load current loop, no matter which decoupling capacitor terminates this loop, remains the same as the total resistance R_{tot} of the system without decoupling capacitors,

$$\begin{aligned} R_{\mathrm{b}}^{C} + R_{\mathrm{b}} + R_{\mathrm{p}} + R_{\mathrm{c}} = \\ R_{\mathrm{p}}^{C} + R_{\mathrm{p}} + R_{\mathrm{c}} = \\ R_{\mathrm{c}}^{C} + R_{\mathrm{c}} = R_{\mathrm{tot}}. \end{aligned} \tag{7.39}$$

If the resistance of the decoupling capacitors is reduced below the values determined by (7.33), (7.35), and (7.37), the resonance behavior degrades the impedance characteristics of the power distribution system.

7.8 Case Study

A case study of a power distribution system for a high performance integrated circuit is presented in this section. This case study is intended to provide a practical perspective to the analytic description of the impedance characteristics of the power distribution systems developed in the previous section.

Consider a microprocessor consuming 60 W from a 1.2 V power supply. The average power current of the microprocessor is 50 A. The maximum frequency of interest is assumed to be 20 GHz, corresponding to the shortest gate switching time of approximately 17 ps. The objective is to limit the power supply variation to approximately 8 % of the nominal 1.2 V power supply level under a 50 A load. This objective results in a target impedance specification of $0.08 \times 1.2\,\mathrm{V}/50\,\mathrm{A} = 2\ \mathrm{m}\Omega$ over the frequency range from DC to 20 GHz. Three stages of decoupling capacitors are assumed. The parameters of the initial version of the power distribution system are displayed in Table 7.1.

The impedance characteristics of the overall power distribution system are shown in Fig. 7.23. The resonant modes of this system are significantly underdamped. The resulting impedance peaks exceed the target impedance specifications. The impedance characteristics improve significantly if the damping of the resonant mode is increased to the near-critical level. The greater damping can be achieved by only manipulating the effective series resistance of the decoupling capacitors. The impedance characteristics approach the target specifications, as shown in Fig. 7.23, as the resistance of the board, package, and on-chip capacitors is increased from 0.1 , 0.2 , and 0.4 mΩ to, respectively, 1 , 1.5 , and 1.5 mΩ. Further improvements in the system impedance characteristics require increasing the decoupling capacitance. Fully compensating each decoupling stage renders the impedance purely resistive. The magnitude of the fully compensated impedance equals the total resistance of the power distribution system, 1.45 mΩ (1 + 0.3 + 0.1 + 0.05), as described in Sect. 7.7.

The resonant frequencies of the case study are representative of typical resonant frequencies encountered in power distribution systems. The board decoupling stage resonates in the kilohertz range of frequencies, while the frequency of the resonant peak due to the package decoupling stage is in the low megahertz

Table 7.1 Parameters of a case study power distribution system

Circuit parameter	Initial system	Near-critical damping	Fully compensated
R_r	1 mΩ	1 mΩ	1 mΩ
L_r	10 nH	10 nH	10 nH
C_b	5 mF	5 mF	10 mF
R_b^C	0.1 mΩ	1 mΩ	1 mΩ
L_b^C	0.3 nH	0.3 nH	0.3 nH
R_b	0.3 mΩ	0.3 mΩ	0.3 mΩ
L_b	0.2 nH	0.2 nH	0.2 nH
C_p	250 μF	250 μF	296 μF
R_p^C	0.2 mΩ	1.5 mΩ	1.3 mΩ
L_p^C	1 pH	1 pH	1 pH
R_p	0.1 mΩ	0.1 mΩ	0.1 mΩ
L_p	1 pH	1 pH	1 pH
C_c	500 nF	500 nF	1020 nF
R_c^C	0.4 mΩ	1.5 mΩ	1.4 mΩ
L_c^C	1 fH	1 fH	1 fH
R_c	0.05 mΩ	0.05 mΩ	0.05 mΩ
L_c	4 fH	4 fH	4 fH

frequencies [137, 139]. The frequency of the chip capacitor resonance, often referred to as the chip-package resonance as this effect includes the inductance of the package interconnect, typically varies from tens of megahertz to low hundreds of megahertz [134, 140, 141].

7.9 Design Considerations

The power current requirements of high performance integrated circuits are rapidly growing with technology scaling. These requirements necessitate a significant reduction in the output impedance of the power distribution system over a wider range of frequencies for new generations of circuits. Design approaches to improve the impedance characteristics of next generation power distribution systems are discussed in this section.

As described in previous sections, the inductance of the power distribution interconnect structures is efficiently excluded from the path of high frequency current by the hierarchical placement of decoupling capacitors. The inductance of the power and ground interconnect, however, greatly affects the decoupling capacitance requirements. As indicated by (7.16), (7.18), and (7.20), a lower inductance current path connecting the individual stages of decoupling capacitors relaxes the requirements placed on the decoupling capacitance at each stage of the power distribution network. Lowering the interconnect inductance decreases both

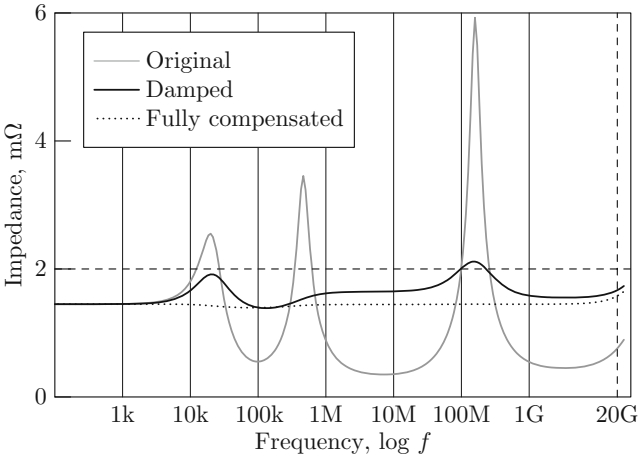


Fig. 7.23 Impedance characteristics of a power distribution system case study. The initial system is significantly underdamped. The resonant impedance peaks exceed the target impedance specification, as shown with a *gray line*. As the effective series resistance of the decoupling capacitors is increased, the damping of the resonant modes also increases, reducing the magnitude of the peak impedance. The damped system impedance effectively satisfies the target specifications, as shown with a *solid black line*. The impedance characteristics can be further improved if the impedance of each decoupling stage is fully compensated, as shown by the *dotted line*

the overall cost of the decoupling capacitors and the effective impedance of the power distribution system. It is therefore desirable to reduce the inductance of the power distribution interconnect.

As indicated by (7.18) and (7.20), the lower bound on the capacitance at each decoupling stage is determined by the effective inductance of the upstream current path $L^C + L^{\text{int}}$, which consists of the inductance of the previous stage decoupling capacitors L^C and the inductance of the interconnect connecting the two stages L^{int}. The impedance characteristics are thereby improved by lowering both the effective series inductance of the decoupling capacitors, as described in Sect. 7.9.1, and the interconnect inductance, as described in Sect. 7.9.2.

7.9.1 Inductance of the Decoupling Capacitors

The series inductance of the decoupling capacitance can be decreased by using a larger number of lower capacity capacitors to realize a specific decoupling capacitance rather than using fewer capacitors of greater capacity. Assume that a specific type of decoupling capacitor is used to realize a decoupling capacitance C, as shown in Fig. 7.7. Each capacitor has a capacity C_1 and a series inductance L_1^C. Placing $N_1 = \frac{C}{C_1}$ capacitors in parallel realizes the desired capacitance C

with an effective series inductance $L^{C_1} = \frac{L_1^C}{N_1}$. Alternatively, using capacitors of lower capacity C_2 requires a larger number of capacitors $N_2 = \frac{C}{C_2}$ to realize the same capacitance C, but results in a lower overall inductance of the capacitor bank $L^{C_2} = \frac{L_2^C}{N_2}$. The efficacy of this approach is enhanced if the lower capacity component C_2 has a smaller form factor than the components of capacity C_1 and therefore has a significantly smaller series inductance L_2^C. Using a greater number of capacitors, however, requires additional board area and incurs higher component and assembly costs. Furthermore, the efficacy of the technique is diminished if the larger area required by the increased number of capacitors necessitates placing some of the capacitors at a greater distance from the load, increasing the inductance of the downstream current path L_2.

The series inductance of the decoupling capacitance L^C, however, constitutes only a portion of the overall inductance $L_2 + L^C$ of the current path between two stages of the decoupling capacitance, referring again to the circuit model shown in Fig. 7.7. Once the capacitor series inductance L^C is much lower than L_2, any further reduction of L^C has an insignificant effect on the overall impedance. The inductance of the current path between the decoupling capacitance and the load therefore imposes an upper limit on the frequency range of the capacitor efficiency. A reduction of the interconnect inductance is therefore necessary to improve the efficiency of the decoupling capacitors.

7.9.2 Interconnect Inductance

Techniques for reducing interconnect inductance can be divided into *extensive* and *intensive* techniques. Extensive techniques lower inductance by using additional interconnect resources to form additional parallel current paths. Intensive techniques lower the inductance of existing current paths by modifying the structure of the power and ground interconnect so as to minimize the area of the current loop. The inductance of a power distribution system can be extensively decreased by allocating more metal layers on a printed circuit board and circuit package for power and ground distribution, increasing the number of package power and ground pins (solder balls in the case of ball grid array mounting) connecting the package to the board, and increasing the number of power and ground solder bumps or bonding wires connecting the die to the package. Extensive methods are often constrained by technological and cost considerations, such as the number and thickness of the metal and isolation layers in a printed circuit board, the total number of pins in a specific package, and the die bonding technology. Choosing a solution with greater interconnect capacity typically incurs a significant cost penalty.

Intensive methods reduce the interconnect inductance without incurring higher manufacturing costs. These methods alter the interconnect structure in order to decrease the area of the power current loop. For example, the inductance of a current

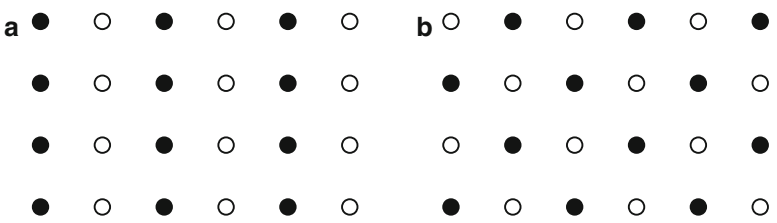


Fig. 7.24 Placement of area array connections for low inductance. The inductance of the area array interconnect strongly depends on the pattern of placement of the power and ground connections. The inductance of pattern (**a**) is relatively high, while pattern (**b**) has the lowest inductance. The power and ground connections are shown, respectively, in *black* and *white*

path formed by two square parallel power and ground planes is proportional to the separation of the planes[3] h,

$$L_{\text{plane}} = \mu_0 h. \tag{7.40}$$

Thus, parallel power and ground planes separated by 1 mil (25.4 μm) have an inductance of 32 pH per square. A smaller separation between the power and ground planes results in a proportionally smaller inductance of the power current loop. Reducing the thickness of the dielectric between the power and ground planes at the board and package levels is an effective means to reduce the impedance of the power distribution network [142, 143].

The same strategy of placing the power and ground paths in close proximity can be exploited to lower the effective inductance of the "vertical conductors," i.e., the conductors connecting the parallel power and ground planes or planar networks. Examples of such connections are pin grid arrays and ball grid arrays connecting a package to a board, a flip-chip area array of solder balls connecting a die to a package, and an array of vias connecting the power and ground planes within a package. In regular pin and ball grid arrays, this strategy leads to a so-called checkerboard pattern [144], as shown in Fig. 7.24.

7.10 Limitations of the One-Dimensional Circuit Model

The one-dimensional lumped circuit model shown in Fig. 7.3 has been used to describe the frequency dependent impedance characteristics of power distribution systems. This model captures the essential characteristics of power distribution

[3] Equation (7.40) neglects the internal inductance of the metal planes, i.e., the inductance associated with the magnetic flux within the planes. Omission of the internal inductance is a good approximation where the thickness of the planes is much smaller than the separation between the planes and at high signal frequencies, where current flow is restricted to the surface of the planes due to a pronounced skin (proximity) effect.

systems over a wide range of frequencies. Due to the relative simplicity, the model is an effective vehicle for demonstrating the primary issues in the design of power distribution systems and related design challenges and tradeoffs. The use of this model for other purposes, however, is limited due to certain simplifications present in the model. Several of these limitations are summarized below.

A capacitor at each decoupling stage is modeled by a single *RLC* circuit. In practice, however, the decoupling capacitance on the board and package is realized by a large number of discrete capacitors. Each capacitor has a distinct physical location relative to the load. The parasitic inductance of the current path between a specific capacitor and the power load is somewhat different for each of the capacitors. The frequency of the resonant impedance peaks of an individual capacitor therefore varies depending upon the location of the capacitor. The overall impedance peak profile of a group of capacitors is therefore spread, resulting in a lower and wider resonant peak.

More significantly, it is common to use two different types of capacitors for board level decoupling. High capacity electrolytic capacitors, typically ranging from a few hundreds to a thousand microfarads, are used to obtain the high capacitance necessary to decouple the high inductive impedance of a voltage regulator. Electrolytic capacitors, however, have a relatively high series inductance of several nanohenrys. Ceramic capacitors with a lower capacity, typically tens of microfarads, and a lower parasitic inductance, a nanohenry or less, are used to extend the frequency range of the low impedance region to higher frequencies. Effectively, there are two decoupling stages on the board, with a very small parasitic impedance between the two stages.

Using lumped circuit elements implies that the wavelength of the signals of interest is much larger than the physical dimensions of the circuit structure, permitting an accurate representation of the impedance characteristics with a few lumped elements connected in series. The current transitions at the power load are measured in tens of picoseconds, translating to thousands of micrometers of signal wavelength. This wavelength is much smaller than the size of a power distribution system, typically of several inches, making the use of a lumped model at first glance unjustified. The properties of power distribution systems, however, support a lumped circuit representation over a wider range of conditions as suggested by the aforementioned simple size criterion.

The design of a power distribution system is intended to restrict the flow of the power current as close to the load as possible. Due to the hierarchy of the decoupling capacitors, the higher the current frequency, the shorter the current loop, confining the current flow closer to the load. As seen from the terminals of the power source, a power distribution system is a multi-stage low pass filter, each stage having a progressively lower cut-off frequency. The spectral content of the current in the on-board power distribution system is limited to several megahertz. The corresponding wavelength is much larger than the typical system dimensions of a few inches, making a lumped model sufficiently accurate. The spectral content of the power current within the package network extends into the hundreds of megahertz frequency range. The signal wavelength remains sufficiently large to use

a lumped model; however, a more detailed network may be required rather than a one-dimensional model to accurately characterize the network impedances.

A one-dimensional model is inadequate to describe an on-chip power distribution network. The on-chip power distribution network is the most challenging element of the power delivery system design problem. The board and package level power distribution networks consist of several metal planes and thousands of vias, pins, and traces as well as several dozens of decoupling capacitors. In comparison, the on-chip power distribution network in a high complexity integrated circuit typically consists of millions of line segments, dramatically exacerbating the complexity of the design and analysis process. The design and analysis of on-chip power distribution networks is the focus of the Chap. 8.

7.11 Summary

The impedance characteristics of power distribution systems with multiple stages of decoupling capacitances have been described in this chapter. These impedance characteristics can be briefly summarized as follows.

- The significant inductance of the power and ground interconnect is the primary obstacle to achieving a low output impedance power distribution system
- The hierarchical placement of decoupling capacitors achieves a low output impedance in a cost effective manner by terminating the power current loop progressively closer to the load as the frequency increases
- The capacitance and effective series inductance determine the frequency range where the decoupling capacitor is effective
- Resonant circuits are formed within the power distribution networks due to the placement of the decoupling capacitors, increasing the output impedance near the resonant signal frequencies
- The effective series resistance of the decoupling capacitors is a critical factor in controlling the resonant phenomena
- The lower the inductance of the power interconnect and decoupling capacitors, the lower the decoupling capacitance necessary to achieve the target impedance characteristics

Chapter 8
On-Chip Power Distribution Networks

The impedance characteristics of a power distribution system are analyzed in the previous chapter based on a one-dimensional circuit model. While useful for understanding the principles of the overall operation of a power distribution system, a one-dimensional model is not useful in describing the distribution of power and ground across a circuit die. The size of an integrated circuit is usually considerably greater than the wavelength of the signals in the power distribution network. Furthermore, the power consumption of on-chip circuitry (and, consequently, the current drawn from the power distribution network) varies across the die area. The voltage across the on-chip power and ground distribution networks is therefore non-uniform. It is therefore necessary to consider the two-dimensional structure of the on-chip power distribution network to ensure that target performance characteristics of a power distribution system are satisfied. The on-chip power distribution network should also be considered in the context of a die-package system as the properties of the die-package interface significantly affect the constraints imposed on the electrical characteristics of the on-chip power distribution network.

The objectives of this chapter is to describe the structure of an on-chip power distribution network as well as review related tradeoffs. Various structural styles of on-chip power distribution networks are described in Sect. 8.1. The influence of the electrical characteristics of the die-package interface on the on-chip power and ground distribution is analyzed in Sect. 8.2. The influence of the on-chip power distribution network on the integrity of the on-chip signals is discussed in Sect. 8.3. The chapter concludes with a summary.

8.1 Styles of On-Chip Power Distribution Networks

Several topological structures are typically used in the design of on-chip power distribution networks. The power network structures range from completely irregular, essentially ad hoc, structures, as in routed power distribution networks, to highly

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0_8

regular and uniform structures, as in gridded power networks and power planes. These topologies and other basic types of power distribution networks are described in Sect. 8.1.1. Design approaches to improve the impedance characteristics of on-chip power distribution networks are presented in Sect. 8.1.2. The evolution of on-chip power distribution networks in the family of Alpha microprocessors is presented in Sect. 8.1.3.

8.1.1 Basic Structure of On-Chip Power Distribution Networks

Several structural types of on-chip power distribution networks are described in this section. Different parts of an on-chip network can be of different types, forming a hybrid network.

Routed Networks

In routed power distribution networks, the local circuit blocks are connected with dedicated routed power trunks to the power I/O pads along the periphery of the die [145], as shown in Fig. 8.1. A power mesh is typically used to distribute

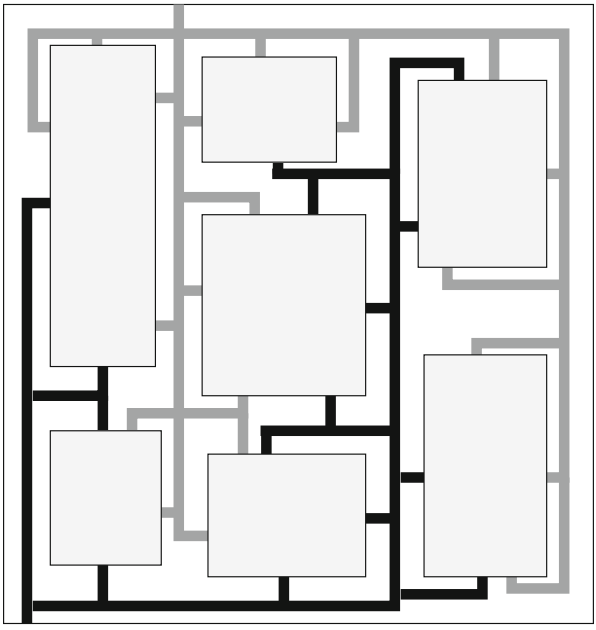


Fig. 8.1 Routed power and ground distribution networks. The on-chip circuit blocks are connected to the I/O power terminals with dedicated power (*black*) and ground (*gray*) trunks. The structure of the power distribution networks within the individual circuit blocks is not shown

the power and ground within a circuit block. The primary advantage of routed networks is the efficient use of interconnect resources, favoring this design approach in circuits with limited interconnect resources. The principal drawback of this topology is the relatively low redundancy of the power network. All of the current supplied to any circuit block is delivered through only a few power trunks. The failure of a single segment in a power distribution network jeopardizes the integrity of the power supply voltage levels in several circuit blocks and, consequently, the correct operation of the entire circuit. Routed power distribution networks are predominantly used in low power, low cost integrated circuits with limited interconnect resources.

Mesh Networks

Improved robustness and reliability are offered by power and ground mesh networks. In mesh structured power distribution networks, parallel power and ground lines in the upper metal layers span an entire circuit or a specific circuit block. These lines are relatively thick and wide, globally distributing the power current. The lines are interconnected by relatively short orthogonal straps in the lower metal layer, forming an irregular mesh, as shown in Fig. 8.2. The power and ground lines in the lower metal layer distribute current in the direction orthogonal to the upper metal layer lines and facilitate the connection of on-chip circuits to the global power distribution network. Mesh networks are used to distribute power in relatively low power circuits with limited interconnect resources [146]. It is often the case in semiconductor processes with only three or four metal layers that a regular power distribution grid in the upper two metal layers cannot be utilized due to an insufficient amount of metal resources and an ensuing large number of routing conflicts. Mesh networks are also used to distribute power within individual circuit blocks.

Grid Structured Networks

Grid structured power distribution networks, shown in Fig. 8.3, are commonly used in high complexity, high performance integrated circuits. Each layer of a power distribution grid consists of many equidistantly spaced lines of equal width. The direction of the power and ground lines within each layer is orthogonal to the direction of the lines in the adjacent layers. The power and ground lines are typically interdigitated within each layer. Each power and ground line is connected by vias to other power and ground lines, respectively, in the adjacent layers at the overlap sites. In a typical integrated circuit, the lower the metal layer, the smaller the width and pitch of the lines. The coarse pitch of the upper metal layer improves the utilization of the metal resources, conforming to the pitch of the I/O pads of the package, while the fine pitch of the lower grid layers brings the power and ground supplies in close proximity to each on-chip circuit, facilitating the connection of these circuits to power and ground.

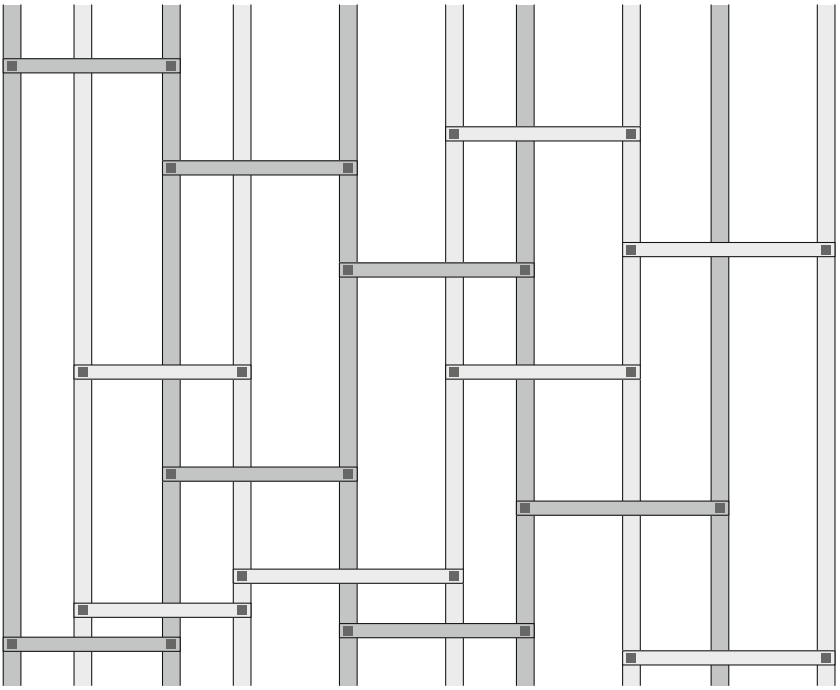


Fig. 8.2 A mesh structured power distribution network. Power (*dark gray*) and ground (*light gray*) lines in the vertically routed metal layer span an entire die or circuit block. These lines are connected by short straps of horizontally routed metal to form a mesh. The lines and straps are connected by vias (*the dark squares*)

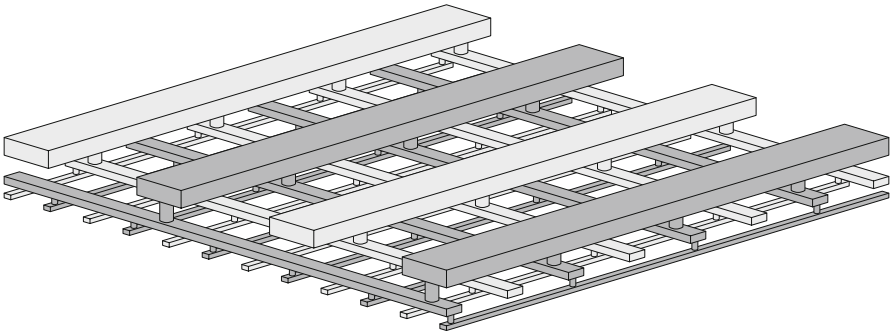


Fig. 8.3 A multi-layer power distribution grid. The ground lines are *light gray*, the power lines are *dark gray*. The pitch, width, and thickness of the lines are smaller in the lower grid layers than in the upper layers

Power distribution grids are significantly more robust than routed distribution networks. Multiple redundant current paths exist between the power terminals of each load circuit and the power supply pads. Due to this property, the power

supply integrity is less sensitive to changes in the power current requirements of the individual circuit blocks. The failure of any single segment of the grid is not critical to delivering power to any circuit block. An additional advantage of power distribution grids is the enhanced integrity of the on-chip data signals due to the capacitive and inductive shielding properties of the power and ground lines. These advantages of power distribution grids, however, are achieved at the cost of a significant share of on-chip interconnect resources. It is not uncommon to use from 20 % to 40 % of the metal resources to build a high density power distribution grid in modern high performance microprocessors [26, 27, 145].

Power and Ground Planes

Dedicated power and ground planes, shown in Fig. 8.4, have also been used in the design of power distribution networks [147]. In this scheme, an entire metal layer is used to distribute power current across a die, as shown in Fig. 8.4. The signal lines above and below the power/ground plane are connected with vias through the holes in the plane. Power planes also provide a close current return path for the surrounding signal lines, reducing the inductance of the signal lines and therefore the signal-to-signal coupling. This advantage, however, diminishes as the interconnect aspect ratios are gradually increased with technology scaling. While power planes provide a low impedance path for the power current and are highly robust, the interconnect overhead is typically prohibitively large, as entire metal layers are unavailable for signal routing.

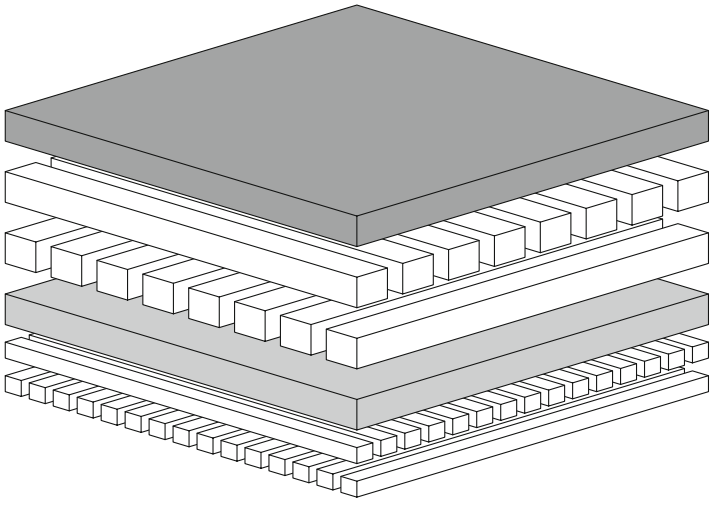


Fig. 8.4 On-chip power distribution scheme using power and ground planes. Two entire metal layers are dedicated to the distribution of power (*dark gray layer*) and ground (*light gray layer*)

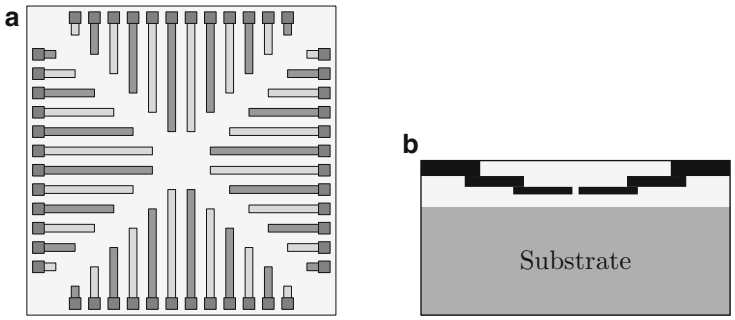


Fig. 8.5 Power distribution network structured as a cascaded power/ground ring; (**a**) cross-sectional view, (**b**) top view

Cascaded Power/Ground Rings

A novel topology for on-chip power distribution networks, called a "cascaded power/ground ring," has been proposed by Lao and Krusius [148] for integrated circuits with peripheral I/O. This approach is schematically illustrated in Fig. 8.5. The power and ground lines are routed from the power supply pads at the periphery of the die toward the die center. The power and ground lines at the periphery of the die, where the power current density is the greatest, are placed on the thick topmost metal layers with the highest current capacity. As the power current decreases toward the die center, the power and ground interconnect is gradually transferred to the thinner lower metal layers.

Hybrid-Structured Networks

Note that the boundaries between these network topologies are not well defined. A routed network with a large number of links looks quite similar to a meshed network, which, in turn, resembles a grid structure if the number of "strapping" links is large. The terms "mesh" and "grid" are often used interchangeably in the literature.

Furthermore, the structure of a power distribution network in a complex circuit often comprises a variety of styles. For example, the global power distribution can be performed through a routed network, while a meshed network is used for the local power distribution. Or the global power distribution network is structured as a regular grid, while the local power network is structured as a mesh network within one circuit block and a routed network within another circuit block. The common style of a *global* power distribution network has, however, evolved from a routed network to a global power grid, as the power requirements of integrated circuits have gradually increased with technology scaling.

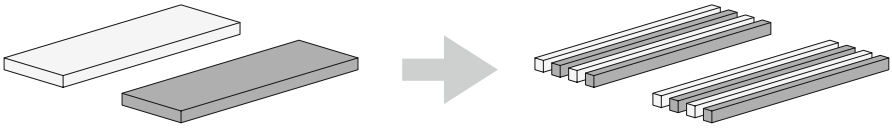


Fig. 8.6 Replacing wide power and ground lines (*left*) with multiple narrow interdigitated lines (*right*) reduces the inductance and characteristic impedance of the power distribution network

8.1.2 Improving the Impedance Characteristics of On-Chip Power Distribution Networks

The on-chip power and ground interconnect carries current from the I/O pads to the on-chip capacitors and from the on-chip capacitors to the switching circuits, acting as a load to the power distribution network. The flow of the power current through the on-chip power distribution network produces a power supply noise proportional to the network impedance, $Z = R + j\omega L$. The primary design objective is to ensure that the resistance and inductance of a power distribution system is sufficiently small so as to satisfy a target noise margin.

Several techniques have been employed to reduce the parasitic impedance of on-chip power distribution networks. The larger width and smaller pitch of the power and ground lines increase the metal area of the power distribution network, decreasing the network resistance. The resistance is effectively lowered by increasing the area of the power lines in the upper metal layers since these layers have a low sheet resistance. It is not uncommon to allocate over half of the topmost metal layer for global power and ground distribution [26, 149].

The inductance of on-chip power and ground lines has traditionally been neglected because the overall inductance of a power distribution network has been dominated by the parasitic inductance of the package pins, planes, vias, and bond wires. This situation is changing due to the increasing switching speed of integrated circuits [128, 129], the lower inductance of advanced flip chip packaging, and the higher on-chip decoupling capacitance which terminates the high frequency current paths. The requirement of achieving a low inductive impedance is in conflict with the requirement of a low resistance, as the use of wide lines to lower the resistance of a global power distribution network increases the network inductance. Replacing a few wide power and ground lines with multiple narrow interdigitated power and ground lines, as shown in Fig. 8.6, reduces the self-inductance of the supply network [150, 151] but increases the resistance. The tradeoffs among the area, resistance, and inductance of on-chip power distribution grids are explored in greater detail in Chap. 28.

8.1.3 *Evolution of Power Distribution Networks in Alpha Microprocessors*

The evolution of on-chip power distribution networks in high speed, high complexity integrated circuits is well illustrated by several generations of Digital Equipment Corporation Alpha microprocessors, as described by Gronowski, Bowhill, and Preston [149]. Supporting the reliable and efficient distribution of rising power currents have required adapting the structure of these on-chip power distribution networks, utilizing a greater amount of on-chip metal resources.

Alpha 21064

The Alpha 21064, the first microprocessor in the Alpha family, is manufactured in a 0.7 μm CMOS process in 1992. The Alpha 21064 consumes 30 W of power at 3.3 V, resulting in 9 A of average power current. Distributing this current across a 16.8 × 13.9 mm die would not have been possible in an existing two metal layer 0.75 μm CMOS process. An additional thick metal layer was added to the process, which is used primarily for the power and clock distribution networks. The power and ground lines in the third metal layer are alternated. All of the power lines are interconnected at one edge of the die with a perpendicular line in metal level three, and all of the ground lines are interconnected at the opposite edge of the die, as shown in Fig. 8.7. The resulting comb-like power and ground global distribution networks are interdigitated. The parallel lines of the global power and ground distribution networks are strapped with the power and ground lines of the second metal layer, forming a mesh-structured power distribution network.

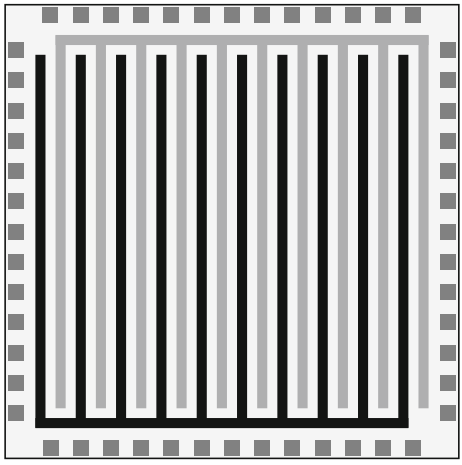


Fig. 8.7 Global power distribution network in Alpha 21064 microprocessor

Alpha 21164

The second generation microprocessor in the series, the Alpha 21164, appeared in 1995 and was manufactured in a 0.5 μm CMOS process. The Alpha 21164 nearly doubled the power current requirements to 15 A, dissipating 50 W from a 3.3 V power supply. The type of power distribution network used in the previous generation could not support these requirements. An additional forth metal layer was therefore added to a new 0.5 μm CMOS process. The power and ground lines in the forth metal layer are routed orthogonally to the lines in the third layer, forming a two layer global power distribution grid.

Alpha 21264

The Alpha 21264, the third generation of Alpha microprocessors, was introduced in 1998 in a 0.35 μm CMOS process. The Alpha 21264 consumes 72 W from a 2.2 V power supply, requiring 33 A of average power current distributed with reduced power noise margins. Utilization of conditional clocking techniques to reduce the power dissipation of the circuit increased the cycle-to-cycle variation in the power current to 25 A, exacerbating the overall power distribution problem [149]. The two layer global power distribution grid used in the 21164 could not provide the necessary power integrity characteristics in an integrated circuit with peripheral I/O. Therefore, two thick metal layers were added to the four layer process to allow the exclusive use of two metal layers as power and ground planes. In 2000, Alpha 21264 microprocessors were fabricated in a 0.18 μm CMOS process, utilizing flip-chip packaging with a high density array of I/O pins.

More recent versions of the Alpha 21264 microprocessor fabricated in newer process technologies utilize flip-chip packaging with a high density area array of I/O contacts. This approach obviates the use of on-chip metal planes for distributing power [152].

Power distribution grids are the design style of choice in most modern high performance integrated circuits [54, 147, 149, 153]. The focus of the material presented herein is therefore on-chip power distribution grids.

8.2 Die-Package Interface

At high frequencies, the impedance of a power distribution system is determined by the impedance characteristics of the on-chip and package power distribution networks, as discussed in Chap. 7. On-chip decoupling is essential to maintain a low impedance power distribution network at the highest signal frequencies of interest, as discussed in Sect. 7.5. The required on-chip decoupling capacitance is determined by the frequency where the package decoupling capacitors become inefficient. This frequency, in turn, is determined by the inductive impedance of the current path

between the package capacitance and the integrated circuit. The required minimum on-chip decoupling capacitance is proportional to this inductance, as expressed by (7.20) (or by the analogous constraints presented in Sect. 7.6). Minimizing this inductance achieves the target impedance characteristics of a power distribution network with the smallest on-chip decoupling capacitance.

Achieving a low impedance connection between the package capacitors and an integrated circuit is, however, difficult. Delivering power is only one of the multiple functions of an IC package, which also include connecting the I/O signals to the outside world, maintaining an acceptable thermal environment, providing mechanical support, and protecting the circuit from the environment. These package functions all compete for physical resources within a small volume in the immediate vicinity of the die. Complex tradeoffs among these package design goals are made in practice, often preventing the realization of a resonance-free die-to-package interface.

8.2.1 Wire Bond Packaging

Maintaining a low impedance die-package interface is particularly challenging in wire bonded integrated circuits. The self-inductance of a wire bond connection typically ranges from 4 to 6 nH. The number of bond wires is limited by the perimeter of the die and the pitch of the wire bond connections. The total number of power and ground connections typically does not exceed several hundred. It is therefore difficult to decrease the inductance of the package capacitor connection to the die significantly below one to two nanohenrys. In wire bond packages, the inductance of the current loop terminated by the package capacitors is not significantly smaller than the inductance of the current loop terminated by the board decoupling capacitors. Under these conditions, the package capacitors do not significantly improve the impedance characteristics of the power distribution system and therefore no appreciable gain in circuit speed is achieved [134, 154]. Decoupling a high inductive impedance at gigahertz frequencies typically requires an impractical amount of on-chip decoupling capacitance (and therefore die area), limiting the operational frequency of a wire bonded circuit.

Providing a low impedance connection between the off-chip decoupling capacitors and a wire bonded integrated circuit requires special components and packaging solutions. For example, a so-called “closely attached capacitor” has been demonstrated to be effective for this purpose in wire bonded circuits [155]. A thin flat capacitor is placed on the active side of an integrated circuit. The dimensions of the capacitor are slightly smaller than the die dimensions. The bonding pads of an integrated circuit and the edge of the capacitor are in close proximity, permitting a connection with a short bond wire, as illustrated in Fig. 8.8. The die-to-capacitor wires are several times shorter than the wire connecting the die to the package. The impedance between the circuit and the off-chip decoupling capacitance is therefore significantly decreased.

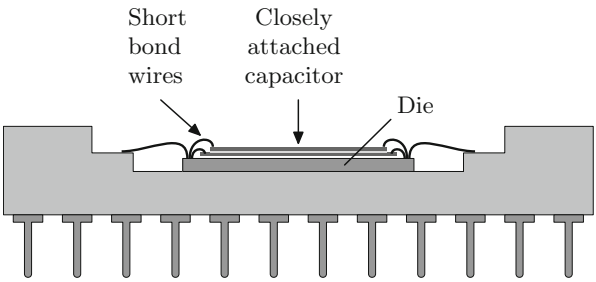


Fig. 8.8 Closely attached capacitor. Stacking a thin flat capacitor on top of a circuit die allows connecting the capacitor and the circuit with bond wires of much shorter length as compared to bond wires connecting the circuit to the package

This wiring technique reduces the power switching noise in CMOS circuits by at least two to three times, as demonstrated by Hashemi et al. [155]. For example, attaching a 12 nF capacitor to a 32-bit microprocessor decreases the internal logic power supply noise from 900 to 150 mV, a sixfold improvement (the microprocessor is packaged in a 169-pin pin grid array package and operates at 20 MHz with a 5 V power supply). Similarly, in a 3.3 V operated bus interface circuit, a 12 nF closely attached capacitor lowers the internal power noise from 215 to 100 mV. A closely attached capacitor is used in the Alpha 21264 microprocessor, as the on-chip decoupling capacitance is insufficient to maintain adequate power integrity [149, 156].

8.2.2 Flip-Chip Packaging

The electrical characteristics of the die-package interface are significantly improved in flip-chip packages. Flip-chip bonding refers to attaching a die to a package with an array of solder balls (or bumps) typically 50–150 μm in diameter. In cost sensitive circuits, the ball connections, similar to wire bond connections, can be restricted to the periphery of the die in order to reduce the interconnect complexity of the package. In high complexity, high speed integrated circuits, however, an area array flip-chip technology is typically used where solder ball connections are distributed across (almost) the entire area of the die. The inductance of a solder ball connection, typically from 0.1 to 0.5 nH, is much smaller than the 4–10 nH typical for a bond wire [114, 126, 157]. Area array flip-chip bonding also provides a larger number of die to package connections as compared to wire bonding. Modern high performance microprocessors have thousands of flip-chip contacts dedicated to the power distribution network [27, 158–160]. A larger number of lower inductance power and ground connectors significantly decreases the overall inductance of the die to package connection.

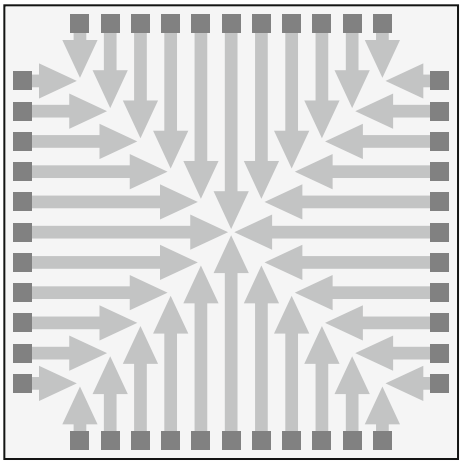


Fig. 8.9 Flow of power current in an integrated circuit with peripheral I/O. The power current is distributed on-chip across a significant distance: from the edge of the die to the die center

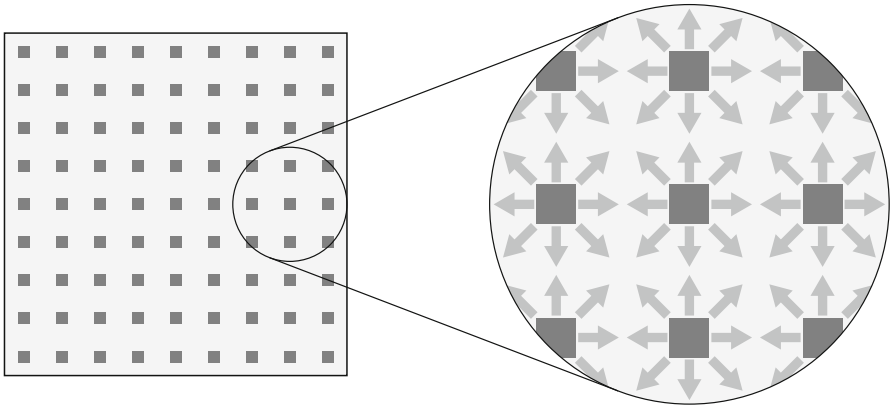


Fig. 8.10 Flow of power current in an integrated circuit with flip-chip I/O. The power current is distributed on-chip over a distance comparable to the pitch of the I/O pads

Also important in integrated circuits with peripheral I/O, the power current is distributed on-chip across a significant distance: from the die edge to the die center, as shown in Fig. 8.9. In integrated circuits with a flip-chip area array of I/O bumps, the power current is distributed on-chip over a distance comparable to the size of the power pad pitch, as shown in Fig. 8.10. This distance is significantly smaller than half the die size. Flip-chip packaging with high density I/O, therefore, significantly reduces the effective resistance and inductance of the on-chip power distribution network, mitigating the resistive [115, 161] and inductive voltage drops.

Flip-chip packaging with high density I/O therefore decreases the area requirements of the on-chip power distribution network, improving the overall performance of a circuit [152, 162–164]. The dependence of the on-chip power voltage drop on the flip-chip I/O pad density and related power interconnect requirements are discussed in Chap. 5.

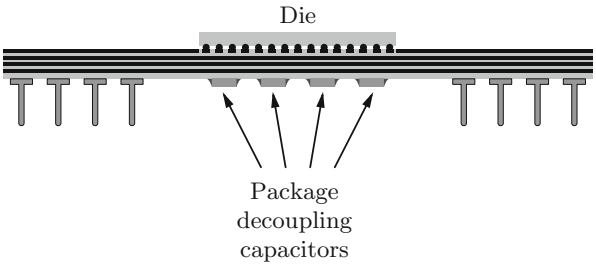


Fig. 8.11 Flip-chip pin grid array package. The package decoupling capacitors are mounted on the bottom side of the package immediately below the die mounted on the top side. In this configuration, the package capacitors are in physical proximity to the die, minimizing the impedance between the capacitors and the circuit

Another advantage of area array flip-chip packaging is the possibility of placing the package decoupling capacitors in close physical proximity to the die, significantly enhancing the efficacy of the capacitors. A bank of low parasitic inductance capacitors can be placed on the underside of the package immediately below the die, as shown in Fig. 8.11. The separation between the capacitors and the on-chip circuitry is reduced to 1–2 mm, minimizing the area of the current loop and associated inductance. The parasitic inductance of a package decoupling capacitor with the package vias and solder bumps connecting the capacitor to the die can be reduced well below 1 nH [154], enhancing the capacitor efficiency at high frequencies. Placing a package decoupling capacitor immediately below the circuit region with the greatest power current requirements further improves the efficiency of the decoupling capacitors of the package [165].

Overall, flip-chip packaging significantly decreases the impedance between the integrated circuit and the package decoupling capacitors, relaxing the constraints on the resistance of the on-chip power distribution network and on-chip decoupling capacitors. Flip-chip packaging therefore can significantly improve the power supply integrity while reducing the die area.

8.2.3 Future Packaging Solutions

The increasing levels of current consumed by CMOS integrated circuits as well as the high switching speeds require power distribution systems with a lower impedance over a wider frequency range. An increasingly lower inductance between the integrated circuit and the package decoupling capacitance is essential to maintain a lower impedance at higher frequencies. Providing a low impedance die-to-package connection remains a challenging task [166]. Future generations of packaging solutions, such as chip-scale and bumpless build-up layer (BBUL) packaging, are addressing this problem with higher density die-to-package contacts, a smaller separation between the power and ground planes, and a lower package height [143, 167, 168].

The electrical characteristics of a package have become one of the primary factors that limit the performance of an integrated circuit [169, 170]. The package design is now crucial in satisfying both the speed and overall cost targets of high performance integrated circuits. Achieving these goals will require explicit co-design of the on-chip global interconnect and the package interconnect networks [141, 169].

8.3 Other Considerations

Dependence of the power supply integrity on the impedance characteristics of the on-chip power distribution network has been discussed previously. The on-chip power distribution network also significantly affects the integrity of the on-chip data and clock signals. The integrity of the on-chip signals depends upon the structure of the power distribution network through two primary mechanisms: inductive interaction among the power and signal interconnect and substrate coupling. These two phenomena are briefly discussed in this section.

8.3.1 Dependence of On-Chip Signal Integrity on the Structure of the Power Distribution Network

The structure of the power distribution grid is one of the primary factors determining the integrity of on-chip signals. The power and ground lines shield adjacent signal lines from capacitive crosstalk. Sensitive signals are typically routed adjacent to the power and ground lines. Co-designing the power and signal interconnect has become an important consideration in the design of high speed integrated circuits [171–173].

Power and ground networks provide a low impedance path for the signal return currents. The structure of the on-chip power distribution network is therefore a primary factor that determines the inductive properties of on-chip signal lines, such as the self- and mutual inductance. Modeling the inductive properties of the power distribution grid is necessary for accurately analyzing high frequency phenomena, such as return current distribution, signal overshoot, and signal delay variations, as demonstrated by on-chip interconnect structures using full-wave partial element equivalent circuit (PEEC) models [174, 175]. These conclusions are also supported by the analysis of commercial microprocessors. Inadequate design of the local power distribution network can lead to significant inductive coupling of the signals, resulting in circuit failure [54].

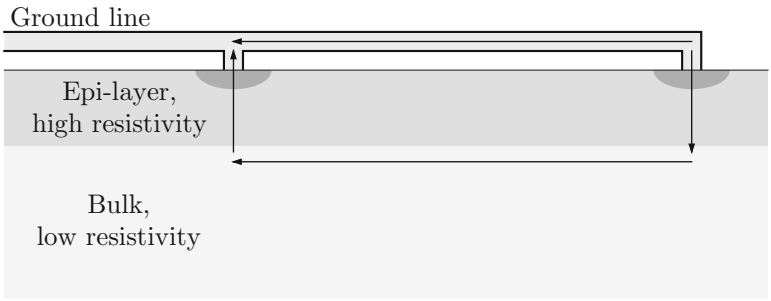


Fig. 8.12 Interaction of the substrate and power distribution network. The low resistivity bulk substrate provides additional current paths where the ground network is connected to the substrate. These current paths are in parallel with the ground distribution network

8.3.2 Interaction Between the Substrate and the Power Distribution Network

In high complexity digital integrated circuits, the ground distribution network is typically connected to the substrate to provide an appropriate body bias for the NMOS transistors. The substrate provides additional current paths in parallel to the ground distribution network, affecting current distribution in the network, as illustrated in Fig. 8.12. This effect is significant in most digital CMOS processes which utilize a low resistivity substrate to prevent device latch-up [176]. A methodology for analyzing power distribution networks together with the silicon substrate requires a complete model, which includes both the power distribution system and the substrate [177]. In [178], a high-level simulation methodology for generating a macromodel of each standard cell is described, reducing the complexity of the process of analyzing the power network and substrate. The efficiency and accuracy of the power distribution network and substrate analysis process can be enhanced by partitioning a power network into voltage domains [179]. To further increase the computational efficiency, the substrate can be characterized by macromodels [180–183]. The substrate significantly reduces the voltage drop in the ground distribution network (assuming an N-well process) by serving as an additional parallel path for the ground current to flow, as demonstrated by an analysis of three Motorola processor circuits [177]. The placement of the substrate contacts also affects the on-chip power supply and substrate noise [177, 184].

8.4 Summary

The structure of the on-chip power distribution network and related design considerations are described in this chapter. The primary conclusions are summarized as follows.

- On-chip power distribution grids are the preferred design style in high speed, high complexity digital integrated circuits
- Constraints placed on the impedance characteristics of the on-chip power distribution networks are greatly affected by the electrical properties of the package
- The high frequency impedance characteristics of a power distribution system are significantly enhanced in packages with an area array of low inductance I/O contacts

Chapter 9
Intelligent Power Networks On-Chip

To facilitate the integration of diverse functionality, architectural, an integrated family of circuit, device, and material level power delivery solutions are required. Per core dynamic voltage and frequency scaling is a primary concern for efficiently managing a power budget, and requires the on-chip integration of compact controllers within hundreds of power domains and thousands of cores, further increasing the design complexity of these power delivery systems. While in-package and on-chip power integration has recently became a primary concern [185, 186], focus remains on developing compact and efficient power supplies. A methodology to design and manage in-package and on-chip power has not been a topic of emphasis. Thus, power delivery in modern ICs is currently dominated by ad hoc approaches. With the increasing number of power domains, greater granularity of the on-chip supply voltages, and domain adaptive power requirements, the design of the power delivery process has greatly increased in complexity, and is impractical without a systematic methodology. The primary objective of this chapter is to describe a systematic methodology for distributed on-chip power delivery and management.

A power network on-chip is a vehicle for distributed on-chip power management. The analogy between a network-on-chip (NoC) and PNoC is illustrated in Fig. 9.1 with simplified NoC and PNoC models. Similar to a NoC, a PNoC decreases the design complexity of a power delivery system, while enhancing the control of the quality of power (QoP) and DVS, and providing a scalable platform for efficient power management.

The rest of the chapter is organized as follows. The principles of the PNoC design methodology are described in Sect. 9.1. In Sect. 26.4, the performance of the PNoC architecture is compared with existing approaches based on the evaluation of several test cases. Design and performance issues of the PNoC architecture are also discussed. Some concluding remarks are offered in Sect. 9.3.

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0_9

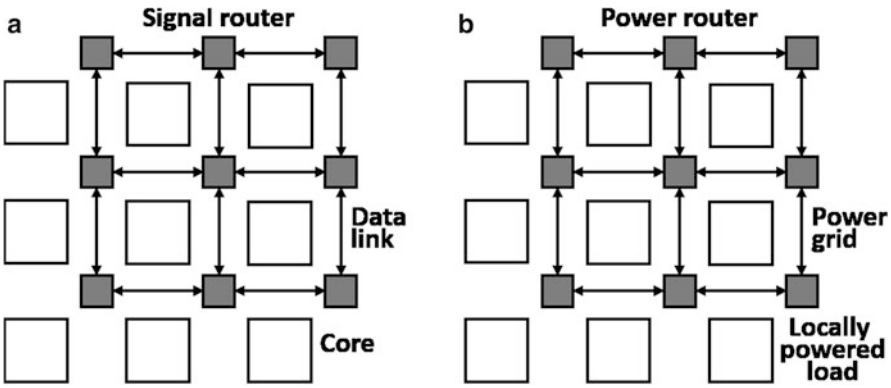


Fig. 9.1 On-chip networks based on the approach of separation of functionality, (**a**) network-on-chip, and (**b**) power network-on-chip

9.1 Power Network-On-Chip Architecture

The key concept in systemizing the design of power delivery is to convert the power off-chip, in-package, and/or on-chip with multiple power efficient but large switching power supplies, deliver the power to on-chip voltage clusters, and regulate the power with hundreds of linear low dropout regulators at the point-of-load [187]. A power network-on-chip is a systematic solution to on-chip power delivery that leverages distributed point-of-load power delivery within a fine grained power management framework. The PNoC architecture is a mesh of power routers and locally powered loads, as depicted in Fig. 9.1. The power routers are connected through power switches, distributing current to those local loads with similar voltage requirements. An example PNoC is illustrated in Fig. 9.2 for a single voltage cluster with nine locally powered loads and three different supply voltages, $V_{DD,1}$, $V_{DD,2}$, and $V_{DD,3}$. The power network configuration is shown at two different times, t_1 and t_2.

A power network-on-chip virtually manages the power in SoCs through specialized power routers, switches, and programmable control logic, while supporting scalable power delivery in heterogeneous ICs. A PNoC is comprised of physical links and routers that provide both virtual and physical power routing. This system senses the voltages and currents throughout the system, and manages the POL regulators through power switches. Based on the sensed voltages and currents, a programmable unit makes real-time decisions to apply a new set of configurations to the routers per time slot, dynamically managing the on-chip power delivery process. Novel algorithms are required to dynamically customize the power delivery policies through a specialized microcontroller that routes the power. These algorithms satisfy real-time power and performance requirements.

A PNoC composed of power routers connected to global power grids and locally powered loads is illustrated in Fig. 9.3. Global power from the converters is managed

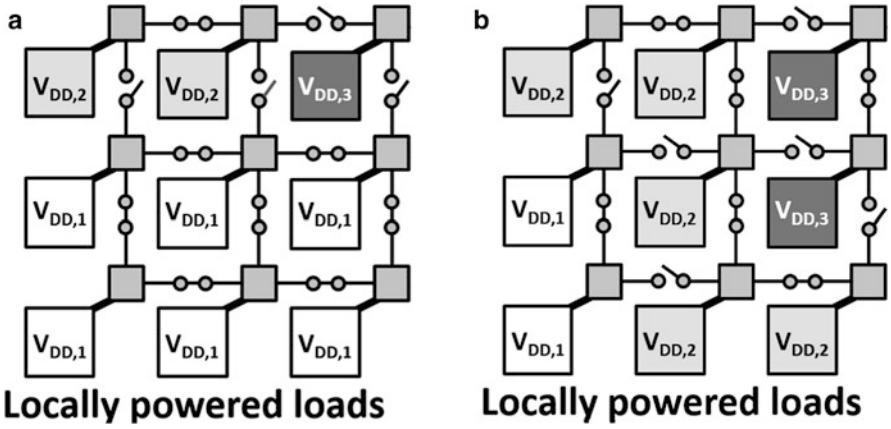


Fig. 9.2 Example of on-chip power network with multiple locally powered loads and three supply voltage levels (**a**) PNoC configuration at time t_1, and (**b**) PNoC configuration at time t_2

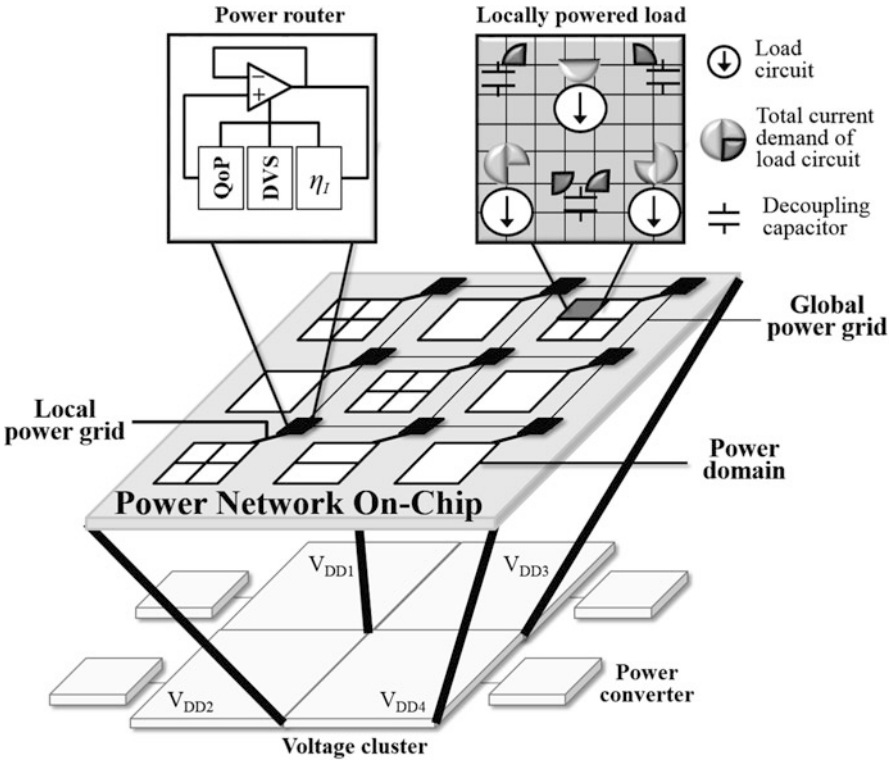


Fig. 9.3 On-chip power network with routers distributing the current over the power grid to the local loads

by the power routers, delivered to individual power domains, and regulated within the locally powered loads. These locally powered loads combine all of the current loads located within a specific on-chip power domain with the decoupling capacitors that supply the local current demand within that region. To support DVS, the power switches within the PNoC are dynamically controlled, (dis)connecting power routers within individual voltage clusters in real time. The current loads are powered at similar voltage levels, and therefore draw current from all of the connected power routers, lessening temporary current variations. Similar to a mesh based clock distribution network [188, 189], the shared power supply lessens the effect of the on-chip parasitic impedances, enhancing voltage regulation and quality of power.

The power routers, local current loads, and power grids are described in the following subsections. Different PNoC topologies and specific design objectives are also considered.

9.1.1 Power Routers

The efficient management of the energy budget is dynamically maintained by the power routers. Each power domain is controlled by a single power router. A router topology ranges from a simple linear voltage regulator, shown in Fig. 9.4a, to a complex power delivery system, as depicted in Fig. 9.4b, with sensors, dynamically adaptable power supplies, switches, and a microcontroller. These structures feature real-time voltage/frequency scaling, adaptable energy allocation, and precise control over the on-chip QoP. With the PNoC routers, the power is managed locally based on specific local current and voltage demands, decreasing the dependence on remotely located loads and power supplies. The scalability of the power delivery process is therefore enhanced with the PNoC approach.

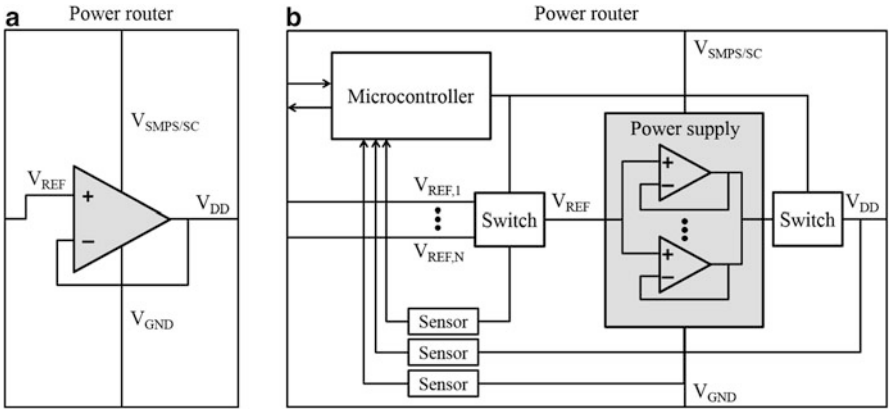


Fig. 9.4 Power routers for PNoC (**a**) Simple topology with linear voltage regulator. (**b**) Advanced topology with dynamically adaptable voltage regulator and microcontroller

9.1.2 Locally Powered Loads

Locally powered loads with different current demands and power budgets can be efficiently managed with a PNoC. The local power grids provide a specific voltage to the nearby load circuits. The highly complex interactions among the multiple power supplies, decoupling capacitors, and load circuits need to be considered, where the interactions among the nearby components are typically more significant. The effective region for a point-of-load power supply is the overlap of the effective regions of the surrounding decoupling capacitors [190, 191]. Loads within the same effective region are combined into a single equivalent locally powered load regulated by a dedicated LDO. All of the LDO regulators within a power domain are controlled by a single power router. A model and closed-form expressions of the interactions among the power supplies, decoupling capacitors, and current loads are required to efficiently partition an IC with billions of loads into power domains and locally powered loads.

9.1.3 Power Grid

Different configurations of local power grids are considered for distributing power from LDO regulators to locally powered loads. A shared power grid with multiple parallel connected LDO sources is illustrated on the left in Fig. 9.5, delivering power to all of the loads within a power domain. A shared power grid with multiple LDO sources is prone to stability issues due to current sharing, and process, voltage, and temperature variations. Specialized adaptive mechanisms are included within the power routers to stabilize a power delivery system that includes a

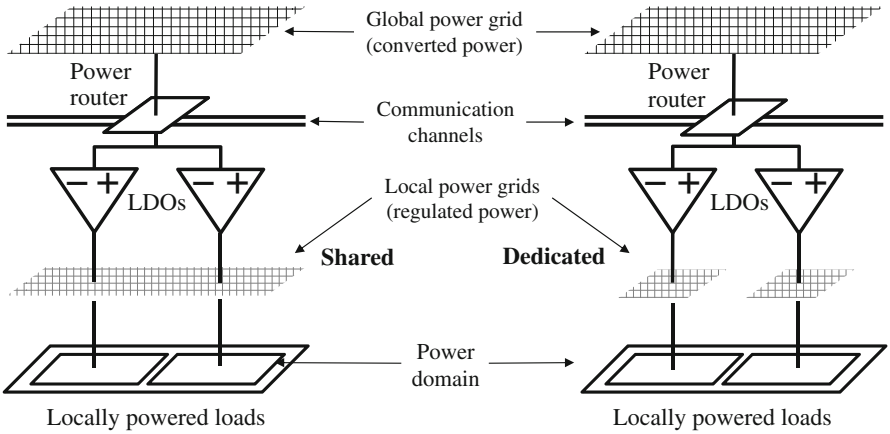


Fig. 9.5 A PNoC power router with two locally powered loads, shared local power grid (*on the left*), and dedicated local power grids (*on the right*)

multi-source shared power grid. Alternatively, to minimize interactions between parallel connected LDO regulators, dedicated power grids each driven by a single LDO should be considered. A topology with dedicated local power grids is illustrated on the right in Fig. 9.5. The dedicated power grids require fine grain distribution of the local power current.

9.2 Case Study

To evaluate the performance of the power router, a PNoC with four power routers is considered, supplying power to four power domains. IBM power grid benchmark circuits [192] model the behavior of the individual power domains. To simulate a dynamic power supply in PNoC, the original IBM voltage profiles are scaled to generate the target power supply voltages between 0.5 and 0.8 V. Target voltage profiles with four voltage levels (0.8, 0.75, 0.7, and 0.65 V) within a PNoC are illustrated in Fig. 9.6. The number of power domains with each of the four supply voltages changes dynamically based on the transient power requirements of the power domains.

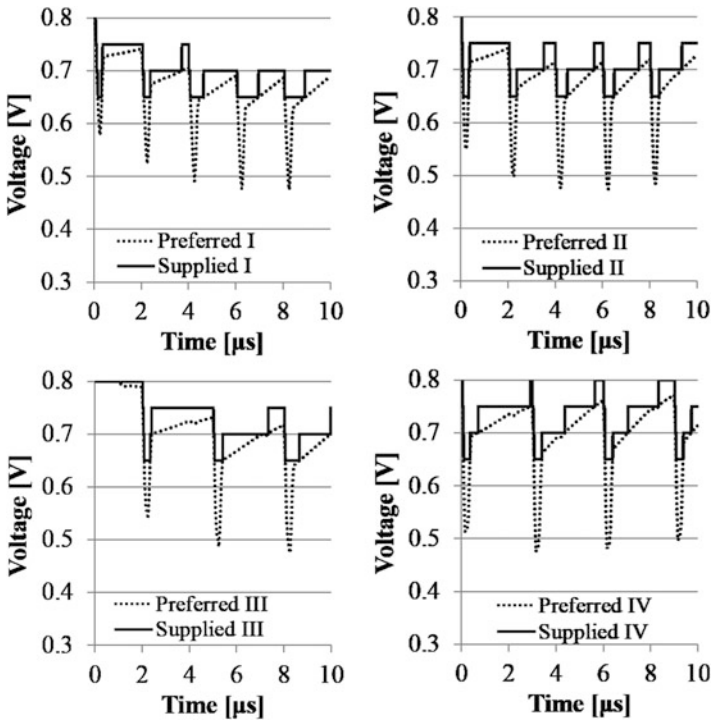


Fig. 9.6 Preferred and supplied voltage levels in PNoC with four power domains

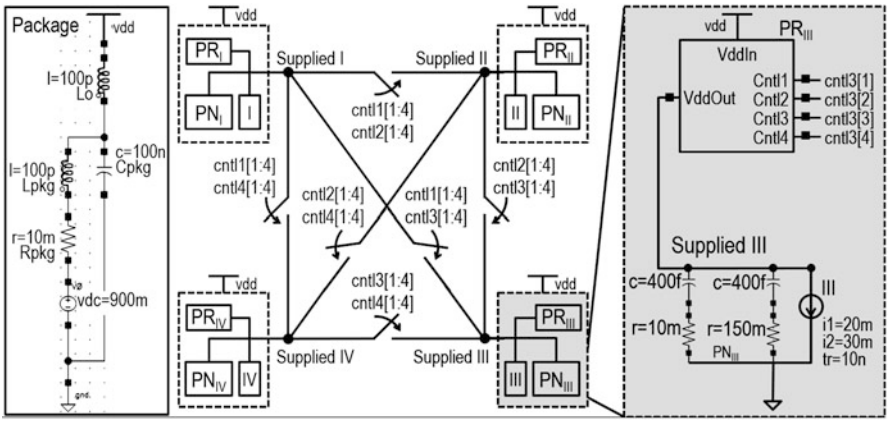


Fig. 9.7 PNoC with four power domains and four power routers connected with control switches

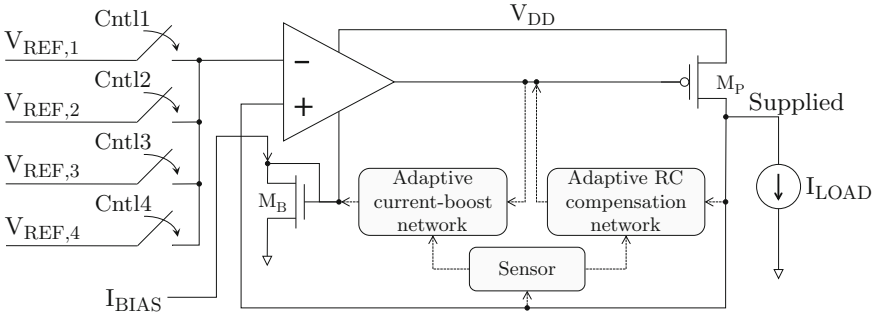


Fig. 9.8 Power router with voltage regulator, load sensor, and adaptive networks

A schematic of a PNoC with four power domains (I, II, III, and IV) and four power routers (PR_I, PR_{II}, PR_{III}, and PR_{IV}) is shown in Fig. 9.7. Each of the power routers is composed of an LDO with four switch controlled reference voltages to support dynamic voltage scaling. In addition, the power routers feature adaptive RC compensation and current boost networks controlled by load sensors to provide quality of power control and optimization, as shown in Fig. 9.8. The adaptive RC compensation network is comprised of a capacitive block connected in series with two resistive blocks, all digitally controlled. These RC impedances are digitally configured to stabilize the LDO within the power routers under a wide range of process variations. The current boost circuit is composed of a sensor block that follows the output voltage at the drain of transistor M_P (see Fig. 9.8), and a current boost block that controls the current through the differential pair within the LDO. When a high slew rate transition at the output of the LDO occurs, the boost mode is activated, raising the tail current of the LDO differential pair. Alternatively, during regular mode, no additional current flows into the differential pair, enhancing the power efficiency of the LDO. A description of the load sensor and adaptive networks are provided in Chap. 18.

Table 9.1 Output load in a PNoC with four power domains

Domain	I	II	III	IV
Minimum output current (mA)	10	75	20	20
Maximum output current (mA)	50	10	30	20
Output transition time (ns)	50	50	10	10

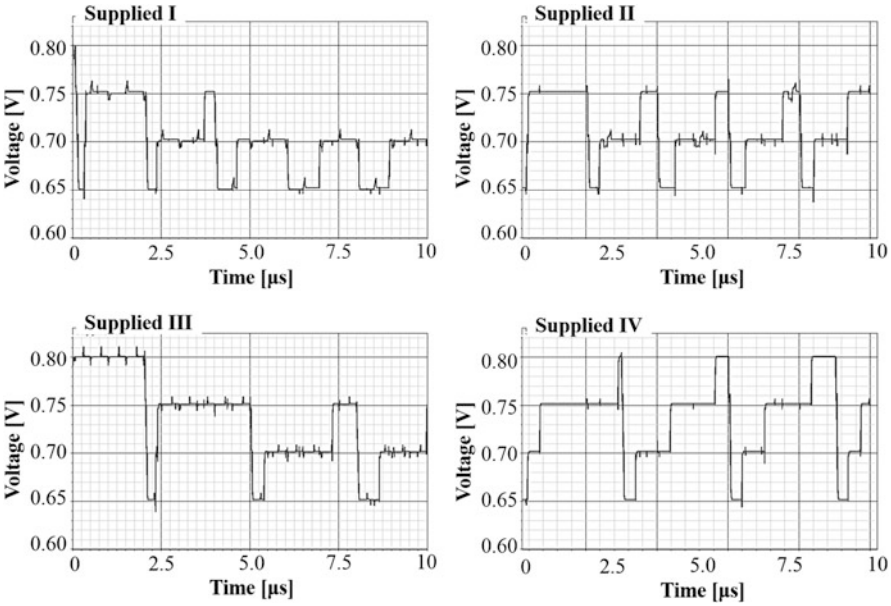


Fig. 9.9 Voltage levels in PNoC with four power domains

The power routers are connected with controlled switches to mitigate load transitions in domains with similar supply voltages. To model the RLC parasitic impedances of the package and power network, non-ideal LDO input and output impedances (PN_I, PN_{II}, PN_{III}, and PN_{IV}) are considered, as shown in Fig. 9.7. The load current characteristics are listed in Table 9.1 for each of the four domains. PNoC SPICE simulation results are shown in Fig. 9.9, exhibiting a maximum error of 0.35 %, 2.0 %, and 2.7 % for, respectively, the steady state dropout voltage, load regulation due to the output current switching, and load regulation due to dynamic PNoC reconfigurations. Good correlation with the required power supply in Fig. 9.6 is demonstrated. The power savings in each of the power domains range between 21.0 % to 31.6 % as compared to a system without dynamic voltage scaling. These average power savings show that the PNoC architecture can control the power supply voltages in real-time, optimizing the power efficiency of the overall power delivery system [193–195].

9.3 Summary

To address the issues of power delivery complexity and quality of power, a power delivery system should provide a scalable modular architecture that supports integration of additional functional blocks and power features (e.g., DVS, adaptive RC compensation, and efficiency optimization with adaptive current boost) without requiring the re-design of the power delivery system. The architecture should also support heterogeneous circuits and technologies.

- The concept and architecture of an on-chip power network PNoC is described in this chapter
- The on-chip network is exploited for systematic power delivery in SoCs to reduce design complexity while increasing scalability
- A methodology that separates power conversion and regulation is provided for efficiently enhancing the quality of power
- The application of local power routing is enabled through a specialized micro-controller for on-chip power management
- Small area power supplies are utilized as point-of-load voltage regulators

Chapter 10
Conclusions

The design of low impedance power distribution networks is a key element in achieving high performance integrated circuits. The inductance of the power grid is a primary obstacle to achieving this goal. Proper allocation of the decoupling capacitors can significantly reduce the network impedance. The effective series resistance and inductance of a decoupling capacitor are key factors in reducing the effectiveness of the decoupling capacitor. The resonant circuit formed within the power network increases the impedance of the network near the resonant frequency. The network impedance can be reduced by a variety of power grid structures, based on area, resistance, and inductance tradeoffs. The package impedance also needs to be considered when designing a power distribution network for high performance integrated circuits.

Centralized power delivery systems have recently been used to dynamically manage power in heterogeneous high performance multicore systems, requiring a long feedback path from the individual power domains to a single power management controller. Additional power is dissipated in centralized power delivery systems due to unnecessary data transport. The slow response from the long feedback path limits real-time control over the energy budget, and all of the power management functions located in one or a few places may not scale. Distributed, locally intelligent power management approaches should therefore be considered to efficiently manage the power budget in real-time. On-chip power networks with locally intelligent power routers and specialized microcontrollers are therefore reviewed in this part.

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0_10

Part III
On-Chip Decoupling Capacitors

Decoupling capacitors are an integral element of the power distribution network design process. Different topics related to decoupling capacitance are discussed in Part III. This part is divided into several chapters, discussing strategies for efficiently placing decoupling capacitors within on-chip power distribution network and the co-design of multiple decoupling capacitors with multiple on-chip power supplies. Each of the following chapters is summarized below.

Decoupling capacitance is described in Chap. 11. A historical perspective of capacitance is provided. The decoupling capacitor is shown to be analogous to a reservoir of charge. A hydraulic analogy of the hierarchical placement of decoupling capacitors is also described. It is demonstrated that the impedance of a power distribution system can be maintained below a target specification over an entire range of operating frequencies by utilizing a hierarchy of decoupling capacitors. Antiresonance in the impedance of a power distribution system with decoupling capacitors is also intuitively explained in this chapter. Different types of on-chip decoupling capacitors are compared. Several allocation strategies for placing on-chip decoupling capacitors are reviewed.

On-chip decoupling capacitors have traditionally been allocated into the available free (or white) space on a die. The efficacy of the on-chip decoupling capacitors however depends upon the impedance of the power/ground lines connecting the capacitors to the current loads and power supplies. A design methodology for placing on-chip decoupling capacitors is presented in Chap. 12. The maximum effective radii of an on-chip decoupling capacitor is determined by the target impedance (during discharge) and the charge time. Two criteria to estimate the minimum required on-chip decoupling capacitance are also presented.

As the minimum feature size continues to scale, additional on-chip decoupling capacitance will be required to support increasing current demands. A larger on-chip decoupling capacitance requires a greater area which cannot conveniently be placed close to the switching load circuits. Moreover, a large decoupling capacitor exhibits a distributed impedance behavior. A lumped model of an on-chip decoupling capacitor, therefore, results in underestimating the capacitance requirements, thereby increasing the power noise. A methodology for efficiently

placing on-chip distributed decoupling capacitors is the subject of Chap. 13. Design techniques to estimate the location and magnitude of a system of distributed on-chip decoupling capacitors are presented. Different tradeoffs in the design of a system of distributed on-chip decoupling capacitors are also evaluated.

Multiple decoupling capacitors with multiple on-chip power supplies is the topic of Chap. 14. The large number of on-chip power supplies and decoupling capacitors inserted throughout an integrated circuit complicates the design and analysis of power distribution networks. Complex interactions among the power supplies, decoupling capacitors, and active load circuitry are evaluated utilizing a computationally efficient analysis methodology. The effect of the physical distance among the power supplies and decoupling capacitors on power supply noise is discussed. A design methodology for simultaneously placing the on-chip power supplies and decoupling capacitors is described. This methodology changes conventional practices where the power distribution network is designed first, followed by placing the decoupling capacitors.

Chapter 11
Decoupling Capacitance

The on-going miniaturization of integrated circuit feature sizes has placed significant requirements on the on-chip power and ground distribution networks. Circuit integration densities rise with each nanoscale technology generation due to smaller devices and larger dies. The on-chip current densities and the total current also increase. Simultaneously, the higher switching speed of smaller transistors produces faster current transients in the power distribution network. Supplying high average currents and continuously increasing transient currents through the high impedance on-chip interconnects results in significant fluctuations of the power supply voltage in scaled CMOS technologies.

Such a change in the supply voltage is referred to as power supply noise. Power supply noise adversely affects circuit operation through several mechanisms, as described in Chap. 1. Supplying sufficient power current to high performance ICs has therefore become a challenging task. Large average currents result in increased IR noise and fast current transients result in increased $L\ dI/dt$ voltage drops (ΔI noise) [23].

Decoupling capacitors are often utilized to manage this power supply noise. Decoupling capacitors can have a significant effect on the principal characteristics of an integrated circuit, i.e., speed, cost, and power. Due to the importance of decoupling capacitors in current and future ICs, significant research has been developed over the past several decades, covering different areas such as hierarchical placement of decoupling capacitors, sizing and placing of on-chip decoupling capacitors, resonant phenomenon in power distribution systems with decoupling capacitors, and static on-chip power dissipation due to leakage current through the gate oxide.

In this chapter, a brief review of the background of decoupling capacitance is provided. In Sect. 11.1, the concept of a decoupling capacitance is described and an historical retrospective is described. A practical model of a decoupling capacitor is also described. In Sect. 11.2, the impedance of a power distribution system with decoupling capacitors is presented. Target specifications of the impedance

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0_11

of a power distribution system are reviewed. Antiresonance phenomenon in a system with decoupling capacitors is intuitively explained. A hydraulic analogy of the hierarchical placement of decoupling capacitors is also presented. Intrinsic and intentional on-chip decoupling capacitances are discussed and compared in Sect. 11.3. Different types of on-chip decoupling capacitors are qualitatively analyzed in Sect. 11.4. The advantages and disadvantages of several types of widely used on-chip decoupling capacitors are also discussed in Sect. 11.4. Enhancing the efficiency of on-chip decoupling capacitors with a switching voltage regulator is presented in Sect. 11.5. Finally, some conclusions are offered in Sect. 11.6.

11.1 Introduction to Decoupling Capacitance

Decoupling capacitors are often used to maintain the power supply voltage within specification so as to provide signal integrity while reducing electromagnetic interference (EMI) radiated noise. In this book, the use of decoupling capacitors to mitigate power supply noise is evaluated. The concept of a decoupling capacitor is described in this section. An historical retrospective is presented in Sect. 11.1.1. A description of a decoupling capacitor as a reservoir of charge is discussed in Sect. 11.1.2. Decoupling capacitors are shown to be an effective way to provide sufficient charge to a switching current load within a short period of time. A practical model of a decoupling capacitor is presented in Sect. 11.1.3.

11.1.1 Historical Retrospective

About 600 BC, Thales of Miletus recorded that the ancient Greeks could generate sparks by rubbing balls of amber on spindles [196]. This is the triboelectric effect [197], the mechanical separation of charge in a dielectric (insulator). This effect is the basis of the capacitor.

In October 1745, Ewald Georg von Kleist of Pomerania invented the first recorded capacitor: a glass jar coated inside and out with metal. The inner coating was connected to a rod that passed through the lid and ended in a metal sphere, as shown in Fig. 11.1 [198]. By layering the insulator between two metal plates, von Kleist dramatically increased the charge density. Before Kleist's discovery became widely known, a Dutch physicist, Pieter van Musschenbroek, independently invented a similar capacitor in January 1746 [199]. It was named the Leyden jar, after the University of Leyden where van Musschenbroek worked.

Benjamin Franklin investigated the Leyden jar and proved that the charge was stored on the glass, not in the water as others had assumed [200]. Originally, the units of capacitance were in "jars." A jar is equivalent to about 1 nF. Early capacitors were also known as *condensors*, a term that is still occasionally used today. The

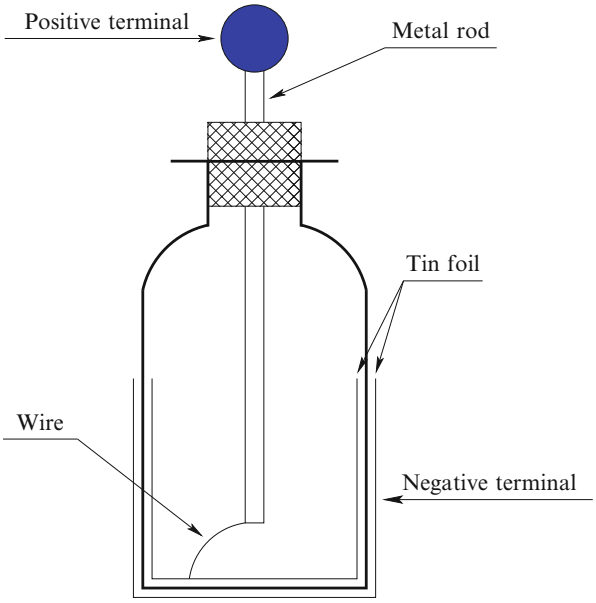


Fig. 11.1 Leyden jar originally developed by Ewald Georg von Kleist in 1745 and independently invented by Pieter van Musschenbroek in 1746. The charge is stored on the glass between two tin foils (capacitor plates) [198]

term condensor was coined by Alessandro Volta in 1782 (derived from the Italian *condensatore*), referencing the ability of a device to store a higher density of electric charge than a normal isolated conductor [200].

11.1.2 Decoupling Capacitor as a Reservoir of Charge

A capacitor consists of two electrodes, or plates, each of which stores an equal amount of opposite charge. These two plates are conductive and are separated by an insulator (dielectric). The charge is stored on the surface of the plates at the boundary with the dielectric. Since each plate stores an equal but opposite charge, the net charge across the capacitor is always zero.

The capacitance C of a capacitor is a measure of the amount of charge Q stored on each plate for a given potential difference (voltage V) which appears between the plates,

$$C = \frac{Q}{V}. \tag{11.1}$$

The capacitance is proportional to the surface area of the conducting plate and inversely proportional to the distance between the plates [201]. The capacitance

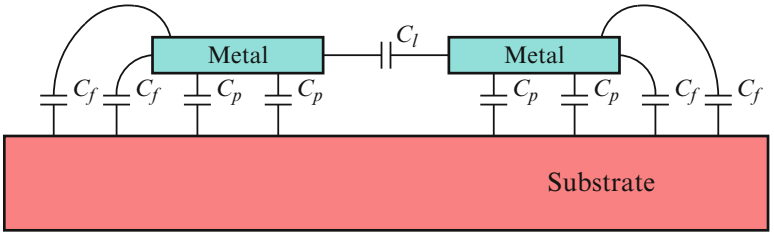


Fig. 11.2 Capacitance of two metal lines placed over a substrate. Three primary components compose the total capacitance of the on-chip metal interconnects. C_l denotes the lateral flux (side) capacitance, C_f denotes the fringe capacitance, and C_p denotes the parallel plate capacitance

is also proportional to the permittivity of the dielectric substance that separates the plates. The capacitance of a parallel plate capacitor is

$$C \approx \frac{\epsilon A}{d}, \tag{11.2}$$

where ϵ is the permittivity of the dielectric, A is the area of the plates, and d is the spacing between the plates. Equation (11.2) is only accurate for a plate area much greater than the spacing between the plates, $A \gg d^2$. In general, the capacitance of the metal interconnects placed over the substrate is composed of three primary components: a parallel plate capacitance, fringe capacitance, and lateral flux (side) capacitance [202], as shown in Fig. 11.2. Accurate closed-form expressions have been developed by numerically fitting a model that describes parallel lines above the plane or between two parallel planes [203–208].

As opposite charge accumulates on the plates of a capacitor across an insulator, a voltage develops across the capacitor due to the electric field formed by the opposite charge. Work must be done against this electric field as more charge is accumulated. The energy stored in a capacitor is equal to the amount of work required to establish the voltage across the capacitor. The energy stored in the capacitor is

$$E_{\text{stored}} = \frac{1}{2}CV^2 = \frac{1}{2}\frac{Q^2}{C} = \frac{1}{2}VQ. \tag{11.3}$$

From a physical perspective, a decoupling capacitor serves as an intermediate storage of charge and energy. The decoupling capacitor is located between the power supply and current load, i.e., electrically closer to the switching circuit. The decoupling capacitor is therefore more efficient in terms of supplying charge as compared to a remote power supply. The amount of charge stored on the decoupling capacitor is limited by the voltage and the capacitance. Unlike a decoupling capacitor, the power supply can provide an almost infinite amount of charge. A hydraulic model of a decoupling capacitor is illustrated in Fig. 11.3. Similar to water stored in a water tank and connected to the consumer through a system of pipes, the charge on the decoupling capacitor stored between the conductive plates

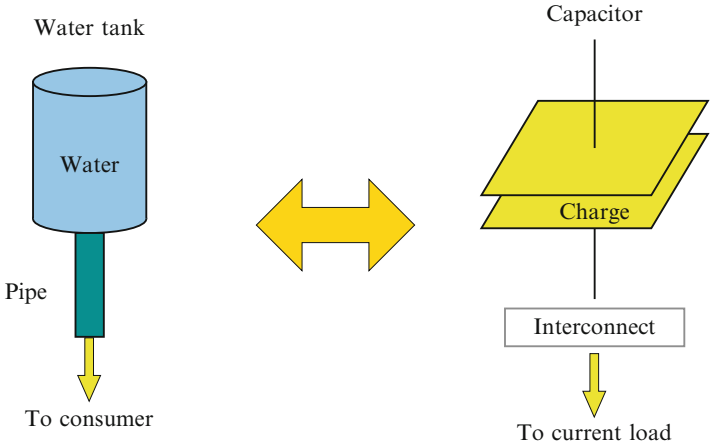


Fig. 11.3 Hydraulic model of a decoupling capacitor as a reservoir of charge. Similar to water stored in a water tank and connected to the consumer through a system of pipes, charge on the decoupling capacitor is stored between the conductive plates connected to the current load through a hierarchical interconnect system

is connected to the current load through a hierarchical interconnect system. To be effective, the decoupling capacitor should satisfy two requirements. First, the capacitor should have sufficient capacity to store a significant amount of energy. Second, to supply sufficient power at high frequencies, the capacitor should be able to release and accumulate energy at a high rate.

11.1.3 Practical Model of a Decoupling Capacitor

Decoupling capacitors are often used in power distribution systems to provide the required charge in a timely manner and to reduce the output impedance of the overall power delivery network [51]. An ideal decoupling capacitor is effective over the entire frequency range: from DC to the maximum operating frequency of a system. Practically, a decoupling capacitor is only effective over a certain frequency range. The impedance of a practical decoupling capacitor decreases linearly with frequency at low frequencies (with a slope of −20 dB/dec in a logarithmic scale). As the frequency increases, the impedance of the decoupling capacitor increases linearly with frequency (with a slope of 20 dB/dec in a logarithmic scale), as shown in Fig. 11.4. This increase in the impedance of a practical decoupling capacitor is due to the parasitic inductance of the decoupling capacitor. The parasitic inductance is referred to as the effective series inductance (ESL) of a decoupling capacitor [126].The impedance of a decoupling capacitor reaches the minimum impedance at the frequency $\omega = \frac{1}{\sqrt{LC}}$. This frequency is known as the resonant

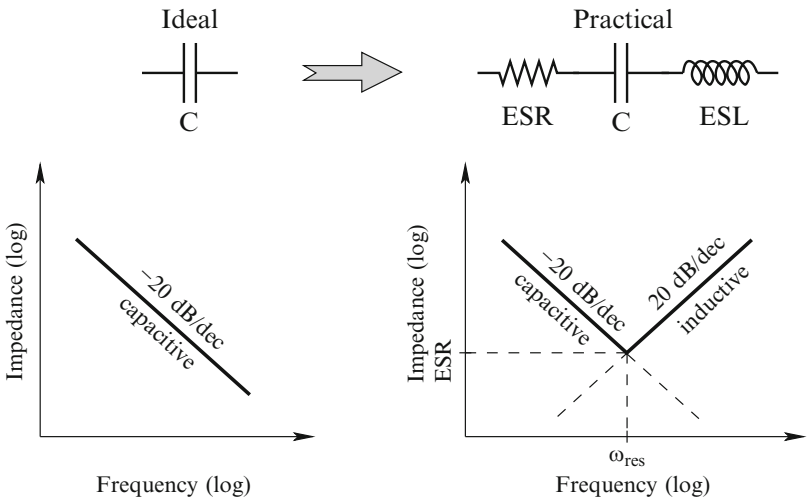


Fig. 11.4 Practical model of a decoupling capacitor. The impedance of a practical decoupling capacitor decreases linearly with frequency, reaching the minimum at a resonant frequency. Beyond the resonant frequency, the impedance of the decoupling capacitor increases linearly with frequency due to the ESL. The minimum impedance is determined by the ESR of the decoupling capacitor

frequency of a decoupling capacitor. Observe that the absolute minimum impedance of a decoupling capacitor is limited by the parasitic resistance, i.e., the effective series resistance (ESR) of a decoupling capacitor.The parasitic resistance of a decoupling capacitor is due to the resistance of the metal leads and conductive plates and the dielectric losses of the insulator. The ESR and ESL of an on-chip metal-oxide-semiconductor (MOS) decoupling capacitor are illustrated in Fig. 11.5. Note that the parasitic inductance of the decoupling capacitor is determined by the area of the current loops, decreasing with smaller area, as shown in Fig. 11.5b [209].

The impedance of a decoupling capacitor depends upon a number of characteristics. For instance, as the capacitance is increased, the capacitive curve moves down and to the right (see Fig. 11.4). Since the parasitic inductance for a particular capacitor is fixed, the inductive curve remains unaffected. As different capacitors are selected, the capacitive curve moves up and down relative to the fixed inductive curve. The primary way to decrease the total impedance of a decoupling capacitor for a specific semiconductor package is to increase the value of the capacitor [211]. Note that to move the inductive curve down, lowering the total impedance characteristics, a number of decoupling capacitors should be connected in parallel. In the case of identical capacitors, the total impedance is reduced by a factor of 2 for each doubling in the number of capacitors [136].

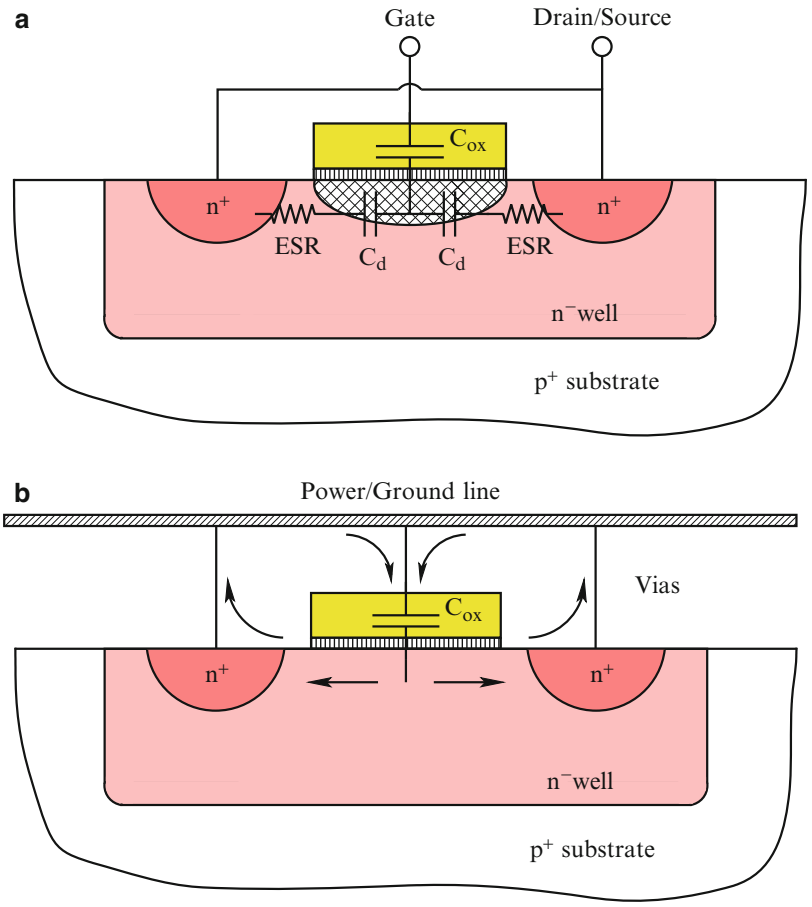


Fig. 11.5 Physical structure of an on-chip MOS decoupling capacitor. (**a**) ESR of an MOS-based decoupling capacitor. The ESR of an on-chip MOS decoupling capacitor is determined by the doping profiles of the n^+ regions and n^- well, the size of the capacitor, and the impedance of the vias and gate material [210]. (**b**) ESL of an MOS-based decoupling capacitor. The ESL of an on-chip MOS decoupling capacitor is determined by the area of the current return loops. The parasitic inductance is lowered by shrinking the area of the current return loops

11.2 Impedance of Power Distribution System with Decoupling Capacitors

As described in Sect. 11.1.2, a decoupling capacitor serves as a reservoir of charge, providing the required charge to the switching current load. Decoupling capacitors are also used to lower the impedance of the power distribution system. The impedance of a decoupling capacitor decreases rapidly with frequency, shunting the high frequency currents and reducing the effective current loop of a power distribution

network. The impedance of the overall power distribution system with decoupling capacitors is the subject of this section. In Sect. 11.2.1, the target impedance of a power distribution system is described. It is shown that the impedance of a power distribution system should be maintained below a target level to guarantee fault-free operation of the entire system. The antiresonance phenomenon is presented in Sect. 11.2.2. A hydraulic analogy of a system of decoupling capacitors is described in Sect. 11.2.3. The analogy is drawn between a water supply system and the hierarchical placement of decoupling capacitors at different levels of a power delivery network.

11.2.1 Target Impedance of a Power Distribution System

To ensure a small variation in the power supply voltage under a significant current load, the power distribution system should exhibit a small impedance as seen from the current load within the frequency range of interest [114]. A circuit network representing the impedance of a power distribution system as seen from the terminals of the current load is shown in Fig. 11.6.

The impedance of a power distribution system is with respect to the terminals of the load circuits. In order to ensure correct and reliable operation of an IC, the impedance of a power distribution system should be maintained below a certain upper bound Z_{target} in the frequency range from DC to the maximum operating frequency f_0 of the system [212–214]. The maximum tolerable impedance of a power distribution system is henceforth referred to as the target impedance. Note that the maximum operating frequency f_0 is determined by the switching time of the on-chip signal transients, rather than by the clock frequency. The shortest signal switching time is typically an order of magnitude smaller than the clock period. The maximum operating frequency is therefore considerably higher than the clock frequency.

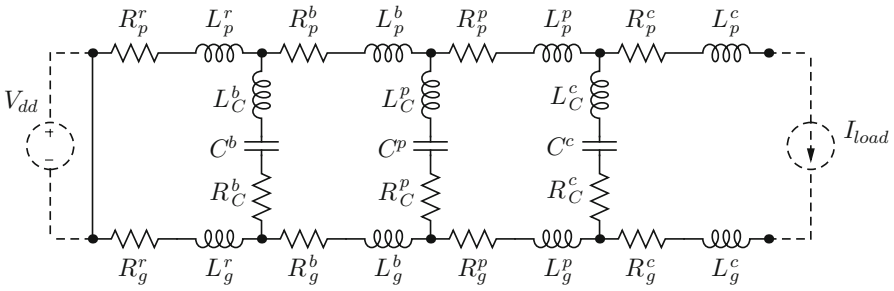


Fig. 11.6 A circuit network representing the impedance of a power distribution system with decoupling capacitors as seen from the terminals of the current load. The ESR and ESL of the decoupling capacitors are also included. Subscript *p* denotes the power paths and subscript *g* denotes the ground path. Superscripts *r*, *b*, *p*, and *c* refer, respectively, to the voltage regulator, board, package, and on-chip power delivery networks

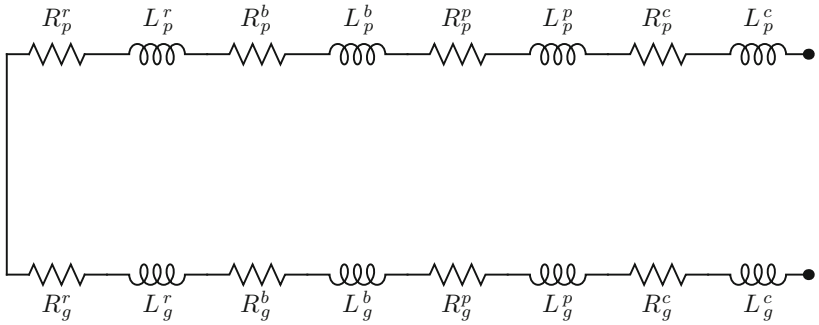


Fig. 11.7 A circuit network representing the impedance of a power distribution system without decoupling capacitors

One primary design objective of an effective power distribution system is to ensure that the output impedance of the network is below a target output impedance level. It is therefore important to understand how the output impedance of the circuit, shown schematically in Fig. 11.6, depends upon the impedance of the comprising circuit elements. A power distribution system with no decoupling capacitors is shown in Fig. 11.7. The power source and load are connected by interconnect with resistive and inductive parasitic impedances. The magnitude of the impedance of this network is

$$|Z_{\text{tot}}(\omega)| = |R_{\text{tot}} + j\omega L_{\text{tot}}|\,, \tag{11.4}$$

where R_{tot} and L_{tot} are the total resistance and inductance of the power distribution system, respectively,

$$R_{\text{tot}} = R_{\text{tot}}^p + R_{\text{tot}}^g, \tag{11.5}$$

$$R_{\text{tot}}^p = R_p^r + R_p^b + R_p^p + R_p^c, \tag{11.6}$$

$$R_{\text{tot}}^g = R_g^r + R_g^b + R_g^p + R_g^c, \tag{11.7}$$

$$L_{\text{tot}} = L_{\text{tot}}^p + L_{\text{tot}}^g, \tag{11.8}$$

$$L_{\text{tot}}^p = L_p^r + L_p^b + L_p^p + L_p^c, \tag{11.9}$$

$$L_{\text{tot}}^g = L_g^r + L_g^b + L_g^p + L_g^c. \tag{11.10}$$

The variation of the impedance with frequency is illustrated in Fig. 11.8. To satisfy a specification at low frequency, the resistance of the power delivery network should be sufficiently low, $R_{\text{tot}} < Z_{\text{target}}$. Above the frequency $f_{L_{\text{tot}}} = \frac{1}{2\pi}\frac{R_{\text{tot}}}{L_{\text{tot}}}$, however, the impedance of the power delivery network is dominated by the inductive reactance $j\omega L_{\text{tot}}$ and increases linearly with frequency, exceeding the target impedance at the frequency $f_{\text{max}} = \frac{1}{2\pi}\frac{Z_{\text{target}}}{L_{\text{tot}}}$.

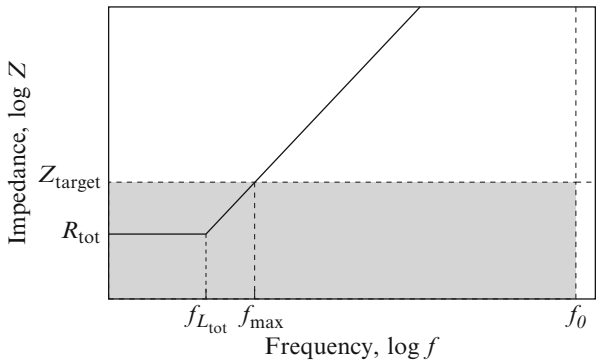


Fig. 11.8 Impedance of a power distribution system without decoupling capacitors. The *shaded area* denotes the target impedance specifications of the overall power distribution system

The high frequency impedance should be reduced to satisfy the target specifications. Opportunities for reducing the inductance of the power and ground paths of a power delivery network are limited [25, 215–218]. The inductance of the power distribution system is mainly determined by the board and package interconnects [219–221]. The feature size of the board and package level interconnect depends upon the manufacturing technology. The output impedance of a power distribution system is therefore highly inductive and is difficult to lower [134].

The high frequency impedance is effectively reduced by placing capacitors across the power and ground interconnections. These shunting capacitors effectively terminate the high frequency current loop, permitting the current to bypass the inductive interconnect, such as the board and package power delivery networks [222–225]. The high frequency impedance of the system as seen from the current load terminals is thereby reduced. Alternatively, at high frequencies, the capacitors decouple the high impedance paths of the power delivery network from the load. These capacitors are therefore referred to as decoupling capacitors [226, 227]. Several stages of decoupling capacitors are typically utilized to maintain the output impedance of a power distribution system below a target impedance [136, 228], as described in Sect. 11.2.3.

11.2.2 Antiresonance

Decoupling capacitors are a powerful technique to reduce the impedance of a power distribution system over a significant range of frequencies. A decoupling capacitor, however, reduces the resonant frequency of a power delivery network, making the system susceptible to resonances. Unlike the classic self-resonance in a series circuit formed by a decoupling capacitor combined with a parasitic resistance and inductance [138, 229] or by an on-chip decoupling capacitor and

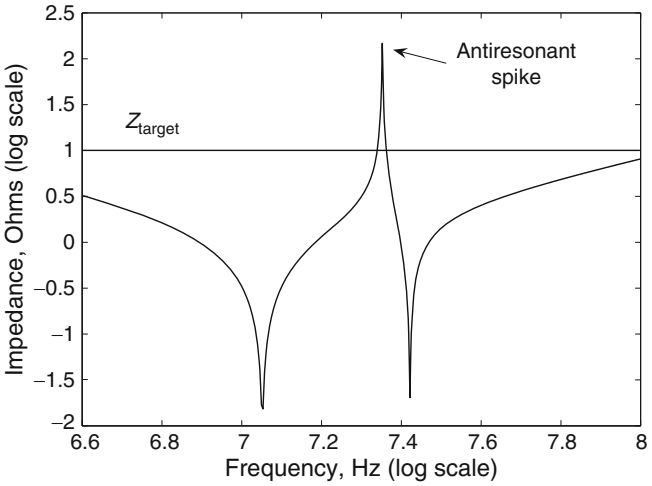


Fig. 11.9 Antiresonance of the output impedance of a power distribution network. Antiresonance results in a distinctive peak, exceeding the target impedance specification

the parasitic inductance of the package (i.e., chip-package resonance) [230, 231], antiresonance occurs in a circuit formed by two capacitors connected in parallel. At the resonant frequency, the impedance of the series circuit decreases in the vicinity of the resonant frequency, reaching the absolute minimum at the resonant frequency determined by the ESR of the decoupling capacitor. At antiresonance, however, the circuit impedance drastically increases, producing a distinctive peak, as illustrated in Fig. 11.9. This antiresonant peak can result in system failures as the impedance of the power distribution system becomes greater than the maximum tolerable impedance Z_{target}. The antiresonance phenomenon in a system with parallel decoupling capacitors is the subject of this section.

To achieve a low impedance power distribution system, multiple decoupling capacitors are placed in parallel. The effective impedance of a power distribution system with several identical capacitors placed in parallel is illustrated in Fig. 11.10. Observe that the impedance of the power delivery network is reduced by a factor of 2 as the number of capacitors is doubled. Also note that the effective drop in the impedance of a power distribution system diminishes rapidly with each additional decoupling capacitor. It is therefore desirable to utilize decoupling capacitors with a sufficiently low ESR in order to minimize the number of capacitors required to satisfy a target impedance specification [136].

A number of decoupling capacitors with different magnitudes is typically used to maintain the impedance of a power delivery system below a target specification over a wide frequency range. Capacitors with different magnitudes connected in parallel, however, result in a sharp antiresonant peak in the system impedance [29]. The antiresonance phenomenon for different capacitive values is illustrated in Fig. 11.11. The antiresonance of parallel decoupling capacitors can be explained as

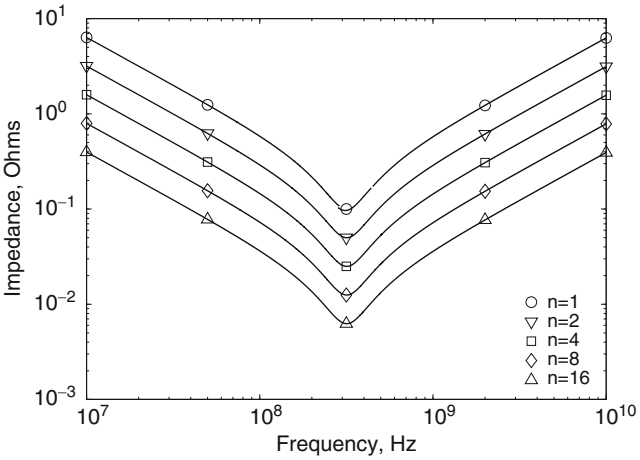


Fig. 11.10 Impedance of a power distribution system with n identical decoupling capacitors connected in parallel. The ESR of each decoupling capacitor is $R = 0.1\,\Omega$, the ESL is $L = 100\,\text{pH}$, and the capacitance is $C = 1\,\text{nF}$. The impedance of a power distribution system is reduced by a factor of 2 as the number of capacitors is doubled

follows. In the frequency range from f_1 to f_2, the impedance of the capacitor C_1 has become inductive whereas the impedance of the capacitor C_2 remains capacitive (see Fig. 11.11). Thus, an LC tank is formed in the frequency range from f_1 to f_2, producing a peak at the resonant frequency located between f_1 and f_2. As a result, the total impedance drastically increases and becomes greater than the target impedance, causing a system to fail.

The magnitude of the antiresonant spike can be effectively reduced by lowering the parasitic inductance of the decoupling capacitors. For instance, as discussed in [136], the magnitude of the antiresonant spike is significantly reduced if board decoupling capacitors are mounted on low inductance pads. The magnitude of the antiresonant spike is also determined by the ESR of the decoupling capacitor, decreasing with larger parasitic resistance. Large antiresonant spikes are produced when low ESR decoupling capacitors are placed on inductive pads. A high inductance and low resistance result in a parallel LC circuit with a high quality factor Q,

$$Q = \frac{L}{R}. \tag{11.11}$$

In this case, the magnitude of the antiresonant spike is amplified by Q. Decoupling capacitors with a low ESR should therefore always be used on low inductance pads (with a low ESL).

Antiresonance also becomes well pronounced if a large variation exists between the capacitance values. This phenomenon is illustrated in Fig. 11.12. In the case of two capacitors with distinctive nominal values ($C_1 \gg C_2$), a significant gap between

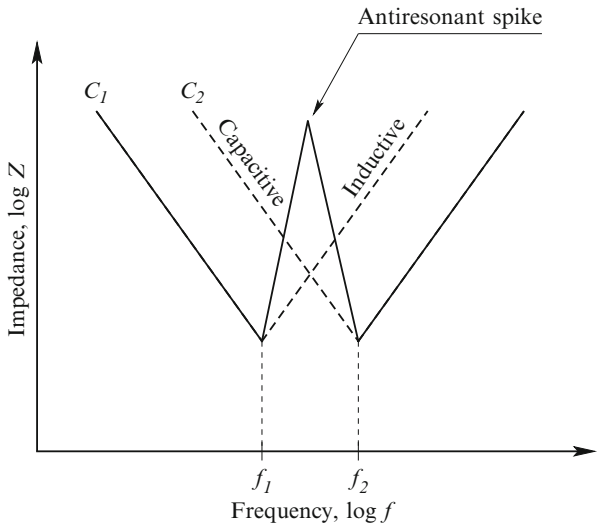


Fig. 11.11 Antiresonance of parallel capacitors, $C_1 > C_2$, $L_1 = L_2$, and $R_1 = R_2$. A parallel LC tank is formed in the frequency range from f_1 to f_2. The total impedance drastically increases in the frequency range from f_1 to f_2 (*the solid line*), producing an antiresonant spike

two capacitances results in a sharp antiresonant spike with a large magnitude in the frequency range from f_1 to f_2, violating the target specification Z_{target}, as shown in Fig. 11.12a. If another capacitor with nominal value $C_1 > C_3 > C_2$ is added, the antiresonant spike is canceled by C_3 in the frequency range from f_1 to f_2. As a result, the overall impedance of a power distribution system is maintained below the target specification over a broader frequency range, as shown in Fig. 11.12b. As described in [232], the high frequency impedance of two parallel decoupling capacitors is only reduced by a factor of 2 (or 6 dB) as compared to a single capacitor. It is also shown that adding a smaller capacitor in parallel with a large capacitor results in only a small reduction in the high frequency impedance. Antiresonances are effectively managed by utilizing decoupling capacitors with a low ESL and by placing a greater number of decoupling capacitors with progressively decreasing magnitude, shifting the antiresonant spike to the higher frequencies (out of the range of the operating frequencies of the circuit) [233].

11.2.3 Hydraulic Analogy of Hierarchical Placement of Decoupling Capacitors

As discussed in Sect. 11.1.2, an ideal decoupling capacitor should provide a high capacity and be able to release and accumulate energy at a sufficiently high rate. Constructing a device with both high energy capacity and high power capability

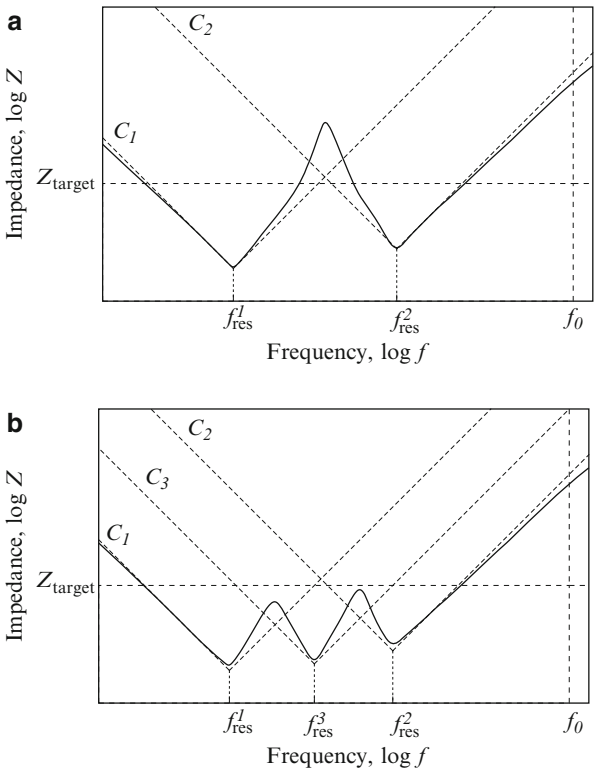


Fig. 11.12 Antiresonance of parallel capacitors. (**a**) A large gap between two capacitances results in a sharp antiresonant spike with a large magnitude in the frequency range from f_1 to f_2, violating the target specification Z_{target}. (**b**) If another capacitor with magnitude $C_1 > C_3 > C_2$ is added, the antiresonant spike is canceled by C_3 in the frequency range from f_1 to f_2. As a result, the overall impedance of the power distribution system is maintained below the target specification over the desired frequency range

is, however, challenging. It is expensive to satisfy both of these requirements in an ideal decoupling capacitor. Moreover, these requirements are typically contradictory in most practical applications. The physical realization of a large decoupling capacitance requires the use of discrete capacitors with a large nominal capacity and, consequently, a large form factor. The large physical dimensions of the capacitors have two implications. The parasitic series inductance of a physically large capacitor is relatively high due to the increased area of the current loop within the capacitors. Furthermore, due to technology limitations, the large physical size of the capacitors prevents placing the capacitors sufficiently close to the current load. A greater physical separation increases the inductance of the current path from the capacitors to the load. A tradeoff therefore exists between the high capacity and low parasitic inductance of a decoupling capacitor for an available component technology.

Gate switching times of a few tens of picoseconds are common in modern high performance ICs, creating high transient currents in the power distribution system. At high frequencies, only those on-chip decoupling capacitors with a low ESR and a low ESL can effectively maintain a low impedance power distribution system. Placing a sufficiently large on-chip decoupling capacitor requires a die area many times greater than the area of a typical circuit. Thus, while technically feasible, a single-tier decoupling solution is prohibitively expensive. A large on-chip decoupling capacitor is therefore typically built as a series of small decoupling capacitors connected in parallel. At high frequencies, a large on-chip decoupling capacitor exhibits a distributed behavior. Only on-chip decoupling capacitors located in the vicinity of the switching circuit can effectively provide the required charge to the current load within the proper time. An efficient approach to this problem is to hierarchically place multiple stages of decoupling capacitors, progressively smaller and closer to the load.

Utilizing hierarchically placed decoupling capacitors produces a low impedance, high frequency power distribution system realized in a cost effective way. The capacitors are placed in several stages: on the board, package, and circuit die. Arranging the decoupling capacitors in several stages eliminates the need to satisfy both the high capacitance and low inductance requirements in the same decoupling stage [30].

The hydraulic analogy of the hierarchical placement of decoupling capacitors is shown in Fig. 11.13. Each decoupling capacitor is represented by a water tank. All of the water tanks are connected to the main water pipe connected to the consumer (current load). Water tanks at different stages are connected to the main pipe through the local water pipes, modeling different interconnect levels. The goal of the water supply system (power delivery network) is to provide uninterrupted water flow to the consumer at the required rate (switching time). The amount of water released by each water tank is proportional to the tank size. The rate at which the water tank is capable of providing water is inversely proportional to the size of the water tank and directly proportional to the distance from the consumer to the water tank.

A power supply is typically treated as an infinite amount of charge. Due to large physical dimensions, the power supply cannot be placed close to the current load (the consumer). The power supply therefore has a long response time. Unlike the power supply, an on-chip decoupling capacitor can be placed sufficiently close to the consumer. The response time of an on-chip decoupling capacitor is significantly shorter as compared to the power supply. An on-chip decoupling capacitor is therefore able to respond to the consumer demand in a much shorter period of time but is capable of providing only a small amount of water (or charge). Allocating decoupling capacitors with progressively decreasing magnitudes and closer to the current load, an uninterrupted flow of charge can be provided to the consumer. In the initial moment, charge is only supplied to the consumer by the on-chip decoupling capacitor. As the on-chip decoupling capacitor is depleted, the package decoupling capacitor is engaged. This process continues until the power supply is activated. Finally, the power supply is turned on and provides the necessary charge with relatively relaxed timing constraints. The voltage regulator, board, package, and

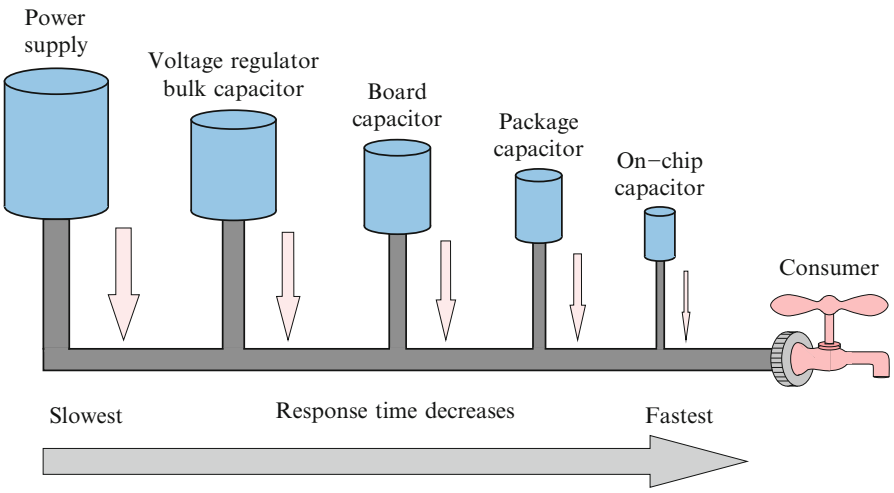


Fig. 11.13 Hydraulic analogy of the hierarchical placement of decoupling capacitors. The decoupling capacitors are represented by the water tanks. The response time is proportional to the size of the capacitor and inversely proportional to the distance from a capacitor to the consumer. The on-chip decoupling capacitor has the shortest response time (located closest to the consumer), but is capable of providing the least amount of charge

on-chip decoupling capacitors therefore serve as intermediate reservoirs of charge, relaxing the timing constraints for the power delivery supply.

A hierarchy of decoupling capacitors is utilized in high performance power distribution systems in order to extend the frequency region of the low impedance characteristics to the maximum operating frequency f_0. The impedance characteristics of a power distribution system with board, package, and on-chip decoupling capacitors (see Fig. 11.6) are illustrated in Fig. 11.14. By utilizing the hierarchical placement of decoupling capacitors, the antiresonant spike is shifted outside the range of operating frequencies (beyond f_0). The overall impedance of a power distribution system is also maintained below the target impedance over the entire frequency range of interest (from DC to f_0).

Fully Compensated System

A special case in the impedance of an *RLC* circuit formed by a decoupling capacitor and the parasitic inductance of the P/G lines is achieved when the zeros of a tank circuit impedance cancel the poles, making the impedance purely resistive and independent of frequency,

$$R_L = R_C = R_0 = \sqrt{\frac{L}{C}}, \tag{11.12}$$

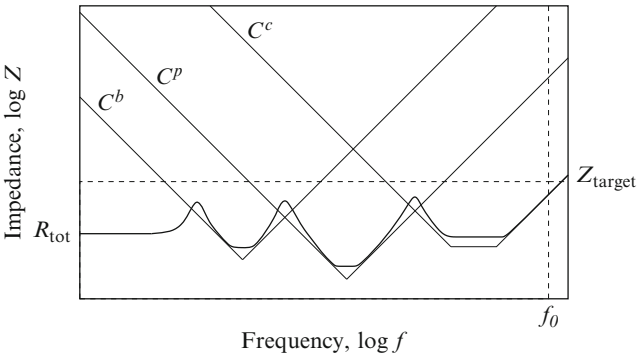


Fig. 11.14 Impedance of a power distribution system with board, package, and on-chip decoupling capacitances. The overall impedance is shown with a *black line*. The impedance of a power distribution system with three levels of decoupling capacitors is maintained below the target impedance (*dashed line*) over the frequency range of interest. The impedance characteristics of the decoupling capacitors are shown by the *thin solid lines*

$$\frac{L}{R_L} = R_C C, \tag{11.13}$$

where R_L and R_C are, respectively, the parasitic resistance of the P/G lines and the ESR of the decoupling capacitor. In this case, the impedance of the *RLC* tank is fully compensated. Equations (11.12) and (11.13) are equivalent to two conditions, i.e., the impedance at the lower frequencies is matched to the impedance at the high frequencies and the time constants of the inductor and capacitor currents are also matched. A constant, purely resistive impedance, characterizing a power distribution system with decoupling capacitors, is achieved across the entire frequency range of interest, if each decoupling stage is fully compensated [137, 234]. The resistance and capacitance of the decoupling capacitors in a fully compensated system are completely determined by the impedance characteristics of the power and ground interconnect and the location of the decoupling capacitors.

The hierarchical placement of decoupling capacitors exploits the tradeoff between the capacity and the parasitic inductance of a capacitor to achieve an economically effective solution. The total decoupling capacitance of a hierarchical scheme $C_{\text{total}} = C^b + C^p + C^c$ is larger than the total decoupling capacitance of a single-tier solution, where C^b, C^p, and C^c are, respectively, the board, package, and on-chip decoupling capacitances. The primary advantage of utilizing a hierarchical placement is that the inductive limit is imposed only on the final stage of decoupling capacitors which constitutes a small fraction of the total required decoupling capacitance. The constraints on the physical dimensions and parasitic impedance of the capacitors in the remaining stages are therefore significantly reduced. As a result, cost efficient electrolytic and ceramic capacitors can be used to provide medium size and high capacity decoupling capacitors [30].

11.3 Intrinsic vs Intentional On-Chip Decoupling Capacitance

Several types of on-chip capacitances contribute to the overall on-chip decoupling capacitance. The *intrinsic* decoupling capacitance is the inherent capacitance of the transistors and interconnects that exists between the power and ground terminals. The thin gate oxide capacitors placed on-chip to solely provide power decoupling are henceforth referred to as an *intentional* decoupling capacitance. The intrinsic decoupling capacitance is described in Sect. 11.3.1. The intentional decoupling capacitance is reviewed in Sect. 11.3.2.

11.3.1 Intrinsic Decoupling Capacitance

An intrinsic decoupling capacitance (or symbiotic capacitance) is the parasitic capacitance between the power and ground terminals within an on-chip circuit structure. The intrinsic capacitance is comprised of three types of parasitic capacitances [235].

One component of the intrinsic capacitance is the parasitic capacitance of the interconnect lines. Three types of intrinsic interconnect capacitances are illustrated in Fig. 11.15. The first type of interconnect capacitance is the capacitance C_1^i between the signal line and the power/ground line. Capacitance C_2^i is the capacitance between signal lines at different voltage potentials. The third type of intrinsic interconnect capacitance is the capacitance C_3^i between the power and ground lines (see Fig. 11.15).

Parasitic device capacitances, such as the drain junction capacitance and gate-to-source capacitance, also contribute to the overall intrinsic decoupling capacitance where the terminals of the capacitance are connected to power and ground. For example, in the simple inverter circuit depicted in Fig. 11.16, if the input is one (high) and the output is zero (low), the NMOS transistor is turned on, connecting C_p from V_{dd} to G_{nd}, providing a decoupling capacitance to the other switching circuits,

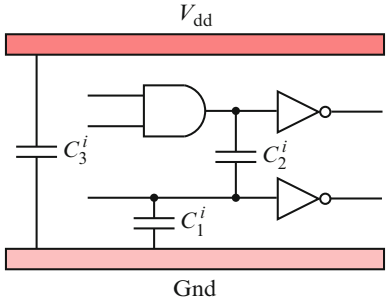


Fig. 11.15 Intrinsic decoupling capacitance of the *interconnect lines*. C_1^i denotes the capacitance between the signal line and the power/ground line. C_2^i denotes the capacitance between signal lines. C_3^i denotes the capacitance between the power and ground lines

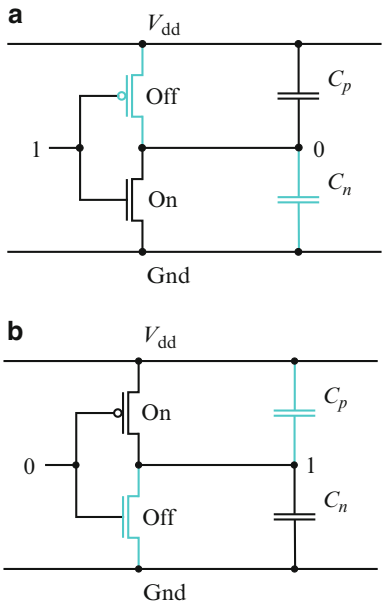


Fig. 11.16 Intrinsic decoupling capacitance of a non-switching circuit; (**a**) inverter input is high, (**b**) inverter input is low

as illustrated in Fig. 11.16a. Alternatively, if the input is zero (low) and the output is one (high), the PMOS transistor is turned on, connecting C_n from G_{nd} to V_{dd}, providing a decoupling capacitance to the other switching circuits, as illustrated in Fig. 11.16b.

Depending upon the total capacitance ($C_p + C_n$) and the switching factor SF, the decoupling capacitance from the non-switching circuits is [236]

$$C_{\text{circuit}} = \frac{P}{V_{\text{dd}}^2 f} \frac{(1 - SF)}{SF}, \tag{11.14}$$

where P is the circuit power, V_{dd} is the power supply voltage, and f is the switching frequency. The time constant for C_{circuit} is determined by $R_{\text{PMOS}} C_n$ or $R_{\text{NMOS}} C_p$ and usually varies in a 0.18 μm CMOS technology from about 50–250 ps [236].

The contribution of the transistor and interconnect capacitance to the overall decoupling capacitance is difficult to determine precisely. The transistor terminals as well as the signal lines can be connected either to power or ground, depending upon the internal state of the digital circuit at a particular time. The transistor and interconnect decoupling capacitance therefore depends on the input pattern and the internal state of the circuit. The input vectors that produce the maximum intrinsic decoupling capacitance in a digital circuit are described in [237].

Another source of intrinsic capacitance is the *p-n* junction capacitance of the diffusion wells. The N-type wells, P-type wells, or wells of both types are implanted into a silicon substrate to provide an appropriate body doping for the PMOS and NMOS transistors. The N-type wells are ohmically connected to the power supply

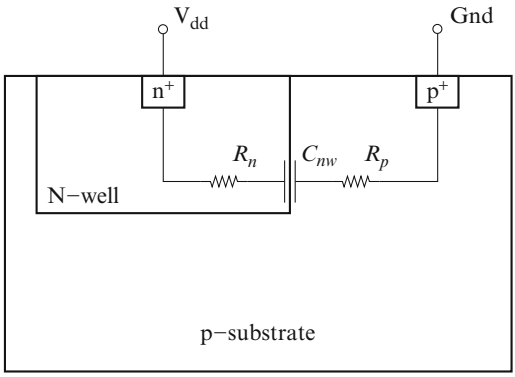


Fig. 11.17 N-well junction intrinsic decoupling capacitance. The capacitor C_{nw} is formed by the reverse-biased *p-n* junction between the N-well and the p-substrate

while the P-type wells are connected to ground to provide a proper body bias for the transistors. The N-well capacitor is the reverse-biased *p-n* junction capacitor between the N-well and p-substrate, as shown in Fig. 11.17. The total on-chip N-well decoupling capacitance C_{nw} is determined by the area, perimeter, and depth of each N-well. Multiplying C_{nw} by the series and contact resistance in the N-well and p-substrate, the time constant $(R_p + R_n + R_{\text{contact}})C_{nw}$ for an N-well capacitor is typically in the range of 250–500 ps in a 0.18 µm CMOS technology [236]. The parasitic capacitance of the wells usually dominates the intrinsic decoupling capacitance of ICs fabricated in an epitaxial CMOS process [177, 238]. The overall intrinsic on-chip decoupling capacitance consists of several components and is

$$C_{\text{intrinsic}} = C_{\text{inter}} + C_{pn} + C_{\text{well}} + C_{\text{load}} + C_{gs} + C_{gb}, \tag{11.15}$$

where C_{inter} is the interconnect capacitance, C_{pn} is the *p-n* junction capacitance, C_{well} is the capacitance of the well, C_{load} is the load capacitance, C_{gs} is the gate-to-source (drain) capacitance, and C_{gb} is the gate-to-body capacitance.

Silicon-on-insulator (SOI) CMOS circuits lack diffusion wells and therefore do not contribute to the intrinsic on-chip decoupling capacitance. A reliable estimate of the contribution of the interconnect and transistors to the on-chip decoupling capacitance is thus particularly important in SOI circuits. Several techniques for estimating the intrinsic decoupling capacitance are presented in [140, 239]. The overall intrinsic decoupling capacitance of an IC can also be determined experimentally. In [240], the signal response of a power distribution system versus frequency is measured with a vector network analyzer. An *RLC* model of the system is constructed to match the observed response. The magnitude of the total on-chip decoupling capacitance is determined from the frequency of the resonant peaks in the response of the power system. Alternatively, the total on-chip decoupling capacitance can be experimentally determined from the package-chip resonance, as described in [231]. The intentional decoupling capacitance placed on-chip during the design process is known within the margins of the process variations. Subtracting the intentional capacitance from the measured overall capacitance yields an estimate of the on-chip intrinsic capacitance.

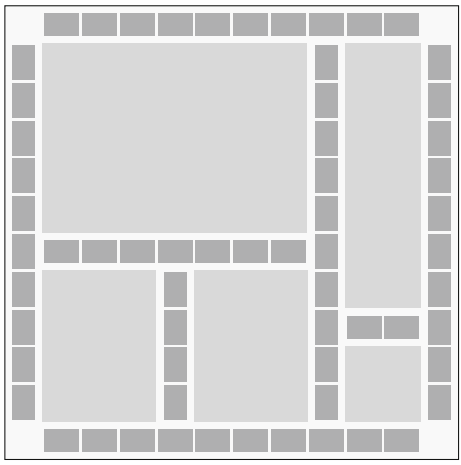


Fig. 11.18 Banks of on-chip decoupling capacitors (*the dark gray rectangles*) placed among circuit blocks (*the light gray rectangles*)

11.3.2 Intentional Decoupling Capacitance

Intentional decoupling capacitance is often added to a circuit during the design process to increase the overall on-chip decoupling capacitance to a satisfactory level. The intentional decoupling capacitance is typically realized as a gate capacitance in large MOS transistors placed on-chip specifically for this purpose. In systems with mixed memory and logic, however, the intentional capacitance can also be realized as a trench capacitance [241, 242].

Banks of MOS decoupling capacitors are typically placed among the on-chip circuit blocks, as shown in Fig. 11.18. The space between the circuit blocks is often referred to as "white" space, as this area is primarily used for global routing and does not contain any active devices. Unless noted otherwise, the term "on-chip decoupling capacitance" commonly refers to the intentional decoupling capacitance. Using more than 20 % of the overall die area for intentional on-chip decoupling capacitance is common in modern high speed integrated circuits [149, 243].

An MOS capacitor uses the thin oxide layer between the N-well and polysilicon gate to provide the additional decoupling capacitance needed to mitigate the power noise, as shown in Fig. 11.19. An optional fuse (or control gate) is typically provided to disconnect the thin oxide capacitor from the rest of the circuit in the undesirable situation of a short circuit due to process defects. As the size and shape of MOS capacitors vary, the $R_n C_{ox}$ time constant typically ranges from 40 to 200 ps in a 0.18 μm CMOS process. Depending upon the switching speed of the circuit, typical on-chip MOS decoupling capacitors are effective for RC time constants below 200 ps [236].

An MOS capacitor is formed by the gate electrode on one side of the oxide layer and the source-drain inversion channel under the gate on the other side of the oxide layer. The resistance of the channel dominates the ESR of the MOS capacitor. Due to the resistance of the transistor channel, the MOS capacitor is modeled as

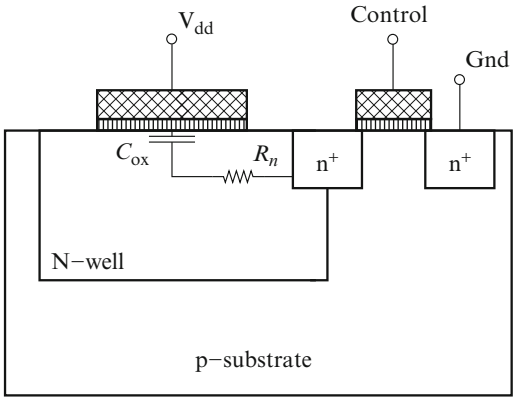


Fig. 11.19 Thin oxide MOS decoupling capacitor

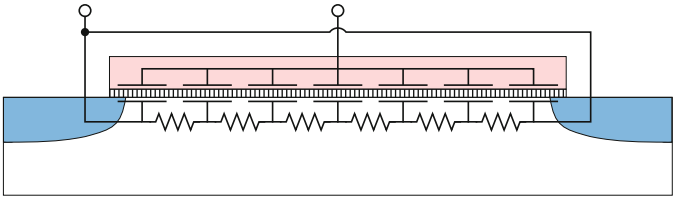


Fig. 11.20 Equivalent *RC* model of an MOS decoupling capacitor

a distributed *RC* circuit, as shown in Fig. 11.20. The impedance of the distributed *RC* structure shown in Fig. 11.20 is frequency dependent, $Z(\omega) = R(\omega) + \frac{1}{j\omega C(\omega)}$. Both the resistance $R(\omega)$ and capacitance $C(\omega)$ decrease with frequency. The low frequency resistance of the MOS capacitor is approximately one twelfth of the source-drain resistance of the MOS transistor in the linear region [244]. The low frequency capacitance is the entire gate-to-channel capacitance of the transistor. At high frequencies, the gate-to-channel capacitance midway between the drain and source is shielded from the capacitor terminals by the resistance of the channel, decreasing the effective capacitance of the MOS capacitor. The higher the channel resistance per transistor width, the lower the frequency at which the capacitor efficiency begins to decrease. Capacitors with a long channel (with a relatively high channel resistance) are therefore less effective at high frequencies as compared to short-channel capacitors. A higher series resistance of the on-chip MOS decoupling capacitor, however, is beneficial in damping the resonance of a die-package *RLC* tank circuit [244].

Long channel transistors, however, are more area efficient. In transistors with a minimum length channel, the source and body contacts dominate the transistor area, while the MOS capacitor stack occupies a relatively small fraction of the total area. For longer channels, the area of the MOS capacitor increases while the area overhead of the source/drain contacts remain constant, increasing the capacitance per total area [30]. A tradeoff therefore exists between the area efficiency and the ESR of the MOS decoupling capacitor. Transistors with a channel length 12 times

greater than the minimum length are a good compromise [244]. In this case, the *RC* time constant is smaller than the switching time of the logic gates, which typically are composed of transistors with a minimum channel length, while the source and drain contacts occupy a relatively small fraction of the total area.

11.4 Types of On-Chip Decoupling Capacitors

Multiple on-chip capacitors are utilized in ICs to satisfy various design requirements. Four types of widely utilized on-chip decoupling capacitors are the subject of this section. Polysilicon-insulator-polysilicon (PIP) capacitors are presented in Sect. 11.4.1. Three types of MOS decoupling capacitors, accumulation, depletion, and inversion, are described in Sect. 11.4.2. Metal-insulator-metal (MIM) decoupling capacitors are reviewed in Sect. 11.4.3. In Sect. 11.4.4, lateral flux decoupling capacitors are described. The design and performance characteristics of the different on-chip decoupling capacitors are compared in Sect. 11.4.5.

11.4.1 Polysilicon-Insulator-Polysilicon (PIP) Capacitors

Both junction and MOS capacitors use diffusion for the lower electrodes. The junction isolating the diffused electrode exhibits substantial parasitic capacitance, limiting the voltage applied across the capacitor. These limitations are circumvented in PIP capacitors, which employ two polysilicon electrodes in combination with either an oxide or an oxide-nitride-oxide (ONO) dielectric [245], as illustrated in Fig. 11.21. Since typical CMOS and BiCMOS processes incorporate multiple polysilicon layers, PIP capacitors do not require any additional masking steps. The gate polysilicon can serve as the lower electrode of the PIP capacitor, while the resistor polysilicon (doped with a suitable implant) can form the upper electrode. The upper electrode is typically doped with either an N-type source/drain (NSD) or P-type source/drain (PSD) implant. The implant resulting in the lowest sheet resistance is preferable, since heavier doping reduces the ESR and minimizes voltage modulation due to polysilicon depletion [245].

PIP capacitors require additional process steps. Even if both of the electrodes consist of existing depositions, the capacitor dielectric is unique to this structure and consequently requires a process extension. The simplest way to form this dielectric is to eliminate the interlevel oxide (ILO) deposition that normally separates the two polysilicon layers and add a thin oxide layer on the lower polysilicon electrode. With this technique, a capacitor can be built between the two polysilicon layers as long as the second polysilicon layer is not used as an interconnection.

Silicon dioxide has a relatively low permittivity. A higher permittivity, and therefore a higher capacitance per unit area, is achieved using a stacked ONO dielectric (see Fig. 11.21b). Observe from Fig. 11.21 that the PIP capacitors normally

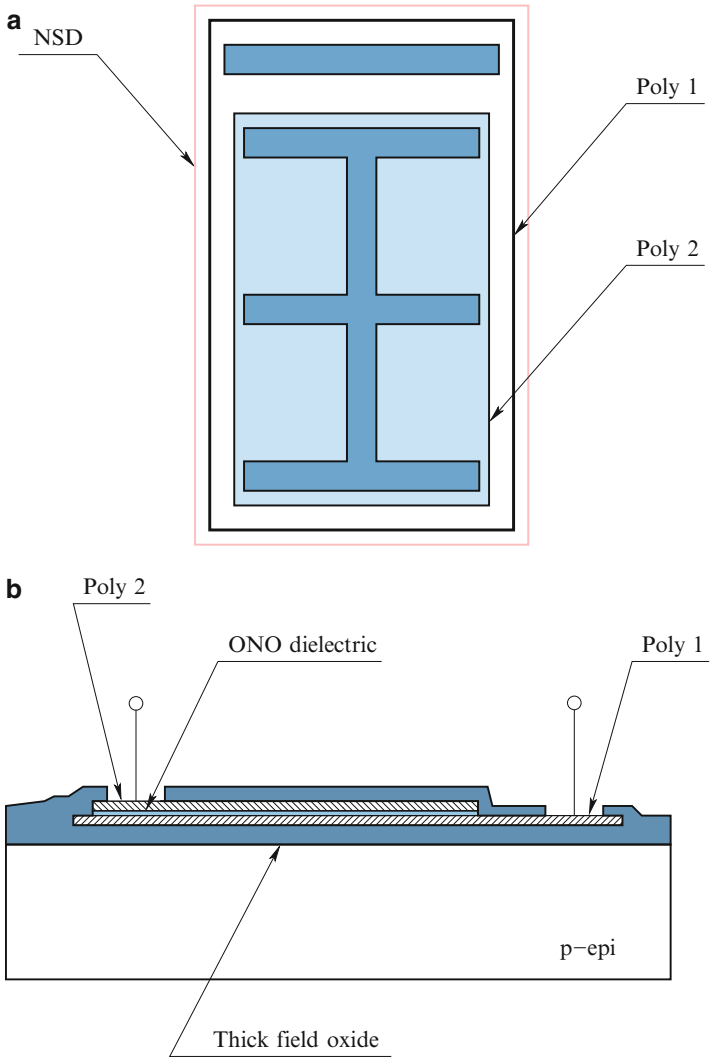


Fig. 11.21 PIP oxide-nitride-oxide (ONO) capacitor. The entire capacitor is enclosed in an N-type source/drain region, reducing the sheet resistance of the polysilicon layer; (**a**) layout, (**b**) cross section

reside over the field oxide. The oxide steps should not intersect the structure, since those steps cause surface irregularities in the lower capacitor electrode, resulting in localized thinning of the dielectric, thereby concentrating the electric field. As a result of the intersection, the breakdown voltage of the capacitor can be severely compromised.

Selecting the dielectric material in a PIP capacitor, several additional issues should be considered. Composite dielectrics experience hysteresis effects at high

frequencies (above 10 MHz) due to the incomplete redistribution of static charge along the oxide-nitride interface. Pure oxide dielectrics are used for PIP capacitors to achieve a relatively constant capacitance over a wide frequency range. Oxide dielectrics, however, typically have a lower capacitance per unit area. Low capacitance dielectrics are also useful for improving matching among the small capacitors.

Voltage modulation of the PIP capacitors is relatively small, as long as both electrodes are heavily doped. A PIP capacitor typically exhibits a voltage modulation of 150 ppm/V [245]. The temperature coefficient of a PIP capacitor also depends on voltage modulation effects and is typically less than 250 ppm/°C [246].

11.4.2 MOS Capacitors

An MOS capacitor consists of a metal-oxide-semiconductor structure, as illustrated in Fig. 11.22. A top metal contact is referred to as the gate, serving as one plate of the capacitor. In digital CMOS ICs, the gate is often fabricated as a heavily doped n^+-polysilicon layer, behaving as a metal. A second metal layer forms an ohmic contact to the back of the semiconductor and is called the bulk contact. The semiconductor layer serves as the other plate of the capacitor. The bulk resistivity is typically 1–10 $\Omega \cdot$cm (with a doping of 10^{15} cm^{-3}).

The capacitance of an MOS capacitor depends upon the voltage applied to the gate with respect to the body. The dependence of the capacitance upon the voltage across an MOS capacitor (a capacitance versus voltage (CV) diagram) is plotted in Fig. 11.23. Depending upon the gate-to-body potential V_{gb}, three regions of operation are distinguished in the CV diagram of an MOS capacitor. In the accumulation mode, mobile carriers of the same type as the body (holes for an NMOS capacitor with a p-substrate) accumulate at the surface. In the depletion mode, the surface is devoid of any mobile carriers, leaving only a space charge (depletion layer). In the inversion mode, mobile carriers of the opposite type of the body (electrons for an NMOS capacitor with a p-substrate) aggregate at the surface, inverting the conductivity type. These three regimes are roughly separated by the

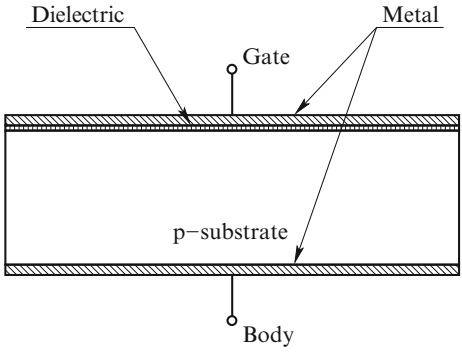


Fig. 11.22 The structure of an n-type MOS capacitor

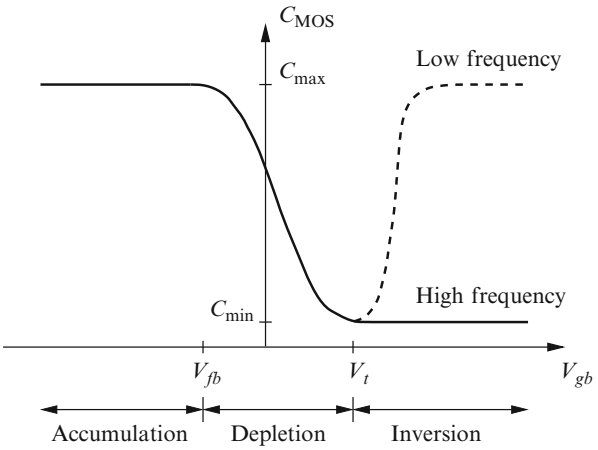


Fig. 11.23 Capacitance versus gate voltage (CV) diagram of an n-type MOS capacitor. The flat band voltage V_{fb} separates the accumulation region from the depletion region. The threshold voltage V_t separates the depletion region from the inversion region

two voltages (see Fig. 11.23). A flat band voltage V_{fb} separates the accumulation regime from the depletion regime. The threshold voltage V_t demarcates the depletion regime from the inversion regime. Based on the mode of operation, three types of MOS decoupling capacitors exist and are described in the following three subsections.

Accumulation

In MOS capacitors operating in accumulation, the applied gate voltage is lower than the flat band voltage ($V_{gb} < V_{fb}$) and induces negative charge on the metal gate and positive charge in the semiconductor. The hole concentration at the surface is therefore above the bulk value, leading to surface accumulation. The charge distribution in an MOS capacitor operating in accumulation is shown in Fig. 11.24. The flat band voltage is the voltage at which there is no charge on the plates of the capacitor (there is no electric field across the dielectric). The flat band voltage depends upon the doping of the semiconductor and any residual charge existing at the interface between the semiconductor and the insulator. In the accumulation mode, the charge per unit area Q_n at the semiconductor/oxide interface is a linear function of the applied voltage V_{gb}. The oxide capacitance per unit area C_{ox} is determined by the slope of Q_n, as illustrated in Fig. 11.25. The capacitance of an MOS capacitor operating in accumulation achieves the maximum value and is

$$C_{\mathrm{MOS}_{\mathrm{accum}}} = C_{\max} = A\, C_{\mathrm{ox}} = A\, \frac{\epsilon_{\mathrm{ox}}}{t_{\mathrm{ox}}}, \tag{11.16}$$

where A is the area of the gate electrode, ϵ_{ox} is the permittivity of the oxide, and t_{ox} is the oxide thickness.

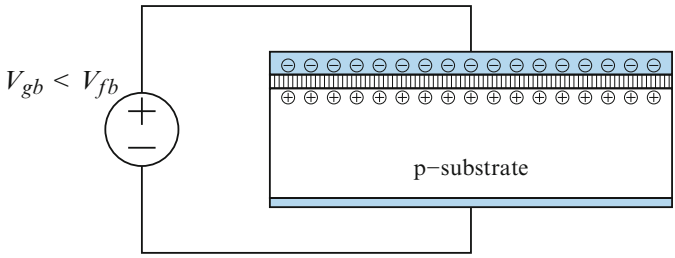


Fig. 11.24 Charge distribution in an NMOS capacitor operating in accumulation ($V_{gb} < V_{fb}$)

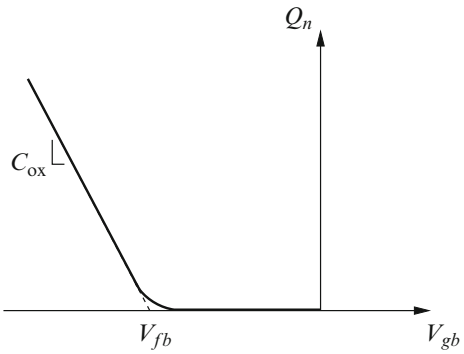


Fig. 11.25 Accumulation charge density as a function of the applied gate voltage. The capacitance per unit area C_{ox} is determined by the slope of the line

Depletion

In MOS capacitors operating in depletion, the applied gate voltage is brought above the flat band voltage and below the threshold voltage ($V_{fb} < V_{gb} < V_t$). A positive charge is therefore induced at the interface between the metal gate and the oxide. A negative charge is induced at the oxide/semiconductor interface. This scenario is accomplished by pushing all of the mobile positive carriers (holes) away, exposing the fixed negative charge from the donors. Hence, the surface of the semiconductor is depleted of mobile carriers, leaving behind a negative space charge. The charge distribution in the MOS capacitor operating in depletion is illustrated in Fig. 11.26.

The resulting space charge behaves like a capacitor with an effective capacitance per unit area C_d. The effective capacitance C_d depends upon the gate voltage V_{gb} and is

$$C_d(V_{gb}) = \frac{\epsilon_{\text{Si}}}{x_d(V_{gb})}, \tag{11.17}$$

where ϵ_{Si} is the permittivity of the silicon and x_d is the thickness of the depletion layer (space charge). Observe from Fig. 11.26 that the oxide capacitance per unit area C_{ox} and depletion capacitance per unit area C_d are connected in series. The capacitance of a MOS structure in the depletion region is therefore

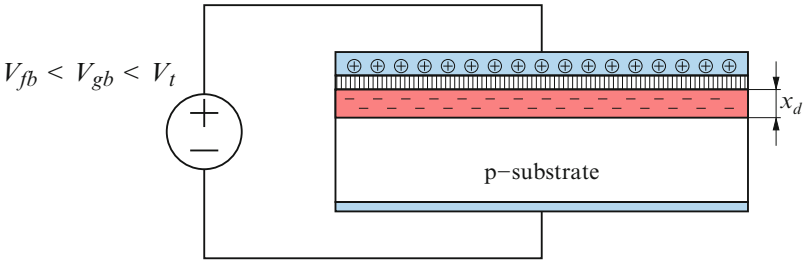


Fig. 11.26 Charge distribution in an NMOS capacitor operating in depletion ($V_{fb} < V_{gb} < V_t$). Under this bias condition, all of the mobile positive carriers (holes) are pushed away, depleting the surface of the semiconductor, resulting in a negative space charge with thickness x_d

$$C_{\mathrm{MOS_{deplet}}} = A \frac{C_{\mathrm{ox}}\, C_d}{C_{\mathrm{ox}} + C_d}. \tag{11.18}$$

Note that the thickness of the silicon depletion layer becomes wider as the gate voltage is increased, since more holes are pushed away, exposing more fixed negative ionized dopants, leading to a thicker space charge layer. As a result, the capacitance of the depleted silicon decreases, reducing the overall MOS capacitance.

Inversion

In MOS capacitors operating in inversion, the applied gate voltage is further increased above the threshold voltage ($V_t < V_{gb}$). The conduction type of the semiconductor surface is inverted (from p-type to n-type). The threshold voltage is referred to as the voltage at which the conductivity type of the surface layer changes from p-type to n-type (in the case of an NMOS capacitor). This phenomenon is explained as follows. As the gate voltage is increased beyond the threshold voltage, holes are pushed away from the Si/SiO_2 interface, exposing the negative charge. Note that the density of holes decreases exponentially from the surface into the bulk. The number of holes decreases as the applied voltage increases. The number of electrons at the surface therefore increases with applied gate voltage and becomes the dominant type of carrier, inverting the surface conductivity. The charge distribution of an MOS capacitor operating in inversion is depicted in Fig. 11.27.

Note that the depletion layer thickness reaches a maximum in the inversion region. The total voltage drop across the semiconductor also reaches the maximum value. Further increasing the gate voltage, the applied voltage drops primarily across the oxide layer. If the gate voltage approaches the threshold voltage, the depleted layer capacitance per unit area C_d^{min} reaches a minimum [247]. In this case, the overall MOS capacitance reaches the minimum value and is

$$C_{\mathrm{MOS_{inv}}} = C_{\mathrm{MOS}}^{\mathrm{min}} = A \frac{C_{\mathrm{ox}}\, C_d^{\mathrm{min}}}{C_{\mathrm{ox}} + C_d^{\mathrm{min}}}, \tag{11.19}$$

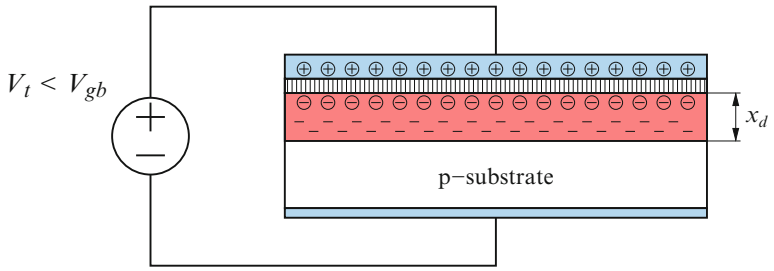


Fig. 11.27 Charge distribution of an NMOS capacitor operating in inversion ($V_t < V_{gb}$). Under this bias condition, a negative charge is accumulated at the semiconductor surface, inverting the conductivity of the semiconductor surface (from p-type to n-type)

where

$$C_d^{\min} = \frac{\epsilon_{\mathrm{Si}}}{x_d^{\max}}. \tag{11.20}$$

Note that at low frequencies (quasi-static conditions), the generation rate of holes (electrons) in the depleted silicon surface layer is sufficiently high. Electrons are therefore swept to the Si/SiO_2 interface, forming a sheet charge with a thin layer of electrons. The inversion layer capacitance under quasi-static conditions therefore reaches the maximum value. At high frequencies, however, the generation rate is not sufficiently high, prohibiting the formation of the electron charge at the Si/SiO_2 interface. In this case, the thickness of the silicon depletion layer reaches the maximum. Hence, the inversion layer capacitance reaches the minimum.

An MOS transistor operated as a capacitor has a substantial ESR, most of which is associated with the lower electrode. This parasitic resistance can be reduced by using a fairly short channel length (25 µm or less) [245]. If the source and drain diffusions are omitted, the backgate contact is typically placed entirely around the gate.

A layout and cross section of an MOS capacitor formed in a BiCMOS process are illustrated in Fig. 11.28. Since the N-type source/drain layer follows the gate oxide growth and polysilicon deposition, the lower plate should consist of some other diffusion (typically deep-n^+). Deep-n^+ has a higher sheet resistance than the N-type source/drain layer (typically 100 Ω/□), resulting in a substantial parasitic resistance of the lower plate. The heavily concentrated n-type doping thickens the gate oxide by 10–30 % through dopant-enhanced oxidation, resulting in higher working voltages but a lower capacitance per unit area. The deep-n^+ is often placed inside the N-well to reduce the parasitic capacitance to the substrate. The N-well can be omitted, however, if the larger parasitic capacitance and lower breakdown voltage of the deep-n^+/p-epi junction can be tolerated.

Regardless of how an MOS capacitor is constructed, the two capacitor electrodes are never entirely interchangeable. The lower plate always consists of a diffusion with substantial parasitic junction capacitance. This junction capacitance

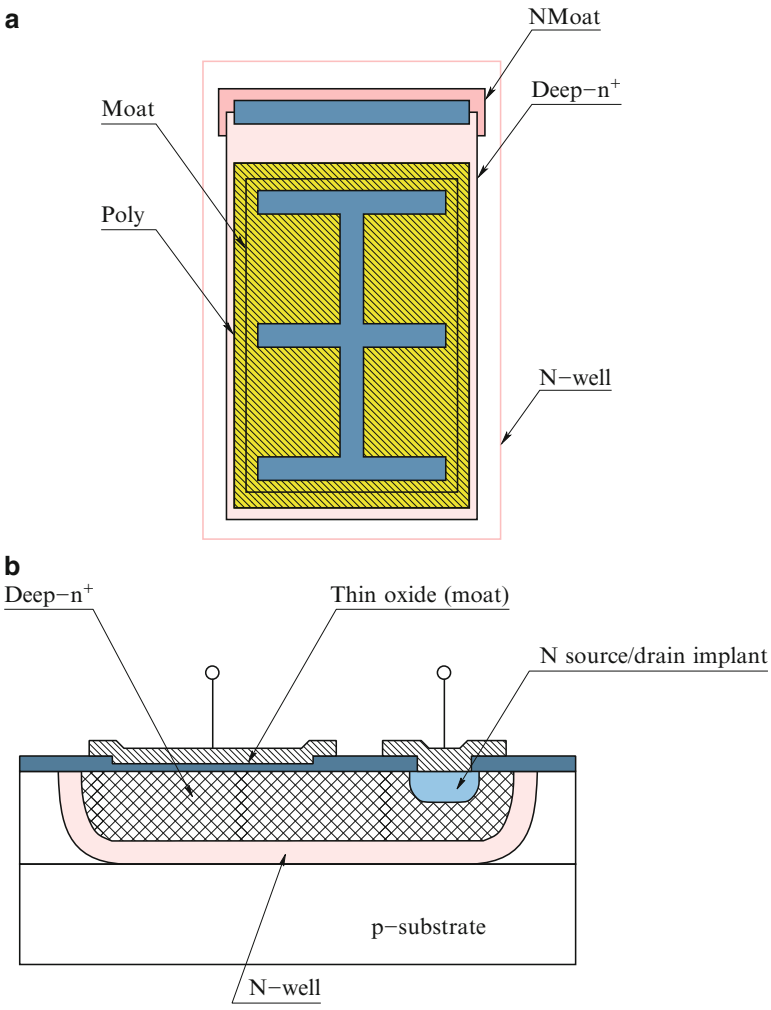


Fig. 11.28 Deep-n^+ MOS capacitor constructed in a BiCMOS process; (**a**) layout, (**b**) cross section

is eliminated by connecting the lower plate of the capacitor to the substrate potential. The upper plate of the MOS capacitor consists of a deposited electrode with a relatively small parasitic capacitance. The lower plate of an MOS capacitor should therefore be connected to the driven node (with the lower impedance). Swapping the two electrodes of an MOS capacitor can load a high impedance node with a high parasitic impedance, compromising circuit performance.

The major benefit of MOS capacitors is the natural compatibility with CMOS technology. MOS capacitors also provide a high capacitance density [248], providing a cost effective on-chip decoupling capacitance. MOS capacitors result in

relatively high matching: the gate oxide capacitance is typically controlled within 5 % error [246]. MOS capacitors, however, are non-linear devices that exhibit strong voltage dependence (more than 100 ppm/V [249]) due to the variation of both the dielectric constant and the depletion region thickness within each plate. The performance of the MOS capacitors is limited at high frequencies due to the large diffusion-to-substrate parasitic capacitance. As technology scales, the leakage currents of MOS capacitors also increase substantially, increasing the total power dissipation. High leakage current is the primary issue with MOS capacitors.

An MOS on-chip capacitance is typically realized as accumulation and inversion capacitors. Note that capacitors operating in accumulation are more linear than capacitors operating in inversion [250]. The MOS capacitance operating in accumulation is almost independent of frequency. Moreover, MOS decoupling capacitors operating in accumulation result in an approximately 15× reduction in leakage current as compared to MOS decoupling capacitors operating in inversion [251]. MOS decoupling capacitors operating in accumulation should therefore be the primary form of MOS decoupling capacitors in modern high performance ICs.

11.4.3 Metal-Insulator-Metal (MIM) Capacitors

A MIM capacitor consists of two metal layers (plates) separated by a deposited dielectric layer. A cross section of a MIM capacitor is shown in Fig. 11.29. A thick oxide layer is typically deposited on the substrate, reducing the parasitic capacitance to the substrate. The parasitic substrate capacitance is also lowered by utilizing the top metal layers as plates of a MIM capacitor. For instance, in comb MIM capacitors [252], the parasitic capacitance to the substrate is less than 2 % of the total capacitance.

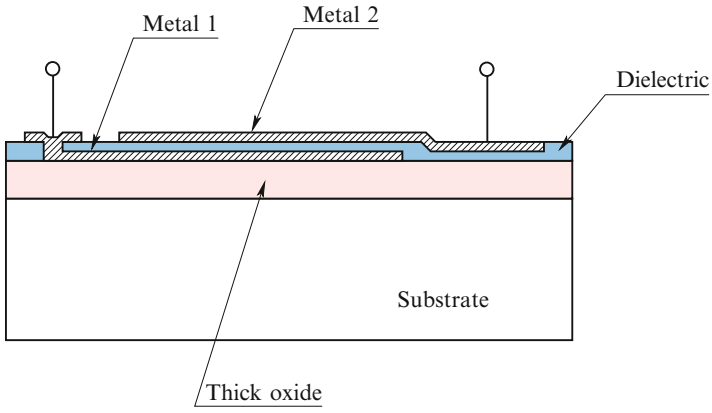


Fig. 11.29 Cross section of a MIM capacitor. A thick oxide (SiO_2) layer is typically deposited on the substrate to reduce the parasitic capacitance to the substrate

Historically, MIM capacitors have been widely used in RF and mixed-signal ICs due to the low leakage, high linearity, low process variations (high accuracy), and low temperature variations [253–255] of MIM capacitors. Conventional circuits utilize SiO_2 as a dielectric deposited between two metal layers. Large MIM capacitors therefore require significant circuit area, prohibiting the use of MIM capacitors as decoupling capacitors in high complexity ICs. The capacitance density can be increased by reducing the dielectric thickness and employing high-k dielectrics. Reducing the dielectric thickness, however, results in a substantial increase in leakage current which is highly undesirable.

MIM capacitors with a capacitance density comparable to MOS capacitors (8–10 $fF/\mu m^2$) have been fabricated using Al_2O_3 and $AlTiO_x$ dielectrics [256], $AlTaO_x$ [257], and HfO_2 dielectric using atomic layer deposition (ALD) [258]. A higher capacitance density (13 $fF/\mu m^2$) is achieved using laminate ALD $HfO_2 - Al_2O_3$ dielectrics [259, 260]. Laminate dielectrics also result in higher voltage linearity and reliability. Recently, MIM capacitors with a capacitance density approximately two times greater than the capacitance density of MOS capacitors have been fabricated [261]. A capacitance density of 17 $fF/\mu m^2$ is achieved using a Nb_2O_5 dielectric with $HfO_2 - Al_2O_3$ barriers.

Unlike MOS capacitors, MIM capacitors require high temperatures for thin film deposition. Integrating MIM capacitors into a standard low temperature ($\leq$400 °C) back-end high complexity digital process is therefore a challenging problem [262]. This problem can be overcome by utilizing MIM capacitors with plasma enhanced chemical vapor deposition (PECVD) nitride dielectrics [263, 264]. Previously, MIM capacitors were unavailable in CMOS technology with copper metallization. Recently, MIM capacitors have been successfully integrated into CMOS and BiCMOS technologies with a copper dual damascene metallization process [265–267]. In [268], a high density MIM capacitor with a low ESR using a plug-in copper plate is described, making MIM capacitors highly efficient for use as a decoupling capacitor.

MIM capacitors are widely utilized in RF and mixed-signal ICs due to low voltage coefficients, good capacitor matching, precision control of capacitor values, small parasitic capacitance, high reliability, and low defect densities [269]. MIM capacitors also exhibit high linearity over a wide frequency range. Additionally, a high capacitance density with lower leakage currents has recently been achieved, making MIM capacitors the best candidate for decoupling power and ground lines in modern high performance, high complexity ICs. For instance, for a MIM capacitor with a dielectric thickness $t_{ox} = 1$ nm, a capacitance density of 34.5 $fF/\mu m^2$ has been achieved [270].

11.4.4 Lateral Flux Capacitors

The total capacitance per unit area can be increased by using more than one pair of interconnect layers. Current technologies offer up to ten metal layers, increasing the capacitance nine times through the use of a sandwich structure. The capacitance

is further increased by exploiting the lateral flux between the adjacent metal lines within a specific interconnect layer. In scaled technologies, the adjacent metal spacing (on the same level) shrinks faster than the spacing between the metal layers (on different layers), resulting in substantial lateral coupling.

A simplified structure of an interdigitated capacitor exploiting lateral flux is shown in Fig. 11.30. The two terminals of the capacitor are shown in *light pink* and *dark pink*. Note that the two plates built in the same metal layer alternate to better exploit the lateral flux. Ordinary vertical flux can also be exploited by arranging the segments of a different metal layer in a complementary pattern [271], as illustrated in Fig. 11.31. Note that a higher capacitance density is achieved by using a lateral flux together with a vertical flux (parallel plate structure).

An important advantage of using a lateral flux capacitor is reducing the bottom plate parasitic capacitance as compared to an ordinary parallel plate structure. This reduction is due to two reasons. First, the higher density of the lateral flux capacitor results in a smaller area for a specific value of total capacitance. Second, some of

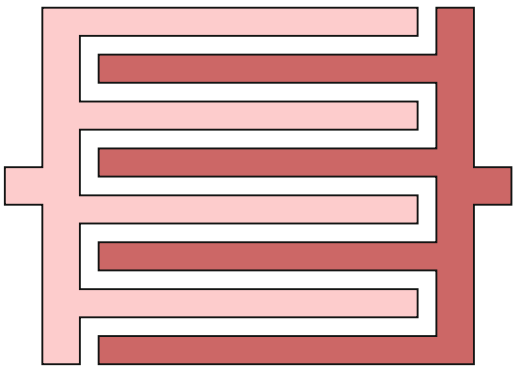


Fig. 11.30 A simplified structure of an interdigitated lateral flux capacitor (top view). Two terminals of the capacitor are shown in *light pink* and *dark pink*

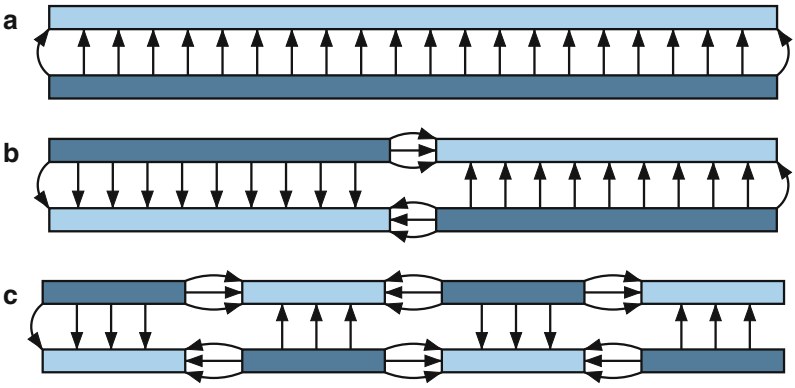


Fig. 11.31 Vertical flux versus lateral flux; (**a**) standard parallel plate structure, (**b**) structure divided by two cross-connected metal layers, (**c**) structure divided by four cross-connected metal layers

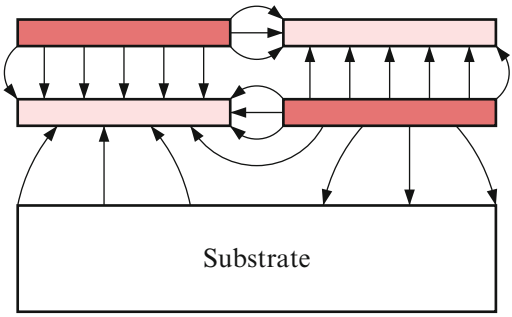


Fig. 11.32 Reduction of the bottom plate parasitic capacitance through flux stealing. Shades of *pink* denote the two terminals of the capacitor

the field lines originating from one of the bottom plates terminate on the adjacent plate rather than the substrate, further reducing the bottom plate capacitance, as shown in Fig. 11.32. Such phenomenon is referred to as flux stealing. Thus, some portion of the bottom plate parasitic capacitance is converted into a useful plate-to-plate capacitance. Three types of enhanced lateral flux capacitors with a higher capacitance density are described in the following three subsections.

Fractal Capacitors

Since the lateral capacitance is dependent upon the perimeter of the structure, the maximum capacitance can be obtained with those geometries that maximize the total perimeter. Fractals are therefore good candidates for use in lateral flux capacitors. A fractal is a structure that encloses a finite area with an infinite perimeter [272]. Although lithography limitations prevent fabrication of a real fractal, quasi-fractal geometries with feature sizes limited by lithography have been successfully fabricated in fractal capacitors [273]. It has been demonstrated that in certain cases, the effective capacitance of fractal capacitors can be increased by more than ten times.

The final shape of a fractal can be tailored to almost any form. The flexibility arises from the characteristic that a wide variety of geometries exists, determined by the fractal initiator and generator [272]. It is also possible to use different fractal generators during each step. Fractal capacitors of any desired form can therefore be constructed due to the flexibility in the shape of the layout. Note that the capacitance per unit area of a fractal capacitor depends upon the fractal dimensions. Fractals with large dimensions should therefore be used to improve the layout density [273].

In addition to the capacitance density, the quality factor Q is important in RF and mixed-signal applications. In fractal capacitors, the degradation in quality factor is minimal, since the fractal structure naturally limits the length of the thin metal sections to a few micrometers, maintaining a reasonably small ESR. Hence, smaller dimension fractals should be used to achieve a low ESR. Alternatively, a tradeoff exists between the capacitance density and the ESR in fractal capacitors.

Existing technologies typically provide tighter control over the lateral spacing of the metal layers as compared to the vertical thickness of the oxide layers (both from wafer to wafer and across the same wafer). Lateral flux capacitors shift the burden of matching from the oxide thickness to the lithography. The matching characteristics are therefore greatly improved in lateral flux capacitors. Furthermore, the pseudo-random nature of the lateral flux capacitors compensate for the effects of nonuniformity in the etching process.

Comparing fractal and conventional interdigitated capacitors, note the inherent parasitic inductance of an interdigitated capacitor. Most fractal geometries randomize the direction of the current flow, reducing the ESL. In an interdigitated capacitor, however, the current flows in the same direction for all of the parallel lines. Also in fractal structures, the electric field concentrates around the sharp edges, increasing the effective capacitance density (about 15 %) [273]. Nevertheless, due to simplicity, interdigitated capacitors are widely used in ICs.

Woven Capacitors

A woven structure is also utilized to achieve high capacitance density. A woven capacitor is depicted in Fig. 11.33. Two orthogonal metal layers are used to construct the plates of the capacitor. Vias connect the metal lines of a specific capacitor plate at the overlap sites. Note that in a woven structure, the current in the adjacent lines flows in the opposite direction. The woven capacitor has therefore much less inherent parasitic inductance as compared to an interdigitated capacitor [202, 274]. In addition, the ESR of a woven capacitor contributed by vias is smaller than the ESR of an interdigitated capacitor. A woven capacitor, however, results in a smaller capacitance density as compared to an interdigitated capacitor with the same metal pitch due to the smaller vertical capacitance.

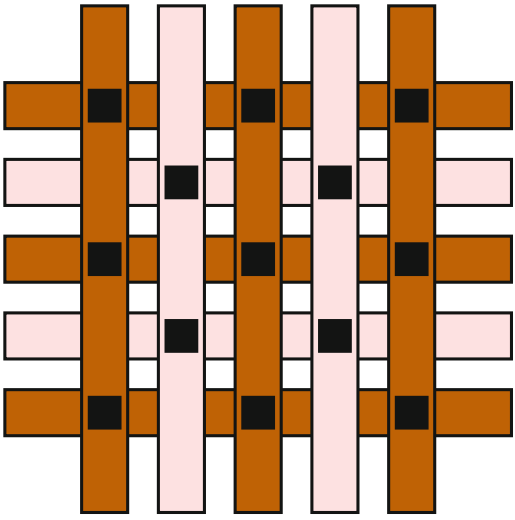


Fig. 11.33 Woven capacitor. The two terminals of the capacitor are shown in *light pink* and *brown*. The vias are illustrated by the *black colored squares*

Vertical Parallel Plate (VPP) Capacitors

Another way to utilize a number of metal layers in modern CMOS technologies is to construct conductive vertical plates out of vias in combination with the interconnect metal. Such a capacitor is referred to as a vertical parallel plate (VPP) capacitor [275]. A VPP capacitor consists of metal slabs connected vertically using multiple vias between the vertical plates. This structure fully exploits lateral scaling trends as compared to fractal structures [274].

11.4.5 Comparison of On-Chip Decoupling Capacitors

On-chip decoupling capacitors can be designed in ICs in a number of ways. The primary characteristics of four common types of on-chip decoupling capacitors, discussed in Sects. 11.4.1, 11.4.2, 11.4.3 and 11.4.4, are listed in Table 11.1. Note that typical MIM capacitors provide a lower capacitance density (1–10 fF/μm^2) than MOS capacitors. Recently, a higher capacitance density (13 fF/μm^2) of MIM capacitors has been achieved using laminate ALD HfO_2–Al_2O_3 dielectrics [259, 260]. A capacitance density of 34.5 fF/μm^2 has been reported in [270] for a MIM capacitor with a dielectric thickness of 1 nm.

Note that the quality factor of the MOS and lateral flux capacitors is limited by the channel resistance and the resistance of the multiple vias. Decoupling capacitors with a low quality factor produce wider antiresonant spikes with a significantly reduced magnitude [276]. It is therefore highly desirable to limit the quality factor of the on-chip decoupling capacitors. Note that in the case of a low ESR (high quality factor), an additional series resistance should be provided, lowering the magnitude of the antiresonant spike. This additional resistance, however, is limited by the target impedance of the power distribution system [28].

Table 11.1 Four common types of on-chip decoupling capacitors in a 90 nm CMOS technology

Feature	PIP capacitor	MOS capacitor	MIM capacitor	Lateral flux capacitor
Capacitance density (fF/μm^2)	1–5	10–20	1–30	10–20
Bottom plate capacitance (%)	5–10	20–30	2–5	1–5
Linearity (ppm/V)	50–150	300–500	10–50	50–100
Quality factor	5–15	1–10	50–150	10–50
Parasitic resistance (mΩ)	500–2000	1000–10,000	50–250	100–500
Leakage current (A/cm^2)	10^{-10}–10^{-9}	10^{-2}–10^{-1}	10^{-9}–10^{-8}	10^{-10}–10^{-9}
Temperature dependence (ppm/°C)	150–250	300–500	50–100	50–100
Process complexity	Extra steps	Standard	Standard	Standard

The parasitic resistance is another important characteristic of on-chip decoupling capacitors. The parasitic resistance characterizes the efficiency of a decoupling capacitor. Alternatively, both the amount of charge released by the decoupling capacitor and the rate with which the charge is restored on the decoupling capacitor are primarily determined by the parasitic resistance [277]. The parasitic resistance of PIP capacitors is mainly determined by the resistive polysilicon layer. MIM capacitors exhibit the lowest parasitic resistance due to the highly conductive metal layers used as the plates of the capacitor. The increased parasitic resistance of the lateral flux capacitors is due to the multiple resistive vias, connecting metal plates at different layers [274]. In MOS capacitors, both the channel resistance and the resistance of the metal plates contribute to the parasitic resistance. The performance of MOS capacitors is therefore limited by the high parasitic resistance.

Observe from Table 11.1 that MOS capacitors result in prohibitively large leakage currents. As technology scales, the leakage power is expected to become the major component of the total power dissipation. Thick oxide MOS decoupling capacitors are often used to reduce the leakage power. Thick oxide capacitors, however, require a larger die area for the same capacity as a thinner oxide capacitance. Note that the leakage current in MOS capacitors increases exponentially with temperature, further exacerbating the problem of heat removal. Also note that leakage current is reduced in MIM capacitors as compared to MOS capacitors by about seven orders of magnitude. The leakage current of MIM capacitors is also fairly temperature independent, increasing twofold as the temperature rises from 25 °C to 125 °C [265].

Note that PIP capacitors typically require additional process steps, adding extra cost. From the information listed in Table 11.1, MIM capacitors and stacked lateral flux capacitors (fractal, VPP, and woven) are the best candidates for decoupling the power and ground lines in modern high performance, high complexity ICs.

11.5 On-Chip Switching Voltage Regulator

The efficiency of on-chip decoupling capacitors can be enhanced by an on-chip switching voltage regulator [278]. The decoupling capacitors reduce the impedance of the power distribution system by serving as an energy source when the power voltage decreases, as discussed in previous sections. The smaller the power voltage variation, the smaller the energy transferred from a decoupling capacitor to the load.

Consider a group of N decoupling capacitors C placed on a die. Where connected in parallel between the power and ground network, the capacitors behave as a single capacitor NC. As the power supply level decreases from the nominal level V_{dd} to a target minimum $V_{dd} - \delta V$, the non-switching decoupling capacitors release only a small fraction k of the stored charge into the network,

$$\frac{\delta Q}{Q_0} = \frac{NCV_{dd} - NC(V_{dd} - \delta V)}{NCV_{dd}} = \frac{\delta V}{V_{dd}} \equiv k. \tag{11.21}$$

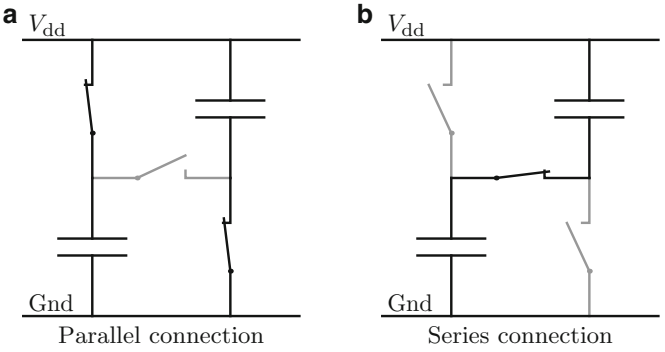


Fig. 11.34 Switching decoupling capacitors from a parallel to a series connection. (**a**) Parallel connection. (**b**) Series connection

A correspondingly small fraction of the total energy E stored in the capacitors is transferred to the load,

$$\frac{\delta E}{E_0} = \frac{V_{dd}^2 - (V_{dd} - \delta V)^2}{V_{dd}^2} \approx \frac{2\delta V}{V_{dd}} = 2k. \tag{11.22}$$

Switching the on-chip capacitors can increase the charge (and energy) transferred from the capacitors to the load as the power voltage decreases below the nominal voltage level [278]. Rather than a fixed connection in parallel as in the traditional non-switching case, the connection of capacitors to the power and ground networks can be changed from parallel to series using switches, as shown in Fig. 11.34 for the case of two capacitors. When the rate of variation in the power supply voltage is relatively small, the capacitors are connected in parallel, as shown in Fig. 11.34a, and charged to V_{dd}. When the instantaneous power supply variation exceeds a certain threshold, the capacitors are reconnected in series, as shown in Fig. 11.34b, transforming the circuit into a capacitor of C/N capacity carrying a charge of CV_{dd}. In this configuration, the circuit can release

$$\delta Q_{sw} = CV_{dd}\left(1 - \frac{1-k}{N}\right) \tag{11.23}$$

amount of charge before the drop in the voltage supply level exceeds the noise margin $\delta V = kV_{dd}$. This amount of charge is greater than the charge released in the non-switching case, as determined by (11.21), if $k < \frac{1}{N+1}$. The effective charge storage capacity of the on-chip decoupling capacitors is thereby enhanced. The area of the on-chip capacitors required to lower the peak resonant impedance of the power network to a satisfactory level is decreased.

This technique is employed in the UltraSPARC III microprocessor, as described by Ang, Salem, and Taylor [278]. In addition to the 176 nF of non-switched on-chip

decoupling capacitance, 134 nF of switched on-chip capacitance is placed on the die. The switched capacitance occupies 20 mm^2 of die area and is distributed in the form of 99 switching regulator blocks throughout the die to maintain a uniform power supply voltage. The switching circuitry is designed to minimize the short-circuit current when the capacitor is switching. Feedback loop control circuitry ensures stable behavior of the switching capacitors. The switching circuitry occupies 0.4 mm^2, a small fraction of the overall regulator area. The regulator blocks are connected directly to the global power distribution grid. In terms of the frequency domain characteristics, the switching regulator lowers the magnitude of the die-package resonance impedance. The switched decoupling capacitors decrease the on-chip power noise by roughly a factor of 2 and increase the operating frequency of the circuit by approximately 20 %.

11.6 Summary

A brief overview of decoupling capacitors has been presented in this chapter. The primary characteristics of decoupling capacitors can be summarized as follows.

- A decoupling capacitor serves as an intermediate and temporary storage of charge and energy located between the power supply and current load, which is electrically closer to the switching circuit
- To be effective, a decoupling capacitor should have a high capacity to store a sufficient amount of energy and be able to release and accumulate energy at a sufficient rate
- In order to ensure correct and reliable operation of an IC, the impedance of the power distribution system should be maintained below the target impedance in the frequency range from DC to the maximum operating frequency
- The high frequency impedance is effectively reduced by placing decoupling capacitors across the power and ground interconnects, permitting the current to bypass the inductive interconnect
- A decoupling capacitor has an inherent parasitic resistance and inductance and therefore can only be effective within a certain frequency range
- Several stages of decoupling capacitors are typically utilized to maintain the output impedance of a power distribution system below a target impedance
- Antiresonances are effectively managed by utilizing decoupling capacitors with low ESL and by placing a large number of decoupling capacitors with progressively decreasing magnitude, shifting the antiresonant spike to a higher frequency
- MIM capacitors and stacked lateral flux capacitors (fractal, VPP, and woven) are preferable candidates for decoupling power and ground lines in modern high speed, high complexity ICs

Chapter 12
Effective Radii of On-Chip Decoupling Capacitors

Decoupling capacitors are widely used to manage power supply noise. A decoupling capacitor acts as a reservoir of charge, which is released when the power supply voltage at a particular current load drops below some tolerable level. Alternatively, decoupling capacitors are an effective way to reduce the impedance of power delivery systems operating at high frequencies [29]. Since the inductance scales slowly [129], the location of the decoupling capacitors significantly affects the design of the P/G network in high performance ICs such as microprocessors. With increasing frequencies, a distributed hierarchical system of decoupling capacitors placed on-chip is needed to effectively manage power supply noise [279].

The efficacy of decoupling capacitors depends upon the impedance of the conductors connecting the capacitors to the current loads and power sources. During discharge, the current flowing from the decoupling capacitor to the current load results in resistive noise (IR drops) and inductive noise ($L\ dI/dt$ drops) due to the parasitic resistances and inductances of the power delivery network. The resulting voltage drop at the current load is therefore always greater than the voltage drop at the decoupling capacitor. Thus, a maximum parasitic impedance between the decoupling capacitor and the current load exists at which the decoupling capacitor is effective. Alternatively, to be effective, a decoupling capacitor should be placed close to a current load during discharge (within the maximum effective distance $d_Z^{\max}$), as shown in Fig. 12.1.

Once the switching event is completed, a decoupling capacitor has to be fully charged before the next clock cycle begins. During the charging phase, the voltage across the decoupling capacitor rises exponentially. The charge time of a capacitor is determined by the parasitic resistance and inductance of the interconnect between the capacitor and the power supply. A design space for a tolerable interconnect resistance and inductance exists, permitting the charge on the decoupling capacitor to be restored within a target charge time. The maximum frequency at which the decoupling capacitor is effective is determined by the parasitic resistance and inductance of the metal lines and the size of the decoupling capacitor. A maximum

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0_12

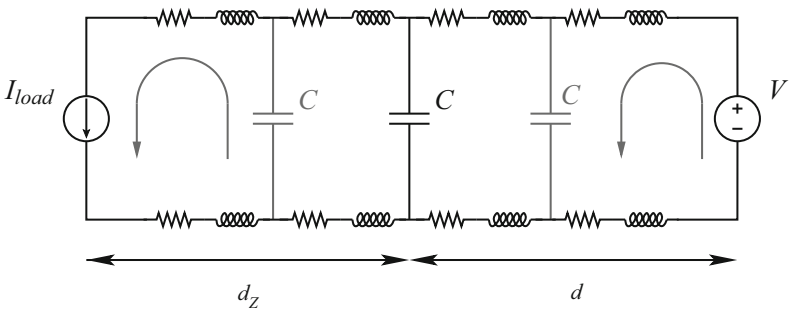


Fig. 12.1 Placement of an on-chip decoupling capacitor based on the maximum effective distance. To be effective, a decoupling capacitor should be placed close to the current load during discharge. During the charging phase, however, the decoupling capacitor should be placed close to the power supply to efficiently restore the charge on the capacitor. The specific location of a decoupling capacitor should therefore be determined to simultaneously satisfy the maximum effective distances $d_Z^{\max}$ during discharge and $d_{\text{ch}}^{\max}$ during charging

effective distance based on the charge time, therefore, exists for each on-chip decoupling capacitor. Beyond this effective distance, the decoupling capacitor is ineffective. Alternatively, to be effective, an on-chip decoupling capacitor should be placed close to a power supply during the charging phase (within the maximum effective distance $d_{\text{ch}}^{\max}$, see Fig. 12.1). The relative location of the on-chip decoupling capacitors is therefore of fundamental importance. A design methodology is therefore required to determine the location of an on-chip decoupling capacitor, simultaneously satisfying the maximum effective distances, $d_Z^{\max}$ and $d_{\text{ch}}^{\max}$. This location is characterized by the effective radii of the on-chip decoupling capacitors and is the primary subject of this chapter. A design methodology to estimate the minimum required on-chip decoupling capacitance is also presented.

This chapter is organized as follows. Existing work on placing on-chip decoupling capacitors is reviewed in Sect. 12.1. The effective radius of an on-chip decoupling capacitor as determined by the target impedance is presented in Sect. 12.2. Design techniques to estimate the minimum magnitude of the required on-chip decoupling capacitance are discussed in Sect. 12.3. The effective radius of an on-chip decoupling capacitor based on the charge time is determined in Sect. 12.4. A design methodology for placing on-chip decoupling capacitors based on the maximum effective radii is presented in Sect. 12.5. A model of an on-chip power distribution network is developed in Sect. 12.6. Simulation results for typical values of on-chip parasitic resistances and inductances are presented in Sect. 12.7. Some circuit design implications are discussed in Sect. 12.8. Finally, some specific conclusions are summarized in Sect. 12.9.

12.1 Background

Decoupling capacitors have traditionally been allocated on a circuit board to control the impedance of a power distribution system and suppress EMI. Decoupling capacitors are also employed to provide the required charge to the switching circuits, enhancing signal integrity. Since the parasitic impedance of a circuit board-based power distribution system is negligible at low frequencies, board decoupling capacitors are typically modeled as ideal capacitors without parasitic impedances. In an important early work by Smith [280], the effect of a decoupling capacitor on the signal integrity in circuit board-based power distribution systems is presented. The efficacy of the decoupling capacitors is analyzed in both the time and frequency domains. Design criteria have been developed, however, which significantly overestimate the required decoupling capacitance. A hierarchical placement of decoupling capacitors has been presented by Smith et al. in [136]. The authors of [136] show that each decoupling capacitor is effective only within a narrow frequency range. Larger decoupling capacitors have a greater form factor (physical dimensions), resulting in higher parasitic impedances [229]. The concept of an effective series resistance and an effective series inductance of each decoupling capacitor is also described. The authors show that by hierarchically placing the decoupling capacitors from the voltage regulator module level to the package level, the impedance of the overall power distribution system can be maintained below a target impedance.

As the signal frequency increases to several megahertz, the parasitic impedance of the circuit board decoupling capacitors becomes greater than the target impedance. The circuit board decoupling capacitors therefore become less effective at frequencies above 10–20 MHz. Package decoupling capacitors should therefore be utilized in the frequency range from several megahertz to several hundred megahertz [136]. In modern high performance ICs operating at several gigahertz, only those decoupling capacitors placed on-chip are effective at these frequencies.

The optimal placement of on-chip decoupling capacitors has been discussed in [281]. The power noise is analyzed assuming an *RLC* network model, representing a multi-layer power bus structure. The current load is modeled by time-varying resistors. The on-chip decoupling capacitors are allocated to only those areas where the power noise is greater than the maximum tolerable level. Ideal on-chip decoupling capacitors are assumed in the algorithm described in [281]. The resulting budget of on-chip decoupling capacitance is therefore significantly overestimated. Another technique for placing on-chip decoupling capacitors has been described in [282]. The decoupling capacitors are placed based on activity signatures determined from microarchitectural simulations. This technique produces a 30 % decrease in the maximum noise level as compared to uniformly placing the on-chip decoupling capacitors. This methodology results in overestimating the capacitance budget due to the use of a simplified criterion for sizing the on-chip decoupling capacitors. Also, since the package level power distribution system is modeled as a single lumped resistance and inductance, the overall power supply noise is greatly underestimated.

An algorithm for automatically placing and sizing on-chip decoupling capacitors in application-specific integrated circuits is described in [283]. The problem is formulated as a nonlinear optimization and solved using a sensitivity-based quadratic programming solver. The algorithm is limited to on-chip decoupling capacitors placed in rows of standard cells (in one dimension). The power distribution network is modeled as a resistive mesh, significantly underestimating the power distribution noise. In [284], the problem of on-chip decoupling capacitor allocation is evaluated. This technique is integrated into a power supply noise-aware floorplanning methodology. Only the closest power supply pins are considered to provide the switching current drawn by the load. Additionally, only the shortest and second shortest paths are considered between a decoupling capacitor and the current load. It is assumed that the current load is located at the center of a specific circuit block. The technique does not consider the degradation in effectiveness of an on-chip decoupling capacitor located at some distance from the current load. Moreover, only the discharge phase is considered. To be effective, a decoupling capacitor should be fully charged before the following switching cycle. Otherwise, the charge on the decoupling capacitor will be gradually depleted, making the capacitor ineffective. The methodology described in [284] therefore results in underestimating the power supply noise and overestimating the required on-chip decoupling capacitance.

The problem of on-chip decoupling capacitor allocation has historically been considered as two independent tasks. The location of an on-chip decoupling capacitor is initially determined. The decoupling capacitor is next appropriately sized to provide the required charge to the current load. As discussed in [277], the size of the on-chip decoupling capacitors is determined by the impedance (essentially, the physical separation) between a decoupling capacitor and the current load (or power supply).

Proper sizing and placement of the on-chip decoupling capacitors however should be determined simultaneously. As shown in this chapter, on-chip decoupling capacitors are only effective in close vicinity to the switching circuit. The maximum effective distance for both the discharge and charging phase is determined. It is also shown that the on-chip decoupling capacitors should be placed both close to the current load to provide the required charge and to the power supply to be fully recharged before the next switching event. A design methodology for placing and sizing on-chip decoupling capacitors based on a maximum effective distance as determined by the target impedance and charge time is presented in this chapter.

12.2 Effective Radius of On-Chip Decoupling Capacitor Based on Target Impedance

Neglecting the parasitic capacitance [285], the impedance of a unit length wire is $Z'(\omega) = r + j\omega l$, where r and l are the resistance and inductance per length, respectively, and ω is an effective frequency, as determined by the rise time of the

current load. The inductance l is the effective inductance per unit length of the power distribution grid, incorporating both the partial self-inductance and mutual coupling among the lines [73]. The target impedance of the metal line of a particular length is therefore

$$Z(\omega) = Z'(\omega) \times d, \tag{12.1}$$

where $Z'(\omega)$ is the impedance of a unit length metal line, and d is the distance between the decoupling capacitor and the current load. Substituting the expression for the target impedance Z_{target} presented in [28] into (12.1), the maximum effective radius $d_Z^{\max}$ between the decoupling capacitor and the current load is

$$d_Z^{\max} = \frac{Z_{target}}{Z'(\omega)} = \frac{V_{\mathrm{dd}} \times Ripple}{I \times \sqrt{r^2 + \omega^2 l^2}}, \tag{12.2}$$

where $\sqrt{r^2 + \omega^2 l^2}$ denotes the magnitude of the impedance of a unit length wire, Z_{target} is the maximum impedance of a power distribution system, resulting in a power noise lower than the maximum tolerable level, and *Ripple* is the maximum tolerable power noise (the ratio of the magnitude of the maximum tolerable voltage drop to the power supply level). Note that the maximum effective radius as determined by the target impedance is inversely proportional to the magnitude of the current load and the impedance of a unit length line. Also note that the per length resistance r and inductance l account for the ESR and ESL of an on-chip decoupling capacitor. The maximum effective radius as determined by the target impedance decreases rapidly with each technology generation (a factor of 1.4, on average, per computer generation), as shown in Fig. 12.2 [286]. Also note that in a meshed structure, multiple paths between any two points are added in parallel. The maximum effective distance corresponding to Z_{target} is, therefore, larger than the maximum effective distance of a single line, as discussed in Sect. 12.7. The maximum effective radius is defined in this chapter as follows.

Definition 1. The effective radius of an on-chip decoupling capacitor is the maximum distance between the current load (power supply) and the decoupling capacitor for which the capacitor is capable of providing sufficient charge to the current load, while maintaining the overall power distribution noise below a tolerable level.

12.3 Estimation of Required On-Chip Decoupling Capacitance

Once the specific location of an on-chip decoupling capacitor is determined as described in Sect. 12.2, the minimum required magnitude of the on-chip decoupling capacitance should be determined, satisfying the expected current demands. Design expressions for determining the required magnitude of the on-chip decoupling

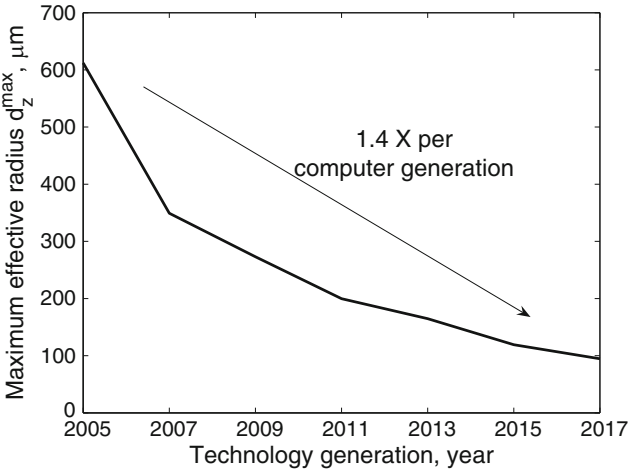


Fig. 12.2 Projection of the maximum effective radius as determined by the target impedance $d_Z^{\max}$ for future technology generations: $I_{\max} = 10\,\text{mA}$, $V_{\text{dd}} = 1\,\text{V}$, and $Ripple = 0.1$. Global on-chip interconnects are assumed, modeling the highly optimistic scenario. The maximum effective radius as determined by the target impedance is expected to decrease at an alarming rate (a factor of 1.4 on average per computer generation)

capacitors based on the dominant power noise are presented in this section. A conventional approach with dominant resistive noise is described in Sect. 12.3.1. Techniques for determining the magnitude of on-chip decoupling capacitors in the case of dominant inductive noise are developed in Sect. 12.3.2. The critical length of the P/G paths connecting the decoupling capacitor to the current load is presented in Sect. 12.3.3.

12.3.1 Dominant Resistive Noise

To estimate the on-chip decoupling capacitance required to support a specific local current demand, the current load is modeled as a triangular current source. The magnitude of the current source increases linearly, reaching the maximum current $I_{\max}$ at peak time t_p. The magnitude of the current source decays linearly, becoming zero at t_f, as shown in Fig. 12.3. The on-chip power distribution network is modeled as a series *RL* circuit. To qualitatively illustrate the methodology for placing on-chip decoupling capacitors based on the maximum effective radii, a single decoupling capacitor with a single current load is assumed to mitigate the voltage fluctuations across the P/G terminals.

The total charge Q_{dis} required to satisfy the current demand during a switching event is modeled as the sum of the area of two triangles (see Fig. 12.3). Since the required charge is provided by an on-chip decoupling capacitor, the voltage across

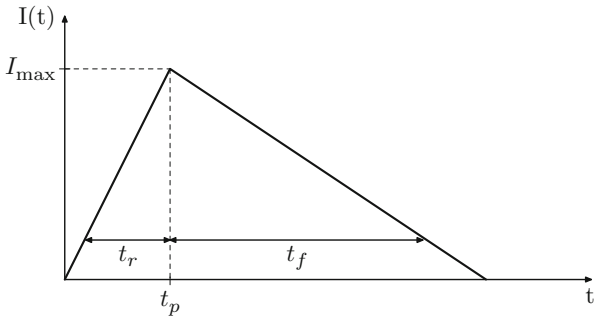


Fig. 12.3 Linear approximation of the current demand of a power distribution network by a current source. The magnitude of the current source reaches the maximum current $I_{\max}$ at peak time t_p. Transition times t_r and t_f denote, respectively, the rise and fall time of the current load

the capacitor during discharge drops below the initial power supply voltage. The required charge during the entire switching event is thus[1]

$$Q^f_{\text{dis}} = \frac{I_{\max} \times (t_r + t_f)}{2} = C_{\text{dec}} \times (V_{\text{dd}} - V^f_C), \tag{12.3}$$

where $I_{\max}$ is the maximum magnitude of the current load of a specific circuit block for which the decoupling capacitor is allocated, t_r and t_f are the rise and fall time, respectively, C_{dec} is the decoupling capacitance, V_{dd} is the power supply voltage, and V^f_C is the voltage across the decoupling capacitor after the switching event. Note that since there is no current after switching, the voltage at the current load is equal to the voltage across the decoupling capacitor.

The voltage fluctuations across the P/G terminals of a power delivery system should not exceed the maximum level (usually 10 % of the power supply voltage [153]) to guarantee fault-free operation. Thus,

$$V^f_C \equiv V^f_{load} \geq 0.9\, V_{\text{dd}}. \tag{12.4}$$

Substituting (12.4) into (12.3) and solving for C_{dec}, the minimum on-chip decoupling capacitance required to support the current demand during a switching event is

$$C^f_{\text{dec}} \geq \frac{I_{\max} \times (t_r + t_f)}{0.2\, V_{\text{dd}}}, \tag{12.5}$$

where C^f_{dec} is the decoupling capacitance required to support the current demand during the entire switching event.

[1] In the general case with an a priori determined current profile, the required charge can be estimated as the integral of $I_{load}(t)$ from 0 to t_f.

12.3.2 Dominant Inductive Noise

Note that (12.5) is applicable only to the case where the voltage drop at the end of the switching event is larger than the voltage drop at the peak time t_p ($IR \gg L\,dI/dt$). Alternatively, the minimum voltage at the load is determined by the resistive drop and the parasitic inductance can be neglected. This phenomenon can be explained as follows. The voltage drop as seen at the current load is caused by current flowing through the parasitic resistance and inductance of the on-chip power distribution system. The resulting voltage fluctuations are the sum of the ohmic IR voltage drop, inductive $L\,dI/dt$ voltage drop, and the voltage drop across the decoupling capacitor at t_p. A critical parasitic RL impedance, therefore, exists for any given set of rise and fall times. Beyond this critical impedance, the voltage drop at the load is primarily caused by the inductive noise ($L\,dI/dt \gg IR$), as shown in Fig. 12.4. The decoupling capacitor should therefore be increased in the case of dominant inductive noise to reduce the voltage drop across the capacitor during the rise time V_C^r, lowering the magnitude of the power noise.

The charge Q_{dis}^r required to support the current demand during the rise time of the current load is equal to the area of the triangle formed by $I_{\max}$ and t_r. The required charge is provided by the on-chip decoupling capacitor. The voltage across the decoupling capacitor drops below the power supply level by ΔV_C^r. The required charge during t_r is[2]

$$Q_{\text{dis}}^r = \frac{I_{\max} \times t_r}{2} = C_{\text{dec}} \times \Delta V_C^r, \tag{12.6}$$

where Q_{dis}^r is the charge drawn by the current load during t_r and ΔV_C^r is the voltage drop across the decoupling capacitor at t_p. From (12.6),

$$\Delta V_C^r = \frac{I_{\max} \times t_r}{2\,C_{\text{dec}}}. \tag{12.7}$$

By time t_p, the voltage drop as seen from the current load is the sum of the ohmic IR drop, the inductive $L\,dI/dt$ drop, and the voltage drop across the decoupling capacitor. Alternatively, the power noise is further increased by the voltage drop ΔV_C^r. In this case, the voltage at the current load is

$$V_{load}^r = V_{\text{dd}} - I \times R - L\frac{dI}{dt} - \Delta V_C^r, \tag{12.8}$$

where R and L are, respectively, the parasitic resistance and inductance of the P/G lines. Linearly approximating the current load, dI is assumed equal to $I_{\max}$ and dt to t_r. Note that the last term in (12.8) accounts for the voltage drop ΔV_C^r across the decoupling capacitor during the rise time of the current at the load.

[2] In the general case with a given current profile, the required charge can be estimated as the integral of $I_{load}(t)$ from 0 to t_r.

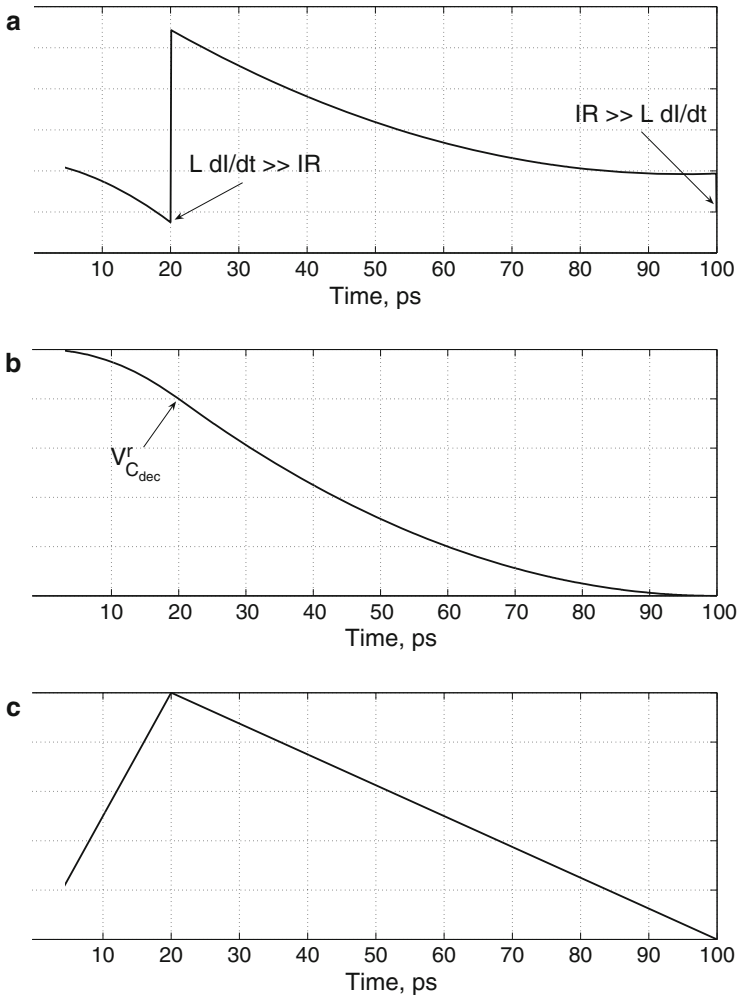


Fig. 12.4 Power distribution noise during discharge of an on-chip decoupling capacitor: $I_{\max} = 100\,\text{mA}$, $V_{\text{dd}} = 1\,\text{V}$, $t_r = 20\,\text{ps}$, $t_f = 80\,\text{ps}$, $R = 100\,\text{m}\Omega$, $L = 15\,\text{pH}$, and $C_{\text{dec}} = 50\,\text{pF}$; (**a**) voltage across the terminals of the current load, (**b**) voltage across the decoupling capacitor, (**c**) current load modeled as a triangular current source. For these parameters, the parasitic impedance of the metal lines connecting the decoupling capacitor to the current load is larger than the critical impedance. The inductive noise therefore dominates the resistive noise and (12.5) underestimates the required decoupling capacitance. The resulting voltage drop on the power terminal of a current load is therefore larger than the maximum tolerable noise

Assuming that $V^r_{load} \geq 0.9\,V_{\text{dd}}$, substituting (12.7) into (12.8), and solving for C_{dec}, the minimum on-chip decoupling capacitance to support the current demand during t_r is

$$C_{\text{dec}}^{r} \geq \frac{I_{\max} \times t_r}{2\left(0.1\,V_{\text{dd}} - I \times R - L\frac{dI}{dt}\right)}. \tag{12.9}$$

Note that if $L\,dI/dt \gg IR$, C_{dec} is excessively large. The voltage drop at the end of the switching event is hence always smaller than the maximum tolerable noise.

Also note that, as opposed to (12.5), (12.9) depends upon the parasitic impedance of the on-chip power distribution system. Alternatively, in the case of the dominant inductive noise, the required charge released by the decoupling capacitor is determined by the parasitic resistance and inductance of the P/G lines connecting the decoupling capacitor to the current load.

12.3.3 Critical Line Length

Assuming the impedance of a single line, the critical line length d_{crit} can be determined by setting C_{dec}^{r} equal to C_{dec}^{f},

$$\frac{I_{\max} \times t_r}{\left(0.1\,V_{\text{dd}} - I\,r\,d_{\text{crit}} - l\,d_{\text{crit}}\frac{dI}{dt}\right)} = \frac{I_{\max} \times (t_r + t_f)}{0.1\,V_{\text{dd}}}. \tag{12.10}$$

Solving (12.10) for d_{crit},

$$d_{\text{crit}} = \frac{0.1\,V_{\text{dd}}\left(1 - \dfrac{t_r}{t_r + t_f}\right)}{I\,r + l\dfrac{dI}{dt}}. \tag{12.11}$$

For a single line connecting a current load to a decoupling capacitor, the minimum required on-chip decoupling capacitor is determined by (12.5) for lines shorter than d_{crit} and by (12.9) for lines longer than d_{crit}, as illustrated in Fig. 12.5. Note that for a line length equal to d_{crit}, (12.5) and (12.9) result in the same required capacitance. Also note that the maximum length of a single line is determined by (12.2). A closed-form solution for the critical line length has not been developed for the case of multiple current paths existing between the current load and a decoupling capacitor. In this case, the impedance of the power grid connecting a decoupling capacitor to a current load is extracted and compared to the critical impedance. Either (12.5) or (12.9) is utilized to estimate the required on-chip decoupling capacitance.

The dependence of the critical line length d_{crit} on the rise time t_r of the current load as determined by (12.11) is depicted in Fig. 12.6. From Fig. 12.6, the critical line length decreases sublinearly with shorter rise times. Hence, the critical line length will decrease in future nanometer technologies as transition times become shorter, significantly increasing the required on-chip decoupling capacitance. Also note that d_{crit} is determined by $\frac{t_r}{t_f}$, increasing with larger fall times.

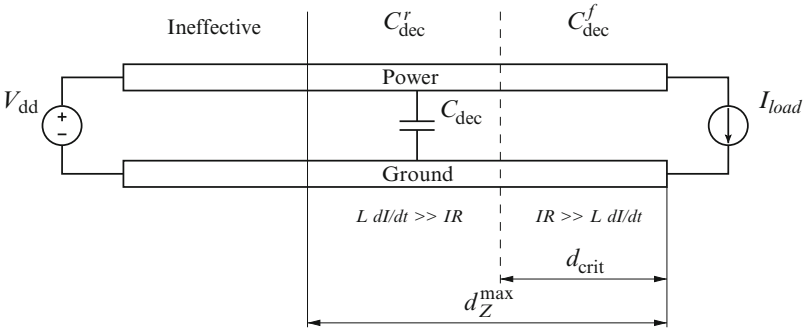


Fig. 12.5 Critical line length of an interconnect between a decoupling capacitor and a current load. The minimum required on-chip decoupling capacitance is determined by (12.5) for lines shorter than d_{crit} and by (12.9) for lines longer than d_{crit}. The decoupling capacitor is ineffective beyond the maximum effective radius as determined by the target impedance d_Z^{max}

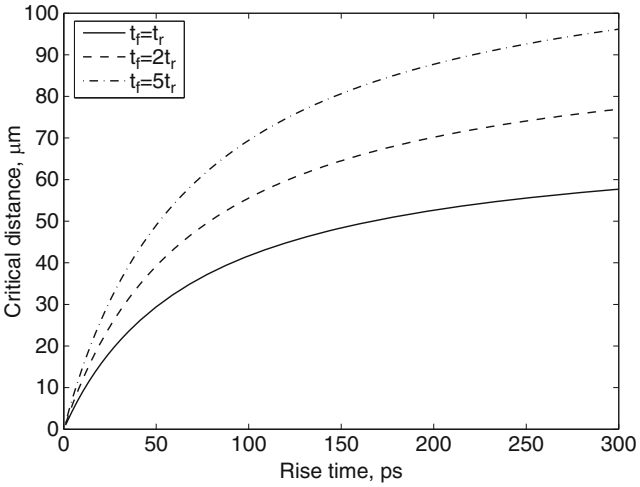


Fig. 12.6 Dependence of the critical line length d_{crit} on the rise time of the current load: $I_{max} = 0.1\,A$, $V_{dd} = 1\,V$, $r = 0.007\,\Omega/\mu m$, and $l = 0.5\,pH/\mu m$. Note that d_{crit} is determined by $\frac{t_r}{t_f}$, increasing with larger t_f. The critical line length will shrink in future nanometer technologies as transition times become shorter

Observe in Fig. 12.5 that the design space for determining the required on-chip decoupling capacitance is broken into two regions by the critical line length. The design space for determining the required on-chip decoupling capacitance (C^r_{dec} and C^f_{dec}) is depicted in Fig. 12.7. For the example parameters shown in Fig. 12.7, the critical line length is 125 µm. Note that the required on-chip decoupling capacitance C^r_{dec} depends upon the parasitic impedance of the metal lines connecting the decoupling capacitor to the current load. Thus, for lines longer than d_{crit}, C^r_{dec} increases exponentially as the separation between the decoupling capacitor and the

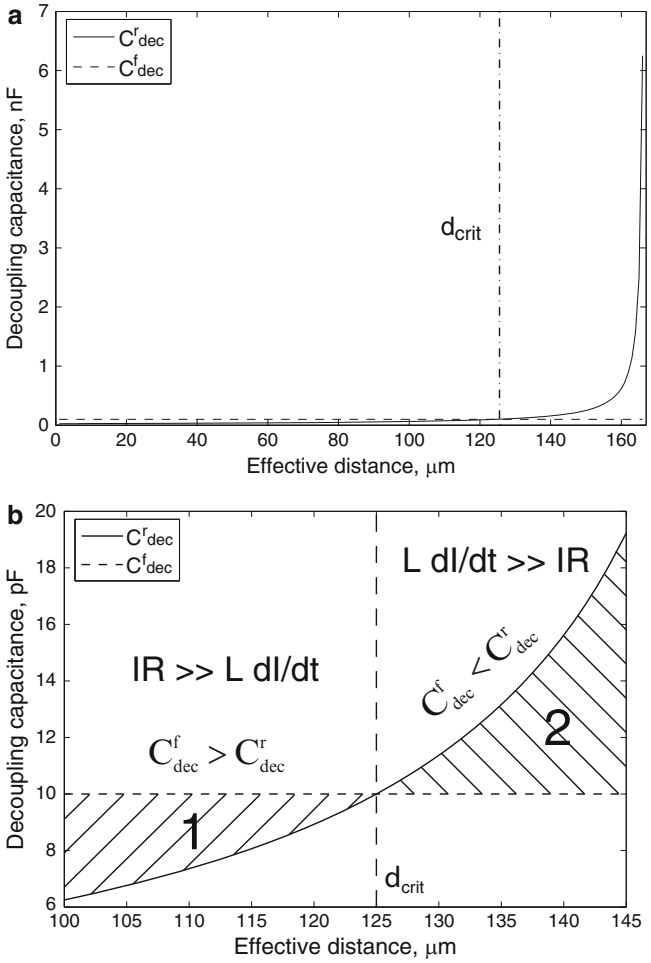


Fig. 12.7 Design space for determining minimum required on-chip decoupling capacitance: $I_{\max} = 50\,\text{mA}$, $V_{\text{dd}} = 1\,\text{V}$, $r = 0.007\,\Omega/\mu\text{m}$, $l = 0.5\,\text{pH}/\mu\text{m}$, $t_r = 100\,\text{ps}$, and $t_f = 300\,\text{ps}$; (**a**) design space for determining the minimum required on-chip decoupling capacitance is broken into two regions by d_{crit}, (**b**) design space around d_{crit}. For the example parameters, the critical line length is 125 μm. In region 1, C^f_{dec} is greater than C^r_{dec} and does not depend upon the parasitic impedance. In region 2, however, C^r_{dec} dominates, increasing rapidly with distance between the decoupling capacitor and the current load

current load increases, as shown in Fig. 12.7a. Also note that for lines shorter than d_{crit}, the required on-chip decoupling capacitance does not depend upon the parasitic impedance of the power distribution grid. Alternatively, in the case of the dominant resistive drop, the required on-chip decoupling capacitance C^f_{dec} is constant and greater than C^r_{dec} (see region 1 in Fig. 12.7b). If $L\,dI/dt$ noise dominates the IR noise (the line length is greater than d_{crit}), the required on-chip decoupling capacitance

C^r_{dec} increases substantially with line length and is greater than C^f_{dec} (see region 2 in Fig. 12.7b). Conventional techniques therefore significantly underestimate the required decoupling capacitance in the case of the dominant inductive noise. Note that in region 1, the parasitic impedance of the metal lines connecting a decoupling capacitor to the current load is not important. In region 2, however, the parasitic impedance of the P/G lines should be considered. A tradeoff therefore exists between the size of C^r_{dec} and the distance between the decoupling capacitor and the current load. As C^r_{dec} is placed closer to the current load, the required capacitance can be significantly reduced.

12.4 Effective Radius as Determined by Charge Time

Once discharged, a decoupling capacitor must be fully charged to support the current demands during the following switching event. If the charge on the capacitor is not fully restored during the relaxation time between two consecutive switching events (the charge time), the decoupling capacitor will be gradually depleted, becoming ineffective after several clock cycles. A maximum effective radius, therefore, exists for an on-chip decoupling capacitor as determined during the charging phase for a target charge time. Similar to the effective radius based on the target impedance presented in Sect. 12.2, an on-chip decoupling capacitor should be placed in close proximity to the power supply (the power pins) to be effective.

To determine the current flowing through a decoupling capacitor during the charging phase, the parasitic impedance of a power distribution system is modeled as a series *RL* circuit between the decoupling capacitor and the power supply, as shown in Fig. 12.8. When the discharge is completed, the switch is closed and the charge is restored on the decoupling capacitor. The initial voltage V_C^0 across the decoupling capacitor is determined by the maximum voltage drop during discharge.

For the circuit shown in Fig. 12.8, the KVL equation for the current in the circuit is [287]

$$L\frac{di_{\text{ch}}}{dt} + R\,i_{\text{ch}} + \frac{1}{C_{\text{dec}}}\int i_{\text{ch}}\,dt = V_{\text{dd}}. \tag{12.12}$$

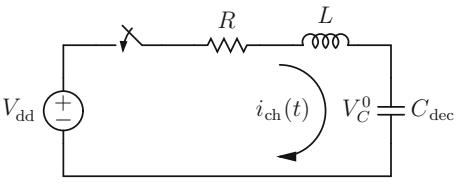


Fig. 12.8 Circuit charging an on-chip decoupling capacitor. The parasitic impedance of the power distribution system connecting the decoupling capacitor to the power supply is modeled by a series *RL* circuit

Differentiating (12.12),

$$L\frac{d^2 i_{\text{ch}}}{dt^2} + R\frac{di_{\text{ch}}}{dt} + \frac{1}{C_{\text{dec}}} i_{\text{ch}} = 0. \tag{12.13}$$

Equation (12.13) is a second order linear differential equation with the characteristic equation,

$$s^2 + \frac{R}{L}s + \frac{1}{LC_{\text{dec}}} = 0. \tag{12.14}$$

The general solution of (12.13) is

$$i_{\text{ch}}(t) = K_1 e^{s_1 t} + K_2 e^{s_2 t}, \tag{12.15}$$

where s_1 and s_2 are the roots of (12.14),

$$s_{1,2} = -\frac{R}{2L} \pm \sqrt{\left(\frac{R}{2L}\right)^2 - \frac{1}{LC_{\text{dec}}}}. \tag{12.16}$$

Note that (12.15) represents the solution of (12.13) as long as the system is overdamped. The damping factor is therefore greater than one, i.e.,

$$\left(\frac{R}{L}\right)^2 > \frac{4}{LC}. \tag{12.17}$$

For a single line, from (12.17), the critical line length resulting in an overdamped system is

$$D > \frac{4\,l}{r^2\,C_{\text{dec}}}, \tag{12.18}$$

where C_{dec} is the on-chip decoupling capacitance, and l and r are, respectively, the per length inductance and resistance. Inequality (12.18) determines the critical length of a line resulting in an overdamped system. Note that for typical values of r and l in a 90 nm CMOS technology, a power distribution system with a decoupling capacitor is overdamped for on-chip interconnects longer than several micrometers. Equation (12.15) is therefore a general solution of (12.13) for a scaled CMOS technology.

Initial conditions are applied to determine the arbitrary constants K_1 and K_2 in (12.15). The current charging the decoupling capacitor during the charging phase is

$$i_{\mathrm{ch}}(t) = \frac{I_{\max}\,(t_r + t_f)}{4\,L\,C_{\mathrm{dec}}\sqrt{\left(\frac{R}{2L}\right)^2 - \frac{1}{LC_{\mathrm{dec}}}}} \times \left\{ \exp\left[\left(-\frac{R}{2L} + \sqrt{\left(\frac{R}{2L}\right)^2 - \frac{1}{LC_{\mathrm{dec}}}}\right) t\right] - \exp\left[\left(-\frac{R}{2L} - \sqrt{\left(\frac{R}{2L}\right)^2 - \frac{1}{LC_{\mathrm{dec}}}}\right) t\right] \right\}. \tag{12.19}$$

The voltage across the decoupling capacitor during the charging phase can be determined by integrating (12.19) from zero to the charge time,

$$V_C(t) = \frac{1}{C_{\mathrm{dec}}} \int_0^{t_{\mathrm{ch}}} i_{\mathrm{ch}}(t)\,dt, \tag{12.20}$$

where t_{ch} is the charge time, and $V_C(t)$ and $i_{\mathrm{ch}}(t)$ are, respectively, the voltage across the decoupling capacitor and the current flowing through the decoupling capacitor during the charging phase. Substituting (12.19) into (12.20) and integrating from zero to t_{ch}, the voltage across the decoupling capacitor during the charging phase is

$$V_{C_{\mathrm{dec}}}(t_{\mathrm{ch}}) = \frac{I_{\max}\,(t_r + t_f)}{4\,C_{\mathrm{dec}}^2\,L\sqrt{\left(\frac{R}{2L}\right)^2 - \frac{1}{LC_{\mathrm{dec}}}}} \times \left\{ \frac{\exp\left[\left(-\frac{R}{2L} + \sqrt{\left(\frac{R}{2L}\right)^2 - \frac{1}{LC_{\mathrm{dec}}}}\right) t_{\mathrm{ch}}\right] - 1}{-\frac{R}{2L} + \sqrt{\left(\frac{R}{2L}\right)^2 - \frac{1}{LC_{\mathrm{dec}}}}} + \frac{1 - \exp\left[\left(-\frac{R}{2L} - \sqrt{\left(\frac{R}{2L}\right)^2 - \frac{1}{LC_{\mathrm{dec}}}}\right) t_{\mathrm{ch}}\right]}{-\frac{R}{2L} - \sqrt{\left(\frac{R}{2L}\right)^2 - \frac{1}{LC_{\mathrm{dec}}}}} \right\}. \tag{12.21}$$

Observe that the criterion for estimating the maximum effective radius of an on-chip decoupling capacitor as determined by the charge time is transcendental. A closed-form expression is therefore not available for determining the maximum

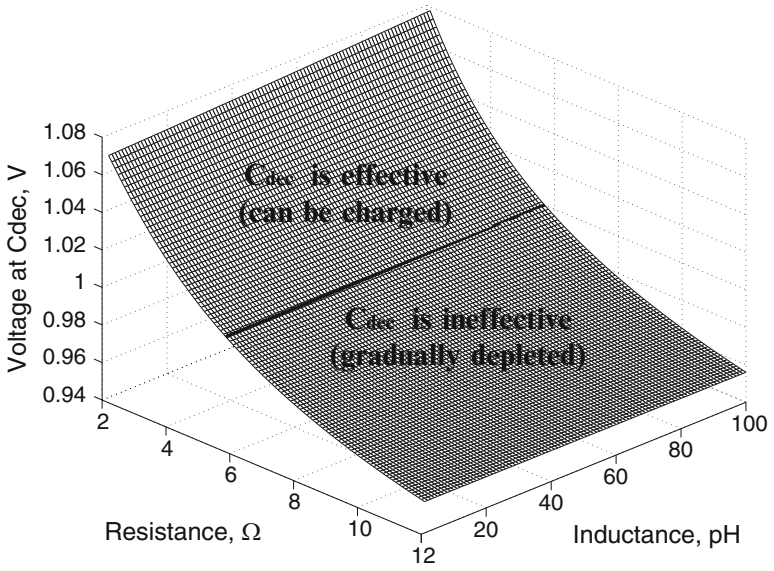


Fig. 12.9 Design space for determining the maximum tolerable parasitic resistance and inductance of a power distribution grid: $I_{\max} = 100\,\text{mA}$, $t_r = 100\,\text{ps}$, $t_f = 300\,\text{ps}$, $C_{\text{dec}} = 100\,\text{pF}$, $V_{\text{dd}} = 1\,\text{V}$, and $t_{\text{ch}} = 400\,\text{ps}$. For a target charge time, the maximum resistance and inductance produce a voltage across the decoupling capacitor that is greater or equal to the power supply voltage (region above the *dark line*). Note that the maximum voltage across the decoupling capacitor is the power supply voltage. A design space that produces a voltage greater than the power supply means that the charge on the decoupling capacitor can be restored within t_{ch}

effective radius of an on-chip decoupling capacitor during the charging phase. Thus from (12.21), a design space can be graphically described in order to determine the maximum tolerable resistance and inductance that permit the decoupling capacitor to be recharged within a given t_{ch}, as shown in Fig. 12.9. The parasitic resistance and inductance should be maintained below the maximum tolerable values, permitting the decoupling capacitor to be charged during the relaxation time.

Note that as the parasitic resistance of the power delivery network decreases, the voltage across the decoupling capacitor increases exponentially. In contrast, the voltage across the decoupling capacitor during the charging phase is almost independent of the parasitic inductance, slightly increasing with inductance. This phenomenon is due to the behavior that an inductor resists sudden changes in the current. Alternatively, an inductor maintains the charging current at a particular level for a longer time. Thus, the decoupling capacitor is charged faster.

12.5 Design Methodology for Placing On-Chip Decoupling Capacitors

A design methodology for placing on-chip decoupling capacitors based on the maximum effective radii is illustrated in Fig. 12.10. The maximum effective radius based on the target impedance is determined from (12.2) for a particular current load (circuit block), power supply voltage, and allowable ripple. The minimum required on-chip decoupling capacitance is estimated to support the required current demand. If the resistive drop is larger than the inductive drop, (12.5) is used to determine the required on-chip decoupling capacitance. If $L\,dI/dt$ noise dominates, the on-chip decoupling capacitance is determined by (12.9). In the case of a single line connecting a decoupling capacitor to a current load, the critical wire length is determined by (12.11).

The maximum effective distance based on the charge time is determined from (12.21). Note that (12.21) results in a range of tolerable parasitic resistance and inductance of the metal lines connecting the decoupling capacitor to the power supply. Also note that the on-chip decoupling capacitor should be placed such that both the power supply and the current load are located inside the effective radius, as shown in Fig. 12.11. If this allocation is not possible, the current load (circuit block) should be partitioned into several blocks and the on-chip decoupling capacitors should be allocated for each block, satisfying both effective radii requirements. The effective radius as determined by the target impedance does not depend upon the decoupling capacitance. In contrast, the effective radius as determined by the charge time is inversely proportional to C_{dec}^2. The on-chip decoupling capacitors should be distributed across the circuit to provide sufficient charge for each functional unit.

12.6 Model of On-Chip Power Distribution Network

In order to determine the effective radii of an on-chip decoupling capacitor and the effect on the noise distribution, a model of a power distribution network is required. On-chip power distribution networks in high performance ICs are commonly modeled as a mesh. Early in the design process, minimal physical information characterizing the P/G structure is available. A simplified model of a power distribution system is therefore appropriate. For simplicity, equal segments within a mesh structure are assumed. The current demands of a particular module are modeled as current sources with equivalent magnitude and switching activities. The current load is located at the center of a circuit module which determines the connection point of the circuit module to the power grid. The parasitic resistance and inductance of the package are also included in the model as an equivalent series resistance R_p and inductance L_p. Note that the parasitic capacitance of the power distribution grid provides a portion of the decoupling capacitance, providing additional charge to the current loads. The on-chip decoupling capacitance intentionally added to the IC is

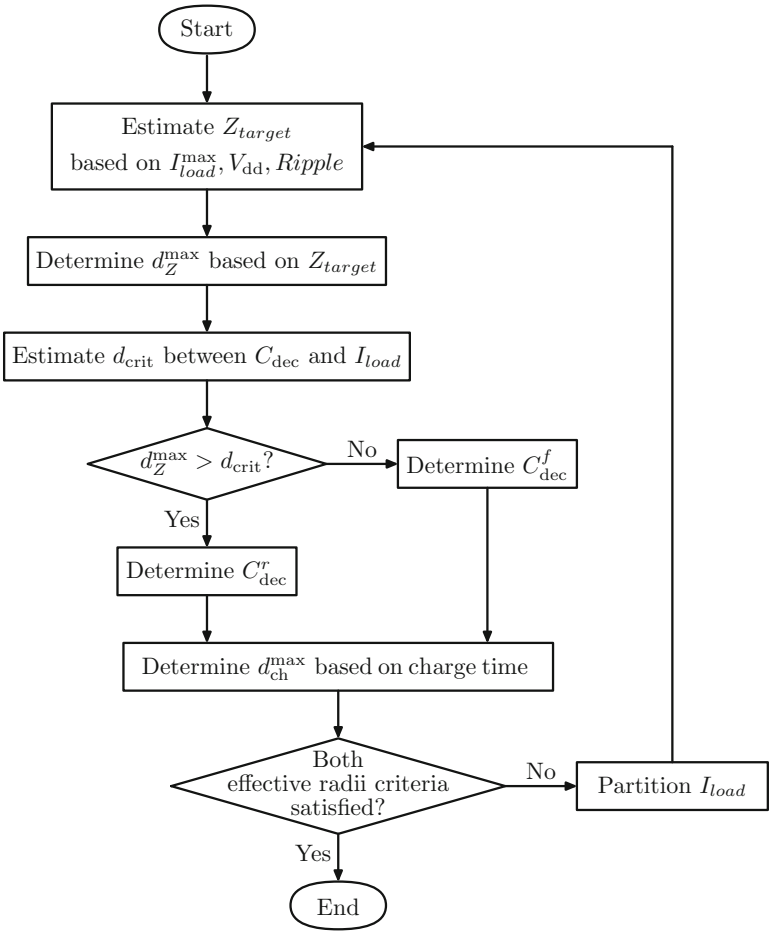


Fig. 12.10 Design flow for placing on-chip decoupling capacitors based on the maximum effective radii

typically more than an order of magnitude greater than the parasitic capacitance of the on-chip power grid. The parasitic capacitance of the power delivery network is, therefore, neglected.

Typical effective radii of an on-chip decoupling capacitor is in the range of several hundred micrometers. In order to determine the location of an on-chip decoupling capacitor, the size of each *RL* mesh segment should be much smaller than the effective radii. In modern high performance ICs such as microprocessors with die sizes approaching 1.5 in. by 1.5 in., a fine mesh is infeasible to simulate. In the case of a coarse mesh, the effective radius is smaller than the size of each segment. The location of each on-chip decoupling capacitor, therefore, cannot be accurately determined. To resolve this dilemma, the accuracy of the capacitor location can be traded off with the complexity of the power distribution network. A hot

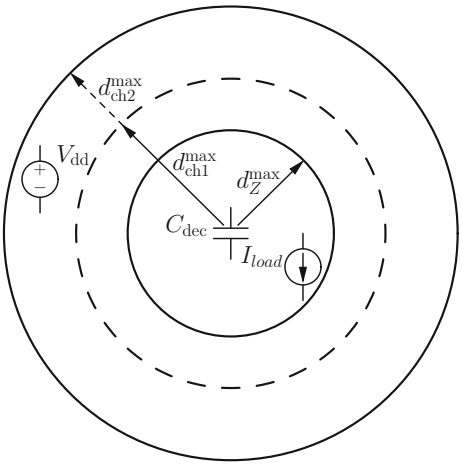


Fig. 12.11 The effective radii of an on-chip decoupling capacitor. The on-chip decoupling capacitor is placed such that both the current load and the power supply are located inside the effective radius. The maximum effective radius as determined by the target impedance d_Z^{max} does not depend on the decoupling capacitance. The maximum effective radius as determined by the charge time is inversely proportional to C_{dec}^2. If the power supply is located outside the effective radius d_{ch1}^{max}, the current load should be partitioned, resulting in a smaller decoupling capacitor and, therefore, an increased effective distance d_{ch2}^{max}

spot (an area where the power supply voltage drops below the minimum tolerable level) is first determined based on a coarse mesh, as shown in Fig. 12.12. A finer mesh is used next within each hot spot to accurately estimate the effective radius of the on-chip decoupling capacitor. Note that in a mesh structure, the maximum effective radius is the Manhattan distance between two points. In disagreement with Fig. 12.11, the overall effective radius is actually shaped more like a diamond, as illustrated in Fig. 12.13.

In modern high performance ICs, up to 3000 I/O pins can be necessary [286]. Only half of the I/O pads are typically used to distribute power. The other half is dedicated to signaling. Assuming an equal distribution of power and ground pads, a quarter of the total number of pads is typically available for power *or* ground delivery. For high performance ICs with die sizes of 1.5 in. by 1.5 in. inside a flip-chip package, the distance between two adjacent power or ground pads is about 1300 μm. By modeling the flip chip area array by a six by six distributed *RL* mesh, the accuracy in determining the effective radii of an on-chip decoupling capacitor is traded off with the computational complexity required to analyze the power delivery network. In this chapter, an on-chip power distribution system composed of the four closest power pins is modeled as an *RL* mesh of forty by forty equal segments to accurately determine the maximum effective distance of an on-chip decoupling capacitor. Note that this approach of modeling a power distribution system is applicable to ICs with both conventional low cost and advanced high performance packaging.

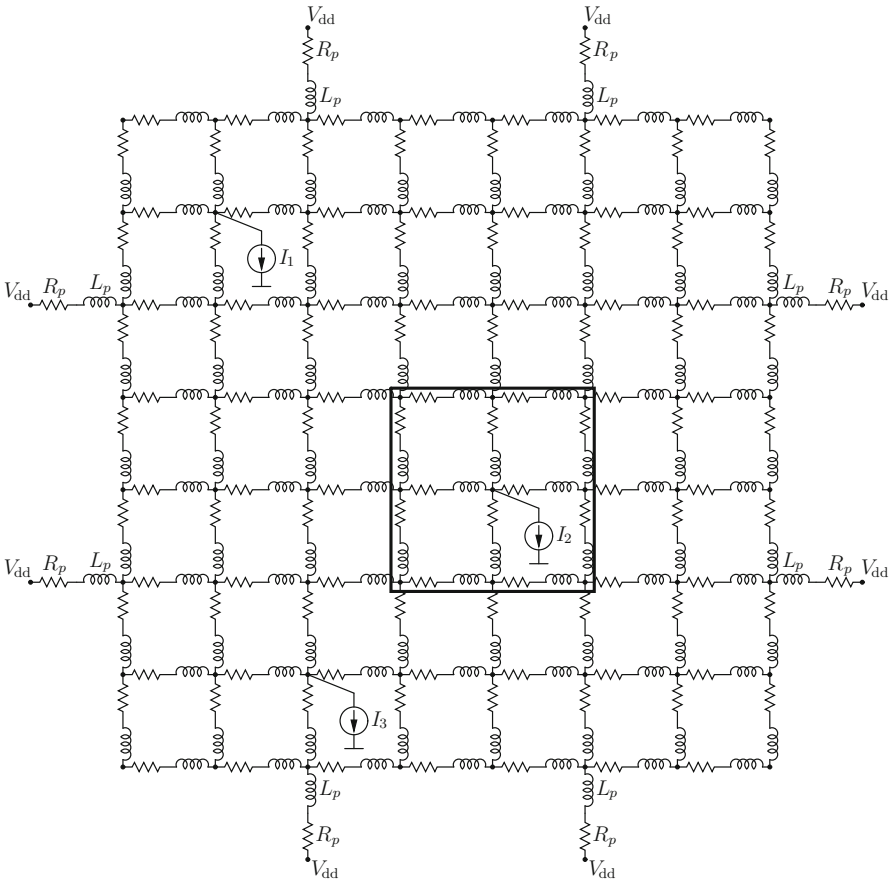


Fig. 12.12 Model of a power distribution network. The on-chip power delivery system is modeled as a distributed *RL* mesh with seven by seven equal segments. The current loads are modeled as current sources with equivalent magnitude and switching activities. Impedances R_p and L_p denote, respectively, the parasitic resistance and inductance of the package. The *rectangle* denotes a "hot" spot—the area where the power supply voltage drops below the minimum tolerable level

12.7 Case Study

The dependence of the effective radii of an on-chip decoupling capacitor on a power distribution system is described in this section to quantitatively illustrate these concepts. The load is modeled as a triangular current source with a 100 ps rise time and 300 ps fall time. The maximum tolerable ripple at the load is 10 % of the power supply voltage. The relaxation time between two consecutive switching events (charge time) is 400 ps. Two scenarios are considered for determining the effective radii of an on-chip decoupling capacitor. In the first scenario, an on-chip decoupling capacitor is connected to the current load by a single line (local connectivity).

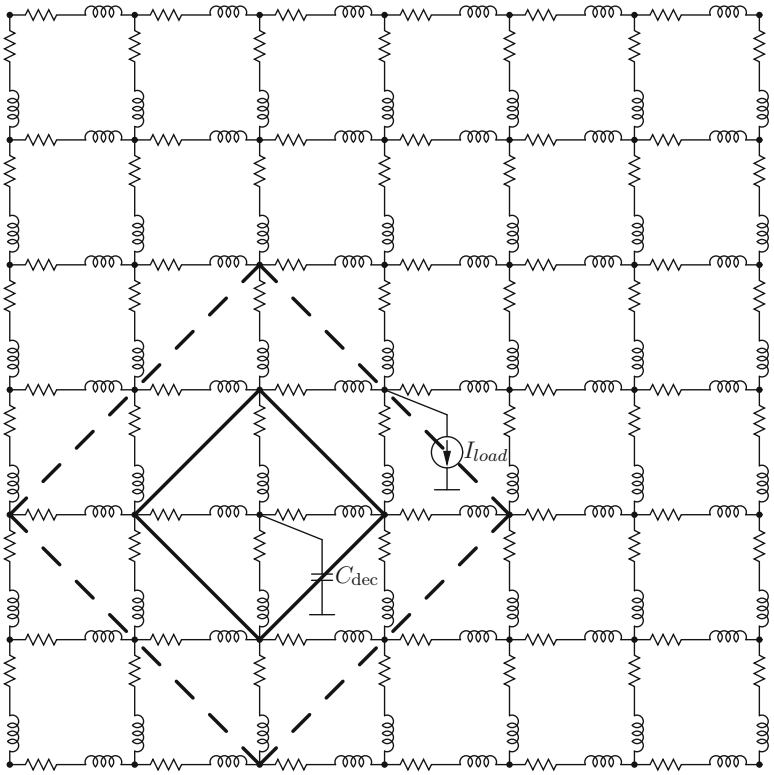


Fig. 12.13 Effective radii of an on-chip decoupling capacitor. For a power distribution system modeled as a distributed *RL* mesh, the maximum effective radius is the Manhattan distance between two points. The overall effective radius is therefore shaped like a *diamond*

In the second scenario, the on-chip decoupling capacitors are connected to the current loads by an on-chip power distribution grid (global connectivity). A flip-chip package is assumed. An on-chip power distribution system with a flip-chip pitch (the area formed by the four closest pins) is modeled as an *RL* distributed mesh of forty by forty equal segments to accurately determine the maximum effective distance of an on-chip decoupling capacitor. The parasitic resistance and inductance of the package (the four closest pins of a flip-chip package) are also included in the model. The methodology for placing on-chip decoupling capacitors provides a highly accurate estimate of the magnitude and location of the on-chip decoupling capacitors. The maximum error of the resulting power noise is less than 0.1 % as compared to SPICE.

For a single line, the maximum effective radii as determined by the target impedance and charge time for three sets of on-chip parasitic resistances and inductances are listed in Table 12.1. These three scenarios listed in Table 12.1 represent typical values of the parasitic resistance and inductance of the top, intermediate, and bottom

Table 12.1 Maximum effective radii of an on-chip decoupling capacitor for a single line connecting a decoupling capacitor to a current load

Metal layer	Resistance (Ω/μm)	Inductance (pH/μm)	I_{load} (A)	C_{dec} (pF)	d_{max} (μm)	
					Z	t_{ch}
Top	0.007	0.5	0.01	20	310.8	1166
	0.007	0.5	0.1	200	31.1	116
	0.007	0.5	1	2000	3.1	11.6
Intermediate	0.04	0.3	0.01	183	226.2	24.2
	0.04	0.3	0.1	1773	22.6	2.4
	0.04	0.3	1	45,454	2.3	0.2
Bottom	0.1	0.1	0.01	50,000	99.8	0
	0.1	0.1	0.1	∞	0	0
	0.1	0.1	1	∞	0	0

$V_{dd} = 1$ V, $V_{ripple} = 100$ mV, $t_r = 100$ ps, $t_f = 300$ ps, $t_{ch} = 400$ ps

layers of on-chip interconnects in a 90 nm CMOS technology [286]. In the case of the top metal layer, the maximum effective distance as determined by the target impedance is smaller than the critical distance as determined by (12.11). Hence, $IR \gg L\, dI/dt$, and the required on-chip decoupling capacitance is determined by (12.5). Note that the decoupling capacitance increases linearly with the current load. For a typical parasitic resistance and inductance of the intermediate and bottom layers of the on-chip interconnects, the effective radius as determined by the target impedance is longer than the critical distance d_{crit}. In this case, the overall voltage drop at the current load is determined by the inductive noise. The on-chip decoupling capacitance can therefore be estimated by (12.9).

In the case of an *RL* mesh, the maximum effective radii as determined by the target impedance and charge time for three sets of on-chip parasitic resistances and inductances are listed in Table 12.2. From (12.11), for the parameters listed in Table 12.2, the critical voltage drop is 75 mV. If the voltage fluctuations at the current load do not exceed the critical voltage, $IR \gg L\, dI/dt$ and the required on-chip decoupling capacitance is determined by (12.5). Note that for the aforementioned three interconnect scenarios, assuming a 10 mA current load, the maximum effective radii of the on-chip decoupling capacitor based on the target impedance and charge time are larger than forty cells (the longest distance within the mesh from the center of the mesh to the corner). The maximum effective radii of the on-chip decoupling capacitor is therefore larger than the pitch size. The decoupling capacitor can therefore be placed anywhere inside the pitch. For a 100 mA current load, the voltage fluctuations at the current load exceed the critical voltage drop. The $L\, dI/dt$ noise dominates and the required on-chip decoupling capacitance is determined by (12.9).

The effective radii of an on-chip decoupling capacitor decreases linearly with current load. The optimal size of an *RL* distributed mesh should therefore be determined for a particular current demand. If the magnitude of the current requirements is low, the mesh can be coarser, significantly decreasing the simulation time.

Table 12.2 Maximum effective radii of an on-chip decoupling capacitor for an on-chip power distribution grid modeled as a distributed *RL* mesh

Metal layer	Resistance (Ω/μm)	Inductance (pH/μm)	I_{load} (A)	C_{dec} (pF)	$d_{\max}$ (cells)	
					Z	t_{ch}
Top	0.007	0.5	0.01	20	>40	>40
	0.007	0.5	0.1	357	2	>40
	0.007	0.5	1	–	<1	–
Intermediate	0.04	0.3	0.01	20	>40	>40
	0.04	0.3	0.1	227	1	<1
	0.04	0.3	1	–	<1	–
Bottom	0.1	0.1	0.01	20	>40	>40
	0.1	0.1	0.1	–	<1	–
	0.1	0.1	1	–	<1	–

$V_{\text{dd}} = 1\,\text{V}$, $V_{ripple} = 100\,\text{mV}$, $t_r = 100\,\text{ps}$, $t_f = 300\,\text{ps}$, $t_{\text{ch}} = 400\,\text{ps}$, cell size is $32.5 \times 32.5\,\mu\text{m}$

For a 10 mA current load, the effective radii as determined from both the target impedance and charge time are longer than the pitch size. Thus, the distributed mesh is overly fine. For a current load of 1 A, the effective radii are shorter than one cell, meaning that the distributed *RL* mesh is overly coarse. A finer mesh should therefore be used to accurately estimate the maximum effective radii of the on-chip decoupling capacitor. In general, the cells within the mesh should be sized based on the current demand and the acceptable computational complexity (or simulation budget). As a rule of thumb, a coarser mesh should be used on the perimeter of each grid pitch. A finer mesh should be utilized around the current loads.

Note that in both cases, C^r_{dec} as determined by (12.9) increases rapidly with the effective radius based on the target impedance, becoming infinite at $d_Z^{\max}$. In this case study, the decoupling capacitor is allocated at almost the maximum effective distance $d_Z^{\max}$, simulating the worst case scenario. The resulting C_{dec} is therefore significantly large. As the decoupling capacitor is placed closer to the current load, the required on-chip decoupling capacitance as estimated by (12.9) can be reduced. A tradeoff therefore exists between the maximum effective distance as determined by the target impedance and the size of the minimum required on-chip decoupling capacitance (if the overall voltage drop at the current load is primarily caused by the inductive $L\,dI/dt$ drop).

The effective radii listed in Table 12.1 are determined for a single line between the current load or power supply and the decoupling capacitor. In the case of a power distribution grid modeled as a distributed *RL* mesh, multiple paths are connected in parallel, increasing the effective radii. For instance, comparing Table 12.1 to Table 12.2, note that the maximum effective radii as determined by the target impedance are increased about three times and two times for the top metal layers with, respectively, a 10 and 100 mA current load. Note also that for typical values of the parasitic resistance and inductance of a power distribution grid, the effective

radius as determined by the target impedance is longer than the radius based on the charge time for intermediate and bottom metal layers. For top metal layers, however, the effective radius as determined by the target impedance is typically shorter than the effective radius based on the charge time.

Also note that the maximum effective radius as determined by the charge time decreases quadratically with the decoupling capacitance. The maximum effective distance as determined by the charge time becomes impractically short for large decoupling capacitances. For the bottom metal layer, the maximum effective radius based on the charge time approaches zero. Note that the maximum effective radius during the charging phase has been evaluated for the case where the decoupling capacitor is charged to the power supply voltage. In practical applications, this constraint can be relaxed, assuming the voltage across the decoupling capacitor is several millivolts smaller than the power supply. In this case, the effective radius of the on-chip decoupling capacitor as determined by the charge time can be significantly increased.

The maximum effective radius as determined by the charge time becomes impractically short for large decoupling capacitors, making the capacitors ineffective. In this case, the decoupling capacitor should be placed closer to the current load, permitting the decoupling capacitance to be decreased. Alternatively, the current load can be partitioned into several blocks, lowering the requirements on a specific local on-chip decoupling capacitance. The parasitic impedance between the decoupling capacitor and the current load and power supply should also be reduced, if possible, increasing the maximum effective radii of the on-chip decoupling capacitors.

12.8 Design Implications

A larger on-chip decoupling capacitance is required to support increasing current demands. The maximum available on-chip decoupling capacitance, which can be placed in the vicinity of a particular circuit block, is limited however by the maximum capacitance density of a given technology, as described in Chap. 13. Large functional units (current loads) should therefore be partitioned into smaller blocks with local on-chip decoupling capacitors to enhance the likelihood of fault-free operation of the entire system. An important concept described in this chapter is that on-chip decoupling capacitors are a *local* phenomenon. Thus, the methodology for placing and sizing on-chip decoupling capacitors results in a greatly reduced budgeted on-chip decoupling capacitance as compared to a uniform (or blind) placement of on-chip decoupling capacitors into any available white space [284].

Typically, multiple current loads exist in an IC. An on-chip decoupling capacitor is placed in the vicinity of the current load such that both the current load and the power supply are within the maximum effective radius. Assuming a uniform distribution of the current loads, a schematic example placement of the on-chip decoupling capacitors is shown in Fig. 12.14. Each decoupling capacitor provides

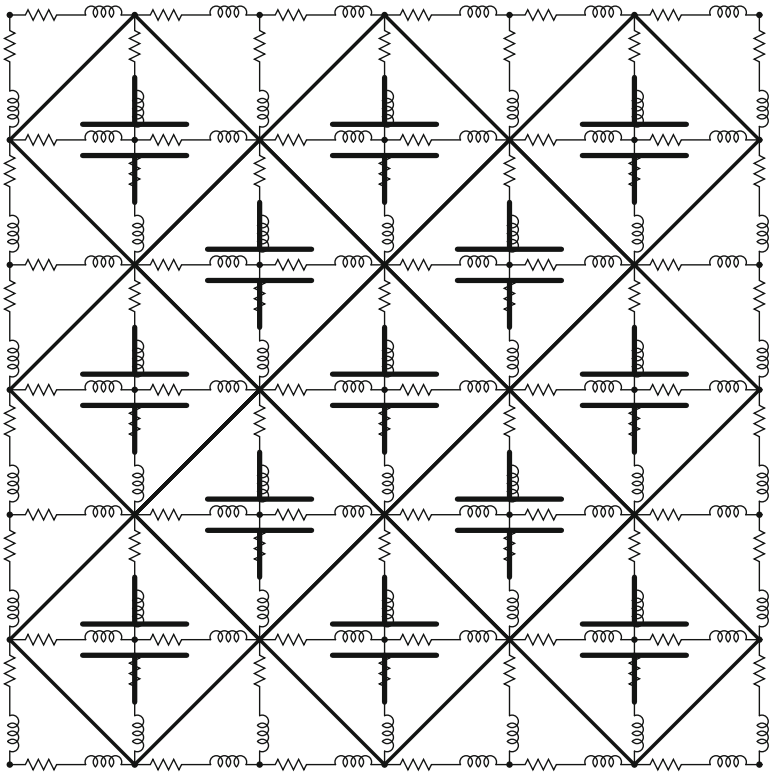


Fig. 12.14 A schematic example allocation of on-chip decoupling capacitors across an IC. Similar current loads are assumed to be uniformly distributed on the die. Each on-chip decoupling capacitor provides sufficient charge to the current load(s) within the maximum effective radius

sufficient charge to the current load(s) within the maximum effective radius. Multiple on-chip decoupling capacitors are placed to provide charge to all of the circuit blocks. In general, the size and location of an on-chip decoupling capacitor are determined by the required charge (drawn by the local transient current loads) and certain system parameters (such as the per length resistance and inductance, power supply voltage, maximum tolerable ripple, and the switching characteristics of the current load).

12.9 Summary

A design methodology for placing and sizing on-chip decoupling capacitors based on effective radii is presented in this chapter and can be summarized as follows.

- On-chip decoupling capacitors have traditionally been allocated into the available white space on a die, i.e., using an unsystematic or ad hoc approach

- On-chip decoupling capacitors behave locally and should therefore be treated as a local phenomenon. The efficiency of on-chip decoupling capacitors depends upon the impedance of the power/ground lines connecting the capacitors to the current loads and power supplies
- Closed-form expressions for the maximum effective radii of an on-chip decoupling capacitor based on a target impedance (during discharge) and charge time (during charging phase) are described
- Depending upon the parasitic impedance of the power/ground lines, the maximum voltage drop is caused either by the dominant inductive $L\,dI/dt$ noise or by the dominant resistive IR noise
- Design expressions to estimate the minimum on-chip decoupling capacitance required to support expected current demands based on the dominant voltage drop are provided
- An expression for the critical length of the interconnect between the decoupling capacitor and the current load is described
- To be effective, an on-chip decoupling capacitor should be placed such that both the power supply and the current load are located inside the appropriate effective radius
- On-chip decoupling capacitors should be allocated within appropriate effective radii across an IC to satisfy local transient current demands

Chapter 13
Efficient Placement of Distributed On-Chip Decoupling Capacitors

Decoupling capacitors are widely used to manage power supply noise [281] and are an effective way to reduce the impedance of power delivery systems operating at high frequencies [28, 29]. A decoupling capacitor acts as a local reservoir of charge, which is released when the power supply voltage at a particular current load drops below some tolerable level. Since the inductance scales slowly [129], the location of the decoupling capacitors significantly affects the design of the power/ground networks in high performance integrated circuits such as microprocessors. At higher frequencies, a distributed system of decoupling capacitors are placed on-chip to effectively manage the power supply noise [279].

The efficacy of decoupling capacitors depends upon the impedance of the conductors connecting the capacitors to the current loads and power sources. As described in [277], a maximum parasitic impedance between the decoupling capacitor and the current load (*or* power source) exists at which the decoupling capacitor is effective. Alternatively, to be effective, an on-chip decoupling capacitor should be placed such that both the power supply and the current load are located inside the appropriate effective radius [277]. The efficient placement of on-chip decoupling capacitors in nanoscale ICs is the subject of this chapter. Unlike the methodology for placing a single lumped on-chip decoupling capacitor presented in Chap. 12, a system of *distributed* on-chip decoupling capacitors is described in this chapter. A design methodology to estimate the parameters of the distributed system of on-chip decoupling capacitors is also presented, permitting the required on-chip decoupling capacitance to be allocated under existing technology constraints.

This chapter is organized as follows. Technology limitations in nanoscale integrated circuits are reviewed in Sect. 13.1. The problem of placing on-chip decoupling capacitors in nanoscale ICs while satisfying technology constraints is formulated in Sect. 13.2. The design of a distributed on-chip decoupling capacitor network is presented in Sect. 13.3. Various design tradeoffs are discussed in Sect. 13.4. A design methodology for placing distributed on-chip decoupling

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0_13

capacitors is presented in Sect. 13.5. Related simulation results for typical values of on-chip parasitic resistances are discussed in Sect. 13.6. Some specific conclusions are summarized in Sect. 13.7.

13.1 Technology Constraints

On-chip decoupling capacitors have traditionally been designed as standard gate oxide CMOS capacitors [288]. As technology scales, leakage current through the gate oxide of an on-chip decoupling capacitor has greatly increased [289–291]. Moreover, in modern high performance ICs, a large portion (up to 40 %) of the circuit area is occupied by the on-chip decoupling capacitance [292, 293]. Conventional gate oxide on-chip decoupling capacitors are therefore prohibitively expensive from an area and yield perspective, as well as greatly increasing the overall power dissipated on-chip [294].

To reduce the power consumed by an IC, MIM capacitors are frequently utilized as decoupling capacitors. The capacitance density of a MIM capacitor in a 90 nm CMOS technology is comparable to the maximum capacitance density of a CMOS capacitor and is typically 10–30 $\text{fF}/\mu\text{m}^2$ [256, 259, 270]. A maximum magnitude of an on-chip decoupling capacitor therefore exists for a specific distance between a current load and a decoupling capacitor (as constrained by the available on-chip metal resources). Alternatively, a minimum achievable impedance per unit length exists for a specified capacitance density of an on-chip decoupling capacitor placed at a specific distance from a circuit module, as illustrated in Fig. 13.1.

Observe from Fig. 13.1 that the available metal area for the second level of a distributed on-chip capacitance is greater than the fraction of metal resources dedicated to the first level of a distributed on-chip capacitance. Capacitor C_2 can therefore be larger than C_1. Note also that a larger capacitor can only be placed farther from the current load. Similarly, the metal resources required by the first level of interconnection (connecting C_1 to the current load) is smaller than the metal resources dedicated to the second level of interconnections. The impedance Z_2 is therefore smaller than Z_1.

13.2 Placing On-Chip Decoupling Capacitors in Nanoscale ICs

Decoupling capacitors have traditionally been allocated into the white space (those areas not occupied by the circuit elements) available on the die based on an unsystematic or ad hoc approach [283, 284], as shown in Fig. 13.2. In this way, decoupling capacitors are often placed at a significant distance from the current load. Conventional approaches for placing on-chip decoupling capacitors result

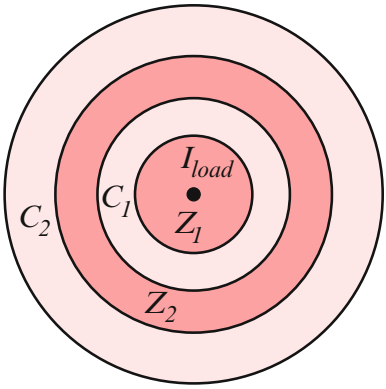


Fig. 13.1 Fundamental limits of on-chip interconnections. Two levels of a distributed on-chip decoupling capacitance are allocated around a current load. The interconnect impedance is inversely proportional to the fraction of metal area dedicated to the interconnect level, decreasing as the decoupling capacitor is farther from the current source ($Z_1 > Z_2$). The decoupling capacitance increases as the capacitor is farther from the current load due to the increased area ($C_1 < C_2$). The two levels of interconnection and distributed decoupling capacitance are shown, respectively, in *dark pink* and *light pink*

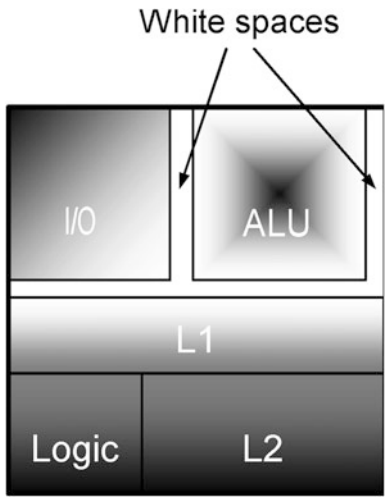


Fig. 13.2 Placement of on-chip decoupling capacitors using a conventional approach. Decoupling capacitors are allocated into the *white space* (those areas not occupied by the circuits elements) available on the die using an unsystematic or ad hoc approach. As a result, the power supply voltage drops below the minimum tolerable level for remote blocks (shown in *dark gray*). Low noise regions are *light gray*

in oversized capacitors. The conventional allocation strategy, therefore, results in increased power noise, compromising the signal integrity of an entire system, as illustrated in Fig. 13.3. This issue of power delivery cannot be alleviated by simply increasing the size of the on-chip decoupling capacitors. Furthermore, increasing the size of more distant on-chip decoupling capacitors results in wasted area, increased power, reduced reliability, and higher cost. A design methodology is therefore required to account for technology trends in nanoscale ICs, such as increasing frequencies, larger die sizes, higher current demands, and reduced noise margins.

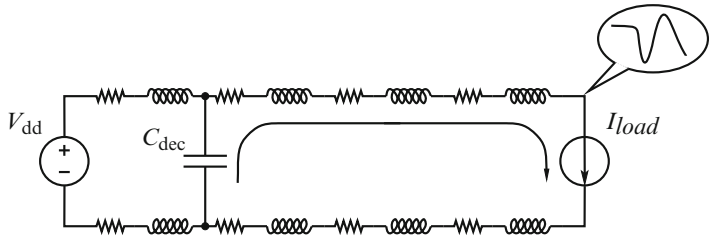


Fig. 13.3 A conventional on-chip decoupling capacitor. Typically, a large decoupling capacitor is placed farther from the current load due to physical limitations. Current flowing through the long power/ground lines results in large voltage fluctuations across the terminals of the current load

To be effective, a decoupling capacitor should be placed physically close to the current load. This requirement is naturally satisfied in board and package applications, since large capacitors are much smaller than the dimensions of the circuit board (or package) [223]. In this case, a lumped model of a decoupling capacitor provides sufficient accuracy [295].

The size of an on-chip decoupling capacitor, however, is directly proportional to the area occupied by the capacitor and can require a significant portion of the on-chip area. The minimum impedance between an on-chip capacitor and the current load is fundamentally affected by the magnitude (and therefore the area) of the capacitor. Systematically partitioning the decoupling capacitor into smaller capacitors solves this issue. A system of distributed on-chip decoupling capacitors is illustrated in Fig. 13.4.

In a system of distributed on-chip decoupling capacitors, each decoupling capacitor is sized based on the impedance of the interconnect segment connecting the capacitor to the current load. A particular capacitor only provides charge to a current load during a short period. The rationale behind this scheme can be explained as follows. The capacitor closest to the current load is engaged immediately after the switching cycle is initiated. Once the first capacitor is depleted of charge, the next capacitor is activated, providing a large portion of the total current drawn by the load. This procedure is repeated until the last capacitor becomes active. Similar to the hierarchical placement of decoupling capacitors presented in [28, 136], this technique provides an efficient solution for providing the required on-chip decoupling capacitance based on specified capacitance density constraints. A system of distributed on-chip decoupling capacitors should therefore be utilized to provide a low impedance, cost effective power delivery network in nanoscale ICs.

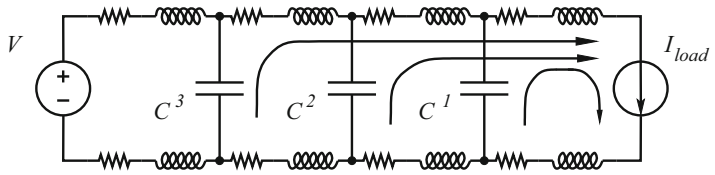


Fig. 13.4 A network of distributed on-chip decoupling capacitors. The magnitude of the decoupling capacitors is based on the impedance of the interconnect segment connecting a specific capacitor to a current load. Each decoupling capacitor is designed to only provide charge during a specific time interval

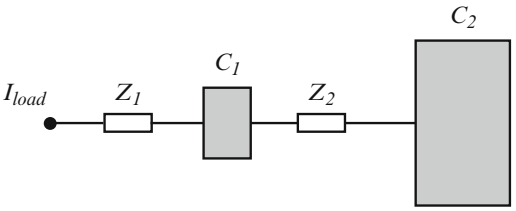


Fig. 13.5 A physical model of a system of distributed on-chip decoupling capacitors. Two capacitors are assumed to provide the required charge drawn by the load. Z_1 and Z_2 denote the impedance of the metal lines connecting, respectively, C_1 to the current load and C_2 to C_1

13.3 Design of a Distributed On-Chip Decoupling Capacitor Network

As described in Sect. 13.2, a system of distributed on-chip decoupling capacitors is an efficient solution for providing the required on-chip decoupling capacitance based on the maximum capacitance density available in a particular technology. A physical model of the technique is illustrated in Fig. 13.5. For simplicity, two decoupling capacitors are assumed to provide the required charge drawn by the current load. Note that as the capacitor is placed farther from the current load, the magnitude of an on-chip decoupling capacitor increases due to relaxed constraints. In the general case, the described methodology can be extended to any practical number of on-chip decoupling capacitors. Note that Z_1 is typically limited by a specific technology (determined by the impedance of a single metal wire) and the magnitude of C_1 (the area available in the vicinity of a circuit block).

A circuit model of a system of distributed on-chip decoupling capacitors is shown in Fig. 13.6. The impedance of the metal lines connecting the capacitors to the current load is modeled as resistors R_1 and R_2. A triangular current source is assumed to model the current load. The magnitude of the current source increases linearly, reaching the maximum current $I_{\max}$ at rise time t_r, i.e., $I_{load}(t) = I_{\max}\frac{t}{t_r}$. The maximum tolerable ripple at the load is 10 % of the power supply voltage.

Note from Fig. 13.6 that since the charge drawn by the current load is provided by the on-chip decoupling capacitors, the voltage across the capacitors during discharge

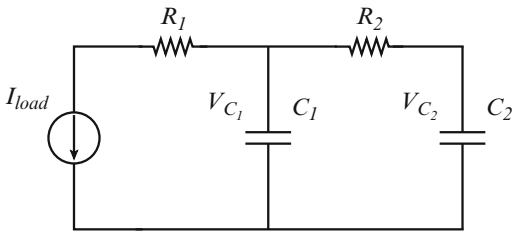


Fig. 13.6 A circuit model of an on-chip distributed decoupling capacitor network. The impedance of the metal lines is modeled, respectively, as R_1 and R_2

drops below the initial power supply voltage. The required charge during the entire switching event is thus determined by the voltage drop across C_1 and C_2.

The voltage across the decoupling capacitors at the end of the switching cycle ($t = t_r$) can be determined from Kirchhoff's laws [287]. Writing KVL and KCL equations for each of the loops (see Fig. 13.6), the system of differential equations describing the voltage across C_1 and C_2 at t_r is

$$\frac{dV_{C_1}}{dt} = \frac{V_{C_2} - V_{C_1}}{R_2 C_1} - \frac{I_{load}}{C_1}, \tag{13.1}$$

$$\frac{dV_{C_2}}{dt} = \frac{V_{C_1} - V_{C_2}}{R_2 C_2}. \tag{13.2}$$

Simultaneously solving (13.1) and (13.2) and applying the initial conditions, the voltage across C_1 and C_2 at the end of the switching activity is

$$\begin{aligned} V_{C_1}|_{t=t_r} = \frac{1}{2(C_1+C_2)^3 t_r}\Bigg[& 2C_1^3 t_r + C_1^2 t_r (6C_2 - I_{\max} t_r) \\ & - C_2^2 t_r \left(2C_2 \left(I_{\max} R_2 - 1\right) + I_{\max} t_r\right) \\ & + 2C_1 C_2 \left(C_2^2 \left(1 - e^{-\frac{(C_1+C_2)t_r}{C_1 C_2 R_2}}\right) I_{\max} R_2^2 \right. \\ & \left. + C_2 \left(3 - I_{\max} R_2\right) t_r - I_{\max} t_r^2\right)\Bigg], \end{aligned} \tag{13.3}$$

$$\begin{aligned} V_{C_2}|_{t=t_r} = \frac{1}{2(C_1+C_2)^3 t_r}\Bigg[& 2C_1^3 t_r + C_2^2 t_r \left(2C_2 - I_{\max} t_r\right) \\ & + 2C_1 C_2 t_r \left(C_2 \left(3 + I_{\max} R_2\right) - I_{\max} t_r\right) \\ & + C_1^2 \left(2C_2^2 \left(e^{-\frac{(C_1+C_2)t_r}{C_1 C_2 R_2}} - 1\right) I_{\max} R_2^2 \right. \\ & \left. + 2C_2 \left(3 + I_{\max} R_2\right) t_r - I_{\max} t_r^2\right)\Bigg], \end{aligned} \tag{13.4}$$

where $I_{\max}$ is the maximum magnitude of the current load and t_r is the rise time.

Note that the voltage across C_1 and C_2 after discharge is determined by the magnitude of the decoupling capacitors and the parasitic resistance of the metal line(s) between the capacitors. The voltage across C_1 after the switching cycle, however, depends upon the resistance of the P/G paths connecting C_1 to a current load and is

$$V_{C_1} = V_{load} + I_{\max} R_1, \tag{13.5}$$

where V_{load} is the voltage across the terminals of a current load. Assuming $V_{load} \geq 0.9V_{\rm dd}$ and $V_{C_1}^{\max} = V_{\rm dd}$ (meaning that C_1 is infinitely large), the upper bound for R_1 is

$$R_1^{\max} = \frac{V_{\rm dd}(1-\alpha)}{I_{\max}}, \tag{13.6}$$

where α is the ratio of the minimum tolerable voltage across the terminals of a current load to the power supply voltage ($\alpha = 0.9$ in this chapter). If $R_1 > R_1^{\max}$, no solution exists for providing sufficient charge drawn by the load. In this case, the circuit block should be partitioned, reducing the current demands ($I_{\max}$).

Note that expressions for determining the voltage across the decoupling capacitors are transcendental functions. No closed-form solution, therefore, exists. From (13.3) and (13.4), the design space can be graphically obtained for determining the maximum tolerable resistance R_2 and the minimum magnitude of the capacitors, maintaining the voltage across the load equal to or greater than the minimum allowable level. The voltage across C_1 after discharge as a function of C_1 and R_2 is depicted in Fig. 13.7.

Observe from Fig. 13.7 that the voltage across capacitor C_1 increases exponentially with capacitance, saturating for large C_1. The voltage across C_1, however, is almost independent of R_2, decreasing slightly with R_2 (see Fig. 13.7a). This behavior can be explained as follows. As a current load draws charge from the decoupling capacitors, the voltage across the capacitors drops below the initial level. The charge released by a capacitor is proportional to the capacitance and the change in voltage. A larger capacitance therefore results in a smaller voltage drop. From Fig. 13.6, note that as resistance R_2 increases, capacitor C_2 becomes less effective (a larger portion of the total current is provided by C_1). As a result, the magnitude of C_1 is increased to maintain the voltage across the load above the minimum tolerable level. Similarly, a larger C_2 results in a smaller C_1. As C_2 is increased, a larger portion of the total current is provided by C_2, reducing the magnitude of C_1. This phenomenon is well pronounced for small R_2, diminishing with larger R_2, as illustrated in Fig. 13.7b.

In general, to determine the parameters of the system of distributed on-chip decoupling capacitors, the following assumptions are made. The parasitic resistance of the metal line(s) connecting capacitor C_1 to the current load is known. R_1 is determined by technology constraints (the sheet resistance) and by design constraints (the maximum available metal resources). The minimum voltage level at the load is $V_{load} = 0.9V_{\rm dd}$. The maximum magnitude of the current load $I_{\max}$ is

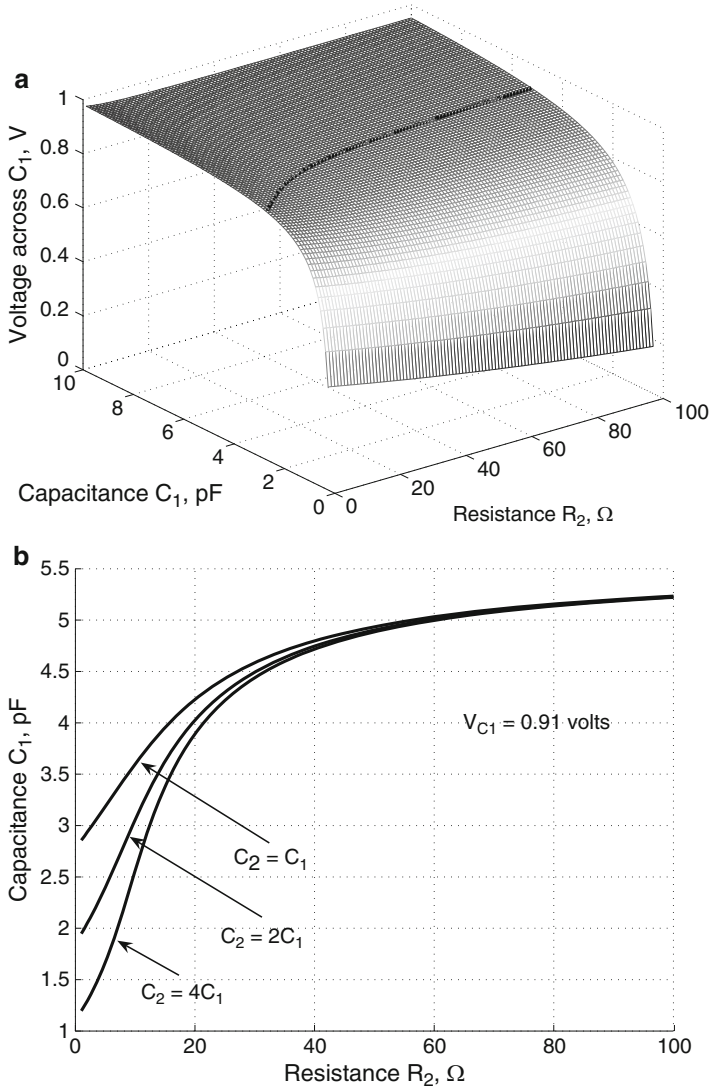


Fig. 13.7 Voltage across C_1 during discharge as a function of C_1 and R_2: $I_{\max} = 0.01$ mA, $V_{\mathrm{dd}} = 1$ V, and $t_r = 100$ ps; (**a**) assuming $C_1 = C_2$ and $R_1 = 10\,\Omega$, the minimum tolerable voltage across C_1, resulting in $V_{load} \geq 0.9V_{\mathrm{dd}}$, is 0.91 V (shown as a *black equipotential line*), (**b**) design space for determining C_1 and R_2 resulting in the voltage across C_1 equal to 0.91 V

0.01 A, the rise time t_r is 100 ps, and the power supply voltage V_{dd} is 1 V. Note that the voltage across C_2 after discharge as determined by (13.4) is also treated as a design parameter. Since the capacitor C_2 is directly connected to the power supply (a shared power rail), the voltage drop across C_2 appears on the global power line,

compromising the signal integrity of the overall system. The voltage across C_2 at t_r is therefore based on the maximum tolerable voltage fluctuations on the P/G line during discharge (the voltage across C_2 at the end of the switching cycle is set to 0.95 V).

The system of equations to determine the parameters of an on-chip distributed decoupling capacitor network as depicted in Fig. 13.6 is

$$V_{load} = V_{C_1} - I_{\max} R_1, \tag{13.7}$$

$$V_{C_1} = f(C_1, C_2, R_2), \tag{13.8}$$

$$V_{C_2} = f(C_1, C_2, R_2), \tag{13.9}$$

$$\frac{I_{\max} t_r}{2} = C_1 \left(V_{\mathrm{dd}} - V_{C_1}\right) + C_2 \left(V_{\mathrm{dd}} - V_{C_2}\right), \tag{13.10}$$

where V_{C_1} and V_{C_2} are the voltage across C_1 and C_2 and determined, respectively, by (13.3) and (13.4). Equation (13.10) states that the total charge drawn by the current load is provided by C_1 and C_2. Note that in the general case with the current load determined a priori, the total charge is the integral of $I_{load}(t)$ from zero to t_r. Solving (13.7) for V_{C_1} and substituting into (13.8), C_1, C_2, and R_2 are determined from (13.8), (13.9), and (13.10) for a specified $V_{C_2}(t_r)$, as discussed in the following section.

13.4 Design Tradeoffs in a Distributed On-Chip Decoupling Capacitor Network

To design a system of distributed on-chip decoupling capacitors, the parasitic resistances and capacitances should be determined based on design and technology constraints. As shown in Sect. 13.3, in a system composed of two decoupling capacitors (see Fig. 13.6) with known R_1; R_2, C_1, and C_2 are determined from the system of Eqs. (13.7), (13.8), (13.9) and (13.10). Note that since this system of equations involves transcendental functions, a closed-form solution cannot be determined. To determine the system parameters, the system of Eqs. (13.7), (13.8), (13.9) and (13.10) is solved numerically [296].

Various tradeoff scenarios are discussed in this section. The dependence of the system parameters on R_1 is presented in Sect. 13.4.1. The design of a distributed on-chip decoupling capacitor network with the minimum magnitude of C_1 is discussed in Sect. 13.4.2. The dependence of C_1 and C_2 on the parasitic resistance of the metal lines connecting the capacitors to the current load is presented in Sect. 13.4.3. The minimum total budgeted on-chip decoupling capacitance is also determined in this section.

Table 13.1 Dependence of the parameters of a distributed on-chip decoupling capacitor network on R_1

	$R_2 = 5\,(\Omega)$		$R_2 = 10\,(\Omega)$	
R_1 (Ω)	C_1 (pF)	C_2 (pF)	C_1 (pF)	C_2 (pF)
1	1.35	7.57	3.64	3.44
2	2.81	5.50	4.63	2.60
3	4.54	3.64	5.88	1.77
4	6.78	1.87	7.56	0.92
5	10.00	0	10.00	0

$V_{\text{dd}} = 1\,\text{V}$, $V_{load} = 0.9\,\text{V}$, $t_r = 100\,\text{ps}$, and $I_{\max} = 0.01\,\text{A}$

13.4.1 Dependence of System Parameters on R_1

The parameters of a distributed on-chip decoupling capacitor network for typical values of R_1 are listed in Table 13.1. Note that the minimum magnitude of R_2 exists for which the parameters of the system can be determined. If R_2 is sufficiently small, the distributed decoupling capacitor network degenerates to a system with a single capacitor (where C_1 and C_2 are combined). For the parameters listed in Table 13.1, the minimum magnitude of R_2 is $4\,\Omega$, as determined from numerical simulations.

Note that the parameters of a distributed on-chip decoupling capacitor network are determined by the parasitic resistance of the P/G line(s) connecting C_1 to the current load. As R_1 increases, the capacitor C_1 increases substantially (see Table 13.1). This increase in C_1 is due to R_1 becoming comparable to R_2, and C_1 providing a greater portion of the total current. Alternatively, the system of distributed on-chip decoupling capacitors degenerates to a single oversized capacitor. The system of distributed on-chip decoupling capacitors should therefore be carefully designed. Since the distributed on-chip decoupling capacitor network is strongly dependent upon the first level of interconnection (R_1), C_1 should be placed as physically close as possible to the current load, reducing R_1. If such an allocation is not practically possible, the current load should be partitioned, permitting an efficient allocation of the distributed on-chip decoupling capacitors under specific technology constraints.

13.4.2 Minimum C_1

In practical applications, the size of C_1 (the capacitor closest to the current load) is typically limited by technology constraints, such as the maximum capacitance density and available area. The magnitude of the first capacitor in the distributed system is therefore typically small. In this section, the dependence of the distributed on-chip decoupling capacitor network on R_1 is determined for minimum C_1. A target magnitude of 1 pF is assumed for C_1. The parameters of a system of distributed

Table 13.2 Distributed on-chip decoupling capacitor network as a function of R_1 under the constraint of a minimum C_1

R_1	$V_{C_2} \neq$ const				$V_{C_2} = 0.95$ V	
(Ω)	R_2 (Ω)	C_2 (pF)	R_2 (Ω)	C_2 (pF)	R_2 (Ω)	C_2 (pF)
1	2	5.59	5	8.69	4.68	8.20
2	2	6.68	5	11.64	3.46	8.40
3	2	8.19	5	17.22	2.28	8.60
4	2	10.46	5	31.70	1.13	8.80
5	2	14.21	5	162.10	–	–

$V_{\text{dd}} = 1$ V, $V_{load} = 0.9$ V, $t_r = 100$ ps, $I_{\max} = 0.01$ A, and $C_1 = 1$ pF

on-chip decoupling capacitors as a function of R_1 under the constraint of a minimum C_1 are listed in Table 13.2. Note that V_{C_2} denotes the voltage across C_2 after discharge.

Note that two scenarios are considered in Table 13.2 to evaluate the dependence of a distributed system of on-chip decoupling capacitors on R_1 and R_2. In the first scenario, the distributed on-chip decoupling capacitor network is designed to maintain the minimum tolerable voltage across the terminals of a current load. In this case, the magnitude of C_2 increases with R_1, becoming impractically large for large R_2. In the second scenario, an additional constraint (the voltage across C_2) is applied to reduce the voltage fluctuations on the shared P/G lines. In this case, as R_1 increases, C_2 slightly increases. In order to satisfy the constraint for V_{C_2}, R_2 should be significantly reduced for large values of R_1, meaning that the second capacitor should be placed close to the first capacitor. As R_1 is further increased, R_2 becomes negligible, implying that capacitors C_1 and C_2 should be merged to provide the required charge to the distant current load. Alternatively, the system of distributed on-chip decoupling capacitors degenerates to a conventional scheme with a single oversized capacitor [297].

Note that simultaneously satisfying both the voltage across the terminals of the current load and the voltage across the last decoupling capacitor is not easy. The system of on-chip distributed decoupling capacitors in this case depends upon the parameters of the first decoupling stage (R_1 and C_1). If C_1 is too small, no solution exists to satisfy $V_{load}^{\min}$ and $V_{C_2}^{\min}$. Sufficient circuit area should therefore be allocated for C_1 early in the design process to provide the required on-chip decoupling capacitance in order to satisfy specific design and technology constraints.

13.4.3 Minimum Total Budgeted On-Chip Decoupling Capacitance

As discussed in Sect. 13.4.1 and 13.4.2, the design of a system of distributed on-chip decoupling capacitors is greatly determined by the parasitic resistance of the metal lines connecting C_1 to the current load and by the magnitude of C_1. Another

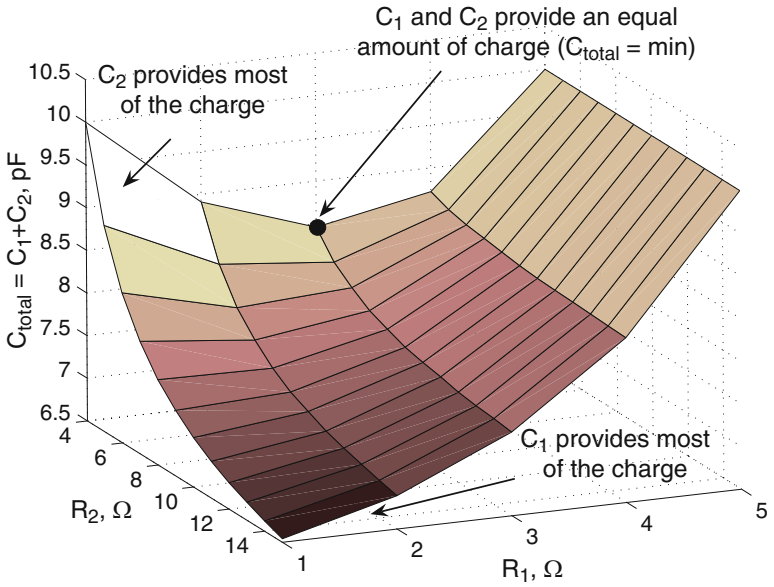


Fig. 13.8 The total budgeted on-chip decoupling capacitance as a function of the parasitic resistance of the metal lines, R_1 and R_2: $I_{\max} = 10\,\text{mA}$, $V_{\text{dd}} = 1\,\text{V}$, $V_{load} = 0.9\,\text{V}$, and $t_r = 100\,\text{ps}$. In the system of distributed on-chip decoupling capacitors, an optimal ratio $\frac{R_2}{R_1}$ exists, resulting in the minimum total budgeted on-chip decoupling capacitance

important design constraint is the total budgeted on-chip decoupling capacitance. Excessive on-chip decoupling capacitance results in increased circuit area and greater leakage currents. Large on-chip decoupling capacitors can also compromise the reliability of the overall system, creating a short circuit between the plates of a capacitor [294]. It is therefore important to reduce the required on-chip decoupling capacitance while providing sufficient charge to support expected current demands.

To estimate the total required on-chip decoupling capacitance, $C_{total} = C_1 + C_2$ is plotted as a function of R_1 and R_2, as depicted in Fig. 13.8. Note that if R_2 is large, C_2 is ineffective and the system of distributed on-chip decoupling capacitors behaves as a single capacitor. Observe from Fig. 13.8 that C_{total} increases with R_1 for large R_2. In this case, C_1 is oversized, providing most of the required charge. C_1 should therefore be placed close to the current load to reduce the total required on-chip decoupling capacitance.

Similarly, if R_2 is reduced with small R_1, C_2 provides most of the charge drawn by the current load. The distributed on-chip decoupling capacitor network degenerates to a conventional system with a single capacitor. As R_1 increases, however, the total required on-chip decoupling capacitance decreases, reaching the minimum (see Fig. 13.8 for $R_1 = 3\,\Omega$ and $R_2 = 4\,\Omega$). In this case, C_1 and C_2 each provide an equal amount of the total charge. As R_1 is further increased (C_1 is placed farther from the current load), C_1 and C_2 increase substantially to compensate for the

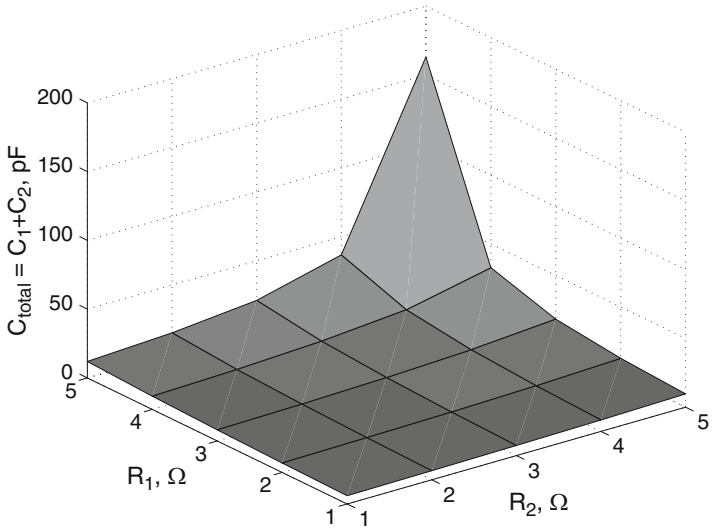


Fig. 13.9 The total budgeted on-chip decoupling capacitance as a function of the parasitic resistance of the metal lines, R_1 and R_2: $I_{\max} = 10\,\text{mA}$, $V_{\text{dd}} = 1\,\text{V}$, $V_{load} = 0.9\,\text{V}$, and $t_r = 100\,\text{ps}$. C_1 is fixed and set to 1 pF. The total budgeted on-chip decoupling capacitance increases with R_1 and R_2. As the parasitic resistance of the metal lines is further increased beyond $4\,\Omega$, C_{total} increases substantially, becoming impractically large

increased voltage drop across R_1. In the system of distributed on-chip decoupling capacitors, an optimal ratio $\frac{R_2}{R_1}$ exists which requires the minimum total budgeted on-chip decoupling capacitance.

Note that in the previous scenario, the magnitude of the on-chip decoupling capacitors has not been constrained. In practical applications, however, the magnitude of the first decoupling capacitor (placed close to the current load) is limited. To determine the dependence of the total required on-chip decoupling capacitance under the magnitude constraint of C_1, C_1 is fixed and set to 1 pF. $C_{total} = C_1 + C_2$ is plotted as a function of R_1 and R_2, as shown in Fig. 13.9. In contrast to the results depicted in Fig. 13.8, the total budgeted on-chip decoupling capacitance required to support expected current demands increases with R_1 and R_2. Alternatively, C_2 provides the major portion of the total charge. Thus, the system behaves as a single distant on-chip decoupling capacitor. In this case, C_1 is too small. A larger area should therefore be allocated for C_1, resulting in a balanced system with a reduced total on-chip decoupling capacitance. Also note that as R_1 and R_2 further increase (beyond $4\,\Omega$, see Fig. 13.9), the total budgeted on-chip decoupling capacitance increases rapidly, becoming impractically large.

Comparing Figs. 13.8 and 13.9, note that if C_1 is constrained, a larger total decoupling capacitance is required to provide the charge drawn by the current load. Alternatively, the system of distributed on-chip decoupling capacitors under a magnitude constraint of C_1 behaves as a single distant decoupling capacitor. As

a result, the magnitude of a single decoupling capacitor is significantly increased to compensate for the *IR* voltage drop across R_1 and R_2. The system of distributed on-chip decoupling capacitors should therefore be carefully designed to reduce the total budgeted on-chip decoupling capacitance. If the magnitude of C_1 is limited, C_2 should be placed close to the current load to be effective, reducing the total required on-chip decoupling capacitance. Alternatively, the parasitic impedance of the P/G lines connecting C_1 and C_2 should be reduced (e.g., utilizing wider lines and/or multiple lines in parallel) [73].

13.5 Design Methodology for a System of Distributed On-Chip Decoupling Capacitors

An overall methodology for designing a distributed system of on-chip decoupling capacitors is illustrated in Fig. 13.10. General differential equations for voltages $V_{C_1}(t)$ and $V_{C_2}(t)$ across capacitors C_1 and C_2 are derived based on Kirchhoff's laws. The maximum parasitic resistance $R_1^{\max}$ between C_1 and the current load is determined from (13.6) for specific parameters of the system, such as the power supply voltage V_{dd}, the minimum voltage across the terminals of the current load V_{load}, the maximum magnitude of the current load $I_{\max}$, and the rise time t_r. If $R_1 > R_1^{\max}$, no solution exists for the system of distributed on-chip decoupling capacitors. Alternatively, the voltage across the terminals of a current load always drops below the minimum acceptable level. In this case, the current load should be partitioned to reduce $I_{\max}$, resulting in $R_1 < R_1^{\max}$.

Simultaneously solving (13.1) and (13.2), the voltage across C_1 and C_2 is estimated at the end of a switching cycle ($t = t_r$), as determined by (13.3) and (13.4). The parameters of the distributed on-chip decoupling capacitor network C_1, C_2, and R_2, are determined from (13.7), (13.8), (13.9) and (13.10). Note that different tradeoffs exist in a system of distributed on-chip decoupling capacitors, as discussed in Sect. 13.4. If the voltage across the terminals of a current load drops below the minimum tolerable level, the total budgeted on-chip decoupling capacitance should be increased. The system of Eqs. (13.7), (13.8), (13.9) and (13.10), is solved for an increased total on-chip decoupling capacitance, resulting in different C_1, C_2, and R_2 until the criterion for the maximum tolerable power noise $V_{load} > V_{load}^{\min}$ is satisfied, as shown in Fig. 13.10.

Note that the system of distributed on-chip decoupling capacitors permits the design of an effective power distribution system under specified technology constraints. The techniques presented in this chapter are also applicable to future technology generations. The methodology also provides a computationally efficient way to determine the required on-chip decoupling capacitance to support expected current demands. In the worst case example presented in this chapter, the simulation time to determine the parameters of the system of on-chip distributed decoupling capacitors is under one second on a Pentium III PC with one gigabyte of RAM. A methodology for efficiently placing on-chip decoupling capacitors can also be

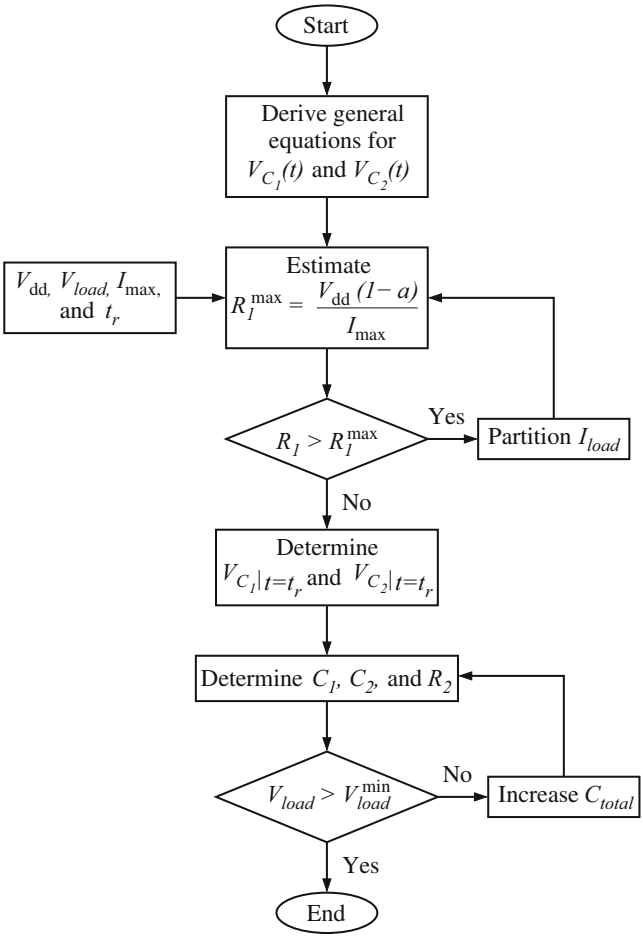


Fig. 13.10 Design flow for determining the parameters of a system of distributed on-chip decoupling capacitors

integrated into a standard IC design flow. In this way, the circuit area required to allocate on-chip decoupling capacitors is estimated early in the design process, significantly reducing the number of iterations and the eventual time to market.

13.6 Case Study

The dependence of the system of distributed on-chip decoupling capacitors on the current load and the parasitic impedance of the power delivery system is described in this section to quantitatively illustrate the previously presented concepts. Resistive

power and ground lines are assumed to connect the decoupling capacitors to the current load and are modeled as resistors (see Fig. 13.6). The load is modeled as a ramp current source with a 100 ps rise time. The minimum tolerable voltage across the load terminals is 90 % of the power supply. The magnitude of the on-chip decoupling capacitors for various parasitic resistances of the metal lines connecting the capacitors to the current load is listed in Table 13.3. The parameters of the distributed on-chip decoupling capacitor network listed in Table 13.3 are determined for two amplitudes of the current load. Note that the values of R_1 and R_2 are typical parasitic resistances of an on-chip power distribution grid for a 90 nm CMOS technology.

The parameters of the system of distributed on-chip decoupling capacitors are analytically determined from (13.7), (13.8), (13.9) and (13.10). The resulting power supply noise is estimated using SPICE and compared to the maximum tolerable level (the minimum voltage across the load terminals $V_{load}^{\min}$). The maximum voltage drop across C_2 at the end of the switching activity is also estimated and compared to $V_{C_2}^{\min}$. Note that the analytic solution produces an accurate estimate of the on-chip decoupling capacitors for typical parasitic resistances of a power distribution grid. The maximum error in this case study is 0.003 %.

From Table 13.3, note that in the case of a large R_2, the distributed decoupling capacitor network degenerates into a system with a single capacitor. Capacitor C_1 is therefore excessively large. Conversely, if C_2 is placed close to C_1 (R_2 is small), C_2 is excessively large and the system again behaves as a single capacitor. An optimal ratio $\frac{R_2}{R_1}$ therefore exists for specific characteristics of the current load that results in a minimum required on-chip decoupling capacitance. Alternatively, in this case, both capacitors provide an equal portion of the total charge (see Table 13.3 for $R_1 = 0.5\,\Omega$ and $R_2 = 10\,\Omega$). Also note that as the magnitude of the current load increases, larger on-chip decoupling capacitors are required to provide the expected current demands.

The parameters of a distributed on-chip decoupling capacitor network listed in Table 13.3 have been determined for the case where the magnitude of the decoupling capacitors is not limited. In most practical systems, however, the magnitude of the on-chip decoupling capacitor placed closest to the current load is limited by technology and design constraints. A case study of a system of distributed on-chip decoupling capacitors with a limited value of C_1 is listed in Table 13.4. Note that in contrast to Table 13.3, where both R_1 and R_2 are design parameters, in the system with a limit on C_1, R_2 and C_2 are determined by R_1. Alternatively, both the magnitude and location of the second capacitor are determined from the magnitude and location of the first capacitor.

The parameters of the distributed on-chip decoupling capacitor network listed in Table 13.4 are determined for two amplitudes of the current load with R_1 representing a typical parasitic resistance of the metal line connecting C_1 to the current load. The resulting power supply noise at the current load and across the last decoupling stage is estimated using SPICE and compared to the maximum tolerable levels, respectively, $V_{load}^{\min}$ and $V_{C_2}^{\min}$. Note that the analytic solution accurately estimates the parameters of the distributed on-chip decoupling capacitor network, producing a worst case error of 0.0001 %.

Table 13.3 The magnitude of the on-chip decoupling capacitors as a function of the parasitic resistance of the power/ground lines connecting the capacitors to the current load

R_1	R_2	$I_{\max}$	C_1	C_2	V_{load} (mV)		Error	V_{C_2} (mV)		Error
(Ω)	(Ω)	(A)	(pF)	(pF)	$V_{load}^{\min}$	SPICE	(%)	$V_{C_2}^{\min}$	SPICE	(%)
0.5	4.5	0.01	0	9.99999	900	899.999	0.0001	950	949.999	0.0001
0.5	6	0.01	1.59747	6.96215	900	899.986	0.002	950	949.983	0.002
0.5	8	0.01	2.64645	4.97091	900	899.995	0.0006	950	949.993	0.0004
0.5	10	0.01	3.22455	3.87297	900	899.997	0.0003	950	949.996	0.0004
0.5	12	0.01	3.59188	3.17521	900	899.998	0.0002	950	949.997	0.0003
0.5	14	0.01	3.84641	2.69168	900	899.998	0.0002	950	949.997	0.0003
0.5	16	0.01	4.03337	2.33650	900	899.999	0.0001	950	949.998	0.0002
0.5	18	0.01	4.17658	2.06440	900	899.998	0.0002	950	949.998	0.0002
0.5	20	0.01	4.28984	1.84922	900	899.999	0.0001	950	949.998	0.0002
0.5	1.5	0.025	0	24.99930	900	899.998	0.0002	950	949.998	0.0002
0.5	2	0.025	4.25092	17.56070	900	899.999	0.0001	950	949.999	0.0001
0.5	3	0.025	7.97609	11.04180	900	899.999	0.0001	950	949.999	0.0001
0.5	4	0.025	9.67473	8.06921	900	899.999	0.0001	950	949.999	0.0001
0.5	5	0.025	10.65000	6.36246	900	899.999	0.0001	950	949.999	0.0001
0.5	6	0.025	11.2838	5.25330	900	899.999	0.0001	950	949.999	0.0001
0.5	7	0.025	11.72910	4.47412	900	899.999	0.0001	950	949.999	0.0001
0.5	8	0.025	12.05910	3.89653	900	899.999	0.0001	950	949.999	0.0001
0.5	9	0.025	12.31110	3.44905	900	899.980	0.002	950	949.973	0.003
1	4	0.01	0	9.99999	900	899.999	0.0001	950	949.999	0.0001
1	6	0.01	2.16958	6.09294	900	899.990	0.001	950	949.988	0.001
1	8	0.01	3.11418	4.39381	900	899.996	0.0004	950	949.994	0.0006
1	10	0.01	3.64403	3.44040	900	899.997	0.0003	950	949.996	0.0004
1	12	0.01	3.98393	2.82871	900	899.998	0.0002	950	949.997	0.0003
1	14	0.01	4.22079	2.40240	900	899.998	0.0002	950	949.997	0.0003
1	16	0.01	4.39543	2.08809	900	899.998	0.0002	950	949.997	0.0003
1	18	0.01	4.52955	1.84668	900	899.998	0.0002	950	949.998	0.0002
1	20	0.01	4.63582	1.65540	900	899.998	0.0002	950	949.998	0.0002
1	1	0.025	0	24.99940	900	899.998	0.0002	950	949.998	0.0002
1	2	0.025	9.08053	11.37910	900	899.999	0.0001	950	949.999	0.0001
1	3	0.025	11.74820	7.37767	900	899.999	0.0001	950	949.999	0.0001
1	4	0.025	13.02600	5.46100	900	899.999	0.0001	950	949.999	0.0001
1	5	0.025	13.77630	4.33559	900	899.999	0.0001	950	949.999	0.0001
1	6	0.025	14.27000	3.59504	900	899.999	0.0001	950	949.999	0.0001
1	7	0.025	14.61950	3.07068	900	899.999	0.0001	950	949.999	0.0001
1	8	0.025	14.88010	2.67987	900	899.999	0.0001	950	949.999	0.0001
1	9	0.025	15.08180	2.37733	900	899.999	0.0001	950	949.999	0.0001

$V_{\rm dd} = 1$ V and $t_r = 100$ ps

Table 13.4 The magnitude of the on-chip decoupling capacitors as a function of the parasitic resistance of the power/ground lines connecting the capacitors to the current load for a limit on C_1

R_1	I_{max}	R_2	C_2	V_{load} (mV)		Error	V_{C_2} (mV)		Error
(Ω)	(A)	(Ω)	(pF)	V_{load}^{min}	SPICE	(%)	$V_{C_2}^{min}$	SPICE	(%)
$C_1 = 0.5$ pF									
1	0.005	10.6123	4.05	900	899.999	0.0001	950	949.999	0.0001
2	0.005	9.3666	4.10	900	899.999	0.0001	950	949.999	0.0001
3	0.005	8.1390	4.15	900	899.999	0.0001	950	949.999	0.0001
4	0.005	6.9290	4.20	900	899.999	0.0001	950	949.999	0.0001
5	0.005	5.7354	4.25	900	899.999	0.0001	950	949.999	0.0001
0.5	0.01	4.8606	9.05	900	899.999	0.0001	950	949.999	0.0001
1	0.01	4.3077	9.10	900	900.000	0.0000	950	949.999	0.0001
2	0.01	3.2120	9.20	900	899.999	0.0001	950	949.999	0.0001
3	0.01	2.1290	9.30	900	899.999	0.0001	950	949.999	0.0001
4	0.01	1.0585	9.40	900	899.999	0.0001	950	949.999	0.0001
$C_1 = 1$ pF									
1	0.005	13.2257	3.1	900	899.999	0.0001	950	949.999	0.0001
2	0.005	11.5092	3.2	900	899.999	0.0001	950	949.999	0.0001
3	0.005	9.8686	3.3	900	899.999	0.0001	950	949.999	0.0001
4	0.005	8.2966	3.4	900	899.999	0.0001	950	949.999	0.0001
5	0.005	6.7868	3.5	900	899.999	0.0001	950	949.999	0.0001
0.5	0.01	5.3062	8.1	900	899.999	0.0001	950	949.999	0.0001
1	0.01	4.6833	8.2	900	899.999	0.0001	950	949.999	0.0001
2	0.01	3.4644	8.4	900	899.999	0.0001	950	949.999	0.0001
3	0.01	2.2791	8.6	900	899.999	0.0001	950	949.999	0.0001
4	0.01	1.1250	8.8	900	899.999	0.0001	950	949.999	0.0001

$V_{dd} = 1$ V and $t_r = 100$ ps

Comparing results from Table 13.4 for two different magnitudes of C_1, note that a larger C_1 results in a smaller C_2. A larger C_1 also relaxes the constraints for the second decoupling stage, permitting C_2 to be placed farther from C_1. The first stage of a system of distributed on-chip decoupling capacitors should therefore be carefully designed to provide a balanced distributed decoupling capacitor network with a minimum total required capacitance, as discussed in Sect. 13.4.3.

On-chip decoupling capacitors have traditionally been allocated during a post-layout iteration (after the initial allocation of the standard cells). The on-chip decoupling capacitors are typically inserted into the available white space. If significant area is required for an on-chip decoupling capacitor, the circuit blocks are iteratively rearranged until the timing and signal integrity constraints are satisfied. Traditional strategies for placing on-chip decoupling capacitors therefore result in increased time to market, design effort, and cost.

The methodology for placing on-chip decoupling capacitors presented in this chapter permits simultaneous allocation of the on-chip decoupling capacitors and

the circuit blocks. In this methodology, a current profile of a specific circuit block is initially estimated [298]. The magnitude and location of the distributed on-chip decoupling capacitors are determined based on expected current demands and technology constraints, such as the maximum capacitance density and parasitic resistance of the metal lines connecting the decoupling capacitors to the current load. Note that the magnitude of the decoupling capacitor closest to the current load should be determined for each circuit block, resulting in a balanced system and the minimum required total on-chip decoupling capacitance. As the number of decoupling capacitors increases, the parameters of a distributed on-chip decoupling capacitor network are relaxed, permitting the decoupling capacitors to be placed farther from the optimal location (permitting the parasitic resistance of the metal lines connecting the decoupling capacitors to vary over a larger range). In this way, the maximum effective radii of a distant on-chip decoupling capacitor is significantly increased [277]. A tradeoff therefore exists between the magnitude and location of the on-chip decoupling capacitors comprising the distributed decoupling capacitor network.

13.7 Summary

A design methodology for placing distributed on-chip decoupling capacitors in nanoscale ICs can be summarized as follows.

- On-chip decoupling capacitors have traditionally been allocated into the available white space using an unsystematic approach. In this way, the on-chip decoupling capacitors are often placed far from the current load
- Existing allocation strategies result in increased power noise, compromising the signal integrity of an entire system
- Increasing the size of the on-chip decoupling capacitors allocated with conventional techniques does not enhance power delivery
- An on-chip decoupling capacitor should be placed physically close to the current load to be effective
- Since the area occupied by the on-chip decoupling capacitor is directly proportional to the magnitude of the capacitor, the minimum impedance between the on-chip decoupling capacitor and the current load is fundamentally affected by the magnitude of the capacitor
- A system of distributed on-chip decoupling capacitors has been described in this chapter to resolve this dilemma. A distributed on-chip decoupling capacitor network is an efficient solution for providing sufficient on-chip decoupling capacitance while satisfying existing technology constraints
- An optimal ratio of the parasitic resistance of the metal lines connecting the capacitors exists, permitting the total budgeted on-chip decoupling capacitance to be significantly reduced

- Simulation results for typical values of the on-chip parasitic resistances are also presented, demonstrating high accuracy of the analytic solution. In the worst case, the maximum error is 0.003 % as compared to SPICE
- A distributed on-chip decoupling capacitor network permits the on-chip decoupling capacitors and the circuit blocks to be simultaneously placed within a single design step

Chapter 14
Simultaneous Co-Design of Distributed On-Chip Power Supplies and Decoupling Capacitors

Multiple power supplies are widely used in high performance integrated circuits to provide current close to the load circuitry in high performance integrated circuits [289]. The number of on-chip power supplies is increasing, requiring innovative design methodologies to satisfy the stringent noise and power constraints of these high complexity integrated circuits [299–302]. Placing the power supply on-chip eliminates losses due to the package parasitic impedances, improving the quality of the delivered power [289].

To provide multiple on-chip power supplies, linear voltage regulators are typically used which require small area with fast load regulation to realize point-of-load voltage delivery [303]. These power supplies alone, however, do not satisfy stringent power and noise constraints. Decoupling capacitors are therefore widely used as a local reservoir of charge which are self-activated and supply current when the power supply level deteriorates [304], as previously discussed in Chap. 14. Inserting decoupling capacitors into the power distribution network is a natural way to lower the power grid impedance at high frequencies [136]. A representative power delivery network with on-chip power supplies, decoupling capacitors, and load circuits is illustrated in Fig. 14.1. Tens of on-chip power supplies, hundreds-to-thousands of on-chip decoupling capacitors, and millions-to-billions of active transistors are anticipated in the design of next generation high performance integrated circuits.

Power supplies and decoupling capacitors exhibit similar characteristics with some important differences such as the response time, decay rate of the capacitor, on-chip area, and power efficiency. On-chip power supplies require greater area, provide limited power efficiency, and exhibit slower response time as compared to decoupling capacitors. Decoupling capacitors, however, should be placed close to a power supply to recharge before the next switching event [304]. Additionally, the placement of the decoupling capacitors should consider the resonance formed by the decoupling capacitor and the power grid inductance which degrades the effectiveness of the decoupling capacitor [305]. Existing design methodologies for

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0_14

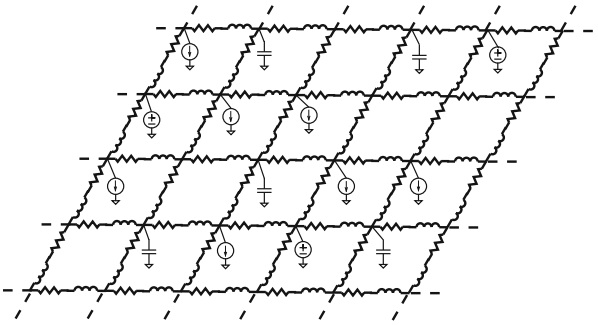


Fig. 14.1 A uniform power distribution network with multiple on-chip power supplies, decoupling capacitors, and current loads. Current loads are used to model the active circuits

placing decoupling capacitors assume one or two on-chip power supplies [304] or one decoupling capacitor interacting with multiple power supplies [306]. These assumptions are inappropriate when voltage is regulated at the point-of-load with multiple on-chip power supplies and decoupling capacitors. Design methodologies are therefore required to simultaneously place multiple on-chip power supplies and decoupling capacitors.

The effective radii of the decoupling capacitors have been developed for a single current path between a current load and a single decoupling capacitor, as described in Chap. 12, and for a mesh structure considering a single decoupling capacitor in [306]. In this chapter, not only the interactions among the power supplies, decoupling capacitors, and load circuitry but also the interactions among the power supplies and between the decoupling capacitors are considered. Those interactions considered in this chapter are schematically illustrated in Fig. 14.2.

These interactions among the circuit components are complicated by the increasing number of components in the power delivery network. In this chapter, interactions among the power supplies, decoupling capacitors, and current loads are evaluated. A methodology is described to simultaneously determine the optimum location of the distributed power supplies and decoupling capacitors within the overall power distribution network, providing an integrated approach to delivering power.

The rest of this chapter is organized as follows. The problem is formulated in Sect. 14.1. Interactions among the on-chip power supplies, decoupling capacitors, and load circuits are analyzed and a methodology for simultaneous power supply and decoupling capacitor placement is presented in Sect. 14.2. Case studies examining the interactions among the power supplies, current loads, and decoupling capacitors are provided in Sect. 14.3. Some specific conclusions are summarized in Sect. 14.4.

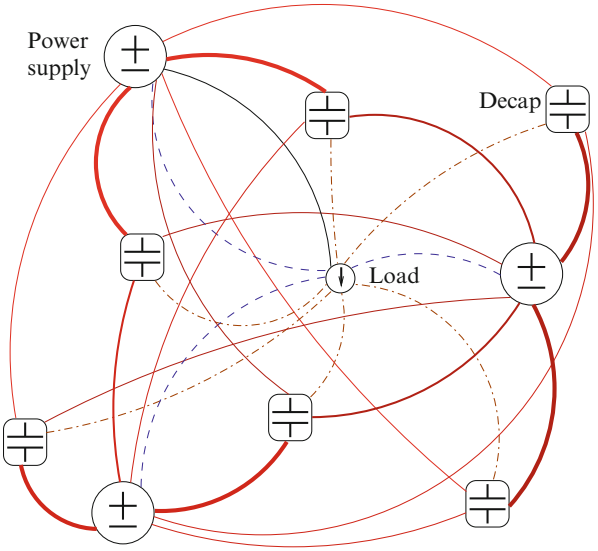


Fig. 14.2 Interactions among the on-chip power supplies, decoupling capacitors, and load circuits. *Thicker lines* represent greater interaction. Note that the effect of a power supply or a decoupling capacitance on the load circuits depends strongly on the physical distance

14.1 Problem Formulation

Power distribution networks are typically modeled as a uniformly distributed *RL* mesh structure. By exploiting the uniform nature of the power grid, the *Euclidean distance* between the circuit components can be used to determine the effective impedance between arbitrary nodes. A closed-form expression for the effective impedance in an infinite resistive mesh is provided by Venezian in [307]. This expression is modified both to produce more accurate results and to include the inductance of the power grid as the power grid impedance depends strongly on the inductance at high frequencies. A closed-form expression to determine the effective impedance between two nodes, N_{x_1,y_1} and N_{x_2,y_2}, is

$$Z_{m,n} = z * \frac{1}{2\pi} * ln(n^2 + m^2) + 0.51469, \tag{14.1}$$

where

$$m = |x_1 - x_2| \text{ and } n = |y_1 - y_2|. \tag{14.2}$$

z is the impedance of one segment of the grid. Applying this effective impedance concept, multiple current paths are considered without increasing the computational complexity of the power grid analysis process.

The complicated power distribution network schematically illustrated in Fig. 14.2 is simplified to a network consisting of only equivalent impedances among the power supplies, decoupling capacitors, and load circuitry. Multiple current paths are efficiently considered using the simplified model in (14.1).

14.2 Simultaneous Power Supply and Decoupling Capacitor Placement

Interactions among a single power supply, decoupling capacitor, and current load are illustrated in Fig. 14.3. The equivalent parasitic resistance and inductance between the power supply and decoupling capacitor, power supply and load circuit, and decoupling capacitor and load circuit are represented, respectively, as R_{vd} and L_{vd}, R_{pl} and L_{pl}, and R_{dl} and L_{dl}. The load current I_l is

$$I_l = i_{dl} + i_{pl}. \tag{14.3}$$

The current supplied from the decoupling capacitor and power supply is represented, respectively, as i_{dl} and i_{pl}. The ratio of the current supplied to the active circuits from the power supplies and decoupling capacitors depends upon the physical distances, the parasitic impedance among these components, and the size of the decoupling capacitors. i_{dl}/i_l increases for greater R_{pl} and L_{pl}, enhancing the effect on the load circuit of the decoupling capacitors as compared to the effect of the power supplies.

The effective region of a decoupling capacitor depends upon the location of those power supplies, decoupling capacitors, and load circuits in close proximity as well as the power grid impedance, and rise and fall times of the load currents. A uniform

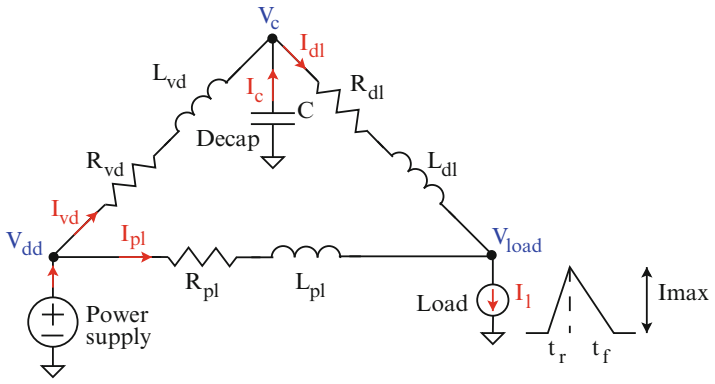


Fig. 14.3 Simplified interactions among a power supply, decoupling capacitor, and load circuit. The components are connected with the corresponding equivalent impedance in the power grid modeled by (14.1). The load current is modeled as a *triangular* current load with a rise time t_r and fall time t_f

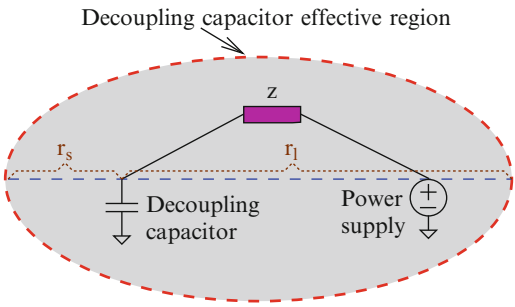


Fig. 14.4 Elliptic structure to illustrate the effective region of a decoupling capacitors with a single power supply. The short radius (r_s) and long radius (r_l) depend upon the size of the capacitor and the effective impedance between the capacitor and power supply

current distribution is assumed in this chapter to simplify this analysis. The analysis can however be generalized to a non-uniform load current distribution. The effective region for a single decoupling capacitor with a single power supply is illustrated in Fig. 14.4. The effective region exhibits an elliptic shape due to the non-uniform location of the power supplies. This elliptic shape can be explained intuitively by examining Fig. 14.4. The current supplied to the load circuit in this elliptic region is provided by the decoupling capacitor and the local power supply. When the load circuit is moved out of this elliptic region, most of the load current is provided either by the local power supply or decoupling capacitor due to the increased parasitic impedance. When the load current is supplied both by the local power supply and the decoupling capacitor, the response is faster and more effectively suppresses the switching noise.

The elliptic shape also depends upon the technology parameters and the noise constraints. The area of the elliptical region is small when the noise constraints of the power distribution network are high. For example, when the maximum target noise of the power distribution system is 5 % of the supply voltage, a smaller ellipse is formed as the effective region around the decoupling capacitor. Alternatively, the area of the elliptic region increases when the noise constraint is increased to 10 % of the supply voltage. Consequently, elliptic equipotential shapes are formed around the decoupling capacitors and power supplies, denoting identical power supply voltage levels.

An elliptic region is described by a long and short radius represented as r_l and r_s, respectively, as shown in Fig. 14.4. r_l can be determined as

$$r_{l_{(n,m)}} = \frac{K * C}{R_{(x_1,y_1)} + k * L_{(x_1,y_1)}}, \tag{14.4}$$

where C is the decoupling capacitance and the effective resistance and inductance between the decoupling capacitor and the power supply are represented, respectively, as $R_{(x_1,y_1)}$ and $L_{(x_1,y_1)}$. The horizontal and vertical distance from the decoupling capacitor to the power supply is represented, respectively, as x_1 and y_1. The effect of the transition time of the load current is embedded into the equation using k. K models the noise constraints, i.e., a smaller K is used for more stringent noise constrained circuits.

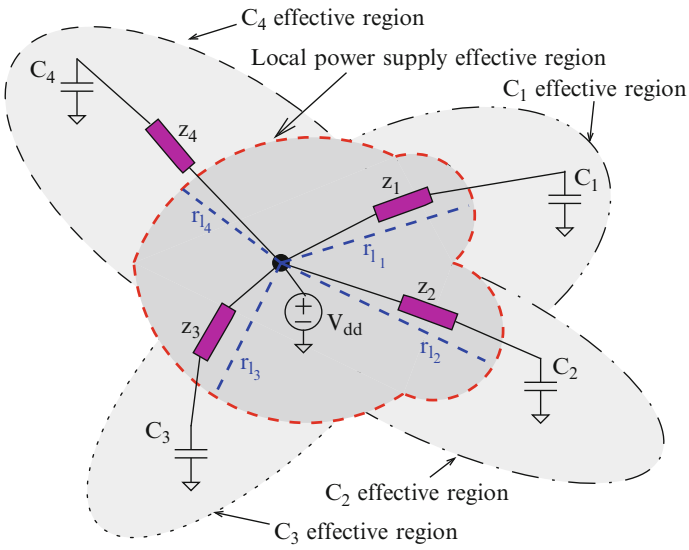


Fig. 14.5 Modified elliptic structure to illustrate the effective regions for a local power supply and decoupling capacitors. Note that the effective region for a local power supply is the overlap of the effective regions for the surrounding decoupling capacitors

The effective region for a local power supply with four decoupling capacitors is illustrated with a dark shaded modified elliptic shape in Fig. 14.5. Since the power supply interacts with four different decoupling capacitors, the effective region is the overlap of the four different elliptic shapes. Since the effect of a decoupling capacitor on r_s is limited, the effective region of the power supply can be described with four different r_l, denoted as r_{l1}, r_{l2}, r_{l3}, and r_{l4}, to represent the effect of, respectively, C_1, C_2, C_3, and C_4. A similar analysis can be performed when the system includes multiple decoupling capacitors and multiple power supplies. Each decoupling capacitor is affected by the remaining decoupling capacitors and power supplies. The effective region of a decoupling capacitor or power supply is therefore described as the overlap of the elliptic equipotential surfaces caused by each power supply and decoupling capacitor, as illustrated in Fig. 14.5.

14.3 Case Study

The method of overlapping elliptic equipotential surfaces to determine the effective region has been verified with SPICE simulations. A uniform *RL* grid structure with 20 horizontal and vertical lines (20 × 20 mesh) is assumed in the analysis. The supply voltage is 1 V and the uniformly distributed current loads switch at a 1 GHz frequency with rise and fall times of, respectively, 100 and 300 ps. A

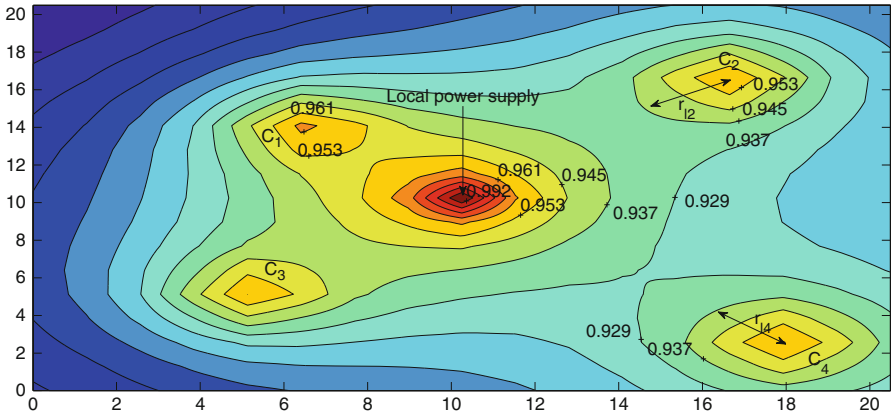


Fig. 14.6 Effective region for a local power supply at $N_{(10,10)}$ with four decoupling capacitors at $N_{(17,17)}$, $N_{(18,2)}$, $N_{(5,5)}$, and $N_{(6,14)}$ in the power distribution network. Note the elliptic shapes around the decoupling capacitors and modified elliptic shape around the local power supply

power distribution system with a local power supply and four decoupling capacitors is initially evaluated. The power supply is placed at $N_{(10,10)}$ and the decoupling capacitors C_1, C_2, C_3, and C_4 are placed, respectively, at nodes $N_{(6,14)}$, $N_{(17,17)}$, $N_{(5,5)}$, and $N_{(18,2)}$. A simulation of the power distribution network is illustrated in Fig. 14.6.

Since the closest decoupling capacitor to the local power supply is C_1, the effective region of the power supply extends towards C_1. Alternatively, the effective region of the local power supply is limited towards C_4 since C_4 is farther from the local power supply. The long radius of the effective regions for C_2 and C_4 is shown in Fig. 14.6 as, respectively, r_{l_2} and r_{l_4}. The ratio r_{l_2}/r_{l_4} is 1.2 as compared to 1.15 from (14.4), exhibiting an error of less than 5 %.

The effect of the transition time of the load current on the effectiveness of the decoupling capacitors and power supplies has also been evaluated. A 20×20 power grid with five decoupling capacitors placed at $N_{(3,3)}$, $N_{(3,17)}$, $N_{(10,10)}$, $N_{(17,3)}$, and $N_{(17,17)}$ and four on-chip power supplies located at $N_{(3,10)}$, $N_{(10,3)}$, $N_{(10,17)}$, and $N_{(16,10)}$ is evaluated. For rise and fall times of the load current of 50 and 150 ps, respectively, the effective region for the decoupling capacitors and power supplies is illustrated in Fig. 14.7a. Note that the effective region of the power supplies and decoupling capacitors exhibits similar behavior as the area of the surrounding equipotential surfaces are approximately the same. No resonance occurs and the decoupling capacitors are sufficiently close to the power supplies to be recharged before the next switching event.

The effective region of the power supplies and decoupling capacitors is depicted in Fig. 14.7b when the rise and fall transition times of the load current are increased to, respectively, 200 and 600 ps. Note that the area of the effective region of the decoupling capacitors becomes smaller. The cause of the reduced effective

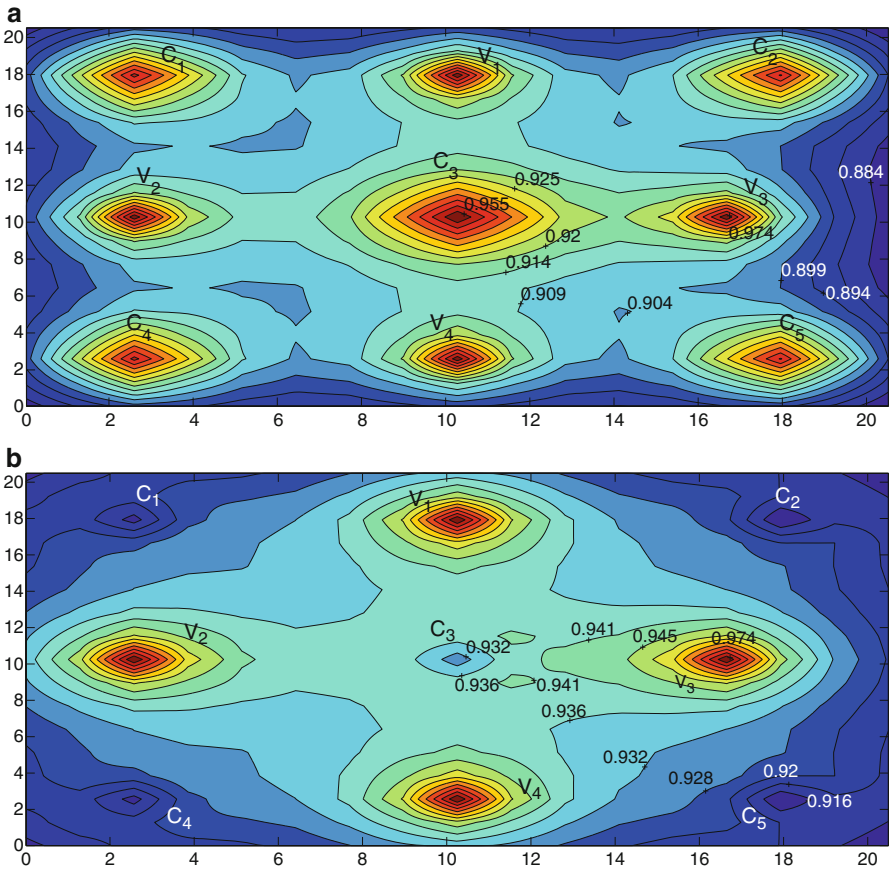


Fig. 14.7 Effective region for multiple decoupling capacitors and power supplies. Four decoupling capacitors, C_1, C_2, C_4, and C_5, are located at the corners and one decoupling capacitor C_3 is located in the middle of the power distribution network. The four power supplies are placed between the decoupling capacitors; (**a**) rise and fall transition times are, respectively, 50 and 150 ps, (**b**) rise and fall transition times are, respectively, 200 and 600 ps

region around the decoupling capacitors is that the equivalent transition times produce a resonance [305] formed by the decoupling capacitor and the power grid inductance. Also, the capacitors cannot fully recover before the next switching event. The resonance phenomenon is considered in (14.4) with k. The effective region around the on-chip power supplies is not significantly affected by the change in transition time. Considering all of these distinct properties (such as the resonance and decay rate of the capacitor) of the decoupling capacitors and power supplies can significantly improve the quality of the power distribution network.

14.4 Summary

The simultaneous co-design of distributed on-chip power supplies and decoupling capacitors is described in this chapter. The primary results are summarized as follows.

- A fundamental change in the design process of on-chip power distribution networks is necessary with the increase in the number of on-chip power supplies
- A closed-form expression to determine the effective impedance between two nodes in a uniform *RL* network is provided
- The effect of the rise and fall transition times on the effective region of the decoupling capacitors is discussed
- A closed-form expression to determine the effective elliptic region of multiple decoupling capacitors is provided
- Highly complex interactions among the power supplies, decoupling capacitors, and load circuitry are evaluated
- The analysis and simultaneous co-placement of on-chip local power supplies and decoupling capacitors are required to provide a more efficient power delivery system

Chapter 15
Conclusions

Several stages of decoupling capacitors are typically placed across the power and ground lines to bypass inductive interconnect. Decoupling capacitors are the focus of Part III. While effective for reducing the high frequency impedance of power distribution system, decoupling capacitors are only useful within a certain frequency range due to inherent parasitic resistances and inductances.

The efficient placement of decoupling capacitors is described in this part. Traditionally, decoupling capacitors have been placed within the available on-chip area. This approach, however, is not effective. Decoupling capacitors need to be placed within a specific distance from the current source to allow the charge to be efficiently transferred from the power supply to the decoupling capacitor, and within a specific distance from the load to achieve efficient charge transfer from the decoupling capacitor to the load. Based on these distances (or radii), the efficient placement of decoupling capacitor is described in Part III.

The power delivery system can be further enhanced by distributing the decoupling capacitors, starting from a large decoupling capacitor placed far from the load and ending with small distributed decoupling capacitors placed close to the load. Signal integrity is also greatly enhanced by utilizing this methodology. Additionally, the co-design of the distributed on-chip power supplies and decoupling capacitors is described in Part III, supporting the simultaneous placement of decoupling capacitors with the on-chip power supplies. These power supplies can be placed in close proximity to the load, since the size of these power supplies is relatively small.

These methodologies are described for different systems under a variety of constraints, exhibiting both computational efficiency and accuracy. The important issue of efficiently placing on-chip decoupling capacitors is the primary topic of this part.

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0_15

Part IV
Power Delivery Circuits

The on-chip integration of multiple low voltage power supplies is a primary concern in high performance ICs. An integrated power system should deliver high quality power to multiple loads in an energy efficient manner. The load regulation and power efficiency of individual power supplies within a power delivery system are particularly important and affect the overall performance of the power delivery process. The physical size of a power supply is also critical for integrating multiple power supplies on-chip. Several power delivery circuits are described in Part IV.

To provide a high quality power delivery system, the power needs to be regulated on-chip with ultra-small locally distributed power efficient converters. Different types of power supplies are reviewed in Chap. 16. To exploit the advantages of existing power supplies, a heterogeneous power delivery system is described. The power efficiency of the system is shown to be a strong function of the clustering of the power supplies—the specific configuration in which the power converters and regulators are co-designed.

The primary design characteristic that affects the development of multiple on-chip power supplies is the on-chip area. A small hybrid on-chip voltage regulator is described in Chap. 17. This active filter based voltage regulator is a combination of a buck converter and a low dropout (LDO) voltage regulator. The performance of the active filter based regulator is compared with other recently developed on-chip voltage regulators.

A fully integrated power delivery system with distributed on-chip LDO regulators developed for voltage regulation in portable mobile devices is described in Chap. 18. The circuit is fabricated in a 28 nm CMOS process. Each LDO employs adaptive bias for fast and power efficient voltage regulation, exhibiting 0.4 ns response time of the regulation loop and 98.45 % current efficiency. An adaptive compensation network is also employed within the distributed power delivery system to maintain a stable system response.

With hundreds of power domains and thousands of cores, DVS and DVFS within each core are primary design objectives for efficiently managing a power budget. In addition, line regulation will become more important when hundreds of power supplies operate together on a single monolithic substrate. Efficient design

techniques and circuits for adaptive control of power lines are therefore important. A digitally controlled current starved pulse width modulator for adaptively changing the voltage of a power converter is described in Chap. 19. Analytic closed-form expressions for the operation of a pulse width modulator are provided. The accuracy and performance of the pulse width modulator is evaluated with 22 nm CMOS predictive technology models under PVT variations. The pulse width modulator is appropriate for dynamic voltage scaling systems due to the small on-chip area and high accuracy under PVT variations.

Chapter 16
Voltage Regulators

Compromises between speed and power are the new semiconductor reality of the twenty first century. To support the demand for decreasing cost per function, a focus on functional diversification, on-chip integration, parallelism, and dynamic control have been adopted, posing significant circuit and system level challenges on power delivery and management in modern ICs. To cope with these challenges, traditional power delivery needs to be revised at the architectural, circuit, device, and material levels. Existing circuit level techniques to convert and regulate power are reviewed in this chapter in terms of the advantages, drawbacks, and compatibility, with a focus on on-chip integration of heterogeneous systems.

One of the basic building blocks of converting and regulating power within a power delivery system is a DC-DC power converter. Power supplies step voltage signals either up (boost) or down (buck) and supply the required current to the load, while regulating output voltages under PVT, line, and load variations. The power efficiency of the conversion process has historically been the defining characteristic of a power supply. With increasing on-chip noise and demand on the quality of the dynamically controlled power, load regulation of the power supply has also become critical. With on-chip integration, the physical area of these power supplies is an additional concern.

Two typical strategies to convert power are to utilize passive storage elements for energy conversion (switching topologies) or to dissipate the excess energy within a resistive element (linear topologies). Historically, large switching power supplies are preferred over compact linear power supplies due to the high, ideally 100 %, power efficiency of switching converters. With on-chip power converters, strict area constraints are imposed on the DC-DC converters, affecting the choice of power supply topology. Compact switching power converters can potentially be designed at higher switching frequencies. The parasitic impedance in these converters however increases, degrading the power efficiency of the power delivery system.

The delivery of high quality power to the on-chip circuitry with minimum energy loss is a fundamental requirement of all ICs. To supply sufficient power,

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0_16

a higher unregulated DC voltage is usually stepped down and regulated within the power delivery system [308]. Power conversion and regulation resources need to be efficiently managed to supply high quality power with minimum energy losses within multiple on-chip voltage domains [289]. The design complexity of a power delivery system increases with greater requirements on the quality of the power supply, limitations of the passive elements, board and package parasitic impedances, and limited number of I/O pins. In a modern system-on-chip, the power supplies provide the required voltage for the ICs within the overall system (CPUs, GPUs, hard disks, storage, sensors, and others), as well as the analog and digital circuit blocks within the ICs. A regulated 12 V output voltage is often derived off-chip from a 48 V battery [308]. The on-chip DC voltages are significantly lower and range from a fraction of a volt in the low power digital blocks to several volts in the input/output buffers, high precision analog blocks, and storage ICs. Furthermore, to effectively exploit power delay tradeoffs, additional power management techniques such as DVS and DVFS are often employed, further increasing the design complexity of the power delivery system. Thus, to efficiently manage the power delivered to a modern SoC, a methodology to distribute and manage the on-chip power supplies is required.

Traditionally, power is managed off-chip with energy efficient voltage converters (see Fig. 16.1a), delivering high quality DC voltage and current to the electrical grid that reliably distributes the on-chip power. The supply voltage, current density, and parasitic impedances, however, scale aggressively with each technology generation, degrading the quality of the power delivered from the off-chip power supplies to the on-chip load circuitry. The power supply in a package (PSiP) approach with partially off-chip yet in package power supplies is considered an intermediate power supply technology with respect to cost, complexity, and performance [309]. The power is regulated on-chip to lower the parasitic impedance of both the board and package (see Fig. 16.1b). To fully integrate a power converter on-chip, advanced passive components, packaging technologies, and circuit topologies are essential. Several power converters suitable for on-chip integration have recently been fabricated [303, 310–328]. Power supply systems with several on-chip power converters/regulators are commonly encountered to improve the quality of the power delivered within an IC [329–337].

On-chip power supply integration is an important cornerstone to the power supply design process. A single on-chip power converter is however not capable

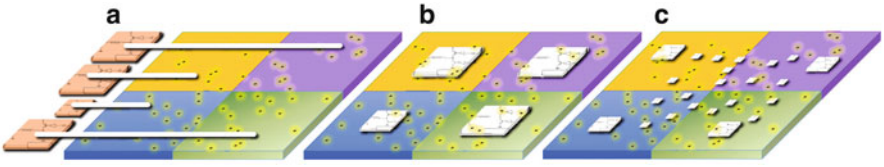


Fig. 16.1 Power delivery system with four voltage domains, (**a**) off-chip, (**b**) integrated on-chip, and (**c**) distributed point-of-load power supplies for voltage conversion and regulation

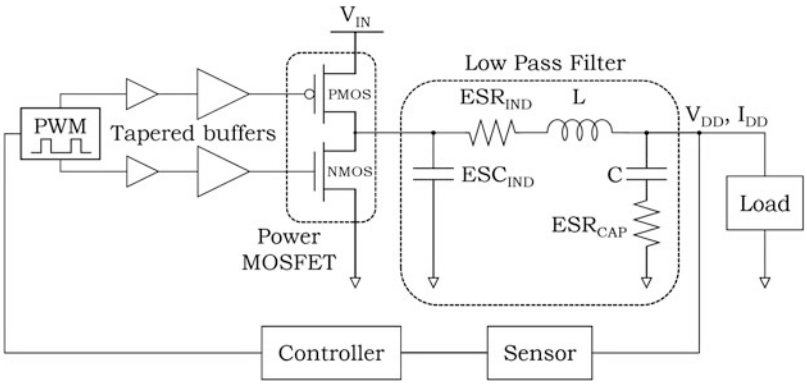


Fig. 16.2 Typical power conversion topologies (**a**) switching-mode power supply, (**b**) switched-capacitor converter, and (**c**) linear regulator

of supplying sufficient, high quality regulated current to the billions of current loads within the tens of on-chip voltage domains. To maintain a high quality power supply despite increasing on-chip parasitic impedances, hundreds of ultra-small power converters should ultimately be integrated on-chip, close to the loads within the individual multiple voltage domains [310–313]. A distributed point-of-load power supply system is illustrated in Fig. 16.1c.

While the quality of the power supply can be efficiently addressed with a distributed multi-voltage domain system, the limited power efficiency of the on-chip converters is a primary concern for the POL approach. The high power efficiency of the off-chip power converters is traded off for small area and locally regulated currents and voltages.

Three typical topologies, switching mode power supplies (SMPS), switched-capacitors (SC), and linear regulators are depicted in Fig. 16.2. Two switching topologies, SMPS and SC, are discussed, respectively, in Sects. 16.1 and 16.2. A linear topology is reviewed in Sect. 16.3. Some conclusions reviewing the preferable choice of on-chip power supply topology are provided in Sect. 16.4, followed by a brief summary of the chapter in Sect. 16.5.

16.1 Switching Mode Power Supplies

A typical switching mode power supply (SMPS) converts an input voltage V_{IN} to an output voltage V_{DD}, supplying the required current I_{DD} to the load circuitry. These converters are operated by a switching signal fed into passive energy storage components through a power MOSFET controlled by a pulse width modulator (PWM). A common step down SMPS converter operating as a buck converter is shown in Fig. 16.3. The stored input energy is restored at the output at the required

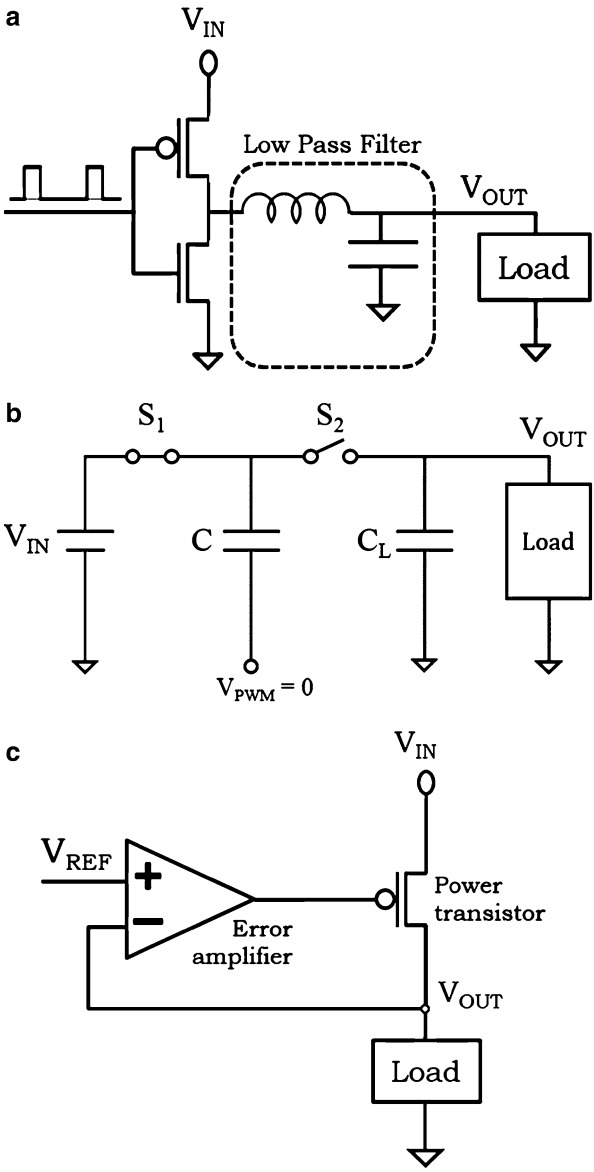


Fig. 16.3 Buck converter circuit

voltage level, maintaining high power efficiency up to a frequency f_s, typically a few megahertz [338]. The operational mode of a buck converter, output voltage, output current, and transient performance are affected by the output LC filter and controller within the feedback loop, as illustrated in Fig. 16.3. The on-chip integration of SMPS converters is greatly complicated due to I/O limitations, and constraints related to the physical size of the passive elements [339]. The area required by

the passive components to achieve a specific impedance is inversely proportional to the frequency, and can be reduced in on-chip converters by operating at ultra-high switching frequencies. An SMPS operating at a high frequency is however more affected by the parasitic impedances, degrading the power efficiency of the converter. The area of a buck converter is dominated by the size of the passive elements and is

$$A_{Buck} \approx \frac{L}{L_{\square}} + \frac{C}{C_{\square}}, \tag{16.1}$$

where $L_{\square}$ and $C_{\square}$ are, respectively, the inductance and capacitance per square micrometer of the LC filter.

Voltage regulation is a primary concern for POL power delivery. In discontinuous conduction mode (DCM) [340], the current ripple $\gamma_i I_{DD}$ within the inductor L exceeds the output current I_{DD}, and the voltage V_{DD} at the output of a converter becomes load dependent, degrading the quality of the delivered power. To support high load regulation, the buck converter is assumed to be loaded with an output current I_{DD} that exceeds the current ripple ($\gamma_i I_{DD} \leq I_{DD}$), yielding expressions for the inductor and capacitor operating in the continuous conduction mode (CCM) [314, 341],

$$L = \frac{V_{IN} - V_{DD}}{2 f_s \gamma_i I_{DD}} \cdot \frac{V_{DD}}{V_{IN}}, \tag{16.2}$$

$$C = \frac{\gamma_i I_{DD}}{8 f_s \gamma_v V_{DD}}, \tag{16.3}$$

where $\gamma_v V_{DD}$ is the voltage ripple at the converter output, and V_{DD} is the voltage at the load. To satisfy tight load regulation specifications, the output voltage ripple is assumed to range up to 10 % of V_{DD} ($\gamma_v = 0.1$). Substituting (16.2) and (16.3) into (16.4), the area of a buck converter is

$$A_{Buck} \approx \left(\frac{(V_{IN} - V_{DD}) V_{DD}}{2 L_{\square} V_{IN} f_s} \right) \frac{1}{\gamma_i I_{DD}} + \left(\frac{1}{8 C_{\square} \gamma_v V_{DD} f_s} \right) \gamma_i I_{DD}. \tag{16.4}$$

At low levels of current ripple, the area of a buck converter is dominated by the inductor and increases with smaller $\gamma_i I_{DD}$. Alternatively, at larger levels of $\gamma_i I_{DD}$, the area of a buck converter is dominated by the size of the capacitor and is proportional to the current ripple. An optimum ripple current $\gamma_{i,OPT} I_{DD}$ therefore exists that minimizes the area of a buck converter for a target output voltage ripple $\gamma_v V_{DD}$, and input and output voltage levels,

$$\gamma_{i,OPT} I_{DD} = \begin{cases} 2\sqrt{\gamma_v \left(1 - \frac{V_{DD}}{V_{IN}}\right) \frac{C_{\square}}{L_{\square}}} \cdot V_{DD} \triangleq \gamma_G G \cdot V_{DD}, \quad \gamma_G G \cdot V_{DD} \leq I_{DD} & (16.5) \\ I_{DD}, \quad \gamma_G G \cdot V_{DD} > I_{DD}, & (16.6) \end{cases}$$

where $\gamma_G G$ is the output conductance ripple and depends upon technology parameters, converted voltages, and the regulation specification. The minimum area of the buck converter is therefore

$$A_{Buck,MIN} \approx \frac{1}{2f_s} \begin{cases} \sqrt{\frac{1}{\gamma_v}\left(1-\frac{V_{DD}}{V_{IN}}\right)\frac{1}{L_\square C_\square}}, & \gamma_G G \cdot V_{DD} \le I_{DD} \quad (16.7) \\ \left[\left(1-\frac{V_{DD}}{V_{IN}}\right)\frac{1}{L_\square}\cdot\frac{V_{DD}}{I_{DD}} + \left(\frac{1}{4\gamma_v}\frac{1}{C_\square}\cdot\frac{V_{DD}}{I_{DD}}\right)\right] \approx & \\ \left(1-\frac{V_{DD}}{V_{IN}}\right)\frac{V_{DD}}{L_\square I_{DD}}, & \gamma_G G \cdot V_{DD} > I_{DD}. \quad (16.8) \end{cases}$$

Thus, in CCM at low current loads ($I_{DD} < \gamma_G G \cdot V_{DD}$), the minimum area of a buck converter is dominated by the inductance characteristics and increases with smaller values of I_{DD}. However, for values of I_{DD} larger than $\gamma_G G \cdot V_{DD}$, the minimum size of a buck converter does not strongly depend on I_{DD}. Alternatively, both the power MOSFET losses and power dissipated in the LC filter are dominant at different frequencies, conversion voltages, and current levels in CCM. The power dissipated in the power MOSFET comprises the MOSFET switching power ($\propto f_s V_{IN}^2$) and the resistive power ($\propto R_{ON} I_{DD}^2$) dissipated by the effective resistor R_{ON} of the MOSFET, yielding

$$P_{Buck,MOS} = \frac{l_{min}^2}{\mu R_{ON}(V_{IN}-V_T)} \cdot f_s V_{IN}^2 + \frac{4}{3} R_{ON} \frac{V_{DD}}{V_{IN}} I_{DD}^2, \quad (16.9)$$

where l_{min} is the minimum channel length, μ is the carrier mobility, and V_T is the threshold voltage [342] of the MOSFET. From (16.9), increasing the effective resistance of the MOSFET reduces the switching power dissipation while increasing the resistive loss. Thus, an optimum resistance R_{ON}^{OPT} exists that minimizes the power dissipated in an MOSFET, yielding

$$R_{ON}^{OPT} = \sqrt{\frac{3}{4}\frac{l_{min}^2}{\mu(V_{IN}-V_T)} \cdot f_s \frac{V_{IN}}{V_{DD}}} \cdot \frac{V_{IN}}{I_{DD}}, \quad (16.10)$$

and

$$P_{Buck,MOS}^{MIN} = 2I_{DD}\sqrt{\frac{4}{3}\frac{l_{min}^2}{\mu(V_{IN}-V_T)} \cdot f_s V_{IN} V_{DD}}. \quad (16.11)$$

The power dissipated in an LC filter [342] comprises the power losses due to the resistive (ESR_{IND}) and capacitive (ESC_{IND}) parasitic impedances of the inductor,

$$P_{Buck,IND} = \frac{4}{3} ESR_{IND} \cdot I_{DD}^2 + ESC_{IND} f_s \cdot V_{IN}^2, \quad (16.12)$$

and the power losses due to the parasitic resistance of the capacitor (ESR_{CAP}),

$$P_{Buck,CAP} = ESR_{CAP}(\gamma_i I_{DD})^2. \quad (16.13)$$

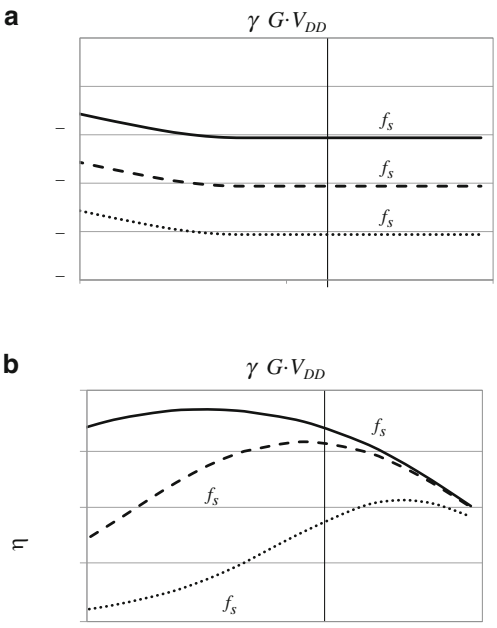


Fig. 16.4 Buck converter (**a**) physical area, and (**b**) power efficiency vs. load current for moderate, high, and ultra-high switching frequencies

The total power dissipation and power efficiency $\frac{P_{Load}}{P_{Load}+P_{Buck}}$ of the buck converter are, respectively,

$$P_{Buck} = \left(\frac{4}{3}ESR_{IND} + ESR_{CAP}\right) \cdot I_{DD}^2 + ESC_{IND} \cdot f_s \cdot V_{IN}^2 + 2\sqrt{\frac{4}{3}\frac{I_{min}^2}{\mu(_{IN}-V_T)} \cdot f_s V_{IN} V_{DD}} \cdot I_{DD}, \tag{16.14}$$

$$\varphi_{Buck} = \frac{P_{Load}}{P_{Load} + P_{Buck}} = \frac{I_{DD}V_{DD}}{I_{DD}V_{DD} + P_{Buck}}. \tag{16.15}$$

Typical passive component parameters, represented by [343–345] and technology parameters [346], are used to demonstrate power and area tradeoffs and trends in buck converters. Current loads from a few milliamperes to several amperes, and input and output voltages of, respectively, 1 and 0.7 V, are considered. The physical area (see (16.7)) and power efficiency (see (16.15)) trends are depicted in Fig. 16.4 for moderate (10 MHz), high (100 MHz), and ultra-high (1 GHz) switching frequencies.

At low current loads, the power losses of a buck converter in CCM are dominated by the parasitic capacitance of the inductor (ESC_{IND}), decreasing the

power efficiency at lower I_{DD} and larger converter size ($A_{Buck} \propto 1/I_{DD}$ for $I_{DD} < \gamma_G G \cdot V_{DD}$). Alternatively, at high current loads, the power efficiency is dominated by the parasitic resistance of the inductor (ESR_{IND}) and capacitor (ESR_{CAP}), increasing the power losses of a buck converter at higher I_{DD}. Thus, a buck converter exhibits a parabolic shaped power efficiency with current in CCM, while the physical size of the converter is reduced at higher currents. Therefore, by targeting high switching frequencies, the preferred current load to convert a voltage with minimum power losses and physical area for a specific switching frequency f_s can be determined. For example, as shown in Fig. 16.4, a preferable current exists for $f_s = 100\,\text{MHz}$ and $f_s = 1\,\text{GHz}$ since the maximum power efficiency is achieved at $I_{DD} > \gamma_G G \cdot V_{DD}$ but not at $f_s = 10\,\text{MHz}$. The minimum power loss in (16.15) is proportional to $\sqrt{f_s}$, significantly degrading the power efficiency at high frequencies. Alternatively, the size of a power converter is proportional to $1/f_s$ and decreases at higher frequencies, exhibiting an undesirable tradeoff between the power efficiency and physical size of a buck converter. The high power efficiency of traditional large power converters operating at low frequencies is therefore traded off for smaller physical size at ultra-high switching frequencies.

16.2 Switched-Capacitor Converters

Another type of switching power supply is a switched-capacitor converter, also referred to as a charge pump (CP) [347]. SC converters utilize capacitors and switches to step up or step down the input voltage based on the principle of charge conservation [289, 308]. The power is typically converted into two non-overlapping time phases, φ_1 and φ_2. During each phase, capacitors are connected in a different configuration, affecting the distribution of charge across the switched capacitors. The operation of a simple one stage CP is illustrated in Fig. 16.5 [347]. During the first phase of the period T_S ($\varphi_1 : 0 \leq t \leq T_S/2$), switches S_1 and S_2 are, respectively, closed and open, and the V_{PWM} signal is low, charging capacitor C through the input source $V_{IN} = V_{DD}$, and discharging the output node by the load current I_{OUT} (see Fig. 16.5a). At the end of φ_1, capacitor C is charged to V_{IN}, and the output node is discharged by $I_{OUT} \cdot T_S/2$. During the second phase ($\varphi_2 : T_S/2 \leq t \leq T_S$), switches S_1 and S_2 change state and the signal V_{PWM} switches from low to high, as shown

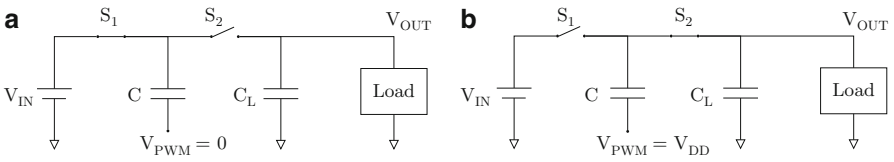


Fig. 16.5 Single stage switched-capacitor converter, (**a**) phase 1 ($\varphi_1 : 0 \leq t \leq T_S/2$) operation, and (**b**) phase 2 ($\varphi_2 : T_S/2 \leq t \leq T_S$) operation

in Fig. 16.5b). The charge stored on C during φ_1 is redistributed between capacitor C and capacitive load C_L, and supplies the load current I_{OUT} during phase φ_2. As a result, the voltage at the output increases during each subsequent cycle up to a final asymptotic steady state voltage [347],

$$V_{OUT}^{(SS)} = 2V_{DD} - \Delta V = 2V_{DD} - \frac{I_{OUT}}{f_s C}, \tag{16.16}$$

where ΔV is the voltage drop due to charge sharing, and $f_s = 1/T_S$ is the switching frequency of the converter. The capacitor C behaves as a charge pump in the SC converter and, for sufficiently high switching frequencies, the output voltage in steady state is maintained close to $2V_{DD}$ [347]. The voltage drop at the output increases with higher values of I_{OUT} (see (16.16)), making load regulation difficult in SC converters with high load currents. To limit the voltage drop at the output of an SC converter, a minimum switching frequency $f_{s,MIN}$ is determined. This voltage drop can also be reduced by increasing the size of the capacitors, trading the larger physical size of the SC converter for improved output voltage regulation. Large power supplies are, however, ineffective for on-chip integration. To enhance the ability of a SC converter to regulate the load, feedback circuitry is added at the expense of lower efficiency of the power conversion process. Alternatively, low current applications that require a low-to-high voltage conversion but not necessarily excellent load regulation (such as wireless monitoring systems, non-volatile memory, and certain mixed-signal systems [348–351]) are natural applications for switched-capacitor converters.

The power efficiency of a typical switched-capacitor DC-DC converter is primarily limited by heat losses incurred when transferring charge between the switched capacitors P_{TRANS}, losses in the power switches P_{SW}, and dynamic power dissipated in the parasitic resistance P_{DYN}. While the theoretical power efficiency of other switching converters (e.g., SMPS) is 100 %, the power loss due to charge transfer P_{TRAN} is unavoidable in switched-capacitor converters, degrading the maximum power efficiency of an SC converter by [308]

$$P_{TRANS} = \frac{f_s}{2} \cdot \frac{CC_L}{C + C_L} \cdot (V_{DD} - \Delta V)^2 . \tag{16.17}$$

The switching and dynamic components of the total power loss are strongly dependent upon the design of the SC switches and are comparable with switching and dynamic power losses in other switching converters. Alternatively, power losses due to the parasitic resistance of the wire P_{TRANS} are specific to the switched-capacitor topology and increase with switching frequency, as described by (16.17). Intuitively, to increase the power efficiency of an SC converter, the capacitors should switch at a lower frequency. This solution is however limited by the minimum frequency constraint $f_s > f_{s,MIN}$ determined from (16.16) which is related to the voltage drop and physical size of the converter. To increase the efficiency of an SC converter over a range of frequencies, dynamically reconfigurable topologies should be considered

[352]. The physical size, load regulation, and power efficiency characteristics are all important design criteria for integrating on-chip power supplies. Due to the unfavorable tradeoff among these characteristics in an SC converter, switched-capacitor converters are not preferred for distributed on-chip integration in modern heterogeneous systems. An alternative topology that exhibits small physical area and excellent load regulation characteristics is described in the following section.

16.3 Linear Converters

To provide a specific voltage V_{DD} and current I_{DD} to the load circuitry, a linear power supply converts an input DC voltage V_{IN} using a resistive voltage divider controlled by feedback from the output. The primary drawback of a linear topology is the resistive power losses that increase with a larger $V_{IN} - V_{DD}$ voltage drop, which limit the power efficiency to V_{DD}/V_{IN}. Alternatively, linear converters require a relatively small area, an important characteristic for on-chip integration. Linear regulators can be either analog or, more recently, digital in nature. Analog and digital voltage regulators are described, respectively, in Sects. 16.3.1 and 16.3.2.

16.3.1 Analog LDO Regulators

A low dropout DC-DC regulator, depicted in Fig. 16.6, is a standard linear converter topology that operates with a low $V_{IN} - V_{DD}$ voltage drop. The current that flows through a linear converter is I_{IN}.

The current supplied by a linear converter comprises the useful LDO current I_{DD} that flows to the load, and the short-circuit $I_{IN} - I_{DD}$ current dissipated

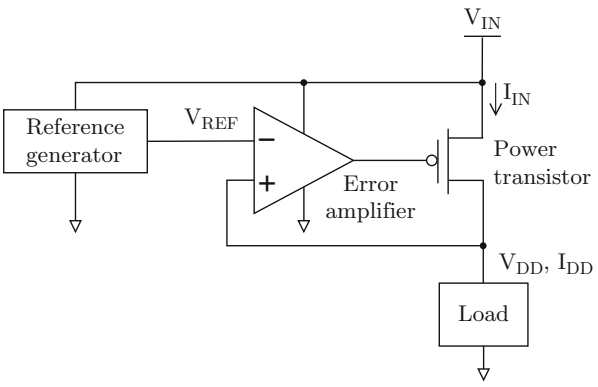


Fig. 16.6 Analog LDO circuit

in the bandgap voltage reference and error amplifier. Power and area efficient voltage references have recently been reported [303, 316–318]. The LDO current is, therefore, dominated by the error amplifier and power transistor currents. To mitigate transient voltage peaks while supporting fast changes in the load current, larger currents should be utilized within the error amplifier, increasing the short-circuit current. Alternatively, to satisfy the current load requirements in modern high performance circuits, high currents of up to several amperes are required by the load circuitry. The current flow within an LDO is therefore dominated by the load current I_{DD}. In this case, both the physical area and power dissipation of a linear converter are primarily dictated by the size and dissipated power of the output power transistor. Thus, the area of an LDO is proportional to the width W of the output transistor, yielding

$$A_{Linear} \propto W l_{min} \propto \frac{I_{DD} \cdot l_{min}^2}{\mu C_{OX}(V_{IN} - V_{AMP} - V_T)^2}, \tag{16.18}$$

where l_{min} is the minimum channel length, μ is the carrier mobility, C_{OX} is the gate oxide capacitance, and V_{AMP} is the output from the error amplifier. To accommodate the effects of the line and load specifications that may significantly affect the physical size of an LDO, a typical area per 1 mA load [303, 310–318] (see Fig. 16.7) is considered for those LDOs with a high current load. A parabolic relationship is exhibited between the physical area and minimum technology dimensions ($A_{Linear}/I_{DD} \propto l_{min}^2$). The ratio $A_{Linear}/I_{DD} = 5 \times 10^{-6}\,\text{mm}^2/\text{mA}$ is based on the 28 nm technology node considered in Fig. 16.8. Typical 28 nm CMOS technology parameters [346], and input and load voltages are assumed in this discussion to demonstrate the need for a large power transistor to supply high current to the load (see Fig. 16.8). The size of the linear converter ranges from $60 \times 60\,\mu\text{m}^2$ for I_{DD} = 0.5 A to $150 \times 150\,\mu\text{m}^2$ for I_{DD} = 3.5 A (see Fig. 16.8), which can be further reduced with technology scaling ($A_{Linear} \propto l_{min}^2$) and advanced design solutions [303, 310–328]. The current can therefore be supplied to the load with an LDO orders of magnitude smaller than a corresponding buck converter.

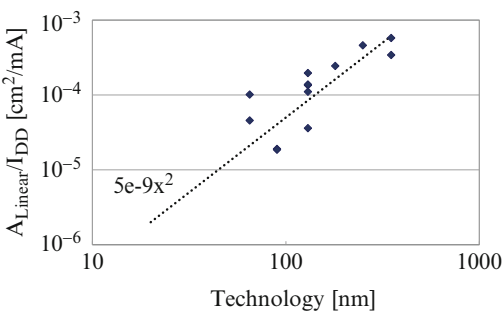


Fig. 16.7 Physical area of LDO per 1 mA load

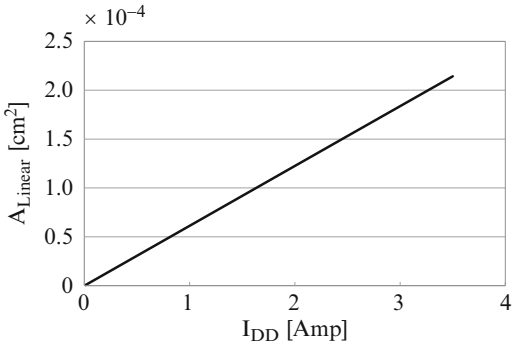


Fig. 16.8 Physical area of LDO for typical current loads at 28 nm CMOS technology

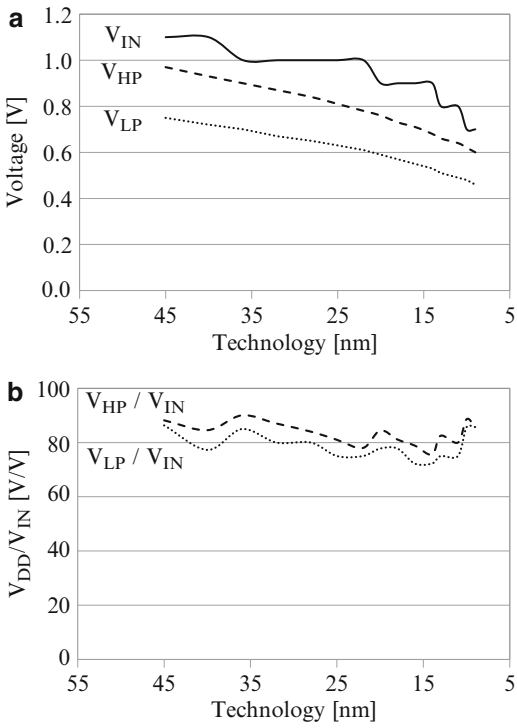


Fig. 16.9 Trends in typical (**a**) high performance (V_{HP}), low power (V_{LP}), and internal core primary (V_{IN}) voltage supplies, and (**b**) voltage conversion ratios (V_{HP}/V_{IN}) and (V_{LP}/V_{IN})

The power dissipation of an LDO is

$$P_{Linear} \approx (V_{IN} - V_{DD})\, I_{DD}. \tag{16.19}$$

Thus, the power loss in a linear converter increases with a higher $V_{IN} - V_{DD}$ drop, degrading the power efficiency of the converter. Recent supply voltage trends are illustrated in Fig. 16.9 for the internal core primary voltage V_{IN}, and typical high and low V_{DD} levels [346], yielding efficiency bounds within the 70 %–90 % range of the V_{DD}/V_{IN} ratio shown in Fig. 16.9. Thus, a moderate LDO power efficiency $\varphi_{Linear} = V_{DD}/V_{IN}$ of at least 70 % is typically expected.

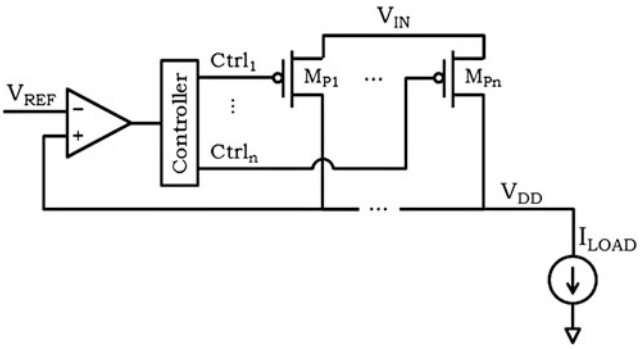


Fig. 16.10 Digital LDO circuit

16.3.2 Digital LDO Regulators

Sub/near threshold computing is a promising technique to reduce the power consumed by an IC [353–357]. To provide a stable supply voltage at sub/near threshold levels, tunable low noise voltage regulation below 0.5 V is required. To provide these output voltages with sufficient power efficiency, an analog LDO regulator should operate with low input voltages (e.g., $V_{IN} < 0.7$ V to provide V_{DD} of 0.5 V with at least 70 % power efficiency). A conventional analog LDO, however, fails to operate at these low input voltages due to the insufficient voltage drop across the individual transistors within the current mirror of the LDO. A digital LDO can however be used to suppress the analog nature of a conventional analog LDO [358, 359].

A digital LDO regulator is typically comprised of a comparator, a digital controller, and an array of MOSFET transistors, as depicted in Fig. 16.10. In this configuration, the output voltage V_{DD} is continually monitored by the comparator, and the number of simultaneously active MOSFET transistors is dynamically determined by the digital controller based on the comparator output. The power transistor within a conventional analog LDO regulator is replaced in a digital LDO by a digitally controlled array of switches. These digital regulators typically exhibit robust ultra-low voltage regulation and high power efficiency. Alternatively, a slow transient response, large output ripple, and greater physical area are the primary concerns in digital LDO regulators [358, 360, 361].

16.4 Comparison of Monolithic Power Supplies

As previously mentioned, the on-chip integration of multiple low voltage power supplies is a primary concern in high performance ICs. An integrated power system should deliver high quality power to multiple loads in an energy efficient manner. The load regulation and power efficiency of individual power supplies within a

power delivery system are particularly important and affect the overall performance of the power delivery process. The physical size of a power supply is also critical for integrating multiple on-chip power supplies. The power efficiency, physical size, and load regulation characteristics are compared in this section for the three classical power supply topologies.

The operation of both switching mode power supplies and switched-capacitor converters is based on a two-phase principle. In both topologies, the energy is stored in passive circuit elements during one phase and restored at the output during the other phase. The capacitors and inductors both increase the physical area of the converter, making integration of an on-chip SMPS problematic. Alternatively, SC converters are comprised of only capacitive elements and therefore require less area, making SC converters more suitable for on-chip integration. The high power efficiency and good load regulation characteristics of an SMPS converter are, however, significantly degraded in an SC topology, making switched-capacitor converters inefficient for certain applications.

Linear regulators exhibit small physical size due to the lack of capacitors and inductors, and excellent load regulation characteristics due to the small output impedance. The power efficiency of a linear regulator is limited by the voltage drop across the output power transistor (dropout voltage), making this topology relatively power inefficient when converting large differences between the input and output voltages. Alternatively, the power efficiency of linear regulators increases with lower dropout voltages. To maintain reasonable efficiency, low dropout regulators should be utilized for on-chip integration.

The primary electrical characteristics of SMPS, SC, and low dropout linear topologies are summarized in Table 16.1. Switched-capacitor converters are not preferred for high performance applications due to the limited power efficiency and difficulty in regulating the output voltage under high load variations. The high power efficiency and good load regulation of SMPS converters have made this topology particularly effective for power conversion in high power, high performance applications, such as microprocessors, DSPs, SRAMs, and hard disks [308]. The large physical area and difficulty to integrate on-chip inductive elements degrade the effectiveness of SMPS converters as the need for high quality, distributed on-chip POL power delivery increases. Alternatively, small LDO regulators are a natural

Table 16.1 Comparison of SMPS, SC, and LDO topologies for power conversion

Power converter	SMPS	SC	LDO
Step up conversion	Yes	Yes	No
Power efficiency	High	Medium	Limited to $\frac{V_{OUT}}{V_{IN}}$
Load regulation	Good	Poor	Good
Physical area	Large	Medium	Small
Historical applications	Microprocessors, DSPs, SRAMs, and hard disks	EEPROM, DRAM, flash, and mixed-signal	DRAM

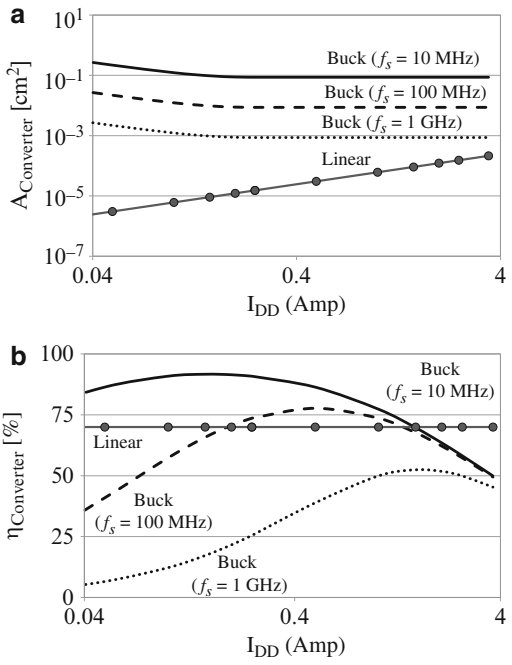


Fig. 16.11 LDO and buck converter (**a**) physical area, and (**b**) power efficiency for moderate, high, and ultra-high switching frequencies

choice for on-chip integration. With continuous scaling of the physical feature size and advanced circuit solutions, thousands of ultra-small LDO regulators can potentially be distributed on-chip, close to the loads [191]. Nevertheless, novel design solutions are required to enhance the overall efficiency of LDO based power delivery systems.

The physical area and power efficiency of an LDO and buck converter are shown in Fig. 16.11. Buck converters are more power efficient than alternative LDOs when operating at low switching frequencies. Alternatively, compact buck converters can operate at high switching frequencies. These buck converters, however, exhibit lower power efficiency and are therefore less effective for on-chip integration. Thus, to deliver high quality power to the load circuitry under typical area constraints, on-chip linear regulators should be considered. The moderate power efficiency of an LDO however becomes a significant constraint when the power consumed by the load increases. For example, converting 2 V into 1 V while delivering 1 μA to the current load results in a 50 % power efficiency and $1\,\text{V} \cdot 1\,\mu\text{A} = 1\,\mu\,\text{W}$ power loss that can possibly be absorbed by the power delivery system. Alternatively, converting 1.25 V into 1 V while delivering 1 mA to the current load results in 80 % power efficiency and a significant $0.25\,\text{V} \cdot 1\,\text{mA} = 250\,\mu\,\text{W}$ power loss that is difficult to tolerate. Thus, for on-chip integration, linear regulators are preferable to switching power supplies, particularly for small input-to-output voltage differences. A heterogeneous power delivery system that efficiently exploits the combined power and area characteristics of linear and switching converters is desirable to enhance the

power quality and efficiency of the overall power delivery system while satisfying on-chip area constraints.

To better exploit the electrical characteristics of existing power converters, different topologies have recently been combined into hybrid power supplies [187, 313, 362]. For example, linear power supplies can be utilized on-chip for local voltage regulation, whereas off-chip or in-package SMPS can mitigate power losses caused by large voltage drops [187]. In this scheme, the small size and excellent regulation characteristics of an LDO are combined with the high power efficiency of a switching converter, enhancing overall performance and quality. Hybrid topologies exhibit promising qualities for heterogeneous on-chip power delivery, while also posing new research challenges. The co-design of different power converters, overall system-wide power efficiency, dynamic control, and scalability of heterogeneous power delivery systems are topics of growing importance which are discussed throughout this book.

16.5 Summary

Switching and linear power supply topologies are reviewed in this chapter. The primary conclusions are summarized as follows.

- The physical size of a buck converter is weakly dependent upon the load current and is smaller at higher frequencies
- A buck converter exhibits a parabolic shaped power efficiency with current when operating in the CCM mode
- The preferred current load can be determined to convert a voltage within a buck converter with minimum power losses and physical area for a specific switching frequency f_s
- The minimum power loss of a buck converter is proportional to $\sqrt{f_s}$, significantly degrading power efficiency at high frequencies
- The high power efficiency of traditional large switching power converters operating at low frequencies is traded off for smaller physical size at ultra-high switching frequencies
- The switching and dynamic components of the total power loss of SC converters are strongly dependent upon the SC switches and are comparable with switching and dynamic power losses in SMPS converters
- Power losses due to the parasitic resistance of the wire are specific to the switched-capacitor topology and increase with switching frequency, limiting the maximum switching frequency of SC converters
- The minimum switching frequency of an SC converter is constrained by the voltage drop and physical size of the converter
- Switched-capacitor converters are less power efficient than SMPS converters, and are less effective for output regulation than SMPS and LDO converters

- To increase the efficiency of an SC converter over a range of frequencies, dynamically reconfigurable topologies should be considered
- The power loss in a linear converter increases with a higher $V_{IN} - V_{DD}$ drop, degrading the power efficiency of the converter
- The efficiency of a linear converter for typical supply voltages is typically within the 70 %–90 % range of the V_{DD}/V_{IN} ratio
- Linear regulators are preferable to switching power supplies for small input-output voltage differences
- Existing power converter topologies exhibit an undesirable tradeoff among power efficiency, load regulation, and physical size
- A heterogeneous power delivery system that efficiently exploits the power and area characteristics of linear and switching converters is desirable to enhance the power quality and efficiency of the entire power delivery system while satisfying on-chip area constraints

Chapter 17
Hybrid Voltage Regulator

A primary issue in the design of a conventional on-chip voltage converter is the physical area. The on-chip passive LC filter within a monolithic buck converter occupies a large area. Multiple on-chip buck converters are therefore infeasible due to the significant area of the passive components (such as in a multi-voltage microprocessor).

A more area efficient voltage converter structure is a low dropout voltage regulator [303, 316–318, 363–366]. These regulators are placed on-chip close to the load for fast and accurate regulation. These regulators require a large output capacitance to achieve fast load regulation. This capacitor occupies significant on-chip area and is therefore generally placed off-chip [316, 317]. The off-chip output capacitor requires dedicated I/O pins and produces high parasitic losses. Alternatively, when the output capacitor is placed on-chip, the output capacitor dominates the total LDO regulator area [303]. Many techniques have been described to eliminate the need for a large off-chip capacitor without sacrificing the stability and performance of an LDO regulator [303, 318, 363–366]. These techniques, however, do not completely negate the need for an output capacitor. Furthermore, the compensation circuitry to produce a dominant pole also requires additional area. Due to these significant area requirements, standard LDO regulators are not appropriate for a system of distributed point-of-load voltage regulators.

An ultra-small area efficient voltage converter is required for the next generation of multi-voltage systems since these systems are highly sensitive to local power/ground noise. The parasitic impedance of the power distribution network is a crucial issue when the voltage converter is far from the load. Voltage converters need to be placed close to the load since $L\ dI/dt$ noise and IR voltage drops have become significant in deeply scaled circuits with aggressively scaled supply voltages [289, 314].

To provide a voltage regulator appropriate for distributed point-of-load voltage generation, the passive LC filter within a buck converter is replaced with a more area efficient active filter circuit [367]. A switching input voltage generates the

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0_17

output voltage and the converter uses a filter structure to produce the desired output voltage. The current supplied to the output node, however, does not originate from the input switching signal; rather, from the operational amplifier (Op Amp) output stage, similar to a linear voltage converter. The voltage converter is therefore a hybrid combination of a switching and linear DC-DC converter. The on-chip area of the hybrid regulator is 0.015 mm^2, significantly smaller than state-of-the-art output capacitorless LDOs. The power efficiency, however, is limited to V_{out}/V_{in}, similar to LDOs.

The rest of the chapter is organized as follows. In Sect. 17.1, different active filter topologies and types such as Butterworth, Chebyshev, and Bessel are considered for the low pass active filter. Several tradeoffs among a number of active filter topologies are discussed. The design requirements of the Op Amp and related tradeoffs are also discussed in this section. The advantages and disadvantages of the voltage regulator as compared to conventional switching and LDO regulators are discussed in Sect. 17.2. Experimental results are provided in Sect. 17.3. A distributed system of point-of-load voltage regulators is described in Sect. 17.4. A summary of the chapter is provided in Sect. 17.5.

17.1 Active Filter Based Switching DC-DC Converter

In the active filter based circuit, the bulky LC filter in a conventional buck converter is replaced with an active filter structure and the tapered buffers are replaced with smaller buffers, as shown in Fig. 17.1. The switching input signal generated at $Node_1$ is filtered by the active filter structure, similar to a buck converter, and a DC voltage is generated at the output. The output voltage

$$V_{out}(t) = V_{out} + V_r(t), \tag{17.1}$$

where V_{out} is the output DC voltage and V_r is the output voltage ripple caused by the non-ideal low pass filter. V_{out} is the average value of the switching voltage at $Node_1$, which is

$$V_{out} = V_{in}\left(D - \frac{t_r - t_f}{2T}\right), \tag{17.2}$$

where D, t_r, t_f, and T are, respectively, the duty cycle, rise time, fall time, and period of the switching voltage. Increasing the duty cycle D of the input switching signal at $Node_1$ increases the generated DC voltage as in (17.2).

Large tapered buffers are required in a conventional buck converter to drive the large PMOS and NMOS power transistors, as shown in Fig. 16.3. The current delivered to the load circuitry is provided by these large power transistors. In the active filter based regulator, however, the current delivered to the load circuitry is supplied by an operational amplifier. Small buffers are therefore sufficient

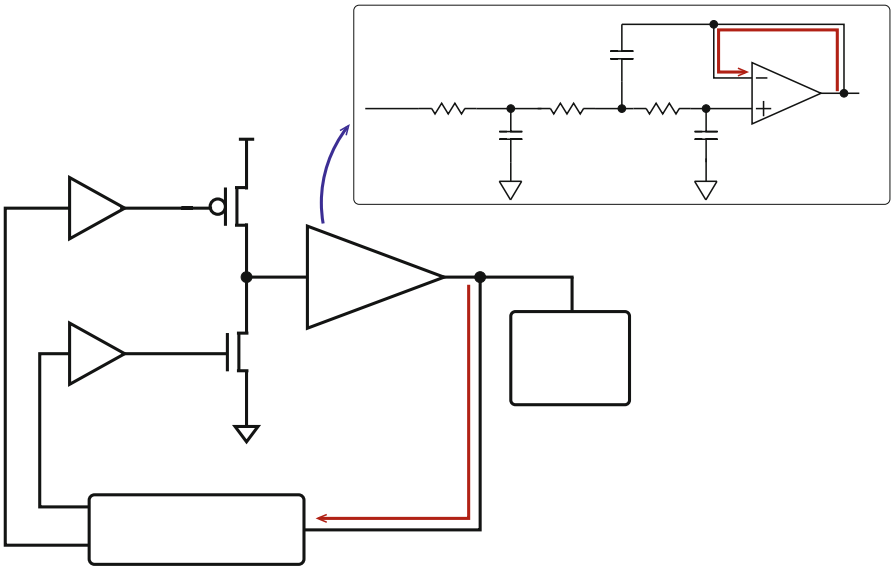


Fig. 17.1 Active filter based DC-DC converter. Note that the passive LC filter is replaced with an active filter and the large tapered buffers are no longer necessary

for driving the active filter. Replacing the tapered buffers with smaller buffers significantly decreases the power dissipated by the input stage. Alternatively, the output buffers within the Op Amp dissipate power within the regulator. Another characteristic of the hybrid regulator is that the feedback required for line and load regulation is satisfied with separate feedback paths, as shown in Fig. 17.1. Feedback_1 is generated by the active filter structure and provides load regulation whereas Feedback_2 is optional and controls the duty cycle of the switching signal for line regulation. In most cases, Feedback_1 is sufficient to guarantee fast and accurate load regulation. When only one feedback path is used, the switching signal is generated by simpler circuitry (e.g., a ring oscillator) and the duty cycle of the switching signal is compensated by a local feedback circuit (a duty cycle adjustor). The primary advantage of a single feedback path is smaller area since Feedback_1 is produced by the active filter and no additional circuitry is required for the compensation structure.

Utilizing active filters within a switching voltage regulator to replace the passive LC filter was first proposed in [367]; however, several important design issues such as the power efficiency, sensitivity of the active filter, importance of the output buffer stage of the Op Amp, and type and topology of the active filter structure were not considered. Additionally, the active filter-based regulator in [367] requires a $10\,\mu\text{F}$ capacitor, which occupies significant on-chip area and is therefore inappropriate for point-of-load voltage regulation. Less than 8 pF capacitance is used within the active filter portion of the voltage regulator for a cutoff frequency of 50 MHz.

Active filters have been well studied over the past several decades [368, 369]. The objective here is to review those properties of active filters that affect the design

of the voltage regulator while providing some relevant background material. Active filter configurations and topologies relevant to the regulator are reviewed in Sect. 17.1.1. In Sect. 17.1.2, the design of the Op Amp circuit is discussed.

17.1.1 Active Filter Design

Active filter structures contain no passive inductors. The filtering function uses capacitors, resistors, and an active circuit (i.e., the Op Amp). Certain design considerations should be considered when utilizing an active filter as a voltage regulator since the appropriate active filter topology depends upon the application. For a voltage regulator, the on-chip area requirement, sensitivity of the active filter to component parameter variations (due to aging, temperature, and process variations), and the power dissipated by the active components should be low. Two topologies are popular for designing an integrated low pass active filter, multiple feedback and Sallen-Key [368]. Multiple feedback low pass filters use capacitive and resistive components within the feedback path from the output to the input. A DC current path exists between the input and output nodes due to the resistive feedback. The DC current increases the power dissipated by the multiple feedback active filter. Multiple feedback active filters are therefore less suitable for an active filter based on-chip voltage regulator. Alternatively, Sallen-Key low pass filters only use capacitive feedback. Hence, the static power dissipation of the Sallen-Key topology is significantly less than the multiple feedback topology.

A third order low pass unity-gain Sallen-Key filter topology is shown in Fig. 17.2. The first section, R_1 and C_1, forms a first order low pass RC filter. The remaining components, R_2, R_3, C_2, C_3, and the Op Amp, form a second order Sallen-Key low pass filter. Note that no DC current path exists between the input and output. The gain of the active filter can be increased by inserting resistive feedback between the non-inverting input and output nodes, forming a DC current path between the output and ground. Since low power dissipation is crucial to a point-of-load voltage regulator, a unity-gain topology is chosen.

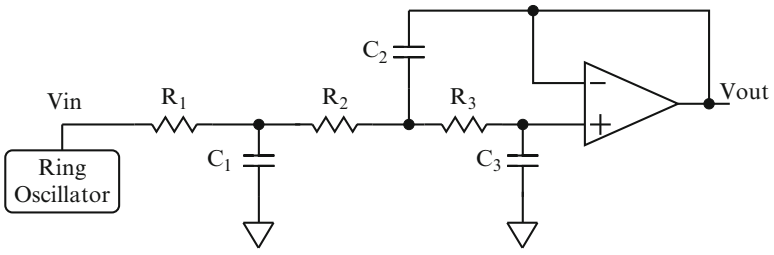


Fig. 17.2 Active low pass Sallen-Key filter circuit. No DC current path exists between the input and output nodes

The transfer function of the active filter shown in Fig. 17.2 is

$$\frac{V_{out}}{V_{in}} = \frac{1}{a_1 s^3 + a_2 s^2 + a_3 s + a_4}, \tag{17.3}$$

where

$$\begin{aligned} a_1 &= R_1 R_2 R_3 C_1 C_2 C_3, \\ a_2 &= R_1 C_1 C_3 (R_2 + R_3) + R_3 C_2 C_3 (R_1 + R_2), \\ a_3 &= R_1 C_1 + C_3 (R_1 + R_2 + R_3), \\ a_4 &= 1. \end{aligned}$$

Various filter types exist in the literature with zeros at infinity, for example, Butterworth, Chebyshev type I, and Bessel [369]. Other filter types such as Elliptic and Chebyshev type II filters exhibit faster transition characteristics. Since the Elliptic and Chebyshev type II filters contain zeros in the transfer function, the Sallen-Key topology depicted in Fig. 17.2 cannot be used to design these filters. Zeros can be produced with more complex feedback structures such as a twin-t or bridged-t circuit [369]. These structures, however, employ resistors connected to ground, increasing the power dissipated by the active filter.

A Chebyshev type I filter is chosen for the active filter due to the steep roll-off factor as compared to those filter structures which do not require resistive components connected to ground to produce finite zeros. The active filter passes the switching signal at a constant frequency and generates a DC output voltage. A third order Chebyshev type I low pass Sallen-Key filter, shown in Fig. 17.2, is utilized in the voltage regulator since no attenuation occurs at DC when the order of the Chebyshev filter is odd. The per cent change in the cutoff frequency and the Q factor of the third order Sallen-Key filter, shown in Fig. 17.2, are listed in Table 17.1 for an increase of 1 % in the value of the individual parameters.

Table 17.1 Sensitivity analysis for a third order Sallen-Key filter. Per cent change in cutoff frequency and Q factor when individual parameter values are increased by 1 %

	R_1	R_2	R_3	C_1	C_2	C_3
Q	0	−0.4	0.4	0	−0.5	0.5
Cut-off frequency	−1	−0.5	−0.5	−1	−0.5	−0.5

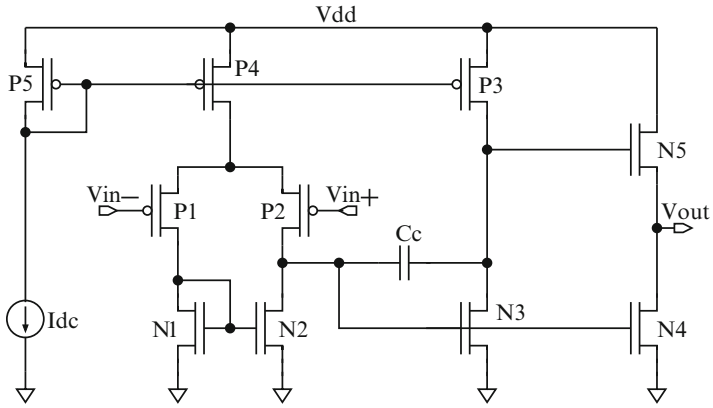


Fig. 17.3 Three stage Op Amp with PMOS input transistors. The PMOS input transistors are used in the first differential input stage. The second stage is a common-source gain stage and the third stage forms the output buffer that supplies the current to the load

17.1.2 Op Amp Design

The performance of an active filter depends strongly on the Op Amp. The gain-bandwidth (GB) product of the Op Amp determines the bandwidth of the active filter. Most of the power loss takes place within the Op Amp structure, since the current provided to the output load is supplied by the Op Amp output stage. Hence, with hundreds of on-chip distributed power regulators, each the Op Amp needs to provide tens of milliamperes to the load while maintaining sufficient performance to reliably operate the active filter.

A three stage classical differential-input single-ended CMOS Op Amp structure is utilized in the voltage regulator, as shown in Fig. 17.3 [370]. The size of the transistors in the output stage is considerably larger than the first two stages to supply sufficient current to the load circuits. The first and second stages are gain stages which provide a cascade gain of greater than 50 dB. The third stage exhibits a gain close to unity, so the overall three stage gain is close to 50 dB with a phase margin of 51°, as depicted in Fig. 17.4.

17.2 Pros and Cons of Active Filter-Based Voltage Regulator

The voltage regulator is a hybrid combination of a switching and linear voltage regulator and exhibits certain advantages and disadvantages from using a combination of a switching and LDO regulator topology.

Voltage regulation: The line and load regulation of the voltage converter is separated into two different feedback paths, as shown in Fig. 17.1. The response

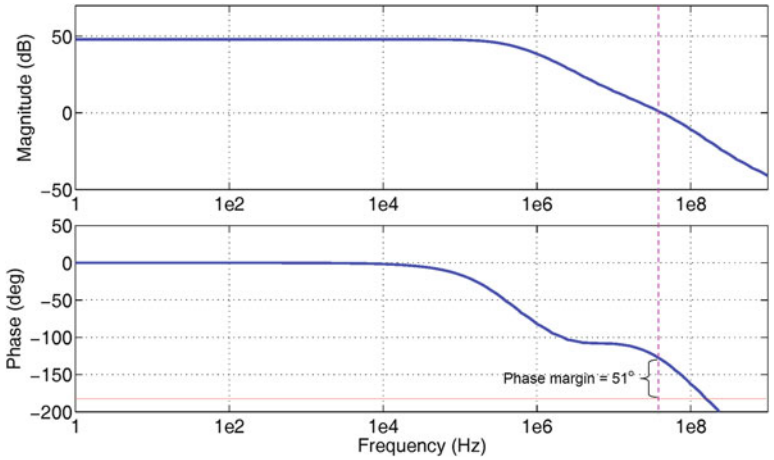


Fig. 17.4 Magnitude and frequency response of the Op Amp in the active filter. The phase margin is 51°

time for abrupt changes in the load current is faster than a switching regulator and similar to an LDO regulator. The line regulation characteristics are, however, similar to a switching voltage regulator where the duty cycle of the input switching signal is altered by the PWM.

On-chip area: The physical area of the voltage regulator is smaller than both a switching and LDO voltage regulator since there is no large output capacitor. The frequency of the input switching signal can be increased without significantly degrading the power efficiency because the buffers delivering this switching signal can be small. With higher switching frequencies, the size of the voltage regulator can be further decreased. The primary advantage of the voltage regulator as compared to other regulator topologies is the small area requirement without significantly degrading the power efficiency.

Power efficiency: The power efficiency of a buck converter can theoretically approach 100 % when the parasitic impedances are negligible. As previously mentioned, for an LDO or the hybrid voltage regulator, the maximum attainable power efficiency is limited to V_{out}/V_{in}.

Maximum load current: The maximum current that can be delivered to the load depends upon the size of the power transistors (PMOS and NMOS shown in Fig. 16.3) driving the LC filter. A higher current can be delivered with larger power transistors. The maximum load current of an LDO regulator depends upon the size of the pass transistor. Similarly, the maximum load current of the hybrid voltage regulator is determined by the size of the output stage of the Op Amp.

17.3 Experimental Results

The active filter based DC-DC voltage converter has been designed and fabricated in a 110 nm CMOS technology. The ultra-small hybrid voltage regulator requires an area smaller than 0.015 mm^2. A significant portion of this area is allocated to the Op Amp, as shown in Fig. 17.5. The active filter, Op Amp, and PWM are placed in the remaining available area. The active filter is designed with a cutoff frequency of approximately 50 MHz. Note that the cutoff frequency increases when the area of the active filter is reduced. The frequency of the input switching signal should be greater than the cutoff frequency of the active filter to not generate high frequency ripple at the output. From simulation results, an 80 MHz input switching signal is sufficiently high to filter out the high frequency harmonics within the input switching signal. An input switching frequency greater than 80 MHz is not preferred since a higher switching frequency would increase the dynamic power dissipation. A ring oscillator supplies a 50 % duty cycle switching signal to the input. Since there is no need for large tapered buffers, the power dissipated by the ring oscillator and output buffers is relatively small. The size of the transistors at the output stage of the Op Amp can be changed for different output voltage or load current demands. The on-chip area of the hybrid voltage regulator therefore depends upon the specific output voltage and load current characteristics. Boost circuitry is not utilized in the regulator at the gate of the NMOS source follower since sufficient margin exists between the input (1.8 V) and output (0.9 V). A charge pump circuit

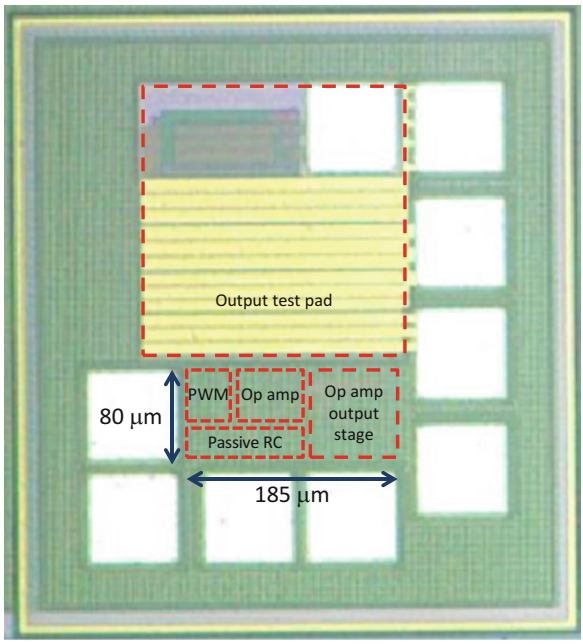


Fig. 17.5 Die microphotograph of the hybrid voltage regulator

can be connected to the gate of the source follower to boost the voltage or, if available, a zero threshold NMOS transistor for the output source follower stage can be used to increase the effective gate voltage.

A 52 % increase in regulator area results in more than a three times increase in the current supplied to the load circuitry or a four times reduction in the load regulation. The on-chip area provides up to 80 mA in less than 0.015 mm^2 ($185 \times 80\,\mu m$), as shown in Fig. 17.5. This on-chip area is significantly less than some recently described LDO regulators [303, 316–318] and SC voltage regulators [371, 372], as listed in Table 17.2.

No capacitor is required at the output node to maintain stability and load regulation, making the hybrid voltage regulator convenient for point-of-load voltage regulation.

Set-up for load transient testing of the voltage regulator with a Teledyne relay is shown in Fig. 17.6 The test board and set-up for the load transient testing is illustrated in Fig. 17.7. A Teledyne GRF303 relay switches the output current of the regulator. The output current is varied between 5 to 70 mA while generating 0.9 V. The experimental results are shown in Fig. 17.8a. A zoomed view of the rise and fall transitions of the output voltage are illustrated, respectively, in Figs. 17.8b and 17.8c. The transition time of the current transients is approximately 70 ns. When the output current demand transitions from 5 to 70 mA and 70 to 5 mA, the output voltage settles in, respectively, 72 and 192 ns. Note that no ringing or overshoot in the output voltage occurs during transient operation, exhibiting highly stable operation of the voltage regulator with abrupt changes in the output current demand.

The hybrid voltage regulator dissipates 0.38 mA quiescent current and delivers up to 80 mA current while generating 0.9 V from a 1.8 input voltage. The current efficiency is over 99 % when the output current demand is greater than 40 mA. When the output current demand changes, a DC voltage shift occurs in the generated voltage, as shown in Fig. 17.9. This DC voltage shift at the output of the regulator is 44 mV when the output current varies between 5 and 70 mA, exhibiting a load regulation of 0.67 mV/ mA. With a 52 % increase in voltage regulator area (i.e., utilizing a larger output buffer), the load regulation can be reduced to $\approx 0.17\,mV/mA$, a fourfold decrease in the DC voltage shift at the regulator output. The amplitude of this output DC voltage shift depends strongly on the current supplied to the load circuitry. When the load current demand increases, the effective voltage across N_5 decreases (see Fig. 17.3). This decrease limits the maximum current that N_5 can supply to the load for a specific output voltage (or limits the output voltage for a specific load current demand). Measurements of the load regulation characteristics of the regulator are illustrated in Fig. 17.10.

A comparison of the performance of the hybrid voltage regulator with other published switching and linear DC-DC converters is listed in Table 17.2. The on-chip area required by the hybrid regulator is significantly less than previously described state-of-the-art buck converters [314, 374], LDO [303, 316–318, 363–366, 375], and SC voltage regulators [371, 372].

A figure of merit (FOM) is described in [318] as

Table 17.2 Performance comparison among different DC-DC converters

	[314]	[373]	[303]	[316]	[317]	[318]	[371]	[372]	This work
Year	2003	1998	2005	2007	2008	2010	2010	2010	2010
Type	Buck	LDO	LDO	LDO	LDO	LDO	SC	SC	Hybrid
Technology [nm]	80	500	90	350	350	90	45	32	110
Response time [ns]	87[a]	150,000	0.054[b]	270	300	3000–5000	120–1200	N/A	72–192
On-chip area [mm^2]	12.6	1	0.098	0.264	0.045[c]	0.019	0.16	0.374	0.015
Output voltage [V]	0.9	2–3.6	0.9	1.8–3.5	1	0.5–1	0.8–1	0.66–1.33	0.9
Input voltage [V]	1.2	5	1.2	2–5.5	1.2	0.75–1.2	N/A	N/A	1.8
Maximum current [mA]	9500	300	100	200	50	100	8	205	80
Maximum current efficiency	N/A	99.8	94	99.8	99.8	99.9	N/A	N/A	99.5
ΔV_{out} [mV]	100	300	90	54	180	114	N/A	N/A	44
Quiescent current [mA]	N/A	10–750	6	0.02–0.34	0.095	0.008	N/A	N/A	0.38
Load regulation [mV/mA]	0.014[a]	0.5	1.8	0.27	0.28	0.1	N/A	N/A	0.67
Transition time [ns]	N/A	N/A	0.1	100	~150	100	N/A	N/A	70
Transition time ratio (K)	N/A	N/A	1	1000	1500	1000	N/A	N/A	700
$\mathrm{FOM}_1 = K\left(\frac{\Delta V_{out} \cdot I_Q}{\Delta I_{out}}\right) \cdot R_t \cdot A$	N/A	N/A	0.029[b]	6.544	6.926[c]	0.893	N/A	N/A	0.518
$\mathrm{FOM}_2 = K\left(\frac{\Delta V_{out} \cdot I_Q}{\Delta I_{out}}\right) \cdot \frac{R_t \cdot A}{T}$	N/A	N/A	3.6[b]	53.4	56.5[c]	110.2	N/A	N/A	42.8

[a] Simulation results (not experimental data)
[b] Mathematical analysis (not experimental data)
[c] An off-chip capacitor of 1 nF to 10 μF is required

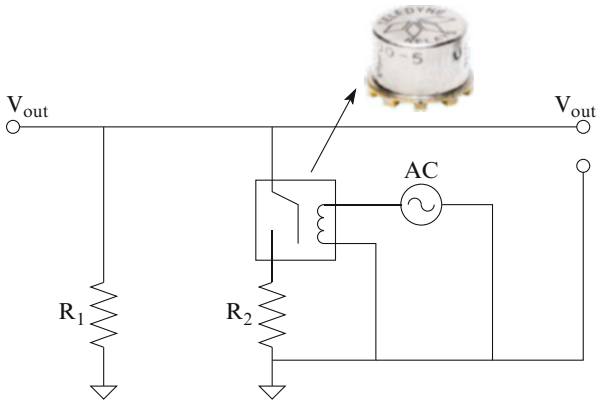


Fig. 17.6 Set-up for load transient testing of the voltage regulator. A Teledyne relay (GRF303 series) is used to switch the output current

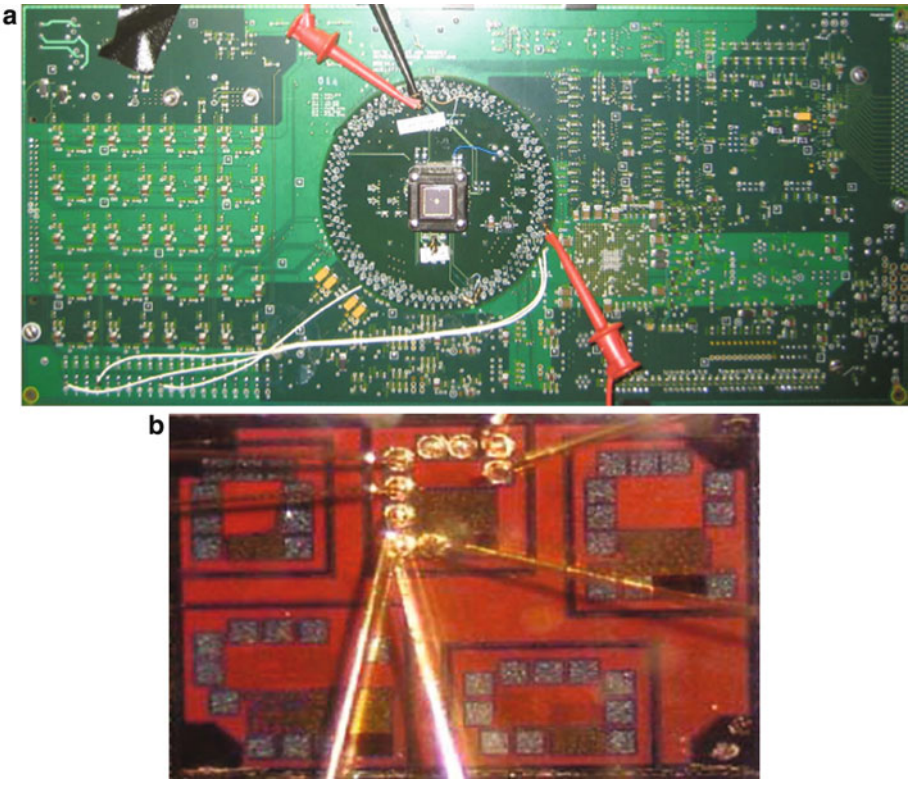


Fig. 17.7 Setup for the test circuit, (**a**) test board, and (**b**) test circuit with wirebonds

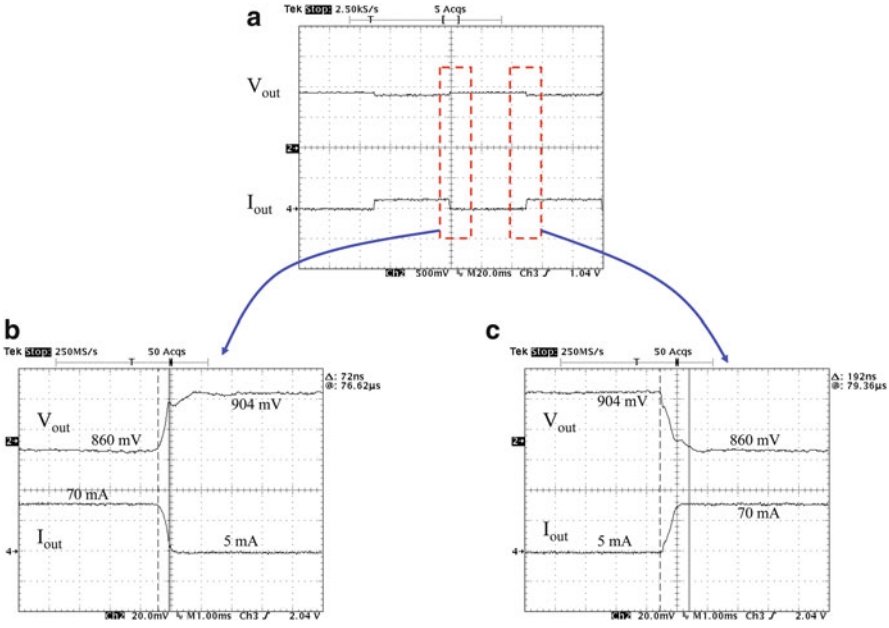


Fig. 17.8 Measured transient response of the active filter based voltage regulator, (**a**) when the output current changes from 5 to 70 mA, and a zoomed view of the transient response when the output current changes from (**b**) 70 to 5 mA and (**c**) 5 to 70 mA. The transition time for the output current is 70 ns

$$\text{FOM}_{guo} = K\left(\frac{\Delta V_{out} \cdot I_Q}{\Delta I_{out}}\right) \qquad (V), \tag{17.4}$$

where K is

$$K = \frac{\Delta t \text{ used in the measurement}}{\text{Smallest } \Delta t \text{ among the compared circuits}}, \tag{17.5}$$

and Δt is the transition time of the load current during test. FOM_{guo} does not however consider the speed of the load regulation which is an important issue in point-of-load voltage regulation.

A second FOM is therefore used that considers the response time and on-chip area of a voltage regulator,

$$\text{FOM}_1 = K\left(\frac{\Delta V_{out} \cdot I_Q}{\Delta I_{out}}\right) \cdot R_t \cdot A \qquad (V\ \mu sec\ mm^2), \tag{17.6}$$

where R_t and A are, respectively, the response time and area of the voltage regulator. Since the required area is technology dependent, the fabrication technology can also be included in the FOM_1, assuming a linear reduction in area with technology.

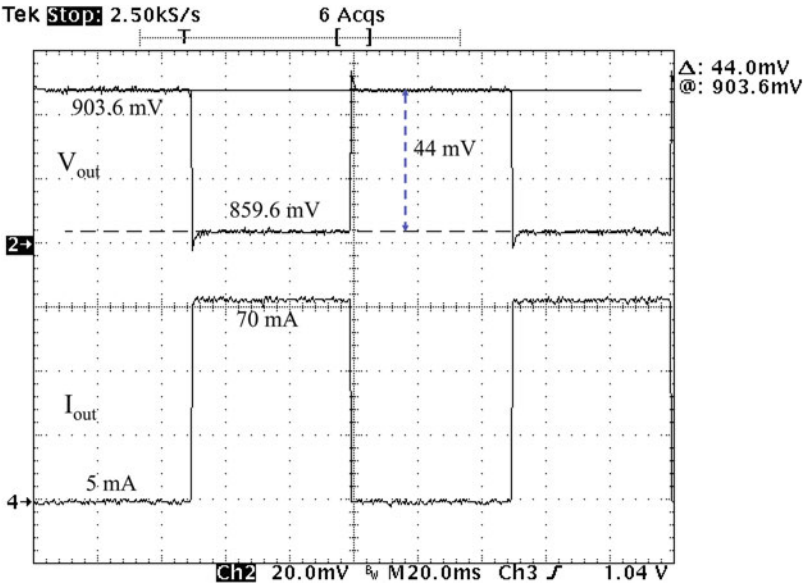


Fig. 17.9 Measured load regulation when the transient output current changes between 5 and 70 mA. The output DC voltage shift is 44 mV. The transition time of the output current is approximately 70 ns

$$\mathrm{FOM}_2 = K\left(\frac{\Delta V_{out}\cdot I_Q}{\Delta I_{out}}\right)\cdot\frac{R_t\cdot A}{T} \qquad (V\ \mu sec), \tag{17.7}$$

where T is the technology node.

A smaller FOM_1 and FOM_2 of a voltage regulator imply a better choice for point-of-load voltage regulation. The regulator described in [303] exhibits the smallest FOMs; however, the response time of [303] is not a measurement result but originates from a mathematical analysis. The voltage regulator presented in this chapter exhibits the smallest FOM among all of the remaining circuits despite the comparably high quiescent current (I_Q). By reducing I_Q, the FOM for the hybrid voltage regulator can be further improved.

The LDO described in [373] has a source follower output stage similar to the hybrid regulator. A large capacitor C_1 and slow control circuitry behaving as a charge pump are connected to the gate of the NMOS transistor in the source follower, as described in [373]. C_1 decouples the gate voltage of the NMOS transistor from the output voltage where voltage variations occur at the source terminal of this transistor. A larger C_1 is therefore needed if the maximum output current demand increases whereas only the size of the output NMOS transistor is increased for the hybrid regulator. To provide additional output current, the area is doubled in [373] as compared to the hybrid regulator.

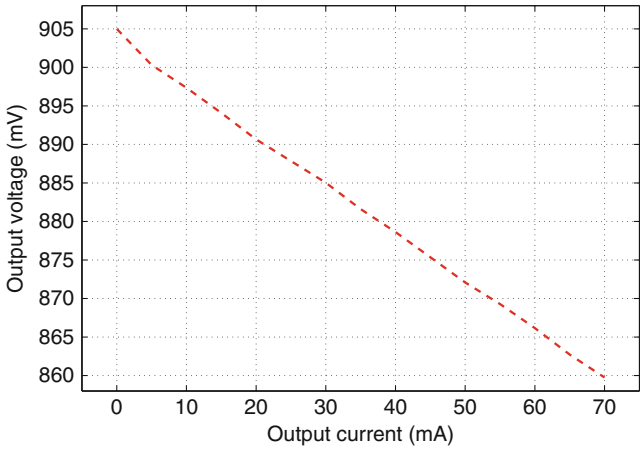


Fig. 17.10 Measured load regulation of the hybrid voltage regulator is approximately 0.67 mV/ mA

The primary disadvantage of the hybrid regulator is that the power efficiency is limited to V_{out}/V_{in} as in a linear voltage regulator. This power loss, however, is somewhat compensated by replacing the large tapered buffers with smaller buffers which drive the active filter. Additionally, the power losses due to the filter inductor and capacitor are eliminated by the active filter structure. The primary advantage of the hybrid regulator is smaller on-chip area. Considering the target application of distributed multi-voltage on-chip power supplies where the local voltage differences are relatively small, this circuit provides a good tradeoff between physical area and power efficiency.

17.4 On-Chip Point-of-Load Voltage Regulation

Optimizing the available resources in a power distribution network has become increasingly more challenging. Multiple distributed supply voltages provide an effective technique to lower the overall power consumed by an integrated circuit [376, 377]. The hybrid voltage regulator is an appropriate power supply for voltage islands operating at different supply voltages and clock frequencies. An active filter-based voltage regulator is a favorable choice for point-of-load voltage regulation due to the small area and flexible drive current to satisfy local current demands. A representative integrated circuit with multiple voltage islands is illustrated in Fig. 17.11. Global power supplies provide the input voltage to the point-of-load voltage regulators. These point-of-load power supplies generate the required voltages within the different voltage islands. The number and size of the voltage regulators depend upon the load current demand and output voltage requirements. The hybrid voltage regulator can also be used to generate a clean supply

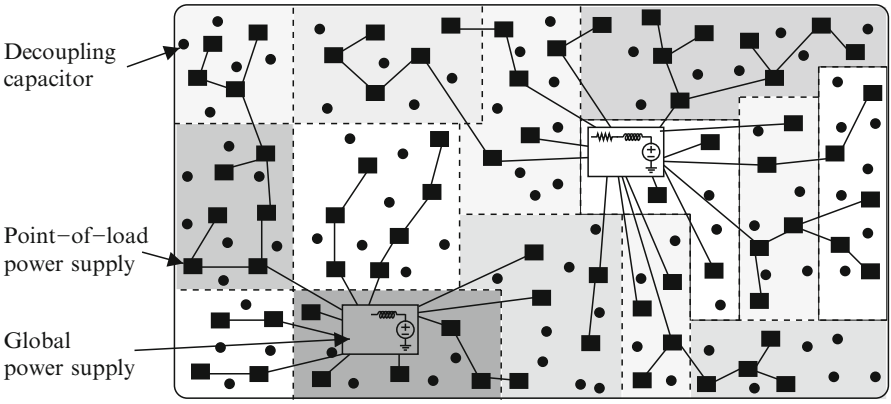


Fig. 17.11 Point-of-load voltage regulators are distributed within different voltage islands to provide a high quality local supply voltage close to the load circuitry

voltage for the noise-sensitive circuit blocks such as the clock generators [378]. In this case, the number and size of the point-of-load voltage regulators also depend upon the load circuitry.

The resistive IR and inductive $L\,dI/dt$ voltage drops are significantly less when the supply voltage is generated close to the load circuitry and reducing the parasitic impedances between the power supply and the load. Additional power savings is also achieved by reducing the supply voltage within the different voltage islands. The disadvantage of the hybrid regulator is the large dropout voltage, reducing the power efficiency. A PMOS output stage, however, can effectively solve this issue without significantly increasing the physical area. In this case, the Op Amp structure should be modified to drive a PMOS output stage.

17.5 Summary

An ultra-small voltage regulator is needed for point-of-load distributed voltage regulation in high performance integrated circuits. An active filter-based hybrid on-chip DC-DC power supply appropriate for point-of-load voltage regulation is described in this chapter.

- The on-chip area of the fully monolithic hybrid voltage regulator is 0.015 mm^2 and provides up to 80 mA output current
- The load regulation is 0.67 mV/ mA and the response time ranges from 72 to 192 ns
- The need for an off-chip capacitor or advanced on-chip compensation techniques to satisfy stability and performance requirements is eliminated in the active filter hybrid regulator

- This circuit provides a means for distributing multiple power supplies close to the load to reduce power/ground noise while enhancing circuit performance by delivering a high quality supply voltage to the load circuitry
- The number and size of the hybrid regulators are based on the application-specific physical area, current demand, and power efficiency requirements
- Less physical area and higher power efficiency is achieved when distributed hybrid voltage regulators are used with lower dropout voltages
- With the hybrid voltage regulator, overall on-chip signal and power integrity is significantly enhanced with the capability of distributing multiple on-chip power regulators

Chapter 18
Distributed Power Delivery with Ultra-Small LDO Regulators

The quality of the power supply in portable electronic systems can be efficiently addressed with POL distributed power delivery [191, 310], which requires the on-chip integration of multiple power supplies. Several low dropout regulators suitable for on-chip integration have recently been fabricated [303, 316, 318, 364–366, 375, 379–386], exhibiting fast load regulation and high current efficiency (i.e., the ratio of the load current and input current). Due to these characteristics, the LDO is a key component in on-chip power management.

To achieve a fast transient response for a load current step of hundreds of milliamperes, the quiescent current of an LDO is typically increased [303, 387], lowering the current efficiency. Dynamically biased shunt feedback, described in [316], has been applied to achieve high current efficiency and system stability over a wide range of load currents. While the impedance-attenuated-buffer in [316] is dynamically biased, the error amplifier in [316] is statically biased, making simultaneous optimization of the LDO speed and current consumption difficult. Alternatively, adaptive biasing techniques have been developed that boost the bias current during fast output transitions [365, 384, 387], yielding a promising technique for fast and power efficient load regulation. The LDO in [365], however, utilizes a 1 μF off-chip capacitor to stabilize the voltage regulation, significantly increasing the response time of the regulation loop. Alternatively, a flipped voltage follower (FVF) LDO compensated by a single Miller capacitor is described in [318] that achieves excellent current efficiency of 99.99 %, good load regulation (0.1 mV/ mA), and moderate regulation speed without an off-chip capacitor [318].

With the increasing number of power domains and high granularity of the on-chip power supply voltages [388], multiple ultra-small voltage regulators will ultimately be integrated on-chip [187, 191]. The physical size of the LDO therefore becomes a primary issue in power management ICs. A nanoscale voltage regulator is expected to exhibit a smaller physical area and improved large- and small-signal characteristics. Alternatively, significant process, voltage, and temperature (PVT) variations pose new stability challenges on the co-design of these ultra-small

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0_18

on-chip voltage regulators. Parallel voltage regulation where multiple regulators are connected to the same power grid has recently attracted significant attention, both from academia [330–334] and industry [335–337]. Satisfying small area, high power efficiency, and stability is however more challenging with parallel voltage regulation. Existing on-chip voltage regulator topologies do not simultaneously overcome these three challenges.

A power delivery and regulation system with six ultra-small 28 nm LDO regulators distributed on-chip is described in this chapter. The distributed power delivery system features an adaptive current boost bias and an adaptive RC compensation network controlled individually within each LDO regulator, increasing the power efficiency and stability of the overall system over a wide range of load currents and PVT variations. As compared to other state-of-the-art LDO regulators providing fast voltage regulation [303, 316, 318, 365], a single LDO within the power delivery system (including all capacitors and a bias generator) is 2.24 times smaller. The power delivery system delivers 3.9–15.8 times more load current, while exhibiting a similar current efficiency. The distributed power delivery system has been tested under a wide range of PVT variations, yielding a stable and fast loop response. Although parallel voltage regulation has previously been demonstrated using eight digital LDO regulators with 77.5 % current efficiency [336], this system is the first successful silicon demonstration of stable parallel analog LDO regulators without off-chip compensation, and exhibits 99.49 % current efficiency.

The rest of the chapter is organized as follows. The power delivery system with six fully integrated LDO regulators with adaptive current boost bias and RC compensation networks is described in Sect. 18.1. Measured performance results are reviewed in Sect. 18.2. The chapter is concluded in Sect. 18.3.

18.1 Power Delivery System

A power delivery system with six fully integrated LDO regulators is described in this section. The system converts any input voltage ranging from 0.9 to 1.1 V into a target output voltage ranging from 0.6 to 0.8 V, supplying up to 788 mA to the load. A model of the power delivery system with six LDO regulators and a distributed power delivery network is shown in Fig. 18.1. The accuracy of the power grid model with n lumped sections, rise time t_r, inductance L, and capacitance C is verified with the $n \geq 5\sqrt{LC}/t_r$ rule of thumb (see [389]), exhibiting less than 2.5 % error in the characteristic impedance for $n = 1$. A single current load is considered to account for the worst case load characteristics with the maximum current step and fastest load transition.

Current sharing is a primary concern in a distributed power delivery system. Each LDO contributes differently to the voltage regulation of a power network based on the position of the active current loads and the level of consumed current. Load sharing among the LDO regulators is illustrated in Fig. 18.2 with a single current load (at the upper right corner of the power network) switching between

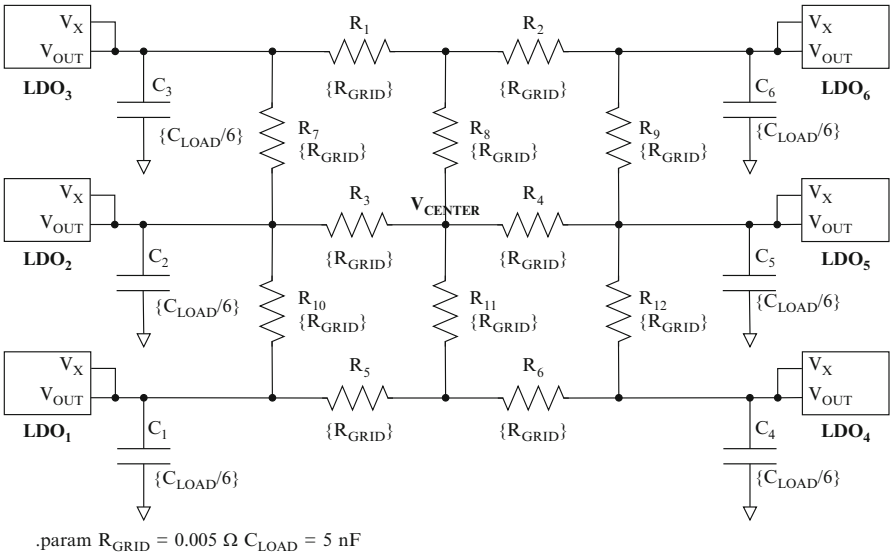


Fig. 18.1 Model of distributed LDO and power delivery system

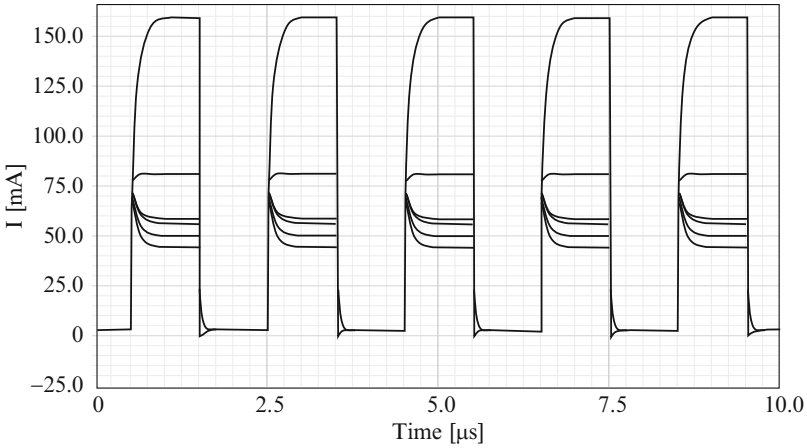


Fig. 18.2 Load sharing in distributed power delivery system

18 and 450 mA. The LDO at the upper right corner (in Fig. 18.1) is located in close proximity with the current load and supplies the largest portion (up to 160 mA) of the total current requirements, which is higher by a factor of 2 than the average current load supplied by a single LDO. Alternatively, the remote LDO at the bottom left corner supplies significantly less current (up to 40 mA), only half of the average LDO load current. In the specific configuration, the LDO in the upper right corner regulates voltage under larger load current steps and exhibits enhanced

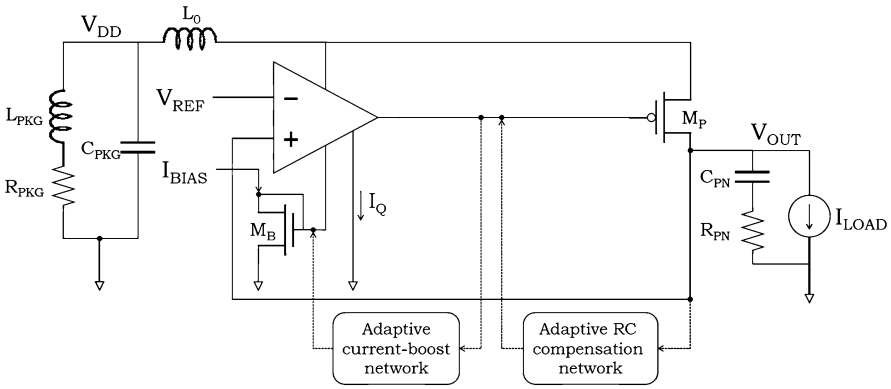


Fig. 18.3 LDO topology

stability characteristics. In modern high performance circuits, the load map may change significantly over time [192] and under PVT variations. Mechanisms are therefore required to co-design the distributed on-chip regulators to dynamically stabilize the power delivery system over time. Adaptive mechanisms are described in this chapter that respond to load variations at the output of each of the LDO regulators, increasing the power efficiency of the system and enhancing overall performance and stability.

An adaptive current boost bias and an adaptive RC compensation network are included within each LDO to, respectively, enhance the slew rate with low power overhead, and stabilize the power regulation over a wide range of load currents and PVT variations. The operation of the dynamic mechanisms is controlled within the individual LDO regulators, providing fine grain regulation of the power voltage. Alternatively, both mechanisms within each LDO are adaptively triggered by the same sensing circuit, exhibiting a more compact power delivery system. The circuit topology of the LDO is shown in Fig. 18.3. The components of the power delivery system are described in the following subsections.

18.1.1 Op Amp Based LDO

The open loop output resistance, load capacitance, and control loop gain and bandwidth are important criteria when developing a fast LDO. To address these challenging transient requirements, a three current mirror OTA topology [383] is used within each LDO, as shown in Fig. 18.4. A linear model that considers the effects of the open loop output resistance, load capacitance, and control loop gain and bandwidth is used to model the behavior of the three current mirror OTA. Miller compensation is used to achieve a dominant pole. The model is shown in Fig. 18.5. The open loop gain of the LDO regulator is

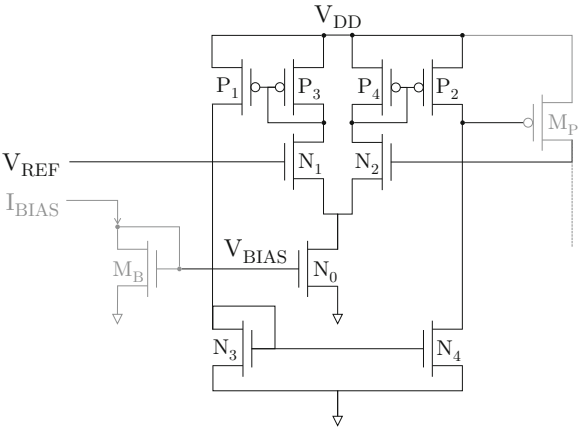


Fig. 18.4 Three current mirror OTA

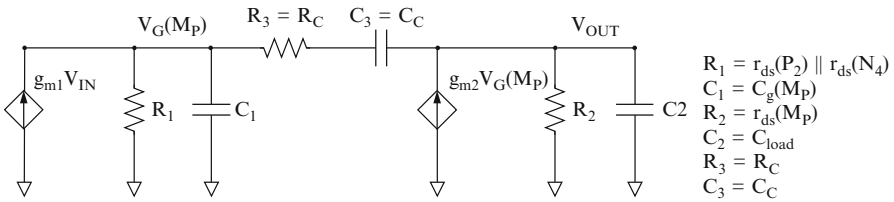


Fig. 18.5 Small signal linear model of LDO

$$A(s) = \frac{V_{OUT}}{V_{IN}} = \frac{-(g_{m1}R_1)\cdot(g_{m2}R_2)(1+N\cdot s)}{1+D_1\cdot s+D_2\cdot s^2+D_3\cdot s^3}, \tag{18.1}$$

where

$$N = (R_3 - \frac{1}{g_{m2}})C_3, \tag{18.2}$$

$$D_1 = R_1C_1 + R_2C_2 + R_3C_3 + R_1C_3 + R_2C_3(1+g_{m2}R_1), \tag{18.3}$$

$$\begin{aligned} D_2 = {} & R_1R_2C_1C_2 + R_1C_1R_3C_3 + R_2C_2R_3C_3 \\ & + R_1R_2(C_1+C_2)C_3, \end{aligned} \tag{18.4}$$

$$D_3 = R_1C_1R_2C_2R_3C_3. \tag{18.5}$$

The zero of the LDO regulator is formed by the compensation network at the frequency $f(z)$,

$$f(z) = \frac{1}{2\pi(R_3 - \frac{1}{g_{m2}})C_3} \approx \frac{1}{2\pi(R_CC_C)}. \tag{18.6}$$

The dominant pole frequency $f(p_1)$ is assumed to be significantly lower than the frequency of the other poles, $f(p_2)$ and $f(p_3)$ ($f(p_1) \ll f(p_2), f(p_3)$). All of the poles are assumed to be real and approximated over a feasible range of g_{mi}, R_i, and C_i components, yielding,

$$f(p_1) \approx \frac{1}{2\pi(g_{m2}R_2)(R_1C_3)} = \frac{1}{2\pi[g_{m2}r_{ds}(M_P)][(r_{ds}(P_2)\|r_{ds}(N_4))C_C]}, \tag{18.7}$$

$$f(p_2) \approx \frac{g_{m2}}{2\pi C_2} = \frac{g_{m2}}{2\pi C_{Load}}, \tag{18.8}$$

$$f(p_3) \approx \frac{1}{2\pi R_3 C_1} = \frac{1}{2\pi R_C C_g(M_P)}. \tag{18.9}$$

The DC gain of the LDO regulator $A_0 = (g_{m1}R_1g_{m2}R_2)$ is listed in Table 18.1, exhibiting an average gain of 57 dB and less than 1 % variations over a wide range of process, temperature, and load variations.

To analyze the stability and compensation of a single LDO regulator, the small signal transconductance and drain source resistance of the output device are assumed to be, respectively, $g_{m2} \propto \sqrt{I_{Load}}$ and $R_2 \propto 1/I_{Load}$. Other parameters are assumed to be approximately independent of the load current in the region of interest. Under these assumptions, the frequency of the first and second poles increases with $\sqrt{I_{Load}}$, while the zero frequency $f(z)$ and third pole frequency $f(p_3)$ are approximately constant under load current variations. The value of R_C is chosen to ensure that the frequency of the third pole $f(p_3)$ is larger than the unity gain frequency in the region of interest, yielding a second order system to enhance stability.

Table 18.1 DC gain over a range of load currents at slow (SS, −30 °C), typical (TT, 25 °C), and fast (FF, 105 °C) corners

Process	Temperature	I_{Load} (mA)	DC gain (dB)	PM (deg)	BW (MHz)
SS	−30 °C	70	58.73	57.05	129.5
		20	60.38	57.65	86.0
		1	61.20	43.35	50.3
TT	25 °C	100	56.73	56.31	136.9
		25	58.38	55.68	98.7
		3	57.20	36.70	86.2
FF	105 °C	150	51.23	61.10	146.7
		100	53.40	60.38	144.6
		70	54.35	57.85	134.4

To increase stability over a wide range of load capacitance, the dominant pole is determined by the compensation capacitor, yielding $f(p_1) < f(p_2)$ and, therefore,

$$C_C > \frac{C_2}{g_{m2}^2 R_2 R_1} > 3\,\text{pF}. \tag{18.10}$$

The maximum phase margin is achieved when the second pole is canceled by the zero, yielding a first order system under the following constraint on the compensation network,

$$R_C = \frac{C_2}{g_{m2} C_C} < 3\,\text{k}\Omega. \tag{18.11}$$

Finally, the first order system exhibits a unity gain at $f_t \approx A_0 \cdot f(p_1) = g_{m1}/2\pi C_C < 130\,\text{MHz}$, fulfilling the requirement $f_t < f(p_3)$ under the constraints, (18.10) and (18.11).

$$R_C < \frac{C_C}{g_{m1} C_1} < 2\,\text{k}\Omega. \tag{18.12}$$

Under the constraints, (18.10), (18.11), and (18.12), the LDO regulator is a first order system with a phase margin between 45° and 90° (the PM is reduced over three decades by 90° due to p_1 and a portion of 45° due to p_3) and a bandwidth $f(p_1)$, as shown in Fig. 18.6.

The transconductance of the output device g_{m2} increases, however, with $\sqrt{I_{Load}}$. Thus, under current load variations, the second pole is shifted from the zero frequency, violating the first order assumption, and degrading the stability of the LDO regulator. The behavior of the PM is, therefore, primarily determined by the variations of the frequency of the second pole, as shown in Fig. 18.7, linearly decreasing with a larger $|\log(f(p_2)/f(z))|$ ratio,

$$\begin{aligned} PM(f(z)) - PM(f(p_2)) &\propto \left|\log\left(\frac{f(p_2)}{f(z)}\right)\right| \\ &= \left|\log\left(\frac{g_{m2} R_C C_C}{C_{Load}}\right)\right|. \end{aligned} \tag{18.13}$$

Note the high accuracy of this linear approximation ($R^2 = 0.9511$).

To maximize the stability of an LDO regulator over a range of load currents, the compensation should be modified with changing transconductance g_{m2}, maintaining $g_{m2} R_C C_C / C_{Load}$ at 1. The phase margin is shown in Fig. 18.8 with two different compensation resistors, $R_C = 0.7\,\text{k}\Omega$ and $R_C = 1.7\,\text{k}\Omega$, and a compensation capacitor of $C_C = 8.5\,\text{pF}$ for a range of low load currents. At low currents of $I_{Load} < 3\,\text{mA}$, compensation with a larger resistor ($R_C = 1.7\,\text{k}\Omega$) results in a higher phase margin. At higher currents of $I_{Load} > 3\,\text{mA}$, a lower compensation resistance

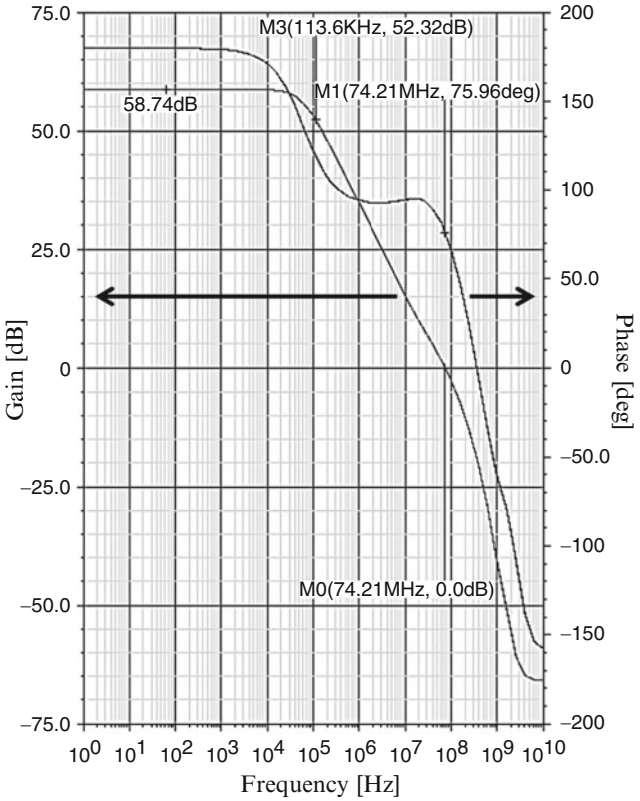


Fig. 18.6 Frequency response of near optimally compensated LDO regulator

($R_C = 0.7\,\text{k}\Omega$) is preferred. Ultimately, the compensation network is adaptively modified with the load current, increasing the phase margin over a wide range of load and PVT variations, as described in Sect. 18.1.4.

The speed of a three current mirror OTA topology, shown in Fig. 18.4, is limited by the bias current that flows into the input differential pair. To produce fast transitions at the load, a higher bias current is preferred. Alternatively, to lower power losses, the OTA should operate under low bias currents. To enhance the loop response while mitigating power dissipation, an adaptive bias is employed in the power delivery system, as described in Sect. 18.1.3.

Distributed power regulators are exploited to regulate the power close to the load, mitigating variations within the power delivery system. The scalability of the power delivery system in terms of the number of distributed LDO regulators is discussed in Sect. 18.1.5.

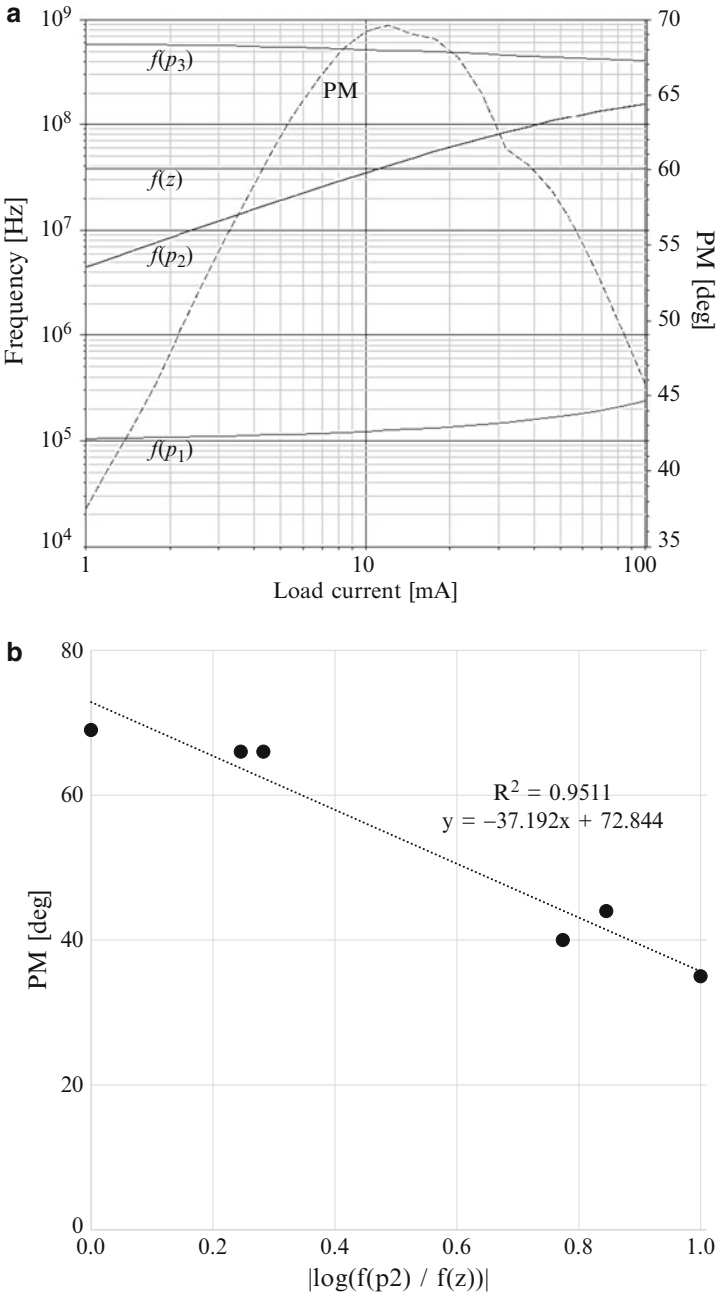


Fig. 18.7 Stability of the LDO regulator as a function of (**a**) load current I_{Load}, and (**b**) compensation accuracy

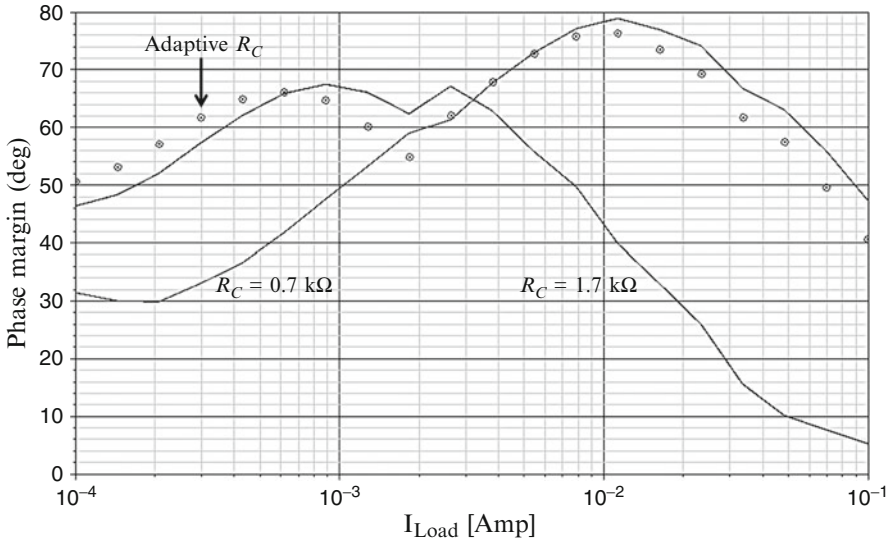


Fig. 18.8 PM with different constant compensation resistors and adaptive compensation

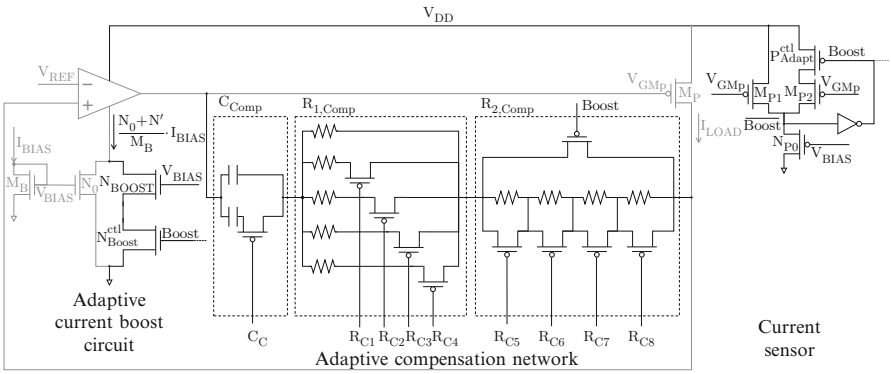


Fig. 18.9 Adaptive bias boost and compensation networks

18.1.2 Current Sensor

A current sensor is used to adaptively control the bias current through the differential pair and RC compensation of the system. The sensor mirrors a portion of the load current, and compares the mirrored current with a threshold current. When the output current is below a threshold current, adaptive mechanisms are activated through the Boost signal, as shown in Fig. 18.9. Hysteresis is employed to prevent the comparator from oscillating at the threshold current. The transient and DC response of the Boost signal is shown in Fig. 18.10. The threshold current is determined

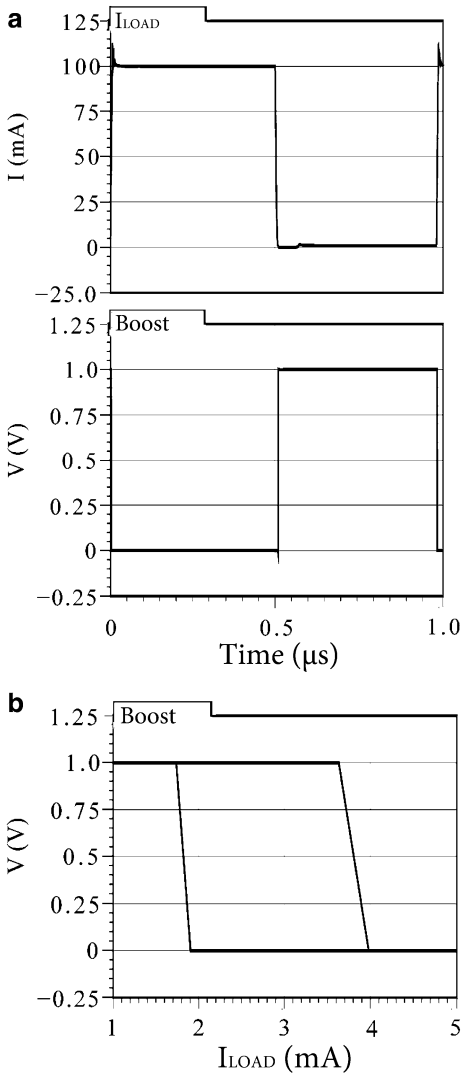


Fig. 18.10 Load current tracking through the Boost signal, (**a**) transient response, and (**b**) DC response

by the width of the transistor N_{P0} (see Fig. 18.9), which is an important design parameter. The system is configured to activate and deactivate the Boost mode at, respectively, I_{LOAD} = 3.8 µA and I_{LOAD} = 1.8 µA (see Fig. 18.10), enhancing the performance (voltage droop, slew rate, and current efficiency) and stability of the system, as described, respectively, in Sects. 18.1.3 and 18.1.4.

18.1.3 Adaptive Bias

A self-adaptive bias current mechanism is described in this section that temporarily boosts the bias current to mitigate fast fluctuations while lowering power losses. The current boost circuit is composed of a sensor block that follows the output voltage at the drain of transistor M_P, and a current boost block that controls the current through the differential pair, as shown in Fig. 18.9. The current boost transistor N_{Boost} is connected in parallel with the bias transistor N_0, and controlled by the Boost line. During the boost mode of operation (the Boost voltage is high), the current into the differential pair is raised, reducing the response time of the LDO. Alternatively, during regular mode (the Boost voltage is low), transistor N_{Boost} is off and no additional current flows into the differential pair, enhancing the power efficiency of the LDO. The Boost line is controlled by the sensor block, and is activated when the load current drops below a threshold current of 3.8 mA, exhibiting faster regulation and a lower voltage droop at the output of the LDO. To prevent oscillations at the threshold current, a hysteretic mechanism is used. The Boost line is therefore deactivated when the load current increases above 1.8 mA, improving the power efficiency of the LDO.

To evaluate the performance of the adaptive biasing technique, the load current is switched in 10 ns from 1 to 70 mA, from 3 to 100 mA, and from 70 to 150 mA at, respectively, the slow, typical, and fast corners. The voltage droop ΔV_{OUT} and quiescent current I_Q are recorded for three different adaptive modes. In the first mode, the current boost mechanism is disabled, limiting the bias current through the differential pair to $(N_0/M_B)I_{BIAS}$. In the second mode, the Boost line is maintained at a high voltage, boosting the current through the differential pair at a constant rate to $((N_0 + N')/M_B)I_{BIAS}$. In the third mode, the Boost line is adaptively boosted in real time by the sensor block, exhibiting a bias current between $(N_0/M_B)I_{BIAS}$ and $((N_0 + N')/M_B)I_{BIAS}$ through the differential pair. Simulation results for each of the modes are shown in Fig. 18.11 for all three corners.

Due to enhanced biasing, the voltage droop is decreased by 35 % (from 43 to 25 mV) at the expense of a significant 137 % increase in current consumption. The bias current is adaptively enhanced under light loads, exhibiting an 18.6 % decrease in voltage droop while avoiding excessive power loss over time.

The power delivery system is designed for modern high performance circuits that draw significant leakage current from the power regulators. Minimum load currents of 1 , 3 , and 70 mA are assumed for a single LDO regulator for, respectively, the slow, typical, and fast corners. Quiescent current simulations for different load currents are listed in Table 18.2 with and without adaptive biasing. Without adaptive biasing, an average quiescent current of 423 µA with less than 2 % variations is demonstrated at 25 °C for all load currents. This current is increased to 1 mA by adaptive biasing at light loads of less than 1 , 1.5 , and 2 mA at, respectively, the slow, typical, and fast corners.

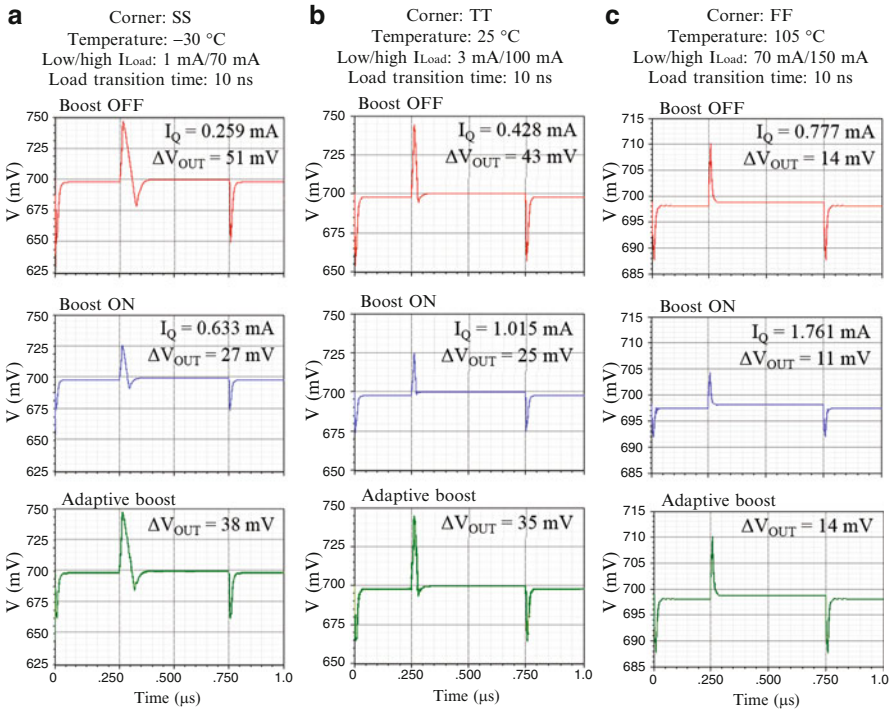


Fig. 18.11 Voltage droop and quiescent current with and without adaptive biasing at (**a**) slow corner, (**b**) typical corner, and (**c**) fast corner

Table 18.2 Quiescent current with and without adaptive biasing

I_{Load} (mA)	−30 °C		25 °C		125 °C	
	Adaptive I_Q (mA)	Constant I_Q (mA)	Adaptive I_Q (mA)	Constant I_Q (mA)	Adaptive I_Q (mA)	Constant I_Q (mA)
0.5	0.841	0.347	1.001	0.416	1.223	0.524
1.0	0.844	0.35	1.005	0.419	1.244	0.529
1.5	0.359	0.359	1.008	0.421	1.253	0.532
2.0	0.353	0.353	0.433	0.433	1.258	0.538
2.5	0.352	0.352	0.424	0.424	0.554	0.554
5.0	0.353	0.353	0.424	0.424	0.543	0.543
100	0.353	0.353	0.424	0.424	0.543	0.543

18.1.4 Adaptive Compensation Network

The large gate capacitance of the pass transistor together with the wide range of possible values of C_{Load} produce a complicated transfer system of poles and zeros. The low frequency non-dominant poles within the unity gain frequency of the

feedback loop create a negative phase shift, degrading the stability of the overall system. To compensate for the negative phase shift, the Miller compensation technique [318] is used. The wide range in load capacitance and currents, and significant PVT variations, however, make compensation with fixed RC values impractical in nanoscale technologies. A digitally configurable compensation network is therefore used that adaptively modifies the dominant pole, maintaining system stability for all values of C_{Load} and I_{Load}. Due to complex interactions among the parallel connected LDO regulators, shared power grid, and current loads, this compensation network is necessary to maintain stability in distributed power delivery systems. Conventional LDO regulators without the compensation network can easily become unstable from device mismatch, offset voltage, and varying load currents when connected in parallel [332, 334, 336, 390–392]. The compensation network is comprised of a capacitive block connected in series with two resistive blocks, as shown in Fig. 18.9. The capacitive (C_{Comp}) and resistive ($R_{1,Comp}$ and $R_{2,Comp}$) blocks are digitally controlled by, respectively, control signals C_C and RC_i, $i = 1, \ldots, 8$. These RC impedances are digitally configured within each of the LDO regulators after fabrication to compensate for process variations. The second resistive block is also controlled by the Boost signal, which is adaptively activated (bypassed) by the current load sensor within each LDO regulator when the individual Boost signal is high (low), compensating for voltage, temperature, and current load variations in run time. During the high-to-low current load transition, the output impedance increases. Thus, the pole introduced by the load is pushed to a lower frequency, degrading the stability of the LDO. Alternatively, the Boost signal is activated during this transition, increasing the compensation impedance, $(R_{1,Comp} + R_{2,Comp}) \cdot C_{Comp}$, to maintain a stable response. At other times, the Boost signal is deactivated, and the LDO is stabilized with $R_{1,Comp} \cdot C_{Comp}$.

To illustrate the effect of compensation on the LDO performance, the phase margin of a single LDO is presented in Fig. 18.12 over a range of C_{Load} values for two load currents, I_{Load} = 1 mA and I_{Load} = 10 mA. For a light load current of 1 mA, the phase margin increases with higher compensation resistance PM(R_C = 1.7 kΩ) > PM(R_C = 0.7 kΩ). Alternatively, for a higher load current of 10 mA, a smaller compensation resistor is preferable. The adaptive compensation illustrated in Fig. 18.12 exhibits a higher phase margin as compared with non-adjustable compensation. The same behavior can be observed in Fig. 18.8, where the compensation network is adaptively reconfigured as a function of the load current, yielding the largest PM as compared to non-adjustable compensation networks.

18.1.5 Distributed Power Delivery

A model of the distributed power delivery system with k LDO regulators is shown in Fig. 18.13. The LDO output devices are connected in parallel at the output node loaded by $k \cdot C_{Load}$, and are driven by the total current from the individual error amplifiers. With equally shared load current $k \cdot I_{Load}$, all of the distributed

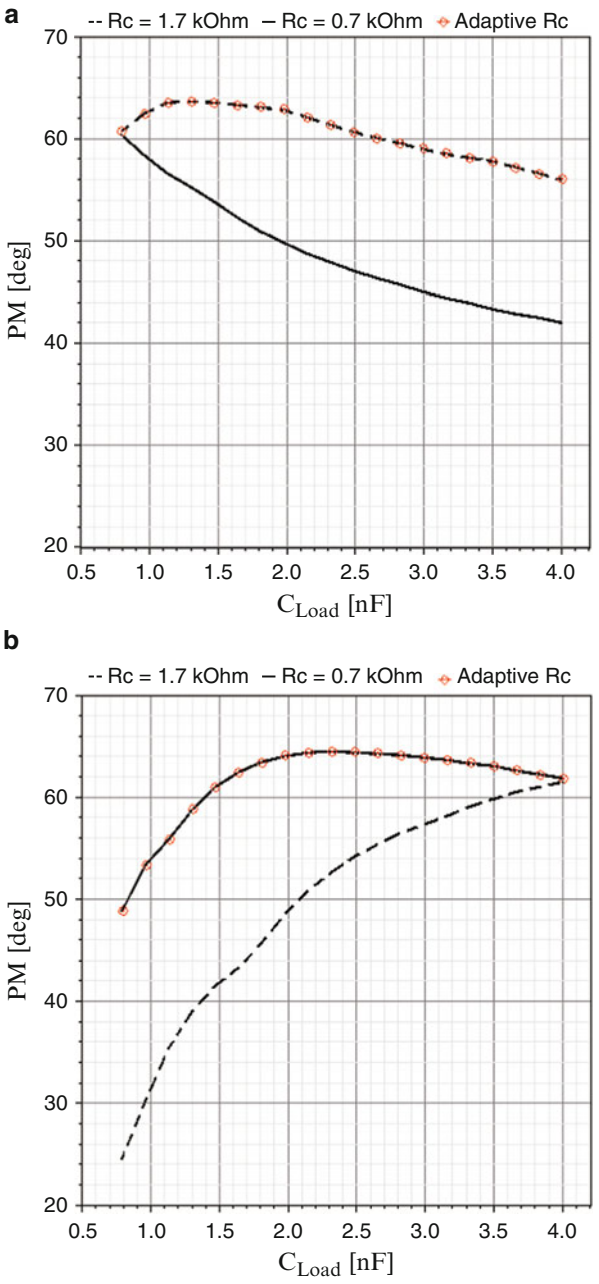


Fig. 18.12 Phase margin with different compensation and load capacitance for (**a**) I_{Load} = 1 mA, and (**b**) I_{Load} = 10 mA

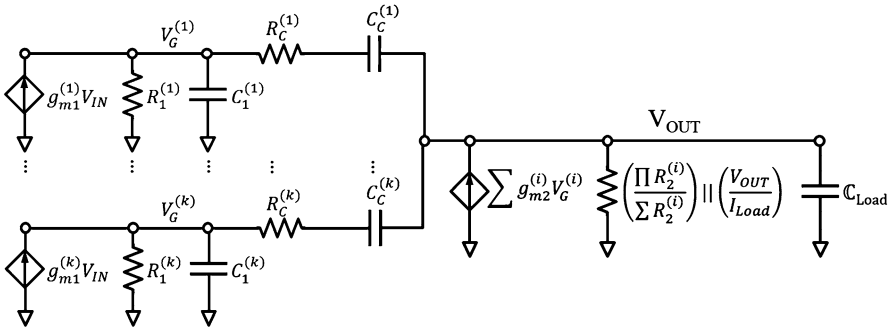


Fig. 18.13 Small signal linear model of distributed power delivery system with k LDO regulators

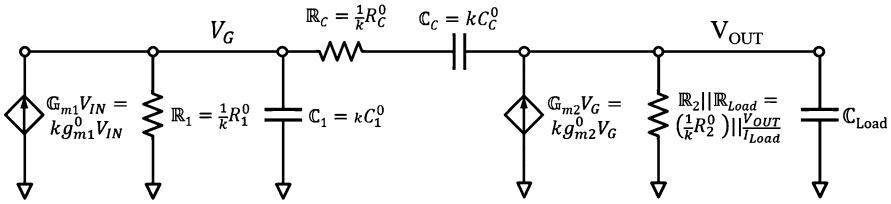


Fig. 18.14 Small signal linear model of distributed power delivery system with k LDO regulators and equally shared load current

LDO regulators exhibit similar behavior, yielding the simplified model shown in Fig. 18.14 with $g_{m1,2}^{(i)} = g_{m1,2}^0$, $R_{1,2,3}^{(i)} = R_{1,2,3}^0$, and $C_{1,2,3}^{(i)} = C_{1,2,3}^0$, $\forall i = 1, \ldots, k$. The phase margin of a distributed power delivery system with k LDO regulators and equally shared load is determined from (18.13) by

$$PM(f(z)) - PM(f(p_2)) \propto \left| \log\left(\frac{\mathbb{G}_{m2}\mathbb{R}_C\mathbb{C}_C}{\mathbb{C}_{Load}}\right)\right| = \left| \log\left(\frac{g_{m2}^0 R_C^0 C_C^0}{C_{Load}^0}\right)\right|, \tag{18.14}$$

exhibiting similar behavior to a power delivery system with a single LDO regulator. Note that increasing a high load current by a factor of k in a single LDO system with load capacitance C_{Load} lowers the phase margin of the system by $\log\left(\sqrt{k}\right) = (1/2)\log(k)$. Alternatively, the same increase in load current in a distributed power delivery system with k LDO regulators and similar load capacitance C_{Load} lowers the phase margin by $\log(k)$. The stability over a wide range of load currents is, therefore, more challenging with parallel load regulation. In addition, the stability of a distributed power delivery system is limited by the lowest PM among all of the LDO regulators, exhibiting a strong dependence on the load current sharing. Under load current variations, the load at a single LDO regulator can be

n times lower/higher than the average load current, decreasing/increasing the output transconductance g_{m2} by a factor of $\sqrt{n}$. The worst case stability of a distributed power delivery system under load current variations is, therefore,

$$PM(f(z)) - PM(f(p_2)) \propto \left|\log\left(\frac{g_{m2}^0 R_C^0 C_C^0}{C_{Load}^0}\right)\right| + \frac{1}{2}\log(n). \tag{18.15}$$

Based on typical load current variations, as shown in Fig. 18.2 and the linear curve fitting in Fig. 18.7b, the distributed power delivery system exhibits current sharing variations of up to $n = 2$, yielding a $37.2\log\sqrt{2} = 5.6°$ degradation in phase margin. To address the worst case current sharing variations and a wide range of PVT variations, each LDO regulator is compensated around $I_{Load} = 10\,\text{mA}$ with $R_C C_C = 700\,\Omega \cdot 6\,\text{pF}$ to provide a stable response with $40° < PM < 70°$ for high load currents of $10\,\text{mA} < I_{Load} < 150\,\text{mA}$. Alternatively, at low load currents, sensed by the current sensor (see Fig. 18.9), the compensation is adaptively increased, enhancing the stability of the system. Due to the distributive nature of the power delivery system, adaptive compensation and bias are activated individually within each LDO regulator based on the specific locally sensed load currents, providing fine grain control over the local adaptive mechanisms. The same load sensing circuit within each LDO regulator triggers both the adaptive compensation and bias mechanisms, producing a more compact power delivery system.

18.2 Test Results

The power delivery system with six LDO regulators has been fabricated in an advanced 28 nm CMOS technology. The on-chip regulators simultaneously drive a power network, delivering power to the on-chip integrated circuits within a commercial mobile device. All of the measurements are performed on LDO regulators within the distributed power delivery system.

Modern integrated circuits exhibit aggressive transient characteristics and are expected to tolerate significant PVT variations. To illustrate the tolerance of voltage and temperature variations in the power delivery system, the load current of the distributed power delivery system with six LDO regulators is stepped from 52 to 441 mA in 10 ns, drawing an average high (low) current of 73.5 mA (8.67 mA) from each LDO regulator. Due to load sharing variations (see Fig. 18.2), the load current of a single LDO regulator can be increased or decreased by a factor of 2 (and more under PVT variations) as compared to the nominal current, exhibiting currents of up to 147 mA and down to 4.3 mA. The magnitude of these extremely fast load changes is therefore limited under voltage and temperature variations as compared to full range operation. The measured transient response for nominal input and output voltages of, respectively, 1.0 and 0.7 V is illustrated in Fig. 18.15 for -25, 25, and 125 °C. To evaluate the system under line variations, the output is tested at 25 °C under ±10 % input voltage variations. The measured transient

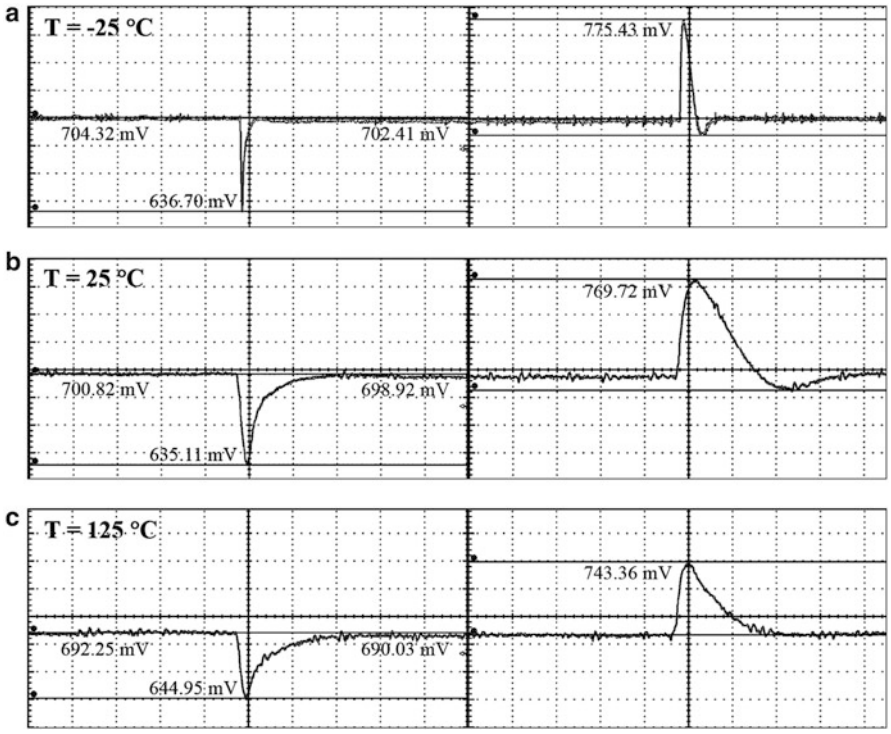


Fig. 18.15 Transient step response at the load for V_{IN} = 1 V, V_{OUT} = 0.7 V, and a load current step from 52 to 441 mA in 10 ns, measured at (**a**) $T = -25$ °C, (**b**) $T = 25$ °C, and (**c**) $T = 125$ °C

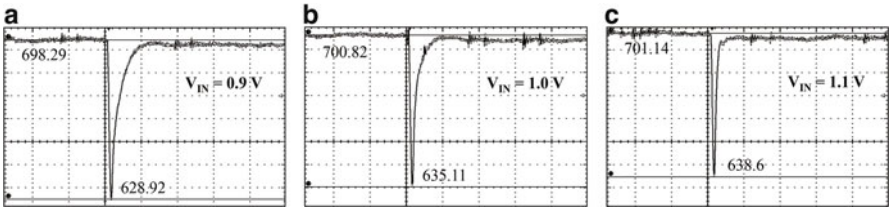


Fig. 18.16 Transient step response at the load at a temperature of 25 °C for V_{OUT} = 0.7 V and a load current step from 52 to 441 mA in 10 ns, measured for (**a**) V_{IN} = 0.9 V, (**b**) V_{IN} = 1.0 V, and (**c**) V_{IN} = 1.1 V

response is illustrated in Fig. 18.16. For both types of variations, the distributed power delivery system exhibits a stable response over a wide range of temperatures with less than 10 % voltage droop.

To demonstrate the stability of the distributed power delivery system under maximum load currents and extremely fast load transitions, the load current of the system is stepped from 52 to 788 mA in 5 ns at 25 °C. The measured transient response for nominal input and output voltages of, respectively, 1.0 and 0.7 V is

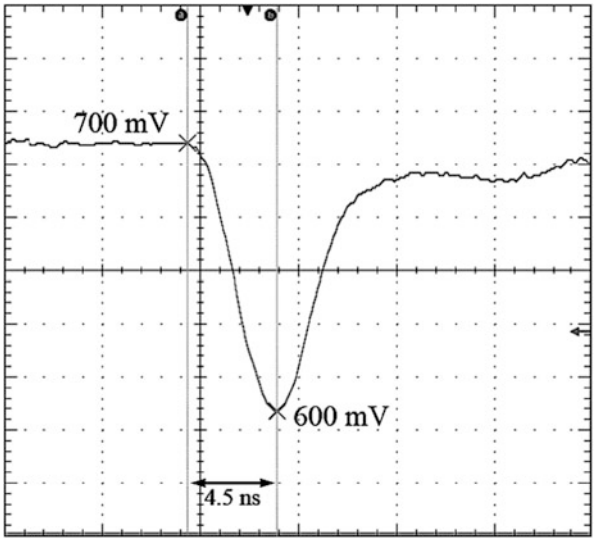


Fig. 18.17 Measured transient response for a load current step from 52 to 788 mA in 5 ns

illustrated in Fig. 18.17, exhibiting a stable response and voltage droop of 0.1 V. The system response time based on the equivalent parasitic capacitance of the load circuit (C_{Load} = 472 pF), maximum load current ($I_{Load,MAX}$ = 788 mA), and voltage droop (ΔV_{OUT} = 100 mV) is evaluated as $T_R = C_{TOT} \cdot \Delta V_{OUT} / I_{Load,MAX} = 0.064$ ns.

The system of parallel LDO regulators yields the shortest transient response time as compared with existing LDO regulators [303, 316, 318, 365]. The voltage droop is a strong function of the magnitude and transition time of the load step. Only the magnitude of the load current is, however, typically considered in T_R. For a fair comparison, T_R is normalized to K, the ratio between the load transition time of the LDO regulator Δt and a 1 ns transition time ($K = \Delta t$ / 1 ns). In addition, the response time of the LDO regulator is normalized to an estimated fan-out of four FO4 delay (T_G), canceling the advantages of technology scaling. The LDO regulators [303, 316, 318, 365] and the power delivery system are compared based on the normalized, technology independent loop response $(T_R)_{Norm} = T_R \times ({}^{\Delta t}/_{1\,\mathrm{ns}})/T_G$. The system exhibits a smaller $(T_R)_{Norm}$ than the LDO regulators described in [316, 318, 365], yielding a faster response time to a load transition, while exhibiting a similar current efficiency (99.49 % vs. 99.99 % in [318]). The loop response time in these regulators is increased by the size of the off-chip capacitor (1 μF in [316, 365]) or by a small bias current [318]. Alternatively, the speedup in the loop response achieved in [303] requires a significant increase in bias current, degrading the power efficiency of the LDO regulator (94 %). The response time, power efficiency, and other primary parameters of the power delivery system, compared to fully integrated on-chip regulators [303, 318], are listed in Table 18.3.

The power delivery system converts an input voltage between 0.9 and 1.1 V into any required output voltage between 0.6 and 0.8 V, while exhibiting a stable response and less than 69.4 mV voltage droop at 25 °C for all of the input and

Table 18.3 Performance summary and comparison with previously published LDO regulators

Parameters	Unit	[303] Hazucha	[318] Guo	This work
Technology	µm	0.09	0.09	0.028
Active area	mm^2	0.008	0.019	0.00357
Input voltage	volt	1.2	0.75–1.2	0.9–1.1
Output voltage	volt	0.9	0.5–1	0.6–0.8
Minimum dropout voltage	volt	0.3	0.2	0.1
Maximum load $I_{Load,MAX}$	mA	100	100	788
Load regulation	mV/mA	1	0.1	0.023–0.027
On-chip capacitance	pF	600	7	5.91–8.37
Load circuit capacitance	pF	0	50[a]	472[a]
Quiescent current I_Q	µA	6000	8	4000
Voltage droop ΔV_{OUT}	mV	90	200	100
Output transition time T_R	ns	0.54	0.114	0.064
Normalized load transition $^{\Delta t}/_{1ns}$	ns/ns	0.1	100	5
FO4 delay T_G	ns	45	45	14
Normalized response time $(T_R)_{Norm}$	ns/ns	1.2	253	23
Current efficiency	%	94.3	99.99	99.49

[a]Estimated based on equivalent parasitic capacitance of the load circuitry

Table 18.4 Voltage droop for different input and output voltage levels

V_{IN} \ V_{OUT}	0.6 V	0.7 V	0.8 V
0.9 V	64.6 mV	69.4 mV	68.6 mV
1.0 V	61.9 mV	65.9 mV	67.9 mV
1.1 V	60.4 mV	62.5 mV	67.9 mV

Table 18.5 Measured quiescent current and current efficiency

Parameter	Comment	Temperature		
		−30 °C	25 °C	105 °C
I_Q (mA)	Distributed system	3.0	4.0	7.0
	Single LDO (average)	0.5	0.8	1.17
Efficiency (%)	$I_{Load,MAX}$ = 788 mA	99.62	99.49	99.11

output voltages within the range and a load step from 52 to 441 mA in 10 ns. The voltage droop for input and output voltages of, respectively, 0.9–1.1 V and 0.6–0.8 V is listed in Table 18.4. Note the measured transient step response at the output of a single LDO regulator for a distributed power delivery system with the input and output voltage of, respectively, 0.9 and 0.8 V, exhibiting a voltage dropout of 0.1 V and a 68.6 voltage droop at the output.

The quiescent current of the power delivery system of six distributed LDO regulators is listed in Table 18.5, yielding up to 99.49 % current efficiency at 25 °C. A die microphotograph of the LDO is shown in Fig. 18.18. The area occupied by

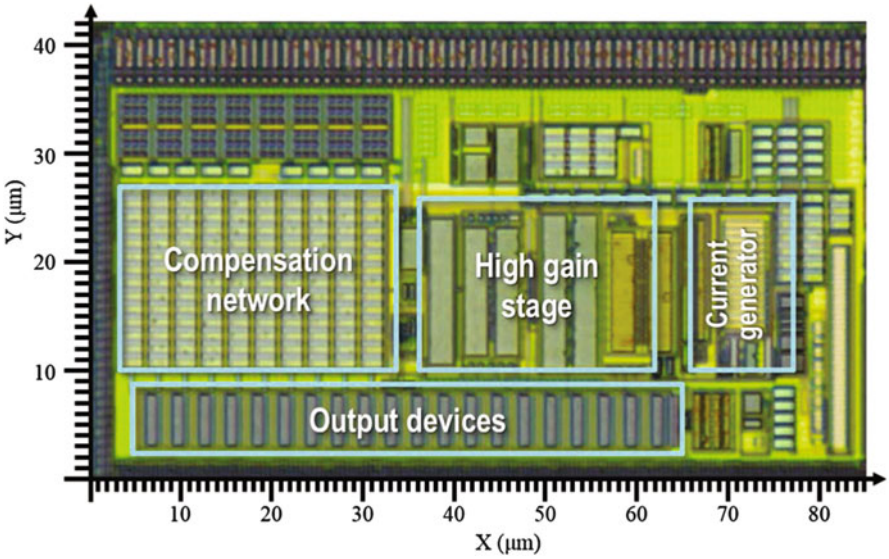


Fig. 18.18 Die microphotograph of 28 nm ultra-small LDO

the LDO with all capacitors is $85 \times 42\,\mu$m, significantly smaller than the LDO regulators described in [303, 316, 318, 365].

18.3 Summary

A distributed power delivery system with six ultra-small fully integrated low dropout regulators is described in this chapter.

- The system is fabricated in a 28 nm CMOS process and exhibits a fast transient response with excellent load regulation under PVT and current sharing variations
- An adaptive bias technique is used to enhance the transient performance and increase the power efficiency by, respectively, boosting and decreasing the bias current
- A voltage droop of less than 10 % and a current efficiency of 99.49 % are measured
- An adaptive compensation network is employed within the power delivery system that supports the co-design of a system of distributed parallel LDO regulators
- A stable system response is measured within −25 to 105 °C and 10 % voltage variations
- Each of the LDO regulators within the adaptive networks and bias current generator occupies $85 \times 42\,\mu\text{m} = 0.003\,57\ \text{mm}^2$; no off-chip capacitors are required

Chapter 19
Pulse Width Modulator for On-Chip Power Management

Voltage controlled oscillators (VCOs) are widely used to generate a switching signal where certain characteristics of this signal can be controlled. These controllable switching signals can be efficiently exploited in different switching and linear regulators to adaptively scale voltages on-chip, providing an important circuit level means for DVS systems. Two types of VCOs, LC oscillators and ring oscillators, are primarily used in high performance integrated circuits. LC oscillators can operate at high frequencies and exhibit superior noise performance. Alternatively, ring oscillators occupy significantly smaller on-chip area with a wider tuning range. Due to these advantages, ring oscillators have found widespread use in modern ICs [310, 393–397].

A conventional ring oscillator consists of an odd number of inverters where the output of the last inverter is fed back to the input of the first inverter, as shown in Fig. 19.1. The delay provided by each inverter in this chain produces a phase shift in the switching signal. The sum of these individual delays (i.e., phase shifts) and the feedback from the last to the first inverter produces a total phase shift of π that causes the circuit to oscillate. The frequency of this oscillation depends upon the sum of the inverter delays within the chain [398].

The duty cycle of the generated switching signal is typically 50 % for conventional ring oscillators where the PMOS and NMOS transistors within the inverters provide the same rise and fall transition times. The duty cycle of a ring oscillator can be tuned by controlling the transition time of the inverters within the ring oscillator. Header and footer circuits are widely used to control the current supplied to the PMOS and NMOS transistors within the ring oscillator inverter chain [399]. Although the header and footer circuits are typically used to control the frequency, these circuits can also control the duty cycle of a ring oscillator.

In this chapter, a digitally controlled pulse width modulator (PWM) comprised of a header circuit, ring oscillator, and duty cycle to voltage (DC2V) converter is described [397, 400]. The duty cycle of the PWM is determined from the closed-form expressions, yielding a simple dependence on the header current. The high

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0_19

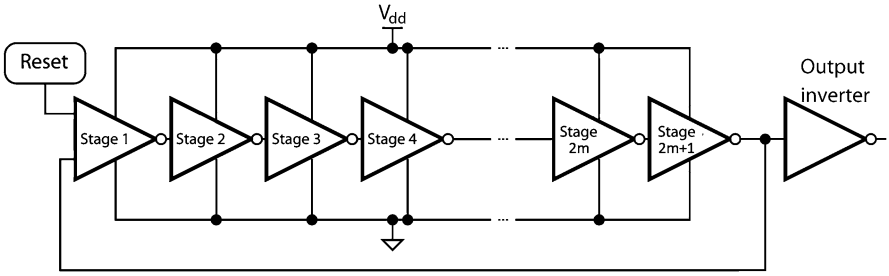


Fig. 19.1 Conventional ring oscillator. Note that an odd number of inverters is required for the system to oscillate

accuracy of these expressions is confirmed by simulation results. The header circuit controls the current delivered to the PMOS transistors within the ring oscillator. Contrary to conventional header circuits, where the header is connected to each of the inverters within the ring oscillator chain, the header circuit is connected to every other inverter stage to dynamically control the pulse width of the output signal. This header circuit provides high granularity control of the duty cycle with a step size of 2 % of the period. An additional header circuit regulates the supply current delivered to the remaining inverter stages, providing improved control while maintaining a constant switching frequency. Additionally, a DC2V converter, based on the frequency to voltage converter described in [401], maintains the accuracy of the PWM under process, voltage, and temperature variations. Under PVT variations, the maximum change in duty cycle is less than 2.7 % of the period.

Owing to the small on-chip area, fast control circuitry, high accuracy under PVT variations, and dynamic duty cycle and frequency control governed by accurate closed-form expressions, the PWM is an effective circuit to dynamically change the duty cycle of the input switching signal for on-chip voltage regulators. This circuit enables high granularity DVS at run time and reduces the response time from milliseconds to nanoseconds.

The remaining part of the chapter is organized as follows. The PWM architecture is described in Sect. 19.1, where the working principle of the header circuitry and DC2V converter is explained, and the analytic expressions for the PWM timing parameters are provided. In Sect. 19.2, the functionality and accuracy of the digitally controlled PWM under PVT variations are validated with predictive technology models at the 22 nm technology node. Some concluding remarks are offered in Sect. 19.3.

19.1 Description of the Digitally Controlled PWM Architecture

A schematic of the PWM is shown in Fig. 19.2. A header circuit is connected to the ring oscillator to current starve every other stage in the ring oscillator chain. Digital control circuitry provides multiple control signals (C_d) to dynamically change the

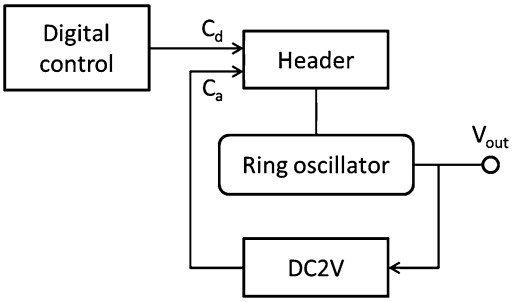


Fig. 19.2 Digitally controlled PWM. The header circuitry has two input control signals, digital control (C_d) and analog control (C_a). C_d is used to dynamically change the individual transistors to provide a high granularity control of the duty cycle whereas C_a maintains a constant current from the header to the ring oscillator under PVT variations

duty cycle, and a DC2V converter ensures the accuracy of the duty cycle under PVT variations by providing an analog signal to the header circuit. The working principles of these circuits are explained in the following subsections.

19.1.1 Header Circuitry

An addition based current source, as shown in Fig. 19.3, is described in [402]. This circuit is used as a header in [403] to compensate for temperature and process variations by maintaining a constant current to the ring oscillator. Note that this header circuit has one input voltage that controls the gate voltage of M_1 and M_2. Alternatively, the gate voltage of M_3 is controlled by the drain of M_2. The current flowing through M_1 (and M_2) is therefore inversely proportional to the current through M_3. For example, if the current passing through M_1 (and M_2) increases, the gate voltage of M_3 also increases, decreasing the current passing through M_3. The sum of these inversely proportional currents is the input current to the ring oscillator, which is approximately constant over a wide range of temperature and process variations [403].

Alternatively, a modified version of this header circuit, as depicted in Fig. 19.4, is described to control the duty cycle by changing the transition time of the PMOS transistors at every other inverter stage within the ring oscillator. Gates M_1 and M_2 are controlled by the analog signal C_a. As opposed to a single transistor M_3 whose gate is connected to a resistor, as shown in Fig. 19.3, multiple parallel PMOS transistors $M_3[i]$, $i = 0, \ldots, n$ are added in place of M_3 in the header circuit. The PMOS transistors are designed with increasing device size to provide both increased dynamic range and dynamic control of the duty cycle with 2 % increments. All of these transistors have the same gate-to-source voltage, but the voltage at the drain terminals is controlled by other switch transistors. Additional PMOS transistors

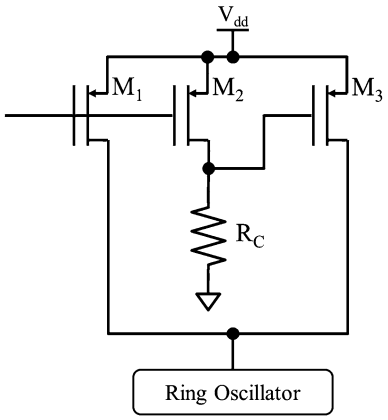


Fig. 19.3 Addition based current source used as a header circuit [403]

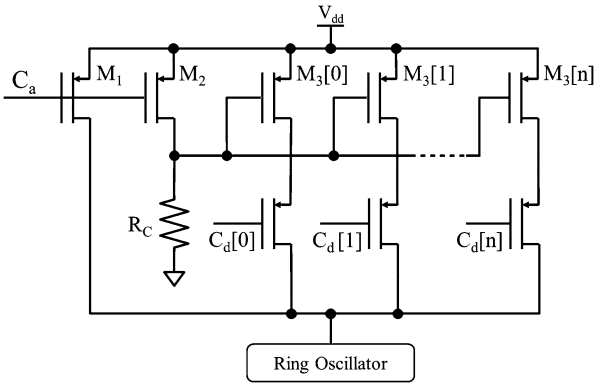


Fig. 19.4 Parallel PMOS transistors replace M_3 to improve the granularity of the current control as well as behave as switch transistors to turn on different sections of the header circuitry

are connected in series behaving as switch transistors. The gate voltage of these switch transistors is controlled by a digital controller that turns on (and off) the individual header stages through control signals $C_d[i]$, $i = 0, \ldots, n$. Turning on all of the header stages produces the maximum current to the ring oscillator which in turn minimizes the duty cycle. Any variations in the leakage current are more prominent when the device size is small. To mitigate these variations, larger than minimum size transistors are used in sub 65 nm technology nodes [404]. The first two transistors in the header circuit (M_1 and M_2) are therefore comparably large to minimize any mismatches. A minimum channel length of 150 nm is used for these two input transistors as opposed to 40 nm for the remaining transistors.

19.1.2 Duty Cycle-to-Voltage Converter

The frequency-to-voltage converter described in [401] operates as a DC2V converter. A circuit schematic of this DC2V converter is shown in Fig. 19.5. There are

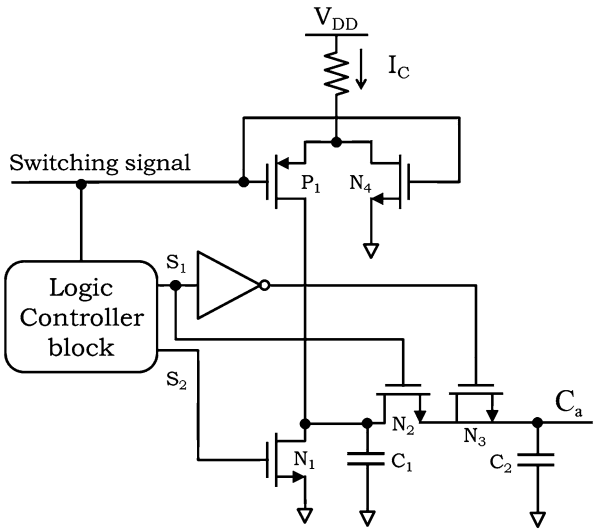


Fig. 19.5 Frequency-to-voltage converter described in [401] operating as a duty cycle-to-voltage converter

primarily three different operational phases of this circuit. During the first phase, capacitor C_1 is charged through transistor P_1. In the second phase, transistors (i.e., switches) N_2 and N_3 are turned on to allow charge sharing between C_1 and C_2. During the last phase, C_1 is discharged through N_1. The charge time of C_1 depends upon the duty cycle of the input switching signal. A signal with a greater duty cycle causes more charge to accumulate on C_1, increasing the output voltage of the DC2V converter. The DC2V converter controls the bias current from the header circuitry through negative feedback, mitigating PVT variations. Intuitively, when the header current is reduced, the duty cycle of the ring oscillator is greater, increasing the output voltage of the DC2V converter. As a result, the voltage at the gate of the M_2 transistor increases and the current I_C through the resistor decreases, pulling down the gate voltage of the active header stages. Thus, the current flow through the header to the ring oscillator is increased, compensating for the initial reduction in current. A more complete explanation of the working principles of this circuit as well as the logic controller block is available in [401].

19.1.3 Ring Oscillator Topology for Pulse Width Modulation

To create a single low-to-high oscillation at the output of the ring oscillator, the signal propagates twice through the entire ring oscillator stages. During the first pass, the PMOS transistor in the odd stages (P_{odd} transistors) and the NMOS transistor in the even stages (N_{even} transistors) are active, contributing to the T_{high} delay of the switching signal at the output of the ring oscillator. Alternatively, during the second round, the PMOS transistor in the even stages (P_{even} transistors) and the

NMOS transistor in the odd stages (N_{odd} transistors) are active, determining the T_{low} delay. A periodic signal that switches between zero and 1 V with duty cycle D and constant frequency $1/P$ is considered. The period and duty cycle of a switching signal are defined, respectively, as

$$P \equiv T_{high} + T_{low}, \tag{19.1}$$

$$D \equiv \frac{T_{high}}{T_{high} + T_{low}}. \tag{19.2}$$

All of the MOSFET transistors exhibit similar rise and fall transition times, contributing equally to the high and low portions of the output signal. The half period of a conventional 50 % duty cycle ($T_{0,high} = T_{0,low}$) ring oscillator with $2m+1$ stages is therefore

$$T_0 = T_{0,high} = T_{0,low} = (2m+1)\frac{C_G \Delta V_{out}}{I_{ave}}, \tag{19.3}$$

where C_G is the input gate capacitance of the next stage, ΔV_{out} is the voltage change at the output during a single rise/fall transition, and I_{ave} is the average current flowing through a single stage active transistor. The period and duty cycle of a conventional ring oscillator are, respectively, $P_0 = 2T_0$ and $D_0 = 1/2$.

The time required to charge the output capacitance of each ring oscillator stage, $C_G \Delta V_{out}/I_{ave}$, depends directly on the current flowing through the stage, affecting the response of the following stage and the frequency of the switching signal at the output. Lower (higher) than I_{ave} current through all of the P_{odd} and/or N_{even} transistors slows down (speeds up) the response of these stages, increasing (decreasing) the $T_{0,high}$ delay, duty cycle, and period of the switching signal at the output. Alternatively, current starvation (enhancement) of all of the P_{even} and/or N_{even} transistors slows down (speeds up) the response of these stages. The $T_{0,low}$ delay is therefore increased (decreased) in this configuration, decreasing (increasing) the duty cycle and increasing (decreasing) the period of the switching signal at the output. By connecting certain ring oscillator stages to a header circuit (in this case, the odd stages), the duty cycle of the ring oscillator can be controlled through starvation/enhancement of the current flowing to the ring oscillator stages. The effect of current starvation/enhancement on the ring oscillator timing behavior is listed in Table 19.1.

Consider a ring oscillator with $2m + 1$ stages with a header connected to $m + 1$ P_{odd} transistors, supplying the current $I_{bias} = \alpha I_{ave}$, as shown in Fig. 19.6. During the first round, only the biased P_{odd} and the affected N_{even} transistors are active, contributing, respectively, T_{bias} and $T_{bias\_affected}$ delays to the $T_{bias,high} = T_{bias} + T_{bias\_affected}$ delay at the output of the ring oscillator. The delay contribution of the $m + 1$ biased stages to the ring oscillator period is

$$T_{bias} = \frac{(m+1)C_G \Delta V_{out}}{I_{bias}} = \frac{m+1}{2m+1}\frac{T_0}{\alpha}. \tag{19.4}$$

Table 19.1 Timing parameters of the current controlled ring oscillator

Controlled stages	P_{odd}	N_{even}	P_{even}	N_{odd}
Affected stages	N_{even}	P_{odd}	N_{odd}	P_{even}
Current at the starved (enhanced) stage	⇓ (⇑)	⇓ (⇑)	⇓ (⇑)	⇓ (⇑)
Transition delay at the controlled and affected stages	⇑ (⇓)	⇑ (⇓)	⇑ (⇓)	⇑ (⇓)
T_{high}	⇑ (⇓)	⇑ (⇓)	Const.	Const.
T_{low}	Const.	Const.	⇑ (⇓)	⇑ (⇓)
Duty cycle	⇑ (⇓)	⇑ (⇓)	⇓ (⇑)	⇓ (⇑)
Period	⇑ (⇓)	⇑ (⇓)	⇑ (⇓)	⇑ (⇓)

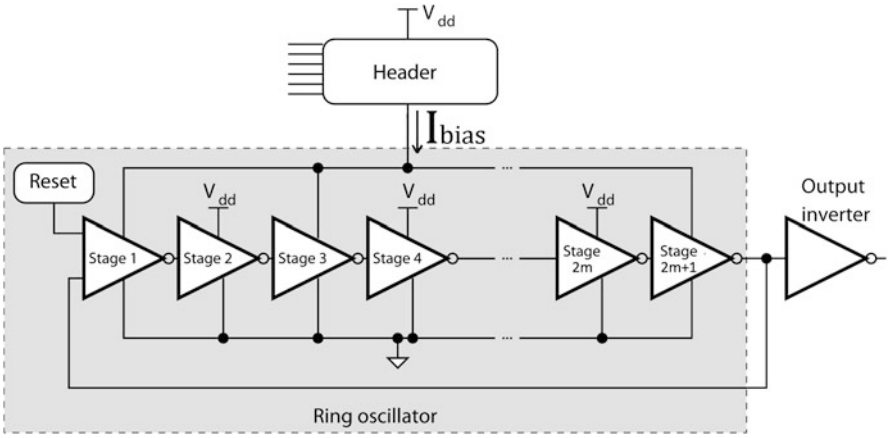


Fig. 19.6 Ring oscillator with current controlled P_{odd} transistors

Limiting the header current ($\alpha < 1$) to the P_{odd} transistors increases the transition delay of these stages, slowing the input transition time of the conventionally connected N_{even} transistors. Under these conditions, the conventionally connected N_{even} transistors switch more slowly. Alternatively, in those configurations where the P_{odd} transistors are enhanced ($\alpha > 1$) rather than starved, the input at the driven N_{even} transistors approaches an ideal step input, yielding faster switching of these conventionally connected stages. The delay of a conventional ring oscillator stage driven by a biased stage is inversely proportional to the bias current. The contribution of the m conventionally connected N_{even} transistors to the period of the biased ring oscillator is therefore

$$T_{bias\_affected} = \frac{m}{(2m+1)} \frac{T_0}{\alpha}. \tag{19.5}$$

During the second round, only the P_{even} and N_{odd} transistors are active, contributing to the $T_{bias,low}$ delay at the output of the ring oscillator. These transistors are not biased and are therefore unaffected by the biased stages of the ring oscillator. The $T_{bias,low}$ delay therefore remains unchanged, $T_{bias,low} = T_{0,low} = T_0$, determining the duty cycle of the ring oscillator,

$$D_{bias} = \frac{T_{bias,high}}{T_{bias,high} + T_{bias,low}} = \frac{1}{1+\alpha}. \tag{19.6}$$

The period of the ring oscillator is therefore

$$P_{bias} = T_{bias,high} + T_{bias,low} = T_0(1 + \frac{1}{\alpha}) = \frac{T_0}{1 - D_{bias}}. \tag{19.7}$$

The duty cycle of a biased ring oscillator is a function of the bias parameter $\alpha = I_{bias}/I_{ave}$ and does not depend on the number of stages $2m + 1$. For $\alpha = 1$, a duty cycle of 50 % in (19.6) corresponds to a duty cycle of a conventional ring oscillator with balanced rise and fall times. Alternatively, the theoretical 100 % duty cycle limit is achieved as $\alpha \to 0$.

This approach permits configuring a ring oscillator with a wide range of duty cycles. The period of a biased ring oscillator, however, depends on α (see (19.7)) and varies with the bias current. Thus, α is constrained by the minimum and maximum period $T_0 < P_{min} \leq P_{bias} \leq P_{max}$,

$$\frac{T_0}{P_{max} - T_0} \leq \alpha \leq \frac{T_0}{P_{min} - T_0}. \tag{19.8}$$

Note that the frequency of the switching signal generated at the output of the ring oscillator changes while varying the duty cycle of the signal. An improved version of the aforementioned ring oscillator with two header circuits is described in the next section to maintain a constant frequency under varying duty cycle ratios.

19.1.4 Ring Oscillator Topology for Pulse Width Modulation with Constant Frequency

The duty cycle of a switching signal can be controlled by changing the current sourced to the odd (or even) stages of a conventional ring oscillator, as demonstrated in Sect. 19.1.3. The period of the switching signal in the digitally controlled PWM topology, however, scales with the duty cycle (19.7), affecting the operational frequency of the ring oscillator. To provide a wide range of duty cycles while maintaining a constant frequency, an additional level of control over the timing parameters of the ring oscillator is required.

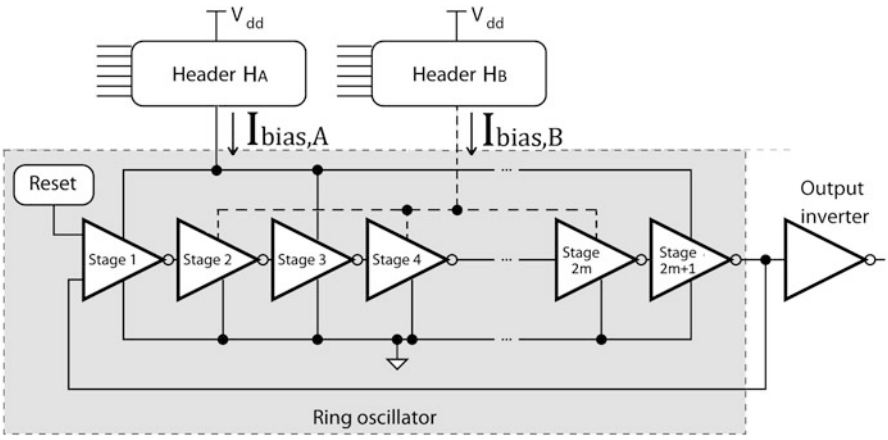


Fig. 19.7 Ring oscillator with current controlled P_{odd} and P_{even} transistors

Consider a ring oscillator with $2m + 1$ stages and two headers H_A and H_B that supply, respectively, current $I_{bias,A} = \alpha I_{ave}$ to the $m + 1$ P_{odd} transistors and $I_{bias,B} = \beta I_{ave}$ to the m P_{even} transistors, as shown in Fig. 19.7. The currents flowing through the P_{odd} and P_{even} transistors affect, respectively, the operating speed of the N_{even} and N_{odd} transistors, as described in Sect. 19.1.3. Alternatively, the P_{odd} and N_{even} transistors are active during the first pass through the ring oscillator independent of the P_{even} and N_{odd} transistors that are active during the second pass. Thus, the timing parameters of the ring oscillator shown in Fig. 19.7 are similar to the parameters used in Sect. 19.1.3, yielding the duty cycle of the ring oscillator,

$$D_{bias} = \frac{1/\alpha}{1/\alpha + 1/\beta} = \frac{1}{1 + \alpha/\beta}, \tag{19.9}$$

and period,

$$P_{bias} = T_0(1/\alpha + 1/\beta). \tag{19.10}$$

To maintain a constant period, the constraint $P^{(2)}_{bias} = 2T_0$ is used in (19.10), yielding $\beta = \alpha/(2\alpha - 1)$ and therefore

$$D_{bias}\,(P = 2T_0) = \frac{1}{2\alpha}. \tag{19.11}$$

Substituting the period, (19.10), and duty cycle, (19.11), of the controlled ring oscillator in Fig. 19.7, and the half period of a conventional ring oscillator with a

50 % duty cycle, (19.3), the expressions for the currents $I_{bias,A}$ and $I_{bias,B}$ shown in Fig. 19.7 are, respectively,

$$I_{bias,A} = \frac{I_{ave}}{2D}, \tag{19.12}$$

$$I_{bias,B} = I_{bias,A} \cdot \frac{D}{1-D} = I_{ave} \cdot \frac{1}{2(1-D)}. \tag{19.13}$$

Thus, to design a switching signal with a specific duty cycle D and period P, the ring oscillator topology shown in Fig. 19.7 should be used with the bias currents $I_{bias,A}$ and $I_{bias,B}$ described by, respectively, (19.12) and (19.13). The currents $I_{bias,A}$ and $I_{bias,B}$ are generated independently, and produce a variation insensitive duty cycle and frequency with properly compensated currents. A constant duty cycle under PVT variations is therefore a useful indicator for PVT mitigation in the digitally controlled PWM.

19.2 Simulation Results

A seven stage ring oscillator is described in this chapter to provide a switching signal with a wide range of duty cycles. The digitally controlled PWM is designed in a 22 nm CMOS predictive technology model [405]. Certain parameters in the technology model file are modified based on [406] to include process corners such as typical-typical (TT), slow-slow (SS), fast-fast (FF), fast-slow (FS), and slow-fast (SF). Simulation results characterizing the accuracy of the PWM are shown in Sect. 19.2.1 for different duty cycle ratios under PVT variations. The effect of the bias current on the duty cycle of the ring oscillator output is discussed in Sect. 19.2.3 without constraints on the period of the output signal, and in Sect. 19.2.2 under a constant period constraint.

19.2.1 Digitally Controlled Pulse Width Modulator Under PVT Variations

To evaluate the effect of PVT variations on the digitally controlled PWM, the current flowing through the P_{odd} transistors in the first, third, fifth, and seventh stages is controlled by the header circuit, as shown in Fig. 19.6 for $m = 3$. The remaining PMOS and NMOS transistors in this section are conventionally connected directly to, respectively, V_{dd} and ground. The supply voltage varies ± 5 % from the nominal 0.95 V and the temperature varies from 27 to 80 °C. The simulations have been performed for TT, SS, FF, FS, and SF process corners for the 22 nm predictive

Table 19.2 Change in the duty cycle of the digitally controlled PWM under PVT variations for the 22 nm predictive CMOS model

V_{dd}	Process	Temperature	Duty cycle			
			55 %	65 %	75 %	85 %
1.0	TT	27	55.03	64.79	74.58	85.79
1.0	TT	80	55.13	65.00	74.02	85.08
1.0	FF	27	55.01	64.86	74.67	85.83
1.0	FF	80	55.29	65.36	74.01	84.08
1.0	SS	27	55.17	65.01	74.87	85.75
1.0	SS	80	55.15	64.88	74.57	85.07
1.0	FS	27	55.15	65.34	75.60	87.06
1.0	FS	80	55.23	65.49	75.71	86.64
1.0	SF	27	55.51	65.37	74.52	84.36
1.0	SF	80	55.51	65.25	74.01	82.85
0.9	TT	27	55.10	65.00	74.77	86.02
0.9	TT	80	55.09	64.92	74.58	85.76
0.9	FF	27	55.00	64.90	74.66	86.28
0.9	FF	80	55.10	65.06	74.74	85.57
0.9	SS	27	55.30	65.34	75.21	86.08
0.9	SS	80	55.21	65.03	74.73	85.44
0.9	FS	27	55.26	65.81	76.18	87.40
0.9	FS	80	55.25	65.68	75.97	87.16
0.9	SF	27	55.61	65.40	74.38	84.26
0.9	SF	80	55.49	65.07	73.88	83.54
Maximum variations (%)			± 0.55	± 0.78	± 1.53	± 2.68

CMOS model [405]. The per cent deviation for different duty cycle ratios is listed in Table 19.2. The deviation of the duty cycle under PVT variations is less than 2.7 % of the targeted duty cycle.

The Monte Carlo simulations that consider process and mismatch variations are shown in Fig. 19.8 for a duty cycle of 55 % and frequency of 55 MHz, yielding a standard variation of, respectively, 0.73 % and 0.69 MHz.

Transistors with smaller dimensions are more sensitive to PVT variations [407, 408] and exhibit greater leakage current variations [404] than wider transistors. The narrower transistors within the header circuitry turn on if a switching signal with a greater duty cycle is required. The effect of PVT variations is therefore more prominent on those signals with a wider duty cycle. This trend can be observed in Table 19.2, where the deviation for signals with a 50 % duty cycle is smaller than for those signals with a 90 % duty cycle.

19.2.2 Duty Cycle Controlled Pulse Width Modulator

The accuracy of the analytic expressions of the duty cycle presented in Sect. 19.1.3 is evaluated in this section for a 25–90 % range of duty cycle. The circuit shown in

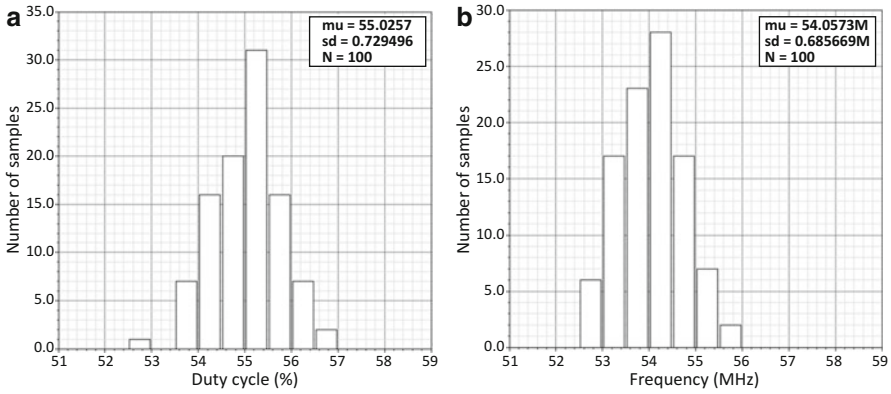


Fig. 19.8 Monte Carlo simulation of (**a**) duty cycle, and (**b**) frequency distribution

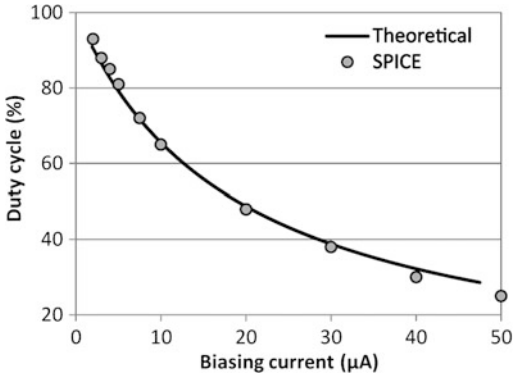


Fig. 19.9 Duty cycle varies between 25 % and 90 % when the header current changes from 50 to 2 μA (*error* < 4.4 %)

Fig. 19.6 is evaluated with an ideal current source replacing the header. The current I_{bias} flowing through the P_{odd} transistors is therefore controlled by an ideal current source. The rise time at the output of the P_{odd} transistors degrades with a longer charge time, increasing the duty cycle of the switching signal at the output of the ring oscillator.

The expressions for the duty cycle in (19.6) are verified with simulations for bias currents between 2 and 50 μA, as shown in Fig. 19.9. Note the good agreement between the analytic expressions and simulation results (*error* < 4.4 %). When the current flowing through the controlled stages is neither starved nor enhanced, the simulated circuit oscillates with a 50 % duty cycle, yielding $\alpha = 1$ where $I_{bias} = I_{ave} = 19\ \mu A$. Using (19.6), the duty cycle can be tuned with a digitally programmable control block.

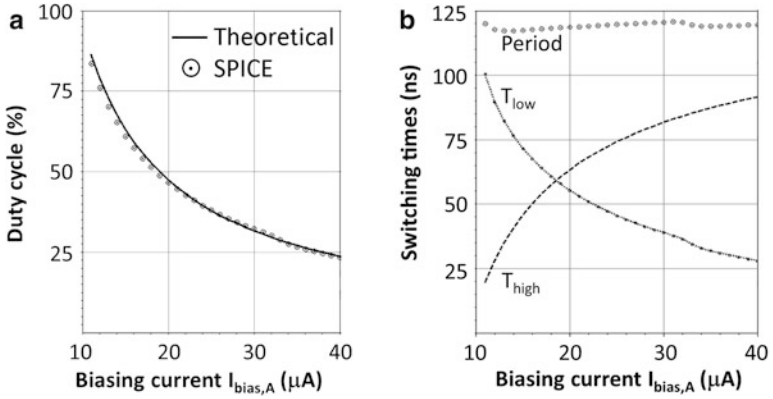


Fig. 19.10 Header current $I_{bias,A}$ changes from 40 to 11 μA. (**a**) Duty cycle varies between 25 % and 90 % (*error* < 3.1 %). (**b**) Period remains approximately constant (*error* < 1.25 %)

19.2.3 Duty Cycle and Frequency Controlled Pulse Width Modulator

The accuracy of the analytic expressions for the duty cycle under the constant frequency constraint (see Sect. 19.1.4) is evaluated in this section at 8.33 MHz for a 25–90 % range of duty cycle. The current supplied to the ring oscillator is controlled with two headers, exhibiting a frequency of 8.33 MHz ± 1.25 % for all values of duty cycle. The circuit shown in Fig. 19.7 is modeled by an ideal current source replacing the headers, H_A and H_B. The currents $I_{bias,A}$ and $I_{bias,B}$ flowing, respectively, through the P_{odd} and P_{even} transistors are assumed to be controlled by ideal current sources. Intuitively, the rise time at the output node of the P_{odd} transistors increases with a longer charge time, increasing the duty cycle of the switching signal. To mitigate the effect of the longer $T_{bias,high}$ delay on the period of the switching signal, the rise time at the output of the P_{even} transistors, based on (19.13), is decreased. The analytic expression for the duty cycle and period, respectively, (19.10) and (19.11), is verified by the simulations for $I_{bias,A}$ currents between 11 and 40 μA, as shown in Fig. 19.10. Note the good agreement between the analytic expressions and simulation results (*error* < 3.1 %).

19.3 Summary

A digitally controlled PWM with a wide pulse width ranging from 25 % to 90 % is described in this chapter. The pulse width modulator provides a means for dynamically changing the voltage in adaptive systems using fast control circuitry,

providing high accuracy under PVT variations and dynamic duty cycle and period control. The primary conclusions can be summarized as follows.

- An enhanced header circuit is presented to provide a greater range of header current
- The header circuit is connected to every other stage of the ring oscillator to significantly improve the dynamic range of the pulse width
- The parallel transistors within the header circuit control the duty cycle with high granularity
- To efficiently control both the duty cycle and the period of the switching signal, an additional header circuit is connected to the remaining ring oscillator stages
- A DC2V converter samples the duty cycle of the output signal and generates an analog voltage to control the header current
- The PVT variations are compensated by the feedback loop generated by the DC2V converter
- Under PVT variations, deviations in the pulse width are less than 2.7 % of the period of the switching signal
- Both the duty cycle and the period of the digitally controlled PWM are analytically determined as a function of the header current, simplifying the control over the PWM timing parameters
- The accuracy of the analytic expressions describing the duty cycle is compared with simulation results, yielding less than 3.1 % and 4.4 % error, respectively, with and without the controlled frequency for the pulse width modulator
- A constant frequency with less than 1.25 % variation is reported for different values of the duty cycle

Chapter 20
Conclusions

To provide a high quality power delivery system, the on-chip power needs to be regulated with ultra-small, locally distributed, power efficient converters. Switching mode, switched-capacitors, and both analog and digital linear voltage regulators are reviewed in this part. Existing power converter topologies exhibit an undesirable tradeoff among power efficiency, load regulation, and physical size. To simultaneously address challenging power requirements and area limitations in modern high performance ICs, heterogeneous power delivery architectures are required that enhance the power quality and efficiency of the overall power delivery system while satisfying on-chip area constraints.

A hybrid combination of a switching and low dropout regulator as a point-of-load power supply for next generation heterogeneous systems is described. The key concept in developing this ultra-small on-chip power supply is to replace the passive LC filter within the buck converter with a more area efficient active filter since the area occupied by a passive LC filter is a primary issue in the design of a monolithic buck converter. This voltage regulator has been successfully designed and manufactured in a commercial 110 nm TSMC CMOS technology. Despite the mature 110 nm technology, the total on-chip area is approximately 0.015 mm^2, significantly smaller than state-of-the-art on-chip voltage regulators. This ultra-small voltage regulator is appropriate for on-chip point-of-load voltage regulation with hundreds of distributed power regulators to support dynamic voltage and frequency scaling.

To demonstrate the feasibility of the evolving concept of distributed, dynamically controllable power delivery systems, several types of power delivery circuits are described. To provide a circuit level means for dynamically scaling the voltage in adaptive systems, a digitally controlled pulse width modulator is described, including closed-form expressions for the duty cycle, and validated under PVT variations. Another key component of distributed power delivery systems is an ultra-small power efficient linear regulator. An ultra-small power efficient linear low dropout regulator, designed and manufactured in a 28 nm CMOS process is

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0_20

described. The regulator exhibits excellent load regulation under PVT variations. To evaluate the performance of these ultra-small LDO regulators within a high performance integrated system, a power delivery system with six distributed LDO regulators has been designed, manufactured in a 28 nm CMOS process, and tested, exhibiting high efficiency and quality of power under a wide range of PVT and load variations.

Part V
Computer-Aided Design of Power Delivery Systems

Computer-aided design (CAD) for power delivery systems is reviewed in Part V. The generation and distribution of power by different types of power supplies and power networks are two primary issues in the power delivery process. The number of nodes in a typical power delivery system may exceed many millions (or billions) of nodes. To cope with this design complexity while maintaining high quality power in these complex power delivery systems, accurate and computationally efficient models, and effective analysis and optimization techniques are required. Existing modeling, analysis, and optimization techniques for power delivery systems with multiple power supplies and decoupling capacitors integrated at the board, package, and circuit levels are the subject of this chapter.

The process of analyzing power distribution networks is the topic of Chap. 21. The flow of computer-aided design processes for on-chip power distribution networks is described. The primary objectives and challenges of power network analysis at each stage of the design process are identified. A description of efficient numerical techniques for analyzing complex power distribution networks closes this chapter.

A multi-feedback system with parallel connected power supplies delivering current to a single grid exhibits significant design complexity and degraded stability due to complex interactions among the power supplies, power distribution network, and current loads. A criterion for evaluating the stability of a distributed power delivery system is described in Chap. 24. A distributed power delivery system with parallel connected LDO regulators is evaluated, exhibiting a stable multi-feedback response if and only if the passivity-based criterion is satisfied.

In Chap. 25, a link breaking methodology is discussed to reduce voltage degradation within a mesh structured power distribution network. The resulting power distribution network is a combination of a single power distribution network to lower the network impedance, and multiple networks to reduce noise coupling among the circuits. Since the sensitivity to supply voltage variations within a power distribution network can vary among different circuits, the methodology reduces the voltage drop at the more sensitive circuits, while penalizing the less sensitive circuits.

Closed-form expressions for the effective resistance of a two layer mesh structure are presented in Chap. 22. These closed-form expressions provide a fast and accurate solution to the effective resistance of a two layer mesh which can be used to solve a variety of problems found in different disciplines. Examples include *IR* voltage drop analysis of integrated circuits, synchronization and localization of sensor networks, the effective chemical distance between bonds, metal mesh interference filters in terahertz physics, and the commute and cover times of undirected graphs.

Closed-form expressions and related algorithms for fast static *IR* voltage drop analysis in a resistive power distribution network is the focus of Chap. 23. Four algorithms are described for non-uniform power supplies and current loads distributed throughout a power grid. The principle of spatial locality is exploited to accelerate the power grid analysis method for locally uniform, globally non-uniform power grids. Since no iterations are necessary for this *IR* drop analysis algorithm, these algorithms are significantly faster than existing methods while exhibiting low error.

To exploit the advantages of existing power supplies, a heterogeneous power delivery system is described with different types of power supplies integrated at different levels of the system hierarchy. The power efficiency of the system is shown to be a strong function of the clustering of the power supplies—the specific configuration in which the power converters and regulators are co-designed. The co-design of power supplies to maximize overall power efficiency is computationally inefficient and impractical with exhaustive clustering approaches in real-time systems. Heterogeneous power delivery and a recursive clustering algorithm with polynomial computational complexity is described in Chap. 26 for providing a real-time power allocation system with low power loss.

Chapter 21
Computer-Aided Design of Power Distribution Networks

The process of computer-aided design and analysis of on-chip power distribution networks is discussed in this chapter. The necessity for designing and analyzing the integrity of the power supply arises at various stages of the integrated circuit design process, as well as during the verification phase. The design and analysis of power distribution networks, however, poses unique challenges and requires different approaches as compared to the design and analysis of logic circuits.

The requirement for analyzing on-chip power distribution networks arises throughout the design process, from the onset of circuit specification to the final verification phase, as discussed in Sect. 21.1. The primary tasks and difficulties in analyzing the power supply vary at different phases of the design process. At the initial and intermediate design phases, the specification of the power distribution network is incomplete. The primary goal of the power supply analysis process is to guide the general design of the on-chip power distribution network based on information characterizing the power current requirements of the on-chip circuits. The information characterizing the power current requirements is limited, giving rise to the principal difficulty of the analysis process: producing efficient design guidance based on data of limited accuracy. The character of the analysis process gradually changes toward the final phases of the design process. The design of both the power distribution network and the on-chip logic circuits becomes more detailed, making a more accurate analysis possible. The principal goal of the analysis process shifts to verifying the design and identifying those locations where the target specifications are not satisfied. The dramatically increased complexity of the analysis process is the primary difficulty, requiring utilization of specialized computational methods.

This chapter is organized as follows. A typical flow of the power distribution network design process is described in Sect. 21.1. An approach for reducing the analysis of power distribution system to a linear problem is presented in Sect. 21.2. The process of constructing circuit models that characterize a power distribution system is discussed in Sect. 21.3. Techniques for characterizing the

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0_21

power current requirements of the on-chip circuits are described in Sect. 21.4. Numerical techniques used in the analysis of power distribution networks are briefly described in Sect. 21.5. Three strategies for allocating on-chip decoupling capacitors are described in Sect. 21.6. The chapter is summarized in Sect. 21.7.

21.1 Design Flow for On-Chip Power Distribution Networks

In high performance circuits, the high level design of the global power distribution network typically begins before the physical design of the circuit blocks. This approach ensures preferential allocation of sufficient metal resources, simplifying the design process. The principal decisions on the structure of the power distribution network are therefore made when little is known about the specific power requirements of the on-chip circuits. The early design of the power grid is therefore based on conservative design tradeoffs and is gradually refined in subsequent phases of the design process.

The design flow for a power distribution grid is shown in Fig. 21.1. As the circuit design becomes better specified, a more accurate characterization of the power requirements is possible and, consequently, the design of the power distribution network becomes more precise. The design process can be roughly divided into three phases: preliminary pre-floorplan design, floorplan-based refinement, and layout-based verification [153, 239]. These phases are described in the rest of this section.

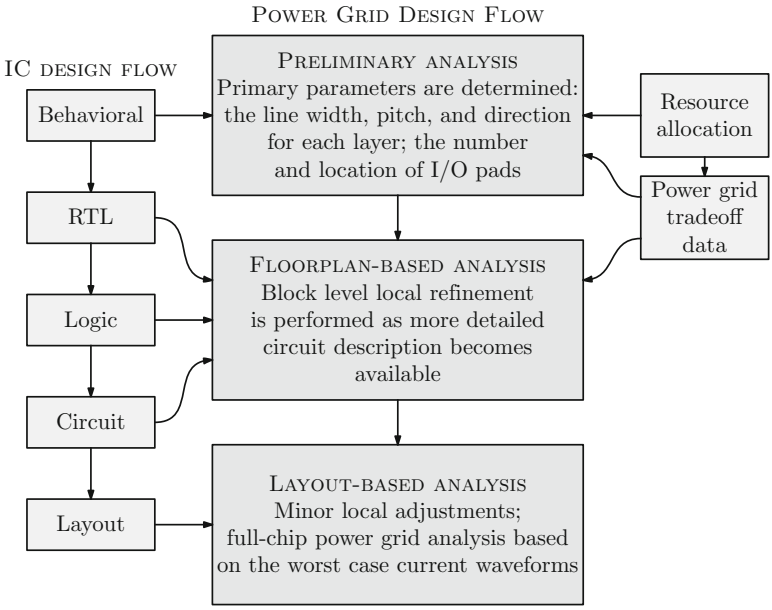


Fig. 21.1 Design flow for on-chip power distribution networks

21.1.1 Preliminary Pre-Floorplan Design

In the initial pre-floorplan phase, little is known about the power current requirements of the circuit. Preliminary estimates of the power current consumed by a circuit are typically made by scaling the power consumption of previously designed circuits considering the target die area, operating frequency, power supply voltage, and other circuit characteristics.

At this phase of the design process, the power distribution grid is often laid out as a regular periodic structure. A preliminary DC analysis of the *IR* drops within a power distribution grid is performed, assuming that the power current requirements are uniform across the die. The average power current requirements used in a DC analysis are increased by three to seven times to estimate the maximum power current. The power distribution network is assumed to be uniformly loaded with constant current sources. Basic parameters of the power distribution grid, such as the width and pitch of the power lines in each metal layer and the location of the power/ground pads, are determined based on a preliminary DC analysis of the network. This initial structure largely determines the tradeoff between robustness and the amount of metal resources used by an on-chip power distribution grid.

21.1.2 Floorplan-Based Refinement

Once the floorplan of a circuit is determined, the initial design of the power distribution grid is refined to better match the local current capacity of the power distribution grid to the power requirements of the individual circuit blocks [145]. The maximum and average power current of each circuit block is determined based on the function of an individual block (e.g., the memory, floating point unit, register file), area, block architecture, and the specific circuit style (e.g., static, dynamic, pass transistor [409], etc.). The current distribution is assumed uniform within the individual circuit blocks.

Block specific estimates of the power current provide an approximation of the non-uniform power requirements across the circuit die. The structure of the power distribution grid is tailored according to a DC analysis of a non-uniform power current distribution. Many of the primary problems in the design of power distribution networks are identified at this phase. Moderate computational requirements permit iterative application of a static analysis of the network. Large scale deficiencies in the coverage and capacity of the power distribution network are detected and repaired.

As the structure of the circuits blocks becomes better specified, the local power consumption of an integrated circuit can be characterized with more detail and accuracy. After the logic structure of the circuits is determined, the accuracy of the current requirements are enhanced based on the number of gates and clocking requirements of the circuit blocks. Gate level simulations provide a per cycle

estimate of the DC power current for a chosen set of input vectors [153]. Cycle-to-cycle variations of the average power current provide an approximation of the temporal variations of the power current, permitting a preliminary dynamic AC analysis of the power distribution system. The accuracy of the dynamic analysis can be improved if more detailed current waveforms are obtained through gate level simulations. The worst case current waveform of each type of gate and circuit structure is precharacterized. The current waveforms of the constituent gates are arranged according to the timing information obtained in the simulations and are combined into an effective power current waveform for an entire circuit block. As the circuit structure and operating characteristics become better specified, the structure of the power distribution grid within each of the circuit blocks is refined to provide sufficient reliability and integrity of the on-chip power supply while minimizing the required routing resources.

As the precise placement of the circuit gates is not known in the pre-layout phase, the spatial resolution of the floorplan-based models is relatively coarse. The die area is divided into a grid of $N \times M$ cells. The power and ground distribution networks within each cell are reduced to a simplified macromodel. These macromodels form a coarse *RC*/*RLC* grid model of the on-chip power distribution network, as shown in Fig. 21.2. The power current of the circuits located in each cell is combined and modeled by a current source connected to the appropriate node of the macromodel. The number of cells in each dimension of the circuit typically varies from several cells to a hundred cells, depending on the size of the circuit and the accuracy of the power consumption estimate. The computational requirements of the analysis process increase with the specificity of the circuit description. The total number of nodes, however, remains relatively small, permitting an analysis with conventional nonlinear circuit simulation tools such as SPICE.

21.1.3 Layout-Based Verification

When the physical design of a circuit is largely completed, a detailed analysis of the power distribution network is performed to verify that the target power supply noise margins are satisfied at the power/ground terminals of each on-chip circuit. A detailed analysis is first performed at the level of the individual circuit blocks. Those areas where the noise margins are violated are identified during this analysis phase. The current capacity of the power distribution grid is locally increased in these areas by widening the existing power lines, adding lines, and placing additional on-chip decoupling capacitance. The detailed verification process is repeated on the modified circuit. The iterative process of analysis and modification is continued until the design targets are satisfied. Finally, the verification process is performed for the entire circuit.

An analysis of an entire integrated circuit is necessary to verify the design of a power distribution network. Analyzing the integrity of the power supply at the circuit block level is insufficient as neighboring blocks affect the flow of current

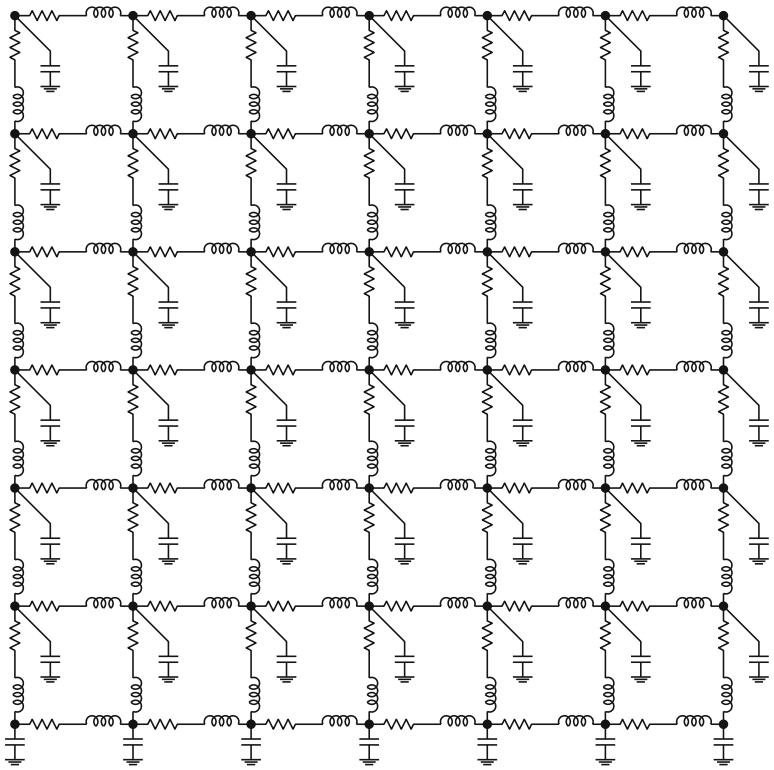


Fig. 21.2 An *RLC* model of an on-chip power distribution network [236]

through the power grid. For example, the design of a power grid within a circuit block drawing a relatively low power current (e.g., a memory block) may appear satisfactory at a block-level analysis. However, it is likely to fail if the block is placed in close proximity to a block drawing high power current. The high power blocks can increase the current flowing through the power network of adjacent low power units [239]. It is therefore necessary to verify the design of the power distribution network at the entire circuit level.

The principal difficulty in verifying an entire power distribution network in a high complexity integrated circuit is the sheer magnitude of the problem. The on-chip power network of a modern high complexity integrated circuit often comprises tens of millions of interconnect line segments and circuit nodes forming a multi-layer power distribution grid, as described in Sect. 8.1. The circuits loading the power distribution network also consist of hundreds of millions of interconnects and transistors. A transistor level circuit simulation of an entire circuit is therefore infeasible due to prohibitive memory and CPU time requirements. Final analysis and verification is therefore one of the most challenging tasks in the design of on-chip high complexity power distribution networks. The remainder of this chapter is

devoted largely to techniques and methodologies to manage the complexity of the analysis and verification process of power distribution networks.

This methodology is successful if the noise margin violations are local and can be corrected with available metal resources. However, if the necessary changes in the power grid require significant changes in the routing of the critical signal lines, the timing and noise performance characteristics of these critical signals can be significantly impaired. The laborious task of signal routing and timing verification of a circuit is repeated, drastically decreasing design productivity and increasing the time to market. This difficulty of making significant changes in the structure of the power distribution grid at late phases of the design process is the primary reason for using a highly conservative approach in the design of on-chip power distribution networks. Worst case scenarios are assumed throughout the design process. The resulting power distribution network is therefore typically overdesigned, significantly increasing the area used by power distribution networks in modern interconnect-limited integrated circuits.

21.2 Linear Analysis of Power Distribution Networks

The process of analyzing a power network consists of building a circuit model of the power and ground networks including the circuits loading the networks. This step is followed by a numerical analysis of the resulting model. The problem is inherently nonlinear as the digital circuits loading the power distribution grid exhibit highly nonlinear behavior. The current drawn by the load circuits from the power distribution network varies nonlinearly with the voltage across the power terminals of the load. Analyzing a network with tens of millions of nodes is infeasible using a nonlinear circuit simulator such as SPICE due to the enormous computational and memory requirements. To permit the use of efficient numerical analysis techniques, the nonlinear part of the problem is separated from the linear part [145, 153, 239], as illustrated in Fig. 21.3. The current drawn from the power distribution network by nonlinear on-chip circuits is characterized assuming a nominal power and ground supply voltage. The load circuits are replaced by time dependent current sources emulating the original power current characteristics. The resulting network consists of power distribution conductors, decoupling capacitors, and time dependent current sources. This network is linear, permitting the use of efficient numerical techniques.

Partitioning the problem into a power current characterization part and a linear system analysis part ignores the negative feedback between the power current and the power supply noise. The current flowing through the power and ground networks causes the power supply levels to deviate from the nominal voltage. In turn, the reduced voltage between the power and ground networks decreases the current drawn by the power load. This typical approach to the power noise analysis process is therefore conservative, overestimating the magnitude of the power supply noise. The relative decrease in the power current due to the reduced power supply voltage is comparable to the relative decrease in the power-to-ground voltage, typically

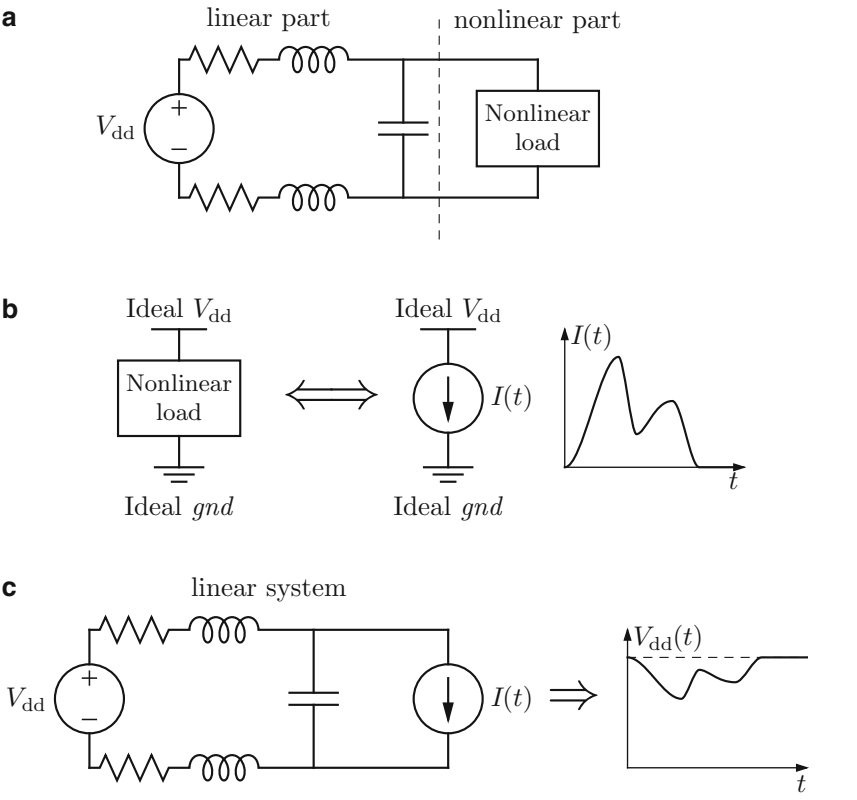


Fig. 21.3 Approximation for analyzing a power distribution network by replacing a nonlinear load with a time-dependent current source. (**a**) The original problem of analyzing a linear power distribution network with a nonlinear load. (**b**) The current requirements of the nonlinear load are characterized under an ideal supply voltage. (**c**) The nonlinear load is replaced with an AC current source. The resulting system can be analyzed with linear methods

maintained below 10 %. The accuracy of these conservative estimates of the power noise is acceptable for most applications. To achieve greater accuracy, the analysis can be performed iteratively. The power current requirements are re-characterized at each iteration based upon the power supply voltage obtained in the previous step. Each iteration yields a more accurate approximation of the power supply voltage across the power distribution network.

The process of power distribution network analysis therefore proceeds in three phases: model construction, load current characterization, and numerical analysis. These tasks are described in greater detail in the following sections.

21.3 Modeling Power Distribution Networks

It is essential that the on-chip power distribution network is considered in the context of the entire power distribution system, including the package and board power distribution networks [185, 410–413]. As discussed in Chap. 7, the package and board power distribution networks determine the impedance characteristics of the overall power distribution system at low and intermediate frequencies. It is therefore important to analyze the entire power distribution system, including the package and board power distribution networks and the decoupling capacitors, to obtain an accurate analysis of the on-chip power supply noise [236].

The complexity of a model depends upon the objectives of the analysis. Models for a DC analysis performed at the preliminary and floorplan-based design phases need to capture only the resistive characteristics of the interconnect structures. The inductive and capacitive circuit characteristics are unimportant in a DC analysis, greatly simplifying the model. As discussed in the previous section, the spatial resolution of these models is typically limited, further simplifying the process of model characterization.

The reactive impedances of the system are, however, essential for an accurate AC analysis of a power distribution system. The capacitance of the board, package, and on-chip decoupling capacitors as well as the inductive properties of the network should be characterized with high accuracy.

The analysis and verification step towards the end of the design process requires highly detailed models that capture the smallest features of a power distribution system. These models are typically constructed through a back annotation process. The complexity of the board and package power distribution networks is relatively moderate, with the number of conductors ranging from hundreds to thousands. The moderate complexity supports the use of relatively sophisticated analysis tools, such as two- and three-dimensional quasi-static electromagnetic field analyzers [134, 140, 236, 414]. Characterizing the on-chip power distribution network is the most difficult part of the modeling process. The on-chip power distribution network comprises tens of millions of nodes and interconnect elements. This level of complexity requires highly efficient algorithms to extract the parasitic impedances of the on-chip circuit structures.

21.3.1 Resistance of the On-Chip Power Distribution Network

The resistance of on-chip interconnect can be efficiently characterized either with simple resistance formulas based on the sheet resistance of a metal layer [153, 414] or well developed shape-based extraction algorithms [415, 416]. The temperature dependence of the interconnect resistance should also be included in the model. If R_{25} is the nominal metal resistance at a room temperature of 25 °C, the metal resistance at the operating temperature of the circuit T_{op} is $R_{25}\big(1 + k_T(T_{op} - 25)\big)$,

where k_T is the temperature coefficient of the metal resistance. For a temperature coefficient of copper doped aluminum metalization of $0.003\,°C^{-1}$ and an operating temperature of 85 °C, the temperature induced per cent increase in the resistance is 18 %, a significant change. Furthermore, the interconnect resistance increases over the circuit lifetime due to electromigration induced defects in the metal structure. This increase in resistance is typically considered in the design process by increasing the nominal metal resistance by a coefficient K_{em}, typically ranging from 10 % to 20 % [414]. The overall resistance of the on-chip metal R_{eff} can therefore be characterized as [236]

$$R_{eff} = R_{25}\big(1 + (T_{op} - 25)\big)\,(1 + K_{em})\,. \tag{21.1}$$

21.3.2 Characterization of the On-Chip Decoupling Capacitance

Characterizing the capacitive impedances within the power distribution system is more difficult as compared to resistance characterization. The intrinsic capacitance of the power and ground lines is dominated by other sources of the decoupling capacitance, i.e., the intrinsic circuit capacitance, well capacitance, and intentional capacitance, as discussed in Sect. 11.3. The capacitance of the power and ground lines can therefore be neglected in this analysis. The intentional and well diffusion decoupling capacitances can be readily characterized by shape-based extraction methods. The intrinsic decoupling capacitance of the on-chip circuits depends upon the state of the digital circuits, making this capacitance difficult to characterize.

The intrinsic circuit decoupling capacitance can be estimated based on the power consumption of the circuit [230, 414]. Assuming that the total power P_0 is dominated by the dynamic switching power $P_{switching}$,

$$P_0 \approx P_{switching} = \alpha C_{total} f_{clk} V_{dd}^2\,, \tag{21.2}$$

where α is the switching factor of the circuit, C_{total} is the total intrinsic capacitance of the circuit, f_{clk} is the clock frequency, and V_{dd} is the supply voltage. The total capacitance C_{total} can therefore be determined from an estimate of the total circuit power: $C_{total} = \frac{P_0}{\alpha f V_{dd}^2}$. The fraction of the total capacitance being switched, i.e., αC_{total} on average, is the load capacitance of the circuit. The rest of the total capacitance, $(1 - \alpha)C_{total}$, is quiescent and effectively serves as a decoupling capacitance. The intrinsic decoupling capacitance of the circuit is, therefore,

$$C_{decap}^{ckt} \approx \frac{P_0}{f V_{dd}^2} \frac{1 - \alpha}{\alpha}\,. \tag{21.3}$$

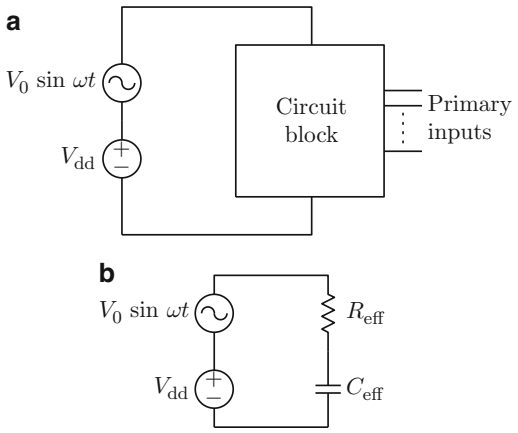


Fig. 21.4 Characterization of the intrinsic decoupling capacitance of the quiescent circuits; (**a**) circuit model to characterize the capacitance, (**b**) equivalent circuit model of the intrinsic decoupling capacitance [140]

An estimate of the intrinsic decoupling capacitance represented by (21.3) is, however, strongly dependent on the switching factor α. The switching factor varies significantly depending upon the specific switching pattern and circuit type. The switching factor is therefore difficult to determine with sufficient accuracy in complex digital circuits.

Alternatively, the decoupling capacitance of quiescent circuits can be characterized by simulating a small number of representative circuit blocks [140]. A complete circuit model of each selected circuit block, including the parasitic impedances of the interconnect, is constructed through a back annotation process. The input terminals of a circuit block are randomly set to either the high or low state. The power terminals of the circuit are biased with the power supply voltage V_{dd}. A sinusoidal AC voltage of relatively small amplitude (5–15 % of V_{dd}) is added to the power terminals of the circuit, modeling the power supply noise, as shown in Fig. 21.4a. The current flowing through the power terminals is obtained and the small signal impedance of the circuit block *as seen from the power terminals* is determined for the specific frequency of the AC excitation. A series *RC* model is subsequently constructed, such that the model impedance approximates the impedance of the original circuit block, as shown in Fig. 21.4b. The model capacitance is scaled by a factor $(1 - \alpha)$ to account for the switching of the circuit capacitance α which does not participate in the decoupling process. The resulting model is an equivalent circuit of the decoupling capacitance of the quiescent circuits, including the decoupling capacitance of both the transistors and interconnect structures. This estimate of the decoupling capacitance is significantly less sensitive to the value of α, as compared to (21.3).

The elements R_{eff} and C_{eff} of the equivalent model depend on the state of the digital circuit and the frequency of the applied AC excitation. Nevertheless, these model parameters typically vary little with the input pattern and the excitation frequency in the range of 0.2–2 times the clock frequency [140]. For example, the model parameters exhibit less than a 3 % variation over all of the input states of an

example circuit block consisting of 240 transistors with ten primary inputs [140]. The decoupling characteristics of a larger circuit block are extrapolated from the characteristics of one or several of the precharacterized blocks, depending upon the circuit structure of the larger block. This technique allows for variations in the intrinsic decoupling capacitance for different circuit types.

In many circuits, however, the circuit decoupling capacitance is dominated by the well diffusion capacitance and the intentional capacitance [177]. The overall accuracy of the power supply noise analysis process in these circuits is only moderately degraded by the inaccuracies in characterizing the circuit decoupling capacitance.

21.3.3 Inductance of the On-Chip Power Distribution Network

The inductance of the on-chip power and ground lines has historically been neglected [140]. The relatively high resistance R of the on-chip interconnect has dominated the inductive impedance ωL, suppressing the inductive behavior, such as signal reflections, oscillations, and overshoots. As the switching time of the on-chip circuits decreases with technology scaling, the spectral content of the on-chip signals has extended to higher frequencies, making on-chip inductive effects more pronounced. The significance of the on-chip inductance has been demonstrated in an investigation of the sensitivity of the power supply noise to various electrical characteristics of the power distribution system [281]. Assuming the package leads provide an ideal nominal voltage of 2.5 V, an *RLC* analysis of the on-chip power grid predicts a minimum on-chip voltage V_{dd} of 2.307, 0.193 V below the nominal level. If the inductance of the on-chip power grid is neglected, the analysis predicts a minimum on-chip power voltage V_{dd} of 2.396 V, underestimating the on-chip power noise by 50 % as compared to a more complete *RLC* model. Including the package model in the analysis further reduces the on-chip power supply to 2.199 V. Modeling the inductive properties of the on-chip power interconnect is therefore necessary to ensure an accurate analysis.

Incorporating the inductive properties of on-chip interconnect into the model of a power distribution network poses two challenges. First, existing techniques for characterizing the inductive properties of complex interconnect structures are computationally expensive, greatly reducing the efficiency of the back annotation process. This issue is further discussed below. Second, including inductance in the model precludes the use of highly efficient techniques for numerically analyzing complex power distribution networks, as discussed in Sect. 21.5.

The inductive properties of power and ground interconnect lines are difficult to characterize. Characterizing the inductance by a conventional method, i.e., determining the loop inductance of the on-chip circuits based on the shape and size of the current loops is difficult as the current path consists of multiple conductors and the path of the current flow is, generally, not known a priori. The inductive properties of regular on-chip power distribution grids can be estimated based on

an electromagnetic analysis of the grid structure [236, 414]. Alternatively, the inductive properties can be extracted in the form of a partial inductance matrix. While extracting the partial inductance matrix of an entire circuit is computationally efficient, this matrix is highly dense. The density of a complete partial inductance matrix drastically degrades the efficiency of the subsequent numerical analysis of the circuit model, as the computational efficiency of the most effective numerical methods is conditioned on the sparsity of the matrices characterizing the system. Techniques for sparsifying partial inductance matrices are an active area of research and several techniques have been described [66–69]. The computational efficiency of these techniques is currently insufficient to make the analysis of multimillion conductor systems practical. A method to characterize the inductance of on-chip power distribution grids is described in Chap. 28.

21.3.4 Exploiting Symmetry to Reduce Model Complexity

The magnitude of the current flowing through the power and ground distribution networks is the same. The power and ground networks have the same electrical requirements and the structures of these networks are often (close to) symmetric, particularly at the initial and intermediate phases of the design process. This symmetry can be exploited to reduce the complexity of the power distribution network model by half [140], as illustrated in Fig. 21.5. The model reduction is achieved by circuit "folding." The original symmetric circuit, as shown in Fig. 21.5a, is transformed into an equivalent circuit, where the sources, loads, and decoupling capacitors are replaced with equivalent symmetric networks, as shown in Fig. 21.5b. The nodes on the axis of symmetry of the circuit (shown with a dashed line in Fig. 21.5b) are equipotential. It is convenient to use the potential of these nodes as a reference potential. These nodes are therefore referred to as a *virtual ground*. The original circuit is transformed into two independent circuits, as shown in Fig. 21.5c. The independent circuits are symmetric; consequently, an analysis of only one circuit is necessary. The currents and voltages in one circuit have an opposite polarity as compared to the currents and voltages in the symmetric circuit. Where the impedances of the power and ground distribution networks are symmetric, the voltages in the power and ground networks (with reference to the virtual ground) are also symmetric. That is, wherever the power voltage is decreased by δV, the ground voltage is increased by δV (thereby decreasing the power rail-to-rail voltage by $2\delta V$).

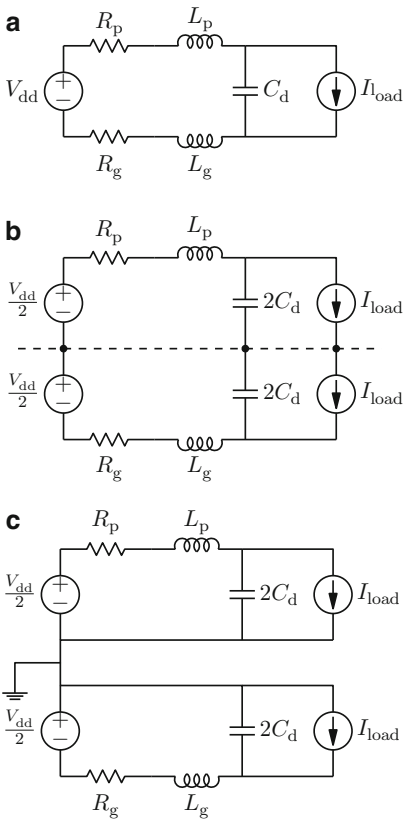


Fig. 21.5 Exploiting the symmetry of the power and ground distribution networks to reduce the model complexity by a factor of 2; (**a**) power and ground current paths in the original model are symmetric, (**b**) virtual ground exists between the power and ground networks in the equivalent circuit, as shown by the *dashed line*, (**c**) resulting circuit model contains two independent symmetric circuits

21.4 Characterizing the Power Current Requirements of On-Chip Circuits

Accurate characterization of the power current requirements is an integral part of the power distribution analysis process, as discussed in Sect. 21.2. A brief overview of the methods for power current characterization is presented in this section. As the structure of an integrated circuit becomes specified in greater detail, the power current characteristics can be specified with greater accuracy. The complexity of the power current characterization process dramatically increases with the complexity of the circuit description.

21.4.1 Preliminary Evaluation of Power Current Requirements

Early estimates of the power current are static. The temporal variation of the power current cannot be assessed in this phase as only the high level structure of the circuit has been developed. Static approaches are based on estimates of the average load currents drawn from the power network [153, 239], permitting static estimates of the *IR* drops and electromigration reliability. Estimating average power current consumption is equivalent to estimating average circuit power, as the power supply voltage is maintained approximately constant. At the onset of the design process, the average power current per circuit area is estimated based on the function of a particular circuit block, circuit style used to design a circuit block, and a scaling analysis of previously designed similar circuits. These estimates can be augmented by estimates of average circuit power based on a circuit description at the behavioral, register transfer, or microarchitectural levels [417, 418].

21.4.2 Gate Level Estimates of the Power Current Requirements

Estimates of the power current requirements are refined once the logic structure of the circuit is determined [419, 420]. The primary difficulty in determining the power requirements of a CMOS circuit with sufficient accuracy is the dependency of the power current on the circuit input pattern [421]. The time of switching and the magnitude of the power current of a particular gate are determined by the temporal and functional relationships with other gates. The switching patterns that produce the greatest variation of the power supply voltage from a nominal specification (i.e., the greatest power supply noise) are referred to as the worst case switching patterns. These worst case switching patterns are difficult to identify.

The worst case power current of small circuit structures, such as individual logic gates and circuit macrocells, are relatively easy to determine as the number of possible switching patterns is small and the patterns can be readily evaluated. The number of possible switching patterns increases exponentially with the number of inputs and internal state variables. The worst case switching patterns of relatively large circuit blocks comprising thousands of circuit gates and macrocells cannot be determined from an exhaustive analysis. The assumption that all of the gates draw the worst case power current at the same time is overly conservative. Incorporating logical dependencies among the logic gates, however, greatly increases the complexity of the analysis process. A tradeoff therefore exists between the accuracy and efficiency of the power current model.

Estimates of the average power current are typically based on a probabilistic or statistical analysis of the average switching activity and the output load (i.e., the switched capacitance) of the gates [421]. Simple estimates of the average load currents can be obtained by determining the saturation current of each gate in a block

and scaling this current to account for the quiescent state of the majority of the gates at any particular time. More accurate estimates of the average current I_{avg} can be obtained through gate level simulations by evaluating the average switching activity P_s and capacitive load C_L of the gates. The average power current of the circuit per switching event is evaluated as $I_{\text{avg}} = \frac{1}{2} P_s f_{\text{clk}} C_L V_{\text{dd}}$, where f_{clk} is the clock frequency of the circuit. Several methods have been developed to determine the upper bound of the power current consumed by the circuit and the associated bound on the power supply noise in an input pattern independent manner [422–425].

21.5 Numerical Methods for Analyzing Power Distribution Networks

The circuit model of a power distribution network is combined with time dependent current sources emulating the worst case load to form a linear model of a power distribution network. The linear model is described by a system of linear differential equations. The system of differential equations is reduced to a system of linear equations, which can be numerically analyzed using a number of efficient linear system solution methods [426]. These linear solution methods are classified into direct and iterative methods [427]. The direct methods rely on factoring the coefficient matrix that characterizes the linear system. Once the matrix decomposition is performed, the system solution at each simulation time step is obtained by forward and backward substitution [428, 429]. Alternatively, iterative methods can be used to obtain the solution through a series of successive approximations [430, 431]. Assuming sufficient memory capacity to store the factorization matrices, the use of direct methods is preferable in analyses requiring a large number of time steps, as the solution at each step is obtained through an efficient substitution procedure. Iterative methods are more efficient in solving large systems with limited memory resources.

Numerical techniques exploiting special properties of the system are commonly employed to enhance the efficiency of the analysis process. The coefficient matrix of a linear system describing a power distribution network is highly sparse, with non-zero elements typically constituting only a 10^{-6}–10^{-8} fraction of the total number of elements [145]. Furthermore, in a modified nodal analysis approach, the matrix is symmetric and, for an *RC* model of a power distribution network, positive definite [145]. Of the direct methods, Cholesky factorization is particularly well suited, requiring moderate memory resources to store the factorization data. Of the iterative methods, the conjugate gradient method is more memory efficient for denser and larger systems [145]. Several techniques to further enhance the efficiency of analyzing power distribution networks are described in the remainder of this section.

21.5.1 Model Partitioning in RC and RLC Parts

These numerical methods are modified if the mutual inductance among the interconnect segments are considered [140]. The matrix describing the system is no longer guaranteed to be positive definite, preventing the use of efficient methods based on this property and forcing the use of more general (and computationally expensive) methods. Including the mutual inductance elements is virtually always necessary to accurately describe the electrical properties of the power distribution networks of a printed circuit board and an integrated circuit package. An *RC*-only model is often adequate for describing the on-chip power distribution network. In many cases, therefore, only a relatively small part of the overall model (describing the package and board power interconnect) contains inductive elements. The computational complexity of the problem can be significantly reduced in these cases [140], as illustrated in Fig. 21.6. A comprehensive model of a power distribution system is partitioned into an *RLC* part containing all of the inductive elements (at the package and board level) and an *RC*-only part (the on-chip network). The *RC*-only part contains the vast majority of elements comprising the overall power distribution system. The complexity of the *RC* part of the system can be reduced by exploiting efficient techniques based on solving a symmetric positive definite system of equations. The *RC* part of the model is replaced with the equivalent admittance at the ports of the interface with the *RLC* part. The resulting system is significantly smaller than the original system and can be solved with general solution methods.

An approach to analyzing power distribution networks composed of *RLC* segments (with no mutual inductance terms) has been described in [432, 433]. An enhanced matrix formulation of the power distribution network problem is described and numerical techniques to solve this formulation are described. As demonstrated on sample networks consisting of several hundred segments, this analysis approach is three to four hundred times faster as compared to SPICE simulations, while maintaining an accuracy within 5 % of SPICE.

21.5.2 Improving the Initial Condition Accuracy of the AC Analysis

The efficiency of the transient analysis can be enhanced by accurate estimates of the steady state condition, i.e., the currents passing through the network inductors and the voltages across the network capacitors. The steady state condition is not known before the analysis. The analysis starts with a rough estimate of the initial conditions, for example, with the voltages and currents determined in a DC analysis. In the beginning of the AC analysis, the initial excitation conditions are maintained and the system is allowed to relax to the AC steady state. After the steady state is reached, the switching pattern of interest can be applied to the circuit inputs, permitting a transient analysis to be initiated. No useful information is produced as

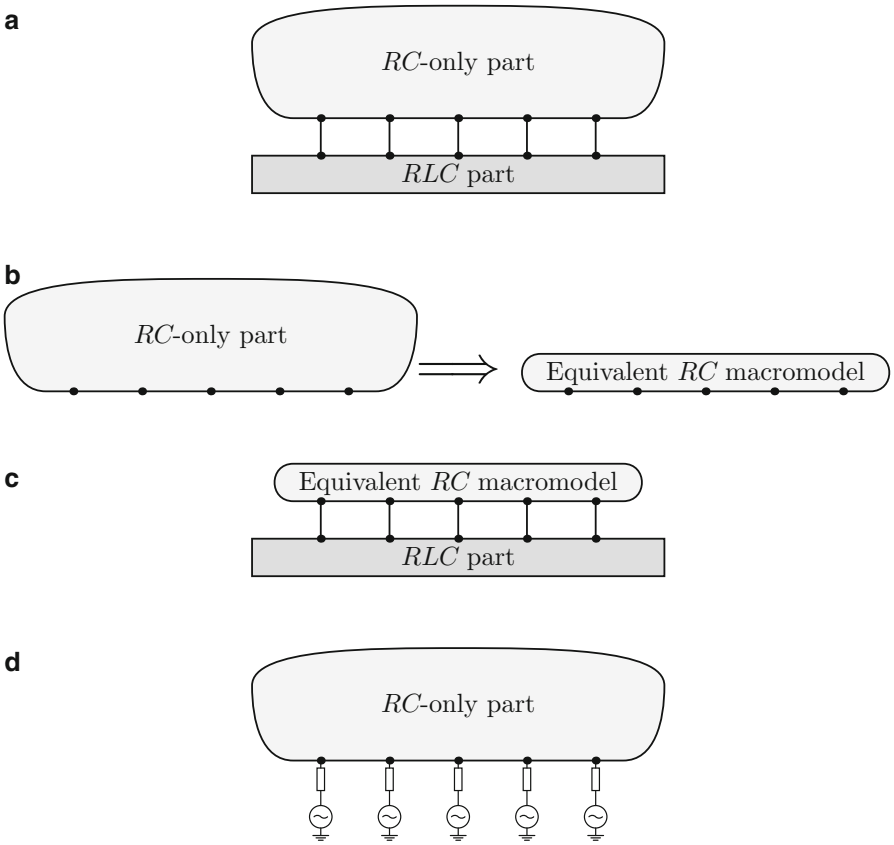


Fig. 21.6 Reducing the computational complexity of the analysis process by separating the analysis of the *RLC* and *RC*-only parts of a power distribution system. (**a**) A circuit model of a power distribution system can be partitioned into a relatively small *RLC* part and an *RC*-only part containing the vast majority of the circuit elements. (**b**) An equivalent admittance macromodel of the *RC*-only part is constructed. (**c**) The *RC* part is replaced with a reduced model and the system is analyzed using robust numerical methods. The voltages and equivalent admittances at the ports of the *RC*-only part are determined. (**d**) The RC-only part is analyzed using efficient numerical methods. The *RLC* part of the model is replaced with equivalent circuits at the appropriate ports, as has been determined in the previous step

the system settles to the AC steady state. In systems with a low damping factor, the time required to reach the steady state can be a substantial portion of the overall time span of the simulation, significantly increasing the computational overhead of the analysis [140].

An accurate estimate of the initial conditions is therefore desirable. This estimate can be efficiently obtained as follows [140]. A simplified circuit model of the power distribution network is constructed. Elements of the simplified model are determined based on the elements of the original network and the worst case

voltage drop obtained from a DC analysis of the original network. The simplified circuit is simulated and the steady state inductor currents and capacitor voltages are determined. These currents and voltages are used as steady state values in the transient analysis of the original network. Using this technique to analyze the power distribution network of a 300 MHz PowerPC microprocessor, the initial conditions are estimated with an accuracy of 6.5 % as compared to a 62 % accuracy based on a DC analysis. The greater accuracy of the initial conditions shortens by a factor of 3 the time required to determine the AC steady state.

21.5.3 Global-Local Hierarchical Analysis

A hierarchical approach to the electrical analysis of an on-chip power distribution network reduces the CPU time and memory requirements as compared to a flat (non-hierarchical) analysis [300, 434–436]. A power network is partitioned into a global grid and many local grids, as depicted in Fig. 21.7. A macromodel is built for every local partition. A macromodel is a linear multi-port network characterized by the same relationship between the port currents and voltages as the original local partition. The power network is simulated with each local partition substituted by the respective macromodel. The problem size is thereby reduced from the total number of nodes in the original power distribution network to the number of nodes in the global partition plus the total sum of the local partition ports. Subsequently, to determine the voltage at the nodes of the local partitions, each local partition can be independently analyzed with the port currents determined during the analysis of the global grid with macromodels. The efficiency of the methodology therefore depends upon judicial partitioning, the computational cost of constructing a macromodel, and the complexity of the macromodel.

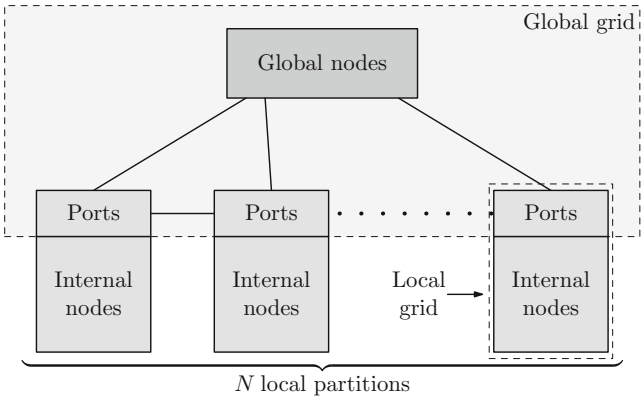


Fig. 21.7 A hierarchical model of a power distribution network. In a global analysis, the local grids are represented by multi-port linear macromodels

The greatest reduction in the complexity of the analysis is achieved if the system is partitioned into subnetworks with the number of internal nodes much larger than the square of the respective number of ports [434]. The macromodel matrices tend to have a higher density than the matrix representation of the original power distribution network. The higher density can limit the choice of suitable numerical solution methods and, therefore, the efficiency of the analysis process. Sparsification of the macromodels can be performed to address this problem [434].

The performance gains of the hierarchical method over a conventional flat analysis have been evaluated on power distribution networks of several industrial DSP and microprocessor circuits [434]. The memory requirements are reduced severalfold. The size of the linear system describing the hierarchical system is approximately ten times smaller then the size of the system in a flat (non-hierarchical) analysis. The memory requirements are reduced by 10–20 times. The one-time overhead of the analysis setup (partitioning, macromodel generation, sparsification, etc., in the case of the hierarchical methodology) is reduced by a factor of 2–5. The run time of the subsequent time steps, however, is greater as compared to a flat analysis. The difference in the runtime decreases with the size of the system, becoming relatively small in networks with ten million nodes or more.

Since each local partition of the network is solved independently, a hierarchical analysis has two other desirable properties [434]. First, the hierarchical analysis is easily amenable to parallel computation, permitting additional speedup in the subsequent time steps. The parallel hierarchical analysis is two to five times faster than a flat analysis. Second, the mutual independence of the macromodels makes the hierarchical analysis quite flexible. Local changes in the circuit structure necessitate regeneration of only a single macromodel. The rest of the setup can be reused, permitting an efficient incremental analysis. Alternatively, if a detailed power analysis of a specific block is only of interest, the local solution of other partitions can be omitted, accelerating the analysis process while preserving the effect of these partitions on the partition of interest.

21.5.4 Random Walk Based Technique

Another approach to analyze a power grid is the use of random walks to determine the voltage at a particular node [437, 438]. In [437], a parallel is drawn between a random walk algorithm in an undirected connected graph and voltage drop analysis of a power grid. In the random walk algorithm, a person walks from an arbitrary location and travels to a nearby house with some probability of arriving at that location. Specifically, if the walker reaches his/her house, the walk ends and the walker is awarded with a certain amount of money. Otherwise, the walker pays some money to stay at a hotel. The objective of the random walk algorithm is to determine the expected amount of money that a walker has at the end of a random walk. This random walk algorithm is run iteratively to minimize the error. A mathematical equivalent of the random walk problem is constructed for a power

grid [437]. In this application, every power supply is modeled as a house targeted by the random walker and the load circuits are modeled as hotels. The probabilities are equated to the power grid impedance between adjacent nodes. The load current determines the cost of the corresponding hotel stay. The voltage at a particular node is modeled as the expected amount of money that the walker has at the end of the walk [437]. Random walk based power grid analysis techniques exhibit a good accuracy/runtime tradeoff when the number of power supply connections is large; however, this algorithm becomes less efficient when the number of power supplies or power supply connections is small. These techniques, however, have poor convergence properties as compared to preconditioned iterative methods, such as Krylov-subspace [439].

21.5.5 Multigrid Analysis

The efficient analysis of power distribution grids can be performed through the use of multigrid methods [440, 441]. Power distribution grids are spatially and temporally well behaved (i.e., smooth and damped) systems. General purpose robust techniques are unnecessary to achieve an accurate solution. A system of linear equations describing a well behaved system is analogous to a finite element discretization of a two-dimensional parabolic partial differential equation [441]. Efficient numerical methods developed for parabolic partial differential equations can therefore be exploited to analyze power distribution grids. The multi-grid method is most commonly used to solve parabolic partial differential equations [442]. Using a fixed time step requires only a single inversion of a large and sparse matrix during the numerical analysis process. Multigrid techniques can be described as either algebraic multigrid or geometric multigrid. Algebraic multigrid has become more popular since a predefined grid structure is not required and the analysis of irregular topologies is less complicated as compared to geometric multigrid techniques [443].

21.5.6 Hierarchical Analysis of Networks with Mesh-Tree Topology

The analysis and optimization of power distribution networks structured as a global mesh feeding local trees can be achieved with a specially formulated hierarchical method in [444]. The process of hierarchical analysis proceeds in three stages. First, each tree is replaced with an equivalent circuit model obtained from the passive reduced-order interconnect macromodeling algorithm (PRIMA) [445]. The system is solved to determine all of the nodal voltages in step two. Each tree is analyzed independently based on the voltage at the root of the tree obtained in step two.

The method produces results within 10 % of SPICE with a greater than ten fold computational speedup.

21.5.7 Efficient Analysis of RL Trees

A worst case *IR* and ΔI noise analysis is efficiently performed in power and ground distribution networks structured as *RL* trees originating from a single I/O pad [446]. The worst case power current requirements of each circuit attached to a power distribution tree are approximated by a trapezoidal waveform. The intrinsic and intentional decoupling capacitances are neglected in the analysis, allowing the power voltage to be efficiently calculated at each node of the tree.

A method for the frequency domain analysis of noise in *RL* power distribution trees has also been developed [447]. A frequency domain noise spectrum is computed by analyzing the effective output impedance of the power distribution network at each current source and the spatial correlation among the trees. A time domain noise waveform is obtained by applying an inverse Fast Fourier transform to the frequency domain spectrum. This approach is more than two orders of magnitude faster than HSPICE simulations, while maintaining an accuracy within 10 % as compared to circuit simulation.

21.6 Allocation of On-Chip Decoupling Capacitors

The allocation of on-chip decoupling capacitors is commonly performed iteratively. Each iteration of the allocation process consists of two steps, as shown in Fig. 21.8. In the power noise analysis phase, the magnitude of the power supply noise is determined throughout the circuit. The size and placement of the decoupling capacitors are then modified during the allocation phase based on the results of the noise analysis. This process continues until all of the target power noise constraints are satisfied. Occasionally, the power noise constraints cannot be satisfied for a specific circuit. In this case, the area dedicated to the on-chip decoupling capacitors should be increased. In some cases, large functional blocks should be partitioned, permitting the placement of decoupling capacitors around the smaller circuit blocks.

Although a sufficiently large amount of on-chip decoupling capacitance distributed across an IC will ensure adequate power supply integrity, the on-chip decoupling capacitors consume considerable die area and leak significant amounts of current. Interconnect limited circuits typically contain a certain amount of white space (area not occupied by the circuit) where intentional decoupling capacitors can be placed without increasing the overall die size. After this area is utilized, accommodating additional decoupling capacitors increases the overall circuit area. The amount of intentional decoupling capacitance should therefore be minimized. A strategy guiding the capacitance allocation process is therefore required to achieve

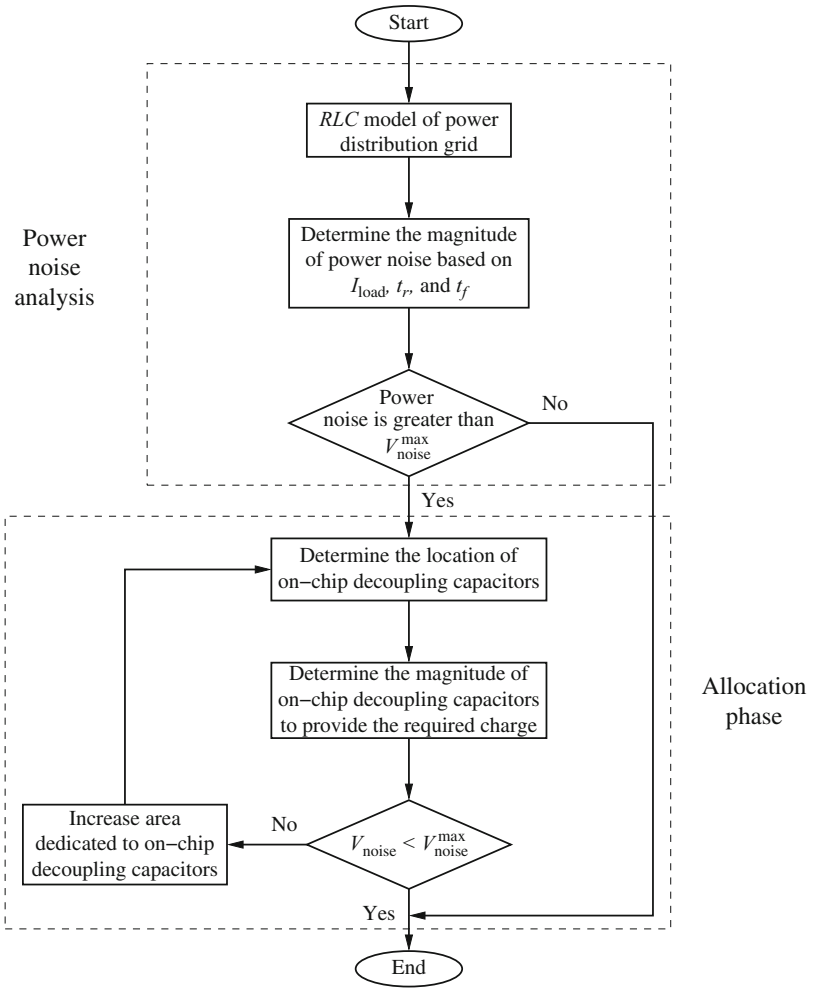


Fig. 21.8 Flow chart for allocating on-chip decoupling capacitors

target specifications with fewer iterations while utilizing the minimum amount of on-chip decoupling capacitance.

Different allocation strategies are the focus of this section. A charge-based allocation methodology is presented in Sect. 21.6.1. An allocation strategy based on an excessive noise amplitude is described in Sect. 21.6.2. An allocation strategy based on excessive charge is discussed in Sect. 21.6.3.

21.6.1 Charge-Based Allocation Methodology

One of the first approaches is based on the average power current drawn by a circuit block [280]. The decoupling capacitance C_i^{dec} at node i is selected to be sufficiently large so as to supply an average power current I_i^{avg} drawn at node i for a duration of a single clock period. To release charge $\delta Q_i = \frac{I_i^{\text{avg}}}{f_{\text{clk}}}$ as the power voltage level varies by a noise margin δV_{dd},

$$C_i^{\text{dec}} = \frac{\delta Q_i}{\delta V_{\text{dd}}} = \frac{I_i^{\text{avg}}}{f_{\text{clk}}\, \delta V_{\text{dd}}}, \tag{21.4}$$

where f_{clk} is the clock frequency.

The rationale behind the approach represented by (21.4) is that the power current during a clock period is provided by the on-chip decoupling capacitors. This allocation methodology is based on two assumptions. First, at frequencies higher than the clock frequency, the on-chip decoupling capacitors are effectively disconnected from the package and board power delivery networks (i.e., at these frequencies, the impedance of the current path to the off-chip decoupling capacitors is much greater than the impedance of the on-chip decoupling capacitors). Second, the on-chip decoupling capacitors are fully recharged to the nominal power supply voltage before the next clock cycle begins.

Both of these assumptions cannot be simultaneously satisfied with high accuracy. The required on-chip decoupling capacitance as determined by (21.4) is neither sufficient nor necessary to limit the power supply fluctuations within the target margin δV_{dd}. If the impedance of the package-to-die interface is sufficiently low, a significant share of the power current during a single clock period is provided by the decoupling capacitors of the package, overestimating the required on-chip decoupling capacitance as determined by (21.4). Conversely, if the impedance of the package-to-die interface is relatively high, the time required to recharge the on-chip decoupling capacitors is greater than the clock period, making the requirement represented by (21.4) insufficient. This inconsistency is largely responsible for the unrealistic dependence of the decoupling capacitance as determined by (21.4) on the circuit frequency, i.e., the required decoupling capacitance decreases with frequency. Certain assumptions concerning the impedance characteristics of the power distribution network of the package and package-die interface should therefore be considered to accurately estimate the required on-chip decoupling capacitance.

The efficacy of the charge-based allocation strategy has been evaluated on the Pentium II and Alpha 21264 microprocessors using microarchitectural estimation of the average current drawn by a circuit block [282, 448, 449]. The characteristics of the power distribution network based on (21.4) are evaluated and compared in both the frequency and time domains to three other cases: no decoupling capacitance is added, decoupling capacitors are placed at the center of each functional unit, and a uniform distribution of the decoupling capacitors. The AC current requirements of functional units within the microprocessor are estimated based on the average

power current obtained from architectural simulations. The charge-based allocation strategy has been demonstrated to result in the lowest impedance power distribution system in the frequency domain and the smallest peak-to-peak magnitude of the power noise in the time domain.

21.6.2 Allocation Strategy Based on the Excessive Noise Amplitude

More aggressive capacitance budgeting is described in [450, 451] to amend the allocation strategy described by (21.4). In this modified scheme, the circuit is first analyzed without an intentional on-chip decoupling capacitance and the worst case power noise inside each circuit block is determined. No additional decoupling capacitance is allocated to those blocks where the power noise target specifications have already been achieved. The intrinsic decoupling capacitance of these circuit blocks is sufficient. In those circuit blocks where the maximum power noise V_{noise} exceeds the target margin δV_{dd}, the amount of additional decoupling capacitance is

$$C_{\text{dec}} = \frac{V_{\text{noise}} - \delta V_{\text{dd}}}{V_{\text{noise}}} \frac{\delta Q}{\delta V_{\text{dd}}}, \tag{21.5}$$

where δQ is the charge drawn from the power distribution system by the current load during a single clock period.

The rationale behind (21.5) is that to reduce the power noise from V_{noise} to δV_{dd} (i.e., by a factor of $\frac{V_{\text{noise}}}{\delta V_{\text{dd}}}$), the capacitance C_{dec} should supply a $1 - \frac{\delta V_{\text{dd}}}{V_{\text{noise}}}$ share of the total current. Consequently, the same share of charge as the power voltage is decreased by δV_{dd}, making $C_{\text{dec}}\,\delta V_{\text{dd}} = \frac{V_{\text{noise}} - \delta V_{\text{dd}}}{V_{\text{noise}}}\,\delta Q$. Adding a decoupling capacitance to only those circuit blocks with a noise margin violation, the allocation strategy based on the excessive noise amplitude implicitly considers the effect of the on-chip intrinsic decoupling capacitance and the off-chip decoupling capacitors [30].

The efficacy of a capacitance allocation methodology based on (21.5) has been tested on five MCNC benchmark circuits [284]. For a 0.25 μm CMOS technology, the described methodology requires, on average, 28 % lower overall decoupling capacitance as compared to the more conservative allocation methodology based on (21.4) [280]. A noise aware floorplanning methodology based on this allocation strategy has also been developed [284]. The noise aware floorplanning methodology produces, on average, a 20 % lower peak power noise and a 12 % smaller decoupling capacitance as compared to a post-floorplanning approach. The smaller required decoupling capacitance occupies less area and produces, on average, a 1.2 % smaller die size.

21.6.3 Allocation Strategy Based on Excessive Charge

The allocation strategy presented in Sect. 21.6.2 can be further refined. Note that (21.5) uses only the excess of the power voltage over the noise margin as a metric of the severity of the noise margin violation. This metric does not consider the duration of the voltage disturbance. Longer variations of the power supply voltage have a greater impact on signal timing and integrity. A time integral of the excess of the signal variation above the noise margin is described in [452, 453] as a more accurate metric characterizing the severity of the noise margin violation. According to this approach, a metric of quality of the ground supply at node j is

$$M_j = \int_0^T \max\left[\left(V_j^{\text{gnd}}(t) - \delta V\right), 0\right] dt, \tag{21.6}$$

or, assuming a single peak noise violates the noise margin between times t_1 and t_2,

$$M_j = \int_{t_1}^{t_2} \left(V_j^{\text{gnd}}(t) - \delta V\right) dt, \tag{21.7}$$

where $V_j^{\text{gnd}}(t)$ is the ground voltage at node j of the power distribution grid.

Worst case switching patterns are used to calculate (21.6) and (21.7). This metric is illustrated in Fig. 21.9. The value of the integral in (21.7) equals the area of the shaded region. Note that if the variation of the ground voltage does not exceed the noise margin at any point in time, the metric M_j is zero. The overall power supply quality M is calculated by summing the quality metrics of the individual nodes,

$$M = \sum_j M_j. \tag{21.8}$$

This metric is zero when the power noise margins are satisfied at all times throughout the circuit.

Application of (21.6) and (21.8) to the decoupling capacitance allocation process requires a more complex procedure than (21.4) and (21.5). Note that utilizing (21.6) requires detailed knowledge of the power voltage waveform V_j^{gnd} at each node of the power distribution grid rather than just the peak magnitude of the deviation from the nominal power supply voltage. Computationally expensive techniques are therefore necessary to obtain the power voltage waveform. Furthermore, the metric of power supply quality as expressed in (21.8) does not explicitly determine the distribution of the decoupling capacitance. A multi-variable optimization is required to determine the distribution of the decoupling capacitors that minimizes (21.8). The integral formulation expressed by (21.6) is, fortunately, amenable to efficient optimization algorithms. The primary motivation for the original integral formulation of the

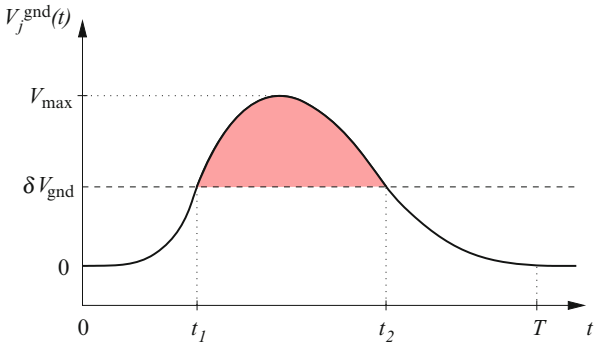


Fig. 21.9 Variation of ground supply voltage with time. The integral of the excess of the ground voltage deviation over the noise margin δV_{gnd} (*the shaded area*) is used as a quality metric to guide the process of allocating the decoupling capacitors

excessive charge metric is, in fact, to facilitate incorporating these noise effects into the circuit optimization process.

The efficacy of the allocation strategy represented by (21.8) in application-specific ICs has been demonstrated in [283, 454]. The distribution of the decoupling capacitance in standard-cell circuit blocks has been analyzed. The total decoupling capacitance within the circuit is determined by the empty space between the standard cells within the cell rows. The total budgeted decoupling capacitance (the amount of empty space) remains constant. As compared to a uniform distribution of the decoupling capacitance across the circuit area, the described methodology results in a significant reduction in both the number of circuit nodes exhibiting noise margin violations and the maximum power supply noise.

21.7 Summary

The process of designing and analyzing on-chip power distribution networks has been presented in this chapter. The primary conclusions of the chapter are summarized as follows.

- The design of on-chip power distribution networks typically begins prior to the physical design of the on-chip circuitry and is gradually refined as the structure of the on-chip circuits is determined
- The primary difficulty in early stages of the design process is accurately assessing the on-chip power current requirements
- The primary challenge shifts to the efficient analysis of the on-chip power distribution network once the circuit structure is specified in sufficient detail
- The complexity of analyzing an entire power distribution network loaded by millions of nonlinear transistors is well beyond the capacity of nonlinear circuit simulators
- Approximating nonlinear loads by time-varying current sources and thereby rendering the problem amenable to the methods of linear analysis is a common approach to manage the complexity of the power distribution network analysis process

- Several techniques have been developed to enhance the efficiency of the numerical analysis process
- The local impedance characteristics of an on-chip power distribution network depend upon the distribution of the decoupling capacitors
- The time integral of the excess of the signal variation above the noise margin is a useful metric for characterizing the severity of a noise margin violation
- Existing capacitance allocation methodologies place large decoupling capacitances near those on-chip circuits with the greater power requirements

Chapter 22
Effective Resistance in a Two Layer Mesh

An on-chip power and ground distribution network is commonly modeled as a resistive mesh structure with different vertical and horizontal unit resistances, as shown in Fig. 22.1a [396, 442, 455, 456], where the thickness and width of the metal lines are typically different in orthogonal metal layers. Power and ground networks are illustrated in Fig. 22.1a with, respectively, dark and light gray lines. A mesh structured power network and the corresponding resistive circuit model are illustrated, respectively, in Fig. 22.1b, c. Since the power and ground distribution networks exhibit similar characteristics, only the power network is considered in this chapter. This approach can also be used to determine the effective resistance in any two layer mesh structure with different horizontal and vertical unit resistances.

The effective resistance of a mesh is used in power grid analysis [457–459], substrate analysis [460], decoupling capacitance allocation [284, 304, 461], power dissipation [462], electrostatic discharge (ESD) analysis [462], and measuring resistance variations in power distribution networks [463]. The effective resistance is used to determine the effective region of a decoupling capacitor [284, 306]. For instance, the effective resistance between hot spots and available white spaces in a circuit floorplan provides a means to evaluate the effectiveness of a decoupling capacitor placed at different locations. A lower effective resistance between a hot spot and decoupling capacitor leads to a faster response time for the decoupling capacitor. Additionally, the effective resistance between a decoupling capacitor and power supply connection provides an estimate of the recharge time of the capacitor. When the effective resistance between two circuit blocks decreases, noise coupling through the power network increases which is quantified by the effective resistance described in this chapter. The effective resistance is also used to determine the coverage and commute times of a random walk in a graph [464]. In an undirected resistive graph, the effective resistance is used to determine the effective chemical distance between bonds, as in [465]. The effective resistance is also used in distributive control and estimation such as synchronization and localization of sensor networks [466].

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0_22

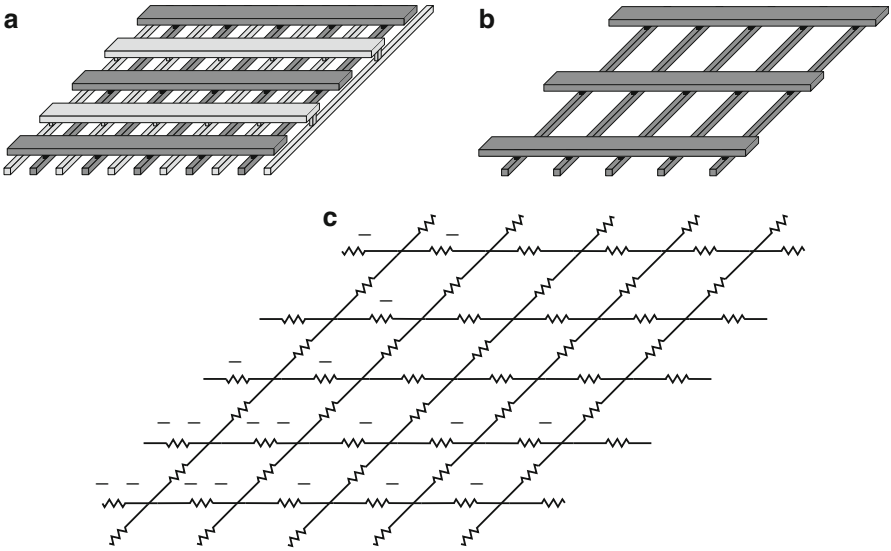


Fig. 22.1 Two layer orthogonal metal lines connected with vias; (**a**) two layer power and ground distribution network where the power and ground lines are illustrated, respectively, with *dark* and *light gray*, (**b**) a two layer power distribution network only, and (**c**) a resistive mesh model of the power distribution network

Venezian in [307] developed a closed-form expression of the resistance of a uniform mesh where the vertical and horizontal unit resistances are the same. The work described in this chapter is inspired by [307], where the effective resistance is generalized for non-isotropic meshes with different vertical and horizontal unit resistances. To determine the effective resistance between nodes n_{x_1,y_1} and n_{x_2,y_1}, where x_1, x_2, and y_1 are, respectively, the horizontal and vertical coordinates of the nodes within an infinite mesh, as shown in Fig. 22.2a, the principal of superposition is applied in two steps [307, 467]. First, current I is sourced at n_{x_1,y_1} and exits the grid at the boundaries (i.e., at infinity), as illustrated in Fig. 22.2b. The current from n_{x_1,y_1} to the adjacent nodes is determined by the resistance between n_{x_1,y_1} and the adjacent nodes. When the mesh is uniform, the currents from (x_1, y_1) to the adjacent nodes are symmetric and $I/4$. Secondly, current I is sourced at infinity and exits the grid at n_{x_2,y_1}, as depicted in Fig. 22.2c. The current from the nodes adjacent to n_{x_2,y_1} is similarly determined. When the mesh is uniform, the current from the adjacent nodes of (x_1, y_1) to (x_2, y_1) are again symmetric and $I/4$. By applying superposition in these two steps, current I is modeled as entering the grid from n_{x_1,y_1} and exiting the grid at n_{x_2,y_1}, as shown in Fig. 22.2d. This current is the sum of the currents in the first and second steps of the superposition process, which is therefore $I/2$. The voltage difference divided by the current is the effective resistance. The effective resistance between n_{x_1,y_1} and n_{x_2,y_1} within a uniform mesh is therefore

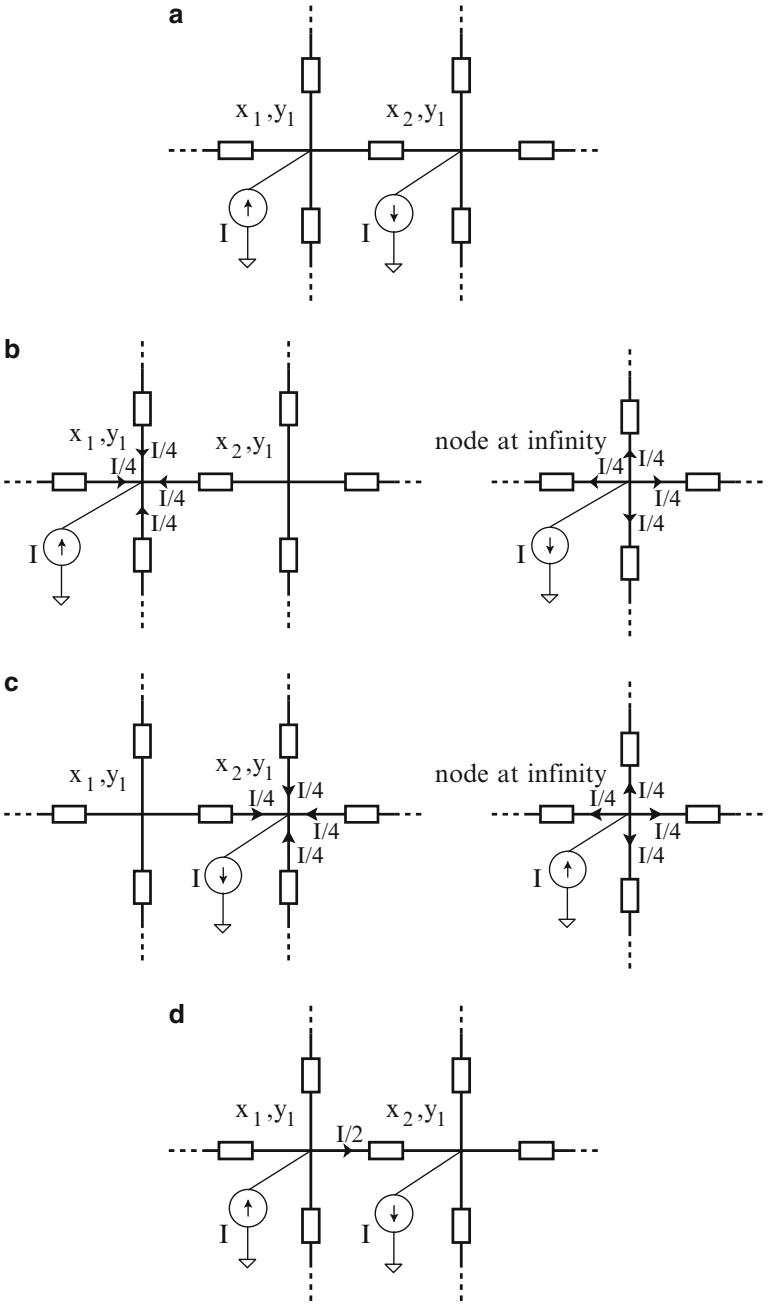


Fig. 22.2 In an infinite mesh structure: (**a**) a current source I is connected to (x_1, y_1) and a current load I is connected to (x_1, y_2) and the effective resistance between these adjacent nodes is determined by applying the principle of superposition in two steps. In the first step, (**b**) the load current is moved to a node at infinity and in the second step, (**c**) the source current is moved to a node at infinity. The current profiles for these two cases are obtained, and (**d**) the current source and load are moved to the original positions and the current during the two superposition steps is summed to determine the effective resistance

$$R_{eff} = 2(V_{x_1,y_1} - V_{x_2,y_1})/I. \tag{22.1}$$

A similar analysis is performed for a non-isotropic mesh structure with different horizontal and vertical resistances to determine closed-form expressions for the effective resistance between two arbitrary nodes.

This chapter is organized as follows. In Sect. 22.1, Kirchhoff's current law is revisited to determine the voltages and currents at a particular node in terms of the neighboring node voltages and resistances. In Sect. 22.2, inhomogeneous differential equations are applied where separation of variables is used to determine the node voltages. The effective resistance between two arbitrary intersections and the corresponding closed-form expressions are described, respectively, in Sects. 22.3 and 22.4. The accuracy of the effective resistance model is discussed in Sect. 22.5. The chapter is summarized in Sect. 22.6. A derivation of the closed-form expression for the effective resistance is offered in Appendix H.

22.1 Kirchhoff's Current Law Revisited

The mesh circuit model considered in this chapter is shown in Fig. 22.1c with horizontal (r_h) and vertical (r_v) resistors. The voltage at node (x,y) $n_{x,y}$ is $V_{x,y}$ and the current from $n_{x,y}$ to ground is $I_{x,y}$. When a current source is connected to $n_{x,y}$, $I_{x,y} = I$. Alternatively, when no current source is connected to $n_{x,y}$, $I_{x,y} = 0$.

The current load at an arbitrary node $n_{x,y}$ can be written in terms of the sum of the current from the four adjacent nodes as

$$\begin{aligned} I_{x,y} = &\frac{V_{x,y} - V_{x,y+1}}{r_v} + \frac{V_{x,y} - V_{x,y-1}}{r_v} \\ &+ \frac{V_{x,y} - V_{x+1,y}}{r_h} + \frac{V_{x,y} - V_{x-1,y}}{r_h}. \end{aligned} \tag{22.2}$$

The vertical resistance between adjacent nodes is r and the horizontal resistance between adjacent nodes is $k * r$, where k is a number $0 < k < \infty$, as

$$r_v = r \tag{22.3}$$

$$r_h = k * r. \tag{22.4}$$

When $I_{x,y} = 0$, the voltage at $n_{x,y}$ is

$$V_{x,y} = \frac{kV_{x,y+1} + kV_{x,y-1} + V_{x+1,y} + V_{x-1,y}}{2k + 2}. \tag{22.5}$$

When a current source is connected to $n_{x,y}$, this current can be described in terms of the adjacent node voltages and corresponding resistors as

$$I_{x,y} = \frac{(2k+2)V_{x,y} - (kV_{x,y+1} + kV_{x,y-1} + V_{x+1,y} + V_{x-1,y})}{k\,r}. \tag{22.6}$$

22.2 Separation of Variables

The difference Eqs. (22.5) and (22.6), can be solved by applying separation of variables [307]. A solution for (22.5) is

$$V_{x,y} = e^{x\alpha + jy\beta}. \tag{22.7}$$

By substituting (22.7) into (22.5), (22.5) can be written as

$$(2k+2)e^{x\alpha+jy\beta} = e^{x\alpha+jy\beta}(ke^{j\beta} + ke^{-j\beta} + e^{\alpha} + e^{-\alpha}), \tag{22.8}$$

$$2k + 2 = k(e^{j\beta} + e^{-j\beta}) + (e^{\alpha} + e^{-\alpha}). \tag{22.9}$$

Using the cosine and sine properties, (22.9) is

$$k + 1 = k\cos\beta + \cosh\alpha. \tag{22.10}$$

When a current source is connected to $n_{0,0}$ and is assumed to exit the system at infinity, the following equations are satisfied due to symmetry of the mesh structure,

$$V_{x,y} = V_{-x,y} = V_{x,-y} = V_{-x,-y}. \tag{22.11}$$

One possible solution to (22.11) is

$$V_{x,y} = e^{-|x|\alpha}\cos y\beta. \tag{22.12}$$

The currents can be described in terms of these voltages. Substituting $x = y = 0$ into (22.6), the current $i_{0,0}$ at $n_{0,0}$ is

$$i_{0,0} = \frac{(2k+2)V_{0,0} - kV_{0,1} - kV_{0,-1} - V_{1,0} - V_{-1,0}}{kr}. \tag{22.13}$$

Substituting (22.12) into (22.13), the current I at $n_{0,0}$ is

$$i_{0,0} = 2(k + 1 - k\cos\beta - e^{-\alpha})/kr. \tag{22.14}$$

Substituting (22.10) into (22.14), the current at $n_{0,0}$ is

$$i_{0,0} = (2\cosh\alpha - 2e^{-\alpha})/kr. \tag{22.15}$$

Using the identities, $\cosh x = 1/2(e^x + e^{-x})$ and $sinhx = 1/2(e^x - e^{-x})$, from Euler's formula [468], the current expression $i_{0,0}$ becomes

$$i_{0,0} = 2sinh\alpha/kr. \tag{22.16}$$

Similarly, when $y \neq 0$, the current $i_{0,y}$ at $n_{0,y}$ is

$$i_{0,y} = \frac{(2k+2)V_{0,y} - kV_{0,y+1} - kV_{0,y-1} - V_{1,y} - V_{-1,y}}{kr}, \tag{22.17}$$

and substituting (22.10) into (22.17), the current can be rewritten as

$$\begin{aligned} i_{0,y} = ((2k+2)\cos y\beta - e^{-\alpha}\cos y\beta - e^{-\alpha}\cos y\beta \\ - k\cos(y+1)\beta - k\cos(y-1)\beta)/kr. \end{aligned} \tag{22.18}$$

After applying certain trigonometric identities and simplifications,

$$i_{0,y} = \left((2k+2-2e^{-\alpha})\cos y\beta - 2k\cos y\beta\ \cos\beta\right)/kr. \tag{22.19}$$

The current $i_{0,y}$ at $n_{0,y}$ is

$$i_{0,y} = 2(k+1-e^{-\alpha} - k\cos\beta)\cos y\beta/kr. \tag{22.20}$$

Substituting (22.10) into (22.20) and applying Euler's formula, the current at $n_{0,y}$ is

$$i_{0,y} = \frac{2sinh\alpha\ \cos y\beta}{kr}. \tag{22.21}$$

22.3 Effective Resistance Between Two Nodes

The voltage at $n_{x,y}$ is a function of α and β where the relationship between these two parameters in (22.10) is in terms of k. The voltage at an arbitrary node $n_{x,y}$ is the sum of all β values,

$$V_{x,y} = \int_0^{\pi} F(\beta)\, v_{x,y}(\beta)\, d\beta, \tag{22.22}$$

where $F(\beta)$ is a function that satisfies a current source at $n_{0,0}$, and no current source at $n_{0,y}$ when $y \neq 0$. Thus, all of the current sources other than at $n_{0,0}$ are effectively eliminated [307]. The corresponding current at $n_{x,y}$ is

$$I_{x,y} = \int_0^{\pi} F(\beta)\, i_{x,y}(\beta)\, d\beta. \tag{22.23}$$

The current at $n_{0,0}$, by substituting (22.16) into (22.23), is

$$I_{0,0} = \int_0^{\pi} F(\beta)\, \frac{2sinh\alpha}{k}\, d\beta, \tag{22.24}$$

and the current at $n_{0,y}$, by substituting (22.21) into (22.23), is

$$I_{0,n} = \int_0^{\pi} F(\beta)\, \frac{2sinh\alpha\, \cos y\beta}{k}\, d\beta. \tag{22.25}$$

From inspection, $F(\beta)$ is

$$F(\beta) = \frac{k\, I\, r}{2\pi\, sinh\alpha}, \tag{22.26}$$

to satisfy (22.22) when only one current source located at $n_{0,0}$ is present within the mesh. Substituting (22.26) and (22.12) into (22.22), the voltage at $n_{x,y}$ is

$$V_{x,y} = \frac{k\, I\, r}{2\pi} \int_0^{\pi} \frac{e^{-|x|\alpha}\, \cos y\beta}{sinh\alpha} d\beta. \tag{22.27}$$

22.4 Closed-Form Expression of the Effective Resistance

The effective resistance of a mesh between $n_{0,0}$ and $n_{x,y}$ is

$$R_{x,y} = 2(V_{0,0} - V_{x,y})/I, \tag{22.28}$$

as discussed previously. Substituting (22.27) into (22.28), the effective resistance between $n_{0,0}$ and $n_{x,y}$ is

$$R_{x,y} = \frac{k\, r}{\pi} \int_0^{\pi} \frac{(2 - e^{-|x|\alpha}\, \cos y\beta)}{sinh\alpha} d\beta. \tag{22.29}$$

$R_{x,y}$ is solved by dividing the integral into two, and writing (22.29) as a sum of two integrals, $R_{x,y}/r = R_{1(x,y)} + R_{2(x,y)}$,

$$R_{x,y}/r = \frac{\sqrt{k}}{\pi}\int_0^{\pi} \frac{(1 - e^{-x\sqrt{k}|\beta|}\cos y\beta)}{\beta} d\beta + \frac{k}{\pi}\int_0^{\pi}\left[\frac{1}{\sqrt{(k+1-k\cos\beta)^2-1}} - \frac{1}{\beta\sqrt{k}}\right] d\beta. \tag{22.30}$$

The first integral $R_{1(x,y)}$ is rewritten in terms of the exponential integral Ein(z) [468],

$$Ein(z) = \int_0^{z} \frac{1 - e^{-t}}{t} d\beta, \tag{22.31}$$

and $R_{1(x,y)}$ is

$$R_{1(x,y)} = (1/\pi k) Re\Big\{Ein[\pi(\sqrt{k}x + iy)]\Big\}. \tag{22.32}$$

Expression (22.32) is numerically solved and $R_{1(x,y)}$ is

$$R_{1(x,y)} = \frac{\sqrt{k}}{2\pi}[ln(x^2 + ky^2) + 2(0.57721 + \ln\pi)], \tag{22.33}$$

while the second integral $R_{2(x,y)}$ is determined assuming $k = n + \epsilon$.

$$R_{2(x,y)} = \frac{k}{\pi}\int_0^{\pi}\left(\left((n+1-n\cos\beta)^2-1\right)^{-1/2} - \frac{1}{\beta\sqrt{n}}\right) d\beta + \frac{k}{\pi}\int_0^{\pi}\left(-\epsilon\frac{(1-\cos\beta)(n+1-n\cos\beta)}{((n+1-n\cos\beta)^2-1)^{3/2}} + \frac{\epsilon}{2\beta n\sqrt{n}}\right) d\beta. \tag{22.34}$$

A derivation of (22.34) is provided in the Appendix. The effective resistance between any two arbitrary nodes $R_{x,y}$ within a mesh when k approaches a different constant is listed in Table 22.1. For instance, the effective resistance when $k \to 1$ is

$$R_{x,y}/r = \frac{\sqrt{k}}{2\pi}[ln(x^2 + k\,y^2) + 3.44388] - 0.033425k - 0.0629k(k-1). \tag{22.35}$$

Table 22.1 Closed-form expressions for $R_{1(x,y)}$ and $R_{2(x,y)}$ where $R_{(x,y)}/r = R_{1(x,y)} + R_{2(x,y)}$ when k approaches a constant

$k \rightarrow$	$R_{1(x,y)}$	$R_{2(x,y)}$
1	$\frac{\sqrt{k}}{2\pi}[ln(x^2 + ky^2) + 3.4439]$	$-0.0334k - 0.0629k(k-1)$
2	$\frac{\sqrt{k}}{2\pi}[ln(x^2 + ky^2) + 3.4439]$	$-0.0692k - 0.0202k(k-2)$
3	$\frac{\sqrt{k}}{2\pi}[ln(x^2 + ky^2) + 3.4439]$	$-0.0829k - 0.0093k(k-3)$
4	$\frac{\sqrt{k}}{2\pi}[ln(x^2 + ky^2) + 3.4439]$	$-0.0896k - 0.0047k(k-4)$
5	$\frac{\sqrt{k}}{2\pi}[ln(x^2 + ky^2) + 3.4439]$	$-0.0932k - 0.0026k(k-5)$
10	$\frac{\sqrt{k}}{2\pi}[ln(x^2 + ky^2) + 3.4439]$	$-0.0964k + 0.00021k(k-10)$
100	$\frac{\sqrt{k}}{2\pi}[ln(x^2 + ky^2) + 3.4439]$	$-0.0657k + 0.00016k(k-100)$

22.5 Experimental Results

The accuracy of the effective resistance model is compared to the exact solution (22.29) in Table 22.2. Although the via resistance r_{via} connecting orthogonal metal layers is neglected in the effective resistance model, for practical values of r_{via} (i.e., when r_{via} is between zero and 5 % of the horizontal or vertical resistance [192]), the effective resistance model is in good agreement with the experimental results. The via resistance r_{via} is modeled as $r_{via} = l \cdot r_v$. The maximum error is less than 5 % for $1 < k < 10$ and $0 < l < 0.05$. The error is maximum when the distance between the two nodes is smallest, and the error decreases with greater separation between the nodes of interest if r_{via} is zero. r_{via} is neglected in the expressions. The approximation error in the expressions converges to zero with greater separation between the nodes of interest. When r_{via} is nonzero, the error exhibits a non-monotonic behavior (i.e., the error does not necessarily decrease with greater separation between the nodes of interest).

Practical mesh structures have finite dimensions. Since an infinite mesh is assumed in the development of these expressions, the error of the expressions is compared to four differently sized mesh structures which is listed in Table 22.3 where $k = 1$ and $l = 0$. With increasing separation between the nodes of interest, the error originating from the infinite grid assumption increases as expected. The error is less than 3 % when the nodes of interest are 20 lines away from the boundary.

22.6 Summary

A closed-form expression for the effective resistance of a two layer mesh structure is presented in this chapter.

- The unit resistance of the horizontal and vertical metal lines within a power grid is often different in adjacent orthogonal metal layers due to the difference in the width and thickness of these metal lines

Table 22.2 Accuracy of the closed-form solution for the effective resistance when $r_v = 1\,\Omega$, $r_h = k\,\Omega$, and $r_{via} = l\,\Omega$

			x = 0 y = 1	x = 1 y = 0	x = 10 y = 0	x = 10 y = 10
k = 1		(22.35)	0.5147	0.5147	1.2476	1.3580
	l = 0	SPICE	0.5	0.5	1.2480	1.3580
		Error (%)	3	3	0	0
	l = 0.01	SPICE	0.5033	0.5033	1.2546	1.3642
		Error (%)	2.2	2.2	0.6	0.5
	l = 0.05	SPICE	0.5167	0.5167	1.2819	1.393
		Error (%)	0.4	0.4	2.7	2.6
k = 2		(22.35)	0.6367	0.7928	1.6733	1.9205
	l = 0	SPICE	0.6082	0.7838	1.6737	1.9206
		Error (%)	4.7	1.1	0	0
	l = 0.01	SPICE	0.6098	0.7886	1.6775	1.9299
		Error (%)	4.2	0.6	0.3	0.5
	l = 0.05	SPICE	0.6168	0.8122	1.6969	1.9677
		Error (%)	3.1	3.6	1.4	2.5
k = 5		(22.35)	0.7596	1.3324	2.399	3.0362
	l = 0	SPICE	0.7322	1.3391	2.399	3.0361
		Error (%)	3.7	1.1	0	0
	l = 0.01	SPICE	0.7331	1.3436	2.4013	3.0485
		Error (%)	3.5	0.8	0.1	0.4
	l = 0.05	SPICE	0.7355	1.382	2.4130	3.0995
		Error (%)	3.2	3.7	0.6	2.1
k = 10		(22.35)	0.769	1.928	3.087	4.294
	l = 0	SPICE	0.805	1.952	3.088	4.294
		Error (%)	4.7	1.2	0	0
	l = 0.01	SPICE	0.8057	1.9464	3.0887	4.307
		Error (%)	4.8	1	0.1	0.3
	l = 0.05	SPICE	0.8066	1.9958	3.0972	4.3684
		Error (%)	4.9	3.5	0.3	1.7

- The closed-form expression presented in this chapter uses a parameter k to model the ratio of the horizontal and vertical resistances
- These closed-form expressions provide a fast and accurate solution to the effective resistance of a two layer mesh which can be used to solve a variety of problems found in different disciplines
- Examples of the use of these expressions include *IR* voltage drop analysis of integrated circuits, synchronization and localization of sensor networks, effective chemical distance between bonds, metal mesh interference filters in terahertz physics, and commute and cover times of undirected graphs

Table 22.3 Error induced by the infinite grid approximation for power grids with different sizes

		x = 0 y = 1	x = 2 y = 3	x = 10 y = 0	x = 10 y = 10
	(22.35)	0.5147	0.924	1.2476	1.358
20 × 20	SPICE	0.5015	0.9425	1.3838	1.665
	Error (%)	2.6	2	10.9	22.6
30 × 30	SPICE	0.5006	0.9324	1.3079	1.486
	Error (%)	2.8	0.9	4.8	9.4
40 × 40	SPICE	0.5004	0.929	1.2815	1.4284
	Error (%)	2.9	0.5	2.7	5.2
80 × 80	SPICE	0.5	0.925	1.252	1.367
	Error (%)	3	0.1	0.4	0.7

Chapter 23
Closed-Form Expressions for Fast *IR* Drop Analysis

Several methods have been described for efficient power grid analysis, as presented in Chap. 21; (1) reduce the size of the linear system, (2) iteratively solve the linear system, and (3) apply advanced linear algebraic techniques to exploit the sparse nature of the power grid. Although these algorithms are faster than conventional linear solvers, significant computational time is required to iteratively apply these algorithms. An accurate closed-form expression would effectively solve this problem.

Although the interactions between the power supplies and load circuitry occur globally, these interactions are more prominent among components close to each other. A power supply connection in a multi-voltage system on one side of an IC has little effect on a circuit block at the other side of the IC. Alternatively, current provided by a power network is generally distributed to nearby circuit blocks. This phenomenon is due to the principle of spatial locality [469]. With this principle, a power grid can be partitioned to enhance the overall power grid analysis process.

Uniform current loads are generally assumed in power distribution networks to exploit symmetry in a linear system. In [470], an *IR* drop analysis algorithm is described for a power grid structure with semi-uniform current loads (e.g., uniform load currents are assumed within each quadrant of the distribution network). Closed-form expressions for the maximum *IR* drop are described in [471], assuming a uniform current distribution. Until these results, no closed-form expressions existed to describe the voltage drop at any point in a locally uniform, globally non-uniform power distribution network with non-uniform current loads and non-uniform voltage supplies.

In this chapter, a novel algorithm, Fast Algorithms based on effective resistance for *IR* drop analysis (FAIR), is provided for locally uniform, globally non-uniform power grids with non-uniform current loads and non-uniform voltage supplies. FAIR exploits the impedance characteristics of the power distribution network and the effective impedance between the active circuit blocks to provide these closed-form expressions. The effective impedance between two points in a uniform

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0_23

grid structure has been considered by Venezian in [307], where he formulated the resistance between any two points in a resistive grid. Since no iterations are required to compute the *IR* drop at any particular node, FAIR outperforms previously described techniques with reasonable error. The principle of locality is also applied in FAIR to accelerate the analysis process.

The rest of this chapter is organized as follows. The power grid model is described and the *effective resistance* concept is explained in Sect. 23.1. In Sect. 23.2, FAIR is reviewed for different power grid conditions. The principle of spatial locality is further explained and exploited to accelerate the power grid analysis process in Sect. 23.3. Experimental results are provided in Sect. 23.4. The chapter is summarized in Sect. 23.5.

23.1 Background of FAIR

IR voltage drop analysis in modern integrated circuits requires a massive amount of simulation time and memory due to the large dimensions of on-chip power grids (i.e., millions of nodes). In this chapter, closed-form expressions for determining the *IR* voltage drop and related algorithms are described that exploit the distance between the voltage sources and current loads. The *IR* voltage drop at an arbitrary node depends upon the distance among the voltage sources, current loads, and analysis nodes. These distances are incorporated into the closed-form expressions by the concept of an effective resistance model since the effective resistance between any two nodes in a uniform grid structure depends upon the Euclidean distance between these two nodes and the power grid resistance. The effective resistance supports the development of closed-form expressions for use within the power grid analysis process [472].

A resistive power grid model, depicted in Fig. 23.1, is considered in this chapter to determine *IR* voltage drops. Each wire segment between adjacent nodes has a resistance *R*. Although the power grid model used to develop FAIR is resistive, the results can be generalized for *RL* power grids with no additional computation. The power distribution model is assumed to be composed of infinite parallel lines within the grid, a reasonable assumption in modern integrated circuits due to the large size of the power grid. Venezian in [307] provides an exact solution for the effective resistance between any two points, $N_1(x_1, y_1)$ and $N_2(x_2, y_2)$, in an infinite grid as

$$R_{m,n} = \int_0^{\pi} \frac{(2 - e^{-|m|\alpha}\cos(n\beta) - e^{-|n|\alpha}\cos(m\beta))}{sinh(\alpha)} d\beta. \tag{23.1}$$

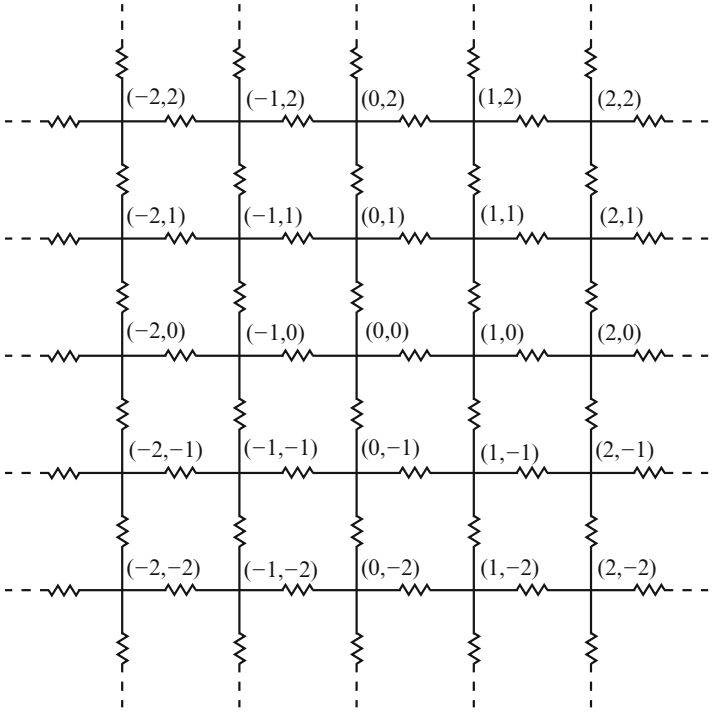


Fig. 23.1 Infinite resistive mesh structure to model a power distribution network

Venezian also provides a closed-form approximation[1] for (23.1) as

$$R_{m,n} = \frac{1}{2\pi} * ln(n^2 + m^2) + 0.51469, \tag{23.2}$$

where

$$m = |x_1 - x_2| \text{ and } n = |y_1 - y_2|. \tag{23.3}$$

α and β are used to rewrite Kirchhoff's node equations as difference equations. The interested reader is urged to read [307] for a complete explanation.

The error of approximation (23.2) is less than 3 % as compared to the exact solution in (23.1). A few examples that demonstrate the validity of (23.2) are listed in Table 23.1. As listed in Table 23.1, the error quickly approaches zero with increasing distance between two points. For instance, the average error when calculating all of the resistances in a 50×50 grid is less than 0.01 %.

[1]The formula has been slightly modified from [307] to produce more accurate results.

Table 23.1 Validity of the effective resistance model

	$R_{1,0}$	$R_{1,1}$	$R_{3,4}$	$R_{5,0}$	$R_{10,10}$
Exact solution (23.1)	0.5	0.636	1.028	1.026	1.358
Approximation (23.2)	0.515	0.625	1.027	1.027	1.358
Error (%)	3	1.8	0.1	0.1	0

23.2 Analytic *IR* Drop Analysis

Four different FAIR-based algorithms are described in this section to determine the *IR* drop at an arbitrary node within a uniform power grid:

- Algorithm I: One power supply and one current load placed arbitrarily within the distribution network.
- Algorithm II: One power supply and multiple current loads placed arbitrarily within the distribution network.
- Algorithm III: Multiple power supplies and one current load placed arbitrarily within the distribution network.
- Algorithm IV: Multiple power supplies and multiple current loads placed arbitrarily within the distribution network.

A simplified model to demonstrate these four cases is illustrated in Fig. 23.2. The voltage supplies and current loads are illustrated as V_{supply} and I_{load}. Algorithm I is the most basic algorithm and is therefore used to explain the other three algorithms. Algorithm IV is the complete algorithm which can be used in the analysis of *IR* drops within practical power grids. The distance between two nodes does not affect the computational complexity of determining the effective impedance between these nodes. The computational complexity of FAIR to determine the *IR* drop at an arbitrary node does not therefore depend upon the size of the power grid.

23.2.1 One Power Supply and One Current Load

In this section, the *IR* voltage drop at an arbitrary node $Node_1$, shown in Fig. 23.3a, is determined when one power supply and one current load exist within the power grid. The power grid model shown in Fig. 23.3a reduces to an effective resistance model, as illustrated in Fig. 23.3b. The effective resistance between N_{supply} and $Node_1$, $Node_1$ and N_{load}, and N_{supply} and N_{load} is denoted, respectively, as R_{sn}, R_{nl}, and R_{sl}. These resistances are determined using (23.2). The voltage at N_{load} is

$$V_{load} = V_{supply} - I_{load} * R_{sl}. \tag{23.4}$$

After determining the voltage at N_{load} (see Fig. 23.3b), the voltage at $Node_1$ can be found as follows. Assume that all of the load current I_{load} flows from N_{supply} to

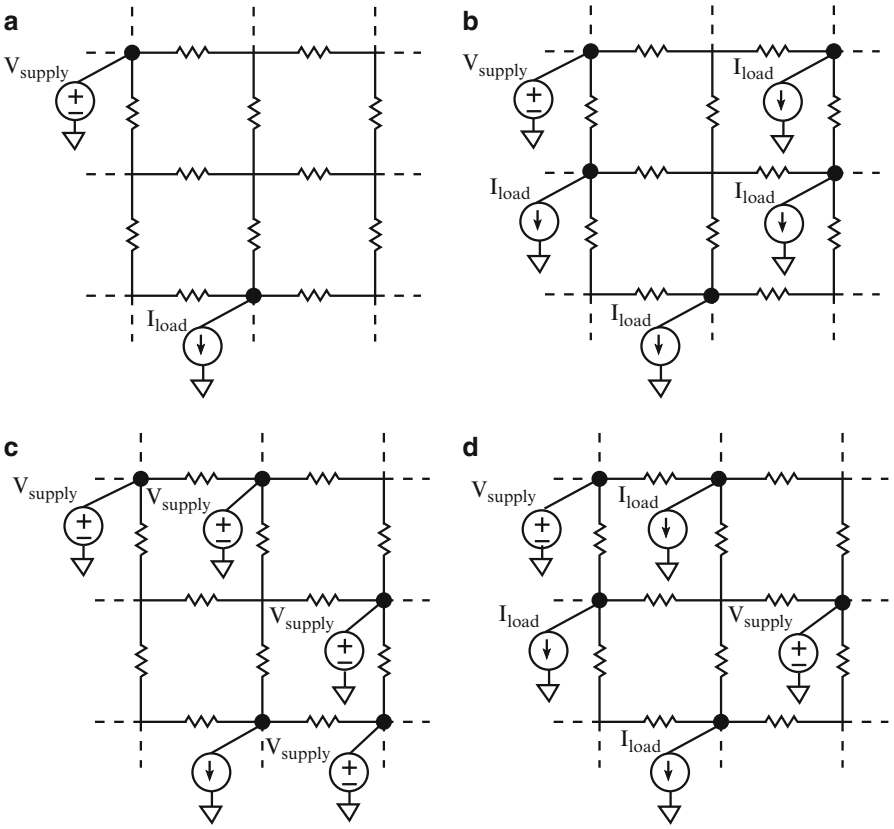


Fig. 23.2 Simplified power grid models for FAIR; (**a**) one voltage source and one current load, (**b**) one voltage source and multiple current loads, (**c**) multiple voltage sources and one current load, (**d**) multiple voltage sources and multiple current loads

N_{load} along the path R_{sn}–Node$_1$–R_{nl}. Since the voltage at N_{supply} and N_{load} is known a priori, the voltage at Node$_1$, V_{Node_1}, can be found with respect to either N_{supply} or N_{load}. V_{Node_1} is

$$V_{Node_1} = V_{supply} - I_{load} * R_{sn} \tag{23.5}$$

with respect to N_{supply} and

$$V_{Node_1} = V_{load} + I_{load} * R_{nl} \tag{23.6}$$

with respect to N_{load}. The voltage at Node$_1$ is calculated with the arithmetic mean of the voltages found using (23.5) and (23.6). The voltage at Node$_1$ is therefore

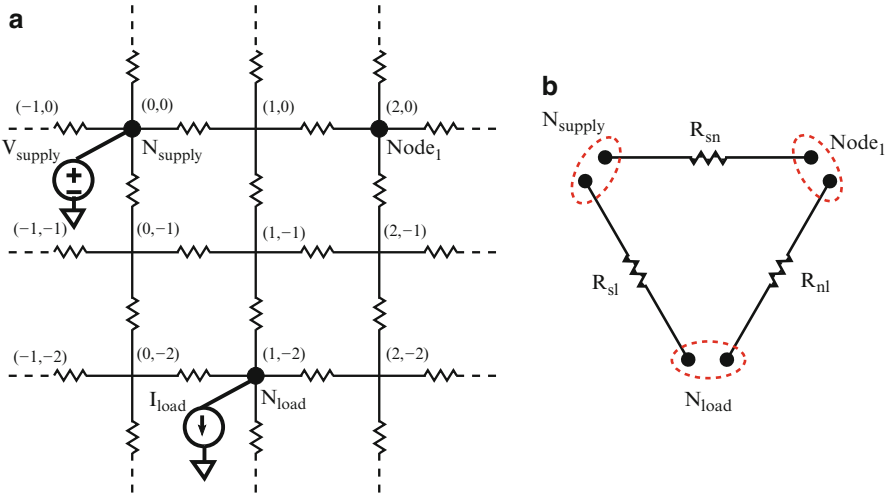


Fig. 23.3 Power distribution grid model for Algorithm I; (**a**) one power supply connected at (0,0) and one current load connected at (1,−2), (**b**) corresponding reduced effective resistance model

$$V_{Node_1} = [V_{supply} + V_{load} + I_{load} * (R_{nl} - R_{sn})]/2. \tag{23.7}$$

Substituting (23.4) into (23.7), the voltage at Node_1 can be written as

$$V_{Node_1} = [2 * V_{supply} + I_{load} * (R_{nl} - R_{sn} - R_{sl})]/2. \tag{23.8}$$

The *IR* voltage drop at Node_1 is equal to $\text{V}_{supply} - -\text{V}_{Node_1}$. The *IR* voltage drop can therefore be written as

$$IR_{Node_1} = I_{load} * (R_{sn} + R_{sl} - R_{nl})/2. \tag{23.9}$$

Pseudo-code of the algorithm to determine the voltage at an arbitrary node within a power grid with one current load and one voltage supply is summarized in Fig. 23.4 (Algorithm I).

23.2.2 One Power Supply and Multiple Current Loads

In this section, the *IR* voltage drop at an arbitrary node within a power distribution network is determined when one power supply and multiple current loads exist within a grid, as shown in Fig. 23.5. Since the current loads are assumed to be ideal current sources, the principle of superposition is performed to provide a closed-form expression for the *IR* voltage drop. Superposition is possible since linear current loads are used to model the active circuit structures. By using superposition for each individual current load, the voltage at Node_1 can be formulated as

IR Drop: One Power Supply and One Current Load
1. Given: Supply voltage (V_{supply}), load current (I_{load})
   Locations of voltage supply (N_{supply}),
   current load (N_{load}), and $Node_1$.
2. Calculate the effective resistances between
   a) N_{supply} and $Node_1$, R_{sn}
   b) $Node_1$ and N_{load}, R_{nl}
   c) N_{supply} and N_{load}, R_{sl}.
3. Calculate the voltage at N_{load}, (23.4).
4. Calculate the voltage at $Node_1$ V_{Node_1}, (23.7).
5. Calculate the *IR* drop at $Node_1$, (23.9).

Fig. 23.4 Algorithm I. *IR* voltage drop at an arbitrary node within a power grid with one power supply and one current load

$$V_{Node_1} = V_{supply} - \frac{1}{2}\sum_{i=1}^{n}[I_{load(i)} * (R_{sn} + R_{sl(i)} - R_{nl(i)})], \tag{23.10}$$

and the corresponding *IR* voltage drop at $Node_1$ is

$$IR_{Node_1} = \frac{1}{2}\sum_{i=1}^{n}[I_{load(i)} * (R_{sn} + R_{sl(i)} - R_{nl(i)})], \tag{23.11}$$

where n is the number of current loads, $I_{load(i)}$ is the ith current load, $R_{sl(i)}$ is the effective resistance between N_{supply} and the ith current load, and $R_{nl(i)}$ is the effective resistance between $Node_1$ and the ith current load within the power grid. Pseudo-code of the algorithm to determine the *IR* voltage drop at an arbitrary point when one voltage supply and multiple current loads are connected to a power distribution grid is provided in Fig. 23.6 (Algorithm II).

23.2.3 Multiple Power Supplies and One Current Load

In this section, the *IR* voltage drop at an arbitrary node within a power distribution network is determined for multiple voltage sources and one current load, as shown in Fig. 23.7a. In Sect. 23.2.2, superposition is used to analyze the voltage drop contribution to $Node_1$ from each individual current load. To consider multiple voltage supplies, superposition cannot be used in a straightforward manner to consider each individual voltage supply because the voltage supplies are replaced with short circuit equivalents whereas the current loads are replaced with open circuit equivalents.

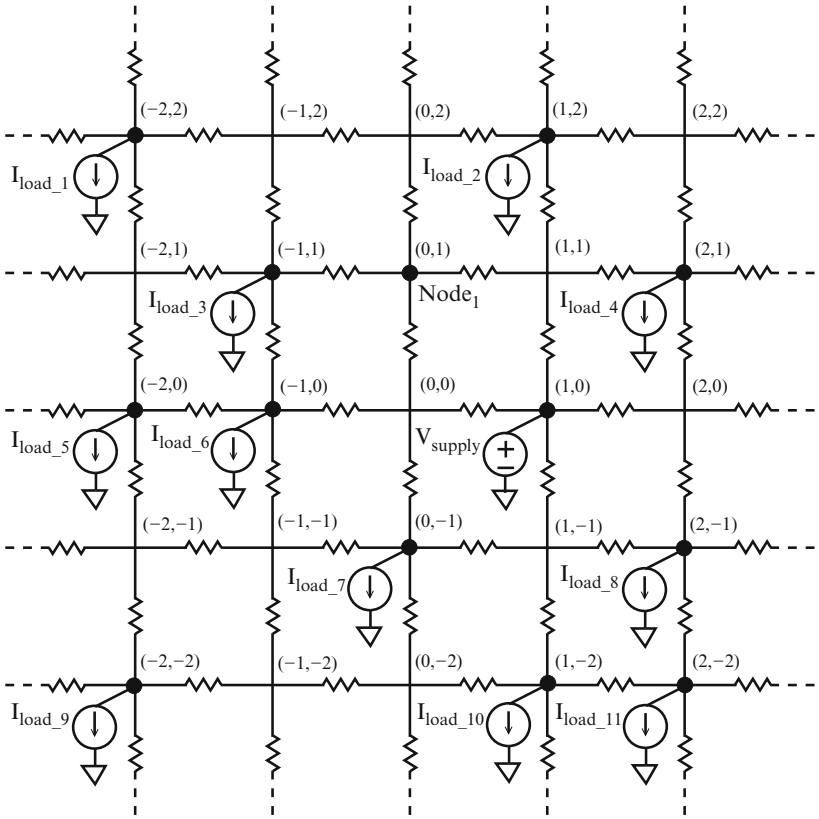


Fig. 23.5 Power distribution grid for Algorithm II, when one power supply is connected to node (1,0) and multiple current loads model the load circuits connected at various nodes within the power distribution grid

To apply superposition to the voltage supplies, the voltage supplies are replaced with equivalent current sources. The current that each individual voltage source contributes to the load depends upon the effective resistance between $N_{supply(i)}$ and N_{load}. Since the location of the voltage supplies and the current load is known a priori, the current delivered by these equivalent current supplies is approximately

$$I_{source}(i) = I_{load} * \frac{Gx_i}{\sum_{i=1}^{n} Gx_i}, \tag{23.12}$$

where

$$Gx_i = 0.26 * G_i^{2.3167} - 0.11527. \tag{23.13}$$

I_{source} is the equivalent current source to replace the ith voltage supply, G_i is the equivalent conductance between the ith voltage supply and current load, and Gx_i is

IR Drop: One Power Supply and Multiple Current Loads

1. Given: Supply voltage (V_{supply}), load currents ($I_{load(i)}$)
   Locations of voltage supply (N_{supply}),
   current loads ($N_{load(i)}$), and $Node_1$.
2. **for** each current load, $I_{load(i)}$, **do**
3. Remove all other $I_{load(k)}$ where k≠i,
4. Calculate the effective resistances between
   a) N_{supply} and $Node_1$, R_{sn}
   b) $Node_1$ and $N_{load(i)}$, $R_{nl(i)}$
   c) N_{supply} and $N_{load(i)}$, $R_{sl(i)}$.
5. Calculate the voltage at $N_{load(i)}$, (23.4).
6. Calculate the *IR* drop at $Node_1$ due to $I_{load(i)}$, (23.9).
7. Calculate the total *IR* drop at $Node_1$ by summing
   all *IR* voltage drops due to all individual current loads, (23.11).
8. Calculate the voltage at $Node_1$, V_{node_1}, (23.10).

Fig. 23.6 Algorithm II. *IR* voltage drop at arbitrary node $Node_1$ within a power grid with one power supply and multiple current loads, as shown in Fig. 23.5

the modified conductance parameter determined from nonlinear least square curve fitting. The maximum root mean square error (RMSE) for the individual current contributions of the voltage sources determined with (23.12) is less than 0.005 as compared to SPICE.

After all but one of the voltage supplies are replaced with equivalent current sources, as illustrated in Fig. 23.7b, the *IR* voltage drop problem becomes similar to the problem discussed in Sect. 23.2.2 where the power grid has one voltage supply and multiple current loads. The primary difference is that the equivalent current sources supply current to the distribution grid whereas, as described in Sect. 23.2.2, all of the current loads demand current from the power grid.

The *IR* voltage drop at an arbitrary node $Node_1$ in the power grid with multiple voltage sources and one current load is

$$\begin{aligned} IR_{Node_1} =& I_{load} * (R_{sn(1)} + R_{sl(1)} - R_{nl})/2 \\ & - \frac{1}{2}\sum_{i=2}^{n}[I_{supply(i)} * (R_{sn(1)} + R_{sl(i)} - R_{nl(i)})], \end{aligned} \tag{23.14}$$

and the voltage at $Node_1$ is

$$\begin{aligned} V_{Node_1} =& V_{supply(1)} - I_{load} * (R_{sn(1)} + R_{sl(1)} - R_{nl})/2 \\ & + \frac{1}{2}\sum_{i=2}^{n}[I_{supply(i)} * (R_{sn(1)} + R_{sl(i)} - R_{nl(i)})]. \end{aligned} \tag{23.15}$$

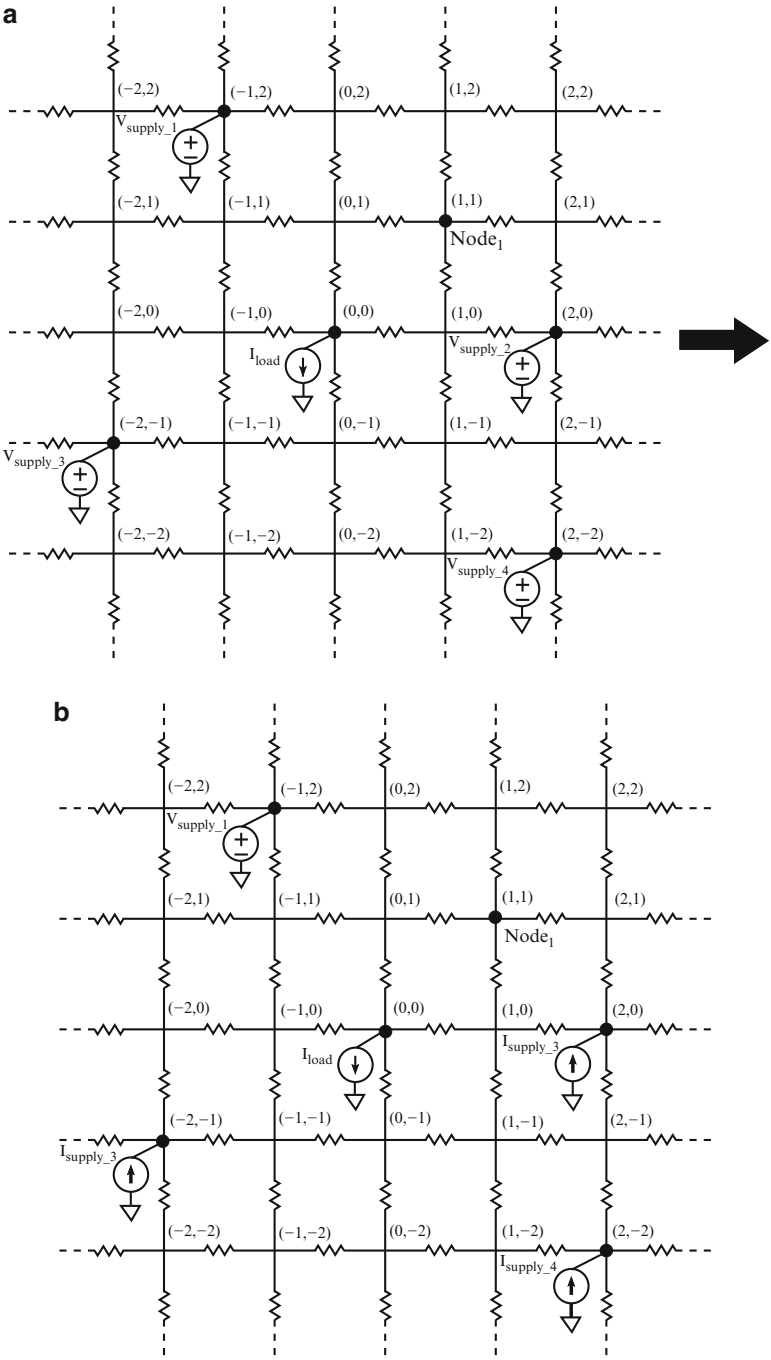


Fig. 23.7 Power distribution grid model for Algorithm III; (**a**) multiple power supplies are connected to several nodes and a current load is connected at (0,0), (**b**) all but one of the voltage sources are replaced with an equivalent current source

IR Drop: Multiple Power Supplies and One Current Load

1. Given: Supply voltage (V_{supply}), load current (I_{load})
   Locations of voltage supplies ($N_{supply(i)}$),
   current load (N_{load}), and $Node_1$.
2. **for** each voltage supply, $V_{supply(i)}$, **do**
3. Calculate the effective resistances between $N_{supply(i)}$ and I_{load}, R_i.
4. **for** each voltage supply, $V_{supply(i)}$, where i≠1, **do**
5. Find the corresponding current source, $I_{supply(i)}$, (23.12).
6. Replace $V_{supply(i)}$ with $I_{supply(i)}$.
7. Remove all current supplies, $I_{supply(i)}$.
8. Calculate the effective resistances between
   a) $N_{supply(1)}$ and $Node_1$, R_{sn}
   b) $Node_1$ and N_{load}, R_{nl}
   c) $N_{supply(1)}$ and N_{load}, R_{sl}.
9. Calculate the *IR* drop at $Node_1$ due to I_{load}, (23.9).
10. **for** each current supplies, $I_{supply(i)}$, **do**
11. Remove all other current supplies, $I_{supply(k)}$, where k≠1.
12. Calculate the effective resistances between
    a) $N_{supply(1)}$ and $Node_1$, R_{sn}
    b) $Node_1$ and $N_{supply(i)}$, $R_{nl(i)}$
    c) $N_{supply(1)}$ and $N_{supply(i)}$, $R_{sl(i)}$.
13. Calculate the voltage difference at $Node_1$ due to $I_{supply(i)}$, (23.9).
14. Calculate the total *IR* drop at $Node_1$ by subtracting
    the result of step 13 from the result of step 9, (23.14).
15. Calculate the voltage at $Node_1$, V_{node_1}, 23.15.

Fig. 23.8 Algorithm III. *IR* voltage drop at arbitrary node $Node_1$ in a power grid with multiple power supplies and one current load, as shown in Fig. 23.7a

Pseudo-code of the algorithm to determine the *IR* voltage drop at an arbitrary node within a power grid with multiple voltage supplies and one current load is summarized in Fig. 23.8 (Algorithm III).

23.2.4 Multiple Power Supplies and Multiple Current Loads

In this section, the *IR* voltage drop at an arbitrary node within a power distribution network is determined when multiple voltage supplies and multiple current loads exist, as shown inFig. 23.9a. To determine the *IR* voltage drop for this system, superposition is applied in two steps. First, the current that each individual voltage supply contributes to each individual current load is determined by removing all but one of the current loads and applying (23.12) to determine the current contribution of each voltage supply to each current load. After determining the individual current contributions, the equivalent current source of a voltage supply is

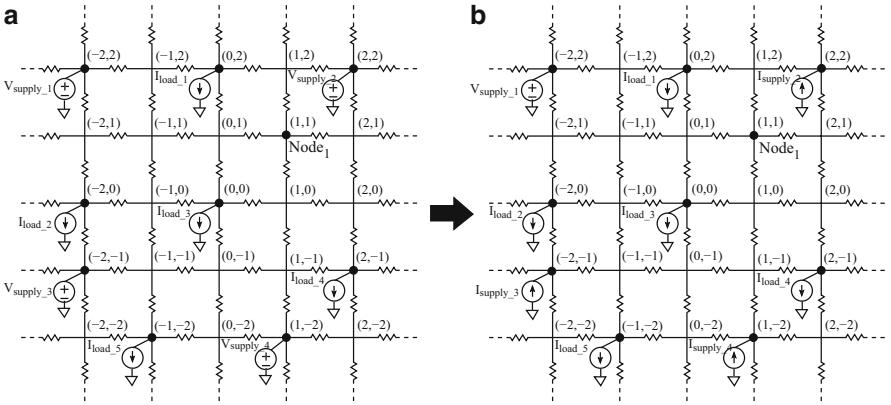


Fig. 23.9 Power distribution grid model for Algorithm IV; (**a**) multiple power supplies and current loads are connected to several nodes, (**b**) all but one of the voltage sources are replaced with an equivalent current source

$$I_{source}(i) = \sum_{j=1}^{m} I_{source(i,j)}, \tag{23.16}$$

where m is the number of current loads, $I_{source(i)}$ is the equivalent current source of the ith voltage supply, and $I_{source(i,j)}$ is the current contribution of the ith voltage supply to the jth current load. Since the total current sourced by the voltage supplies is equal to the total current sunk by the current sources, the following expression is satisfied,

$$\sum_{i=1}^{n} I_{source}(i) = \sum_{j=1}^{m} I_{load(j)}. \tag{23.17}$$

All but one of the voltage supplies are replaced with an equivalent current source, as illustrated in Fig. 23.9b. The *IR* voltage drop at an arbitrary node within a power distribution network is

$$\begin{aligned} IR_{Node_1} =& \frac{1}{2}\sum_{i=1}^{m}[I_{load(i)} * (R_{sn(1)} + R_{sl(1)} - R_{nl})] \\ & - \frac{1}{2}\sum_{i=2}^{n}[I_{supply(i)} * (R_{sn(1)} + R_{sl(i)} - R_{nl(i)})], \end{aligned} \tag{23.18}$$

and the corresponding voltage at Node_1 is

IR Drop: Multiple Power Supplies and Multiple Current Loads
1. Given: Supply voltage (V_{supply}), load currents ($I_{load(j)}$)
   Locations of voltage supplies ($N_{supply(i)}$),
   current loads ($N_{load(j)}$), and $Node_1$.
2. **for** each voltage supply, $V_{supply(i)}$, **do**
3. **for** each current load, $I_{load(j)}$, **do**
4. Calculate the effective resistances between
   $N_{supply(i)}$ and $I_{load(j)}$, $R_{(i,j)}$.
5. **for** each voltage supply, $V_{supply(i)}$, where i≠1, **do**
6. **for** each current load, $I_{load(j)}$, **do**
7. Find the corresponding current, $I_{supply(i,j)}$, (23.12).
8. Sum up $I_{supply(i,j)}$ for all j to calculate $I_{supply(i)}$, (23.16).
9. Replace $V_{supply(i)}$ with $I_{supply(i)}$.
10. **for** each current load, $I_{load(j)}$, **do**
11. Remove all current supplies, $I_{supply(i)}$.
12. Calculate the effective resistances between
    a) $N_{supply(1)}$ and $Node_1$, R_{sn}
    b) $Node_1$ and $N_{load(j)}$, $R_{nl(j)}$
    c) $N_{supply(1)}$ and $N_{load(j)}$, $R_{sl(1,j)}$.
13. Calculate the *IR* drop at $Node_1$ due to all $I_{load(j)}$, (23.9).
14. **for** each current supply, $I_{supply(i)}$, **do**
15. Remove all other current supplies, $I_{supply(k)}$, where k≠1.
16. Remove all current loads, $I_{load(j)}$.
17. Calculate the effective resistances between
    a) $N_{supply(1)}$ and $Node_1$, R_{sn}
    b) $Node_1$ and $N_{supply(i)}$, $R_{nl(i)}$
    c) $N_{supply(1)}$ and $N_{supply(i)}$, $R_{sl(i)}$.
18. Calculate the voltage difference at $Node_1$ due to $I_{supply(i)}$, (23.9).
19. Calculate the total *IR* drop at $Node_1$ by subtracting
    the result of step 18 from the result of step 13, (23.18).
20. Calculate the voltage at $Node_1$, V_{node_1}, (23.19).

Fig. 23.10 Algorithm IV. *IR* voltage drop at an arbitrary node $Node_1$ within a power grid with multiple power supplies and current loads, as shown in Fig. 23.9a

$$V_{Node_1} = V_{supply(1)} - \frac{1}{2}\sum_{i=1}^{m}[I_{load(i)} * (R_{sn(1)} + R_{sl(1)} - R_{nl})] + \frac{1}{2}\sum_{i=2}^{n}[I_{supply(i)} * (R_{sn(1)} + R_{sl(i)} - R_{nl(i)})], \quad (23.19)$$

where m is the number of current loads and n is the number of voltage supplies. Pseudo-code of the algorithm to determine the *IR* voltage drop at an arbitrary node for multiple voltage supplies and current loads is provided in Fig. 23.10 (Algorithm IV).

23.3 Locality in Power Grid Analysis

Practical power grids in high performance integrated circuits can be treated as locally uniform, globally non-uniform. To apply these algorithms to the analysis of practical power grids, the principle of spatial locality [300, 435, 469] is applied. This principle for a resistive power grid is described in Sect. 23.3.1. The effect of utilizing spatial locality on the power grid analysis process is explained in Sect. 23.3.2. In Sect. 23.3.3, the principle of spatial locality is exploited and integrated into FAIR. The advantages of utilizing spatial locality in the power grid analysis process are also explored. An error correction technique is described in Sect. 23.3.4.

23.3.1 Principle of Spatial Locality in a Power Grid

Flip-chip packages are widely used in high performance integrated circuits, increasing the number of voltage supply connections to the integrated circuit. C4 bumps (controlled collapse chip connect) connect the integrated circuit to external circuitry from the top side of the wafer using solder bumps. A large number of power supply connections are provided to the power grid via these C4 bumps. Most of the current to the load devices is provided from those power supply connections in close proximity due to the smaller effective impedance. This phenomenon can be explained using the principle of spatial locality in a power grid [300, 435, 469].

A power grid for a flip-chip package with C4 connections is illustrated in Fig. 23.11. To exemplify the principle of spatial locality in a power grid, two current loads are connected to the power grid as depicted in Fig. 23.11 to analyze the current contributions from each supply connection. With only one current load L_1 connected to the power grid, the current contributed from each of the C4 connections to L_1 is as illustrated in Fig. 23.12. Most of the current is provided by the close power supplies. The current contribution of a supply connection decreases significantly with distance. The current contribution from most of the supply connections within the third ring is less than 1 % of the total load current. The current contribution from each supply connection is also analyzed with only the current load L_2 connected to the power grid. More than 40 % of the total current is provided by the closest power supply connection, V_{21}. The current contribution of all of the connections is illustrated in Fig. 23.13. Most of the power supply connections within the third ring contribute less than 1 % of the current to the load. When the load circuit is close to the boundary of the power supply ring, the current provided by some power supply connections within the outer ring can be higher than the current contributed by the connections forming the inner ring. For instance, since L_2 is close to the first ring boundary, the current contribution from V_{27} which is in the second ring is higher than the current contributed by V_{16} which is in the first ring. The reason is that V_{27} is physically closer to L_2 than V_{16}. The principle of locality is therefore applicable to power grids with multi-power supply connections such as flip-chip

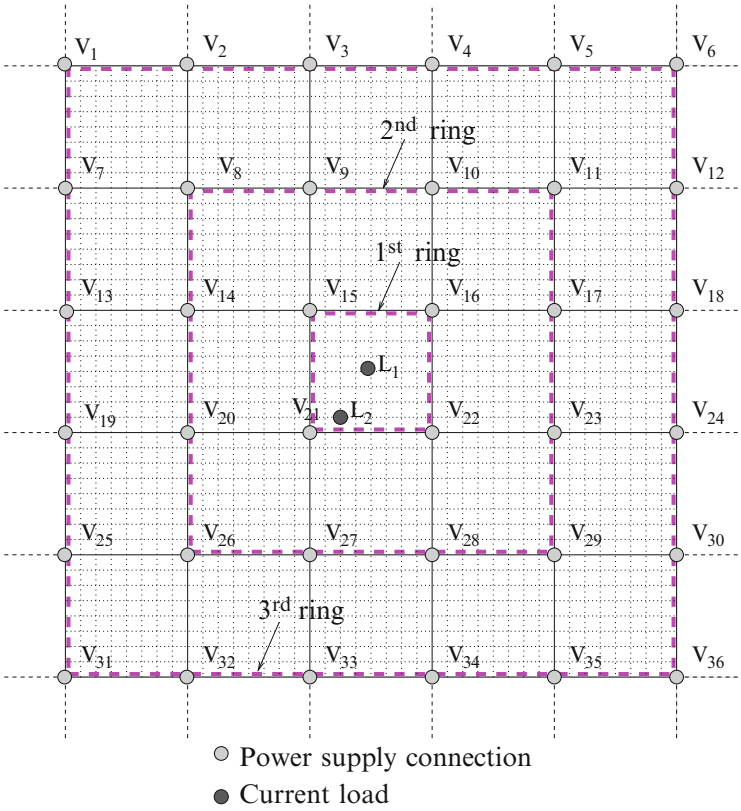


Fig. 23.11 A portion of a typical power grid with C4 bumps illustrated with *light dots* and load devices with *dark dots*. Most of the current to the load devices, L_1 and L_2, is provided by the supply connections forming the first ring. Power supply connections within the third ring contribute less than 1 % of the total current to these load devices

packages. Locality can also be applied to power distribution networks with tens of on-chip voltage regulators. In this case, most of the current is supplied by the closest on-chip power supplies rather than the closest C4 connections.

23.3.2 Effect of Spatial Locality on Computational Complexity

The computational complexity of the power grid analysis process can be significantly reduced by introducing spatial locality since the voltage fluctuations at a specific node are primarily determined by the power grid impedance and placement of those supply connections in close proximity [469]. The complex global interactions among distant circuit components, which typically have a negligible effect on the *IR* drop, is not considered with spatial locality.

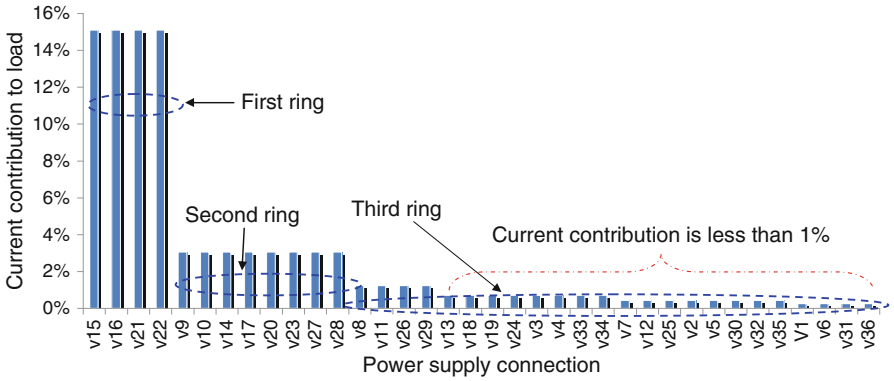


Fig. 23.12 Per cent current provided to the current load L_1 placed in the middle of a uniform power grid from the power supplies, as illustrated in Fig. 23.11. Note that most of the current is provided by the power supplies within the closest two rings whereas the current provided by the power supplies within the third ring is less than 1 %

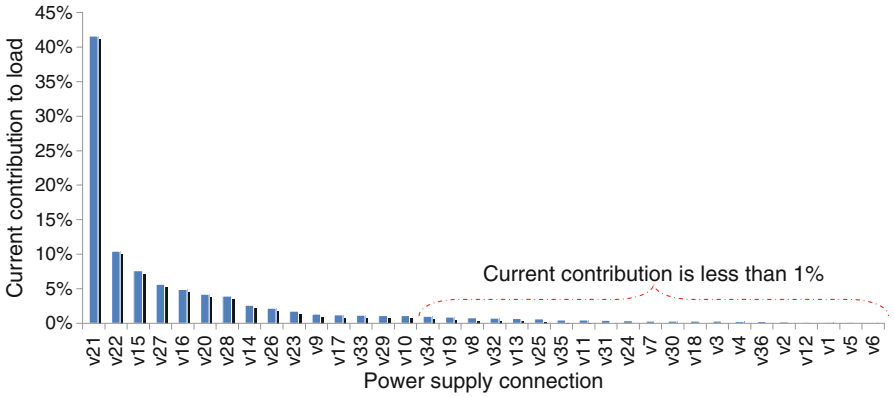


Fig. 23.13 Per cent current provided to the current load L_2 placed within a uniform power grid via the power supply connections, as illustrated in Fig. 23.11. Note that more than 40 % of the current is provided by the closest power supply connection, V_{21}. The current contribution of a supply connection is significantly lower with distance

23.3.3 Exploiting Spatial Locality in FAIR

An infinite grid is assumed to determine the effective resistances of a finite power grid in these algorithms. This assumption introduces a significant approximation error to the power grid analysis process with small power grids. When the size of the power grid increases, the error converges to zero. The maximum error for various grid sizes is illustrated in Fig. 23.14. When the grid size is larger than 30×30, the

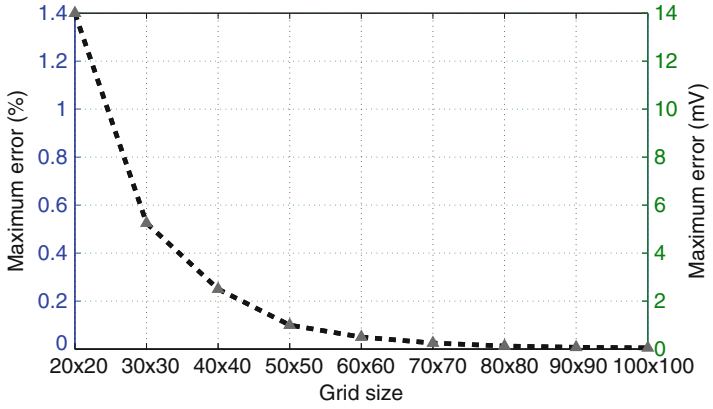


Fig. 23.14 Maximum error for different grid size. The per cent error in terms of the supply voltage and the absolute error is shown, respectively, in the *left* and *right axes*. Note that the error decreases significantly with increasing grid size

error is less than 0.5 % of the supply voltage. Modern power grids can contain more than a million nodes. The size of these grids typically exceed 1000×1000, making the approximation error effectively zero.

A power grid should be divided into smaller partitions [435, 469] to exploit the principle of spatial locality. Each partition is analyzed individually and a complete solution is obtained by combining the results of each partition. The ideal solution is obtained with only one partition (i.e., no partitioning) considering all of the interactions between each power supply connection and load circuit. This approach suffers from long computational time. The fastest solution is obtained when the power grid is divided into the smallest possible partitions. This analysis, however, introduces significant error. A tradeoff in partition size therefore exists between computational complexity and accuracy.

For each partition, the error is smallest in the middle of the partition and increases towards the boundaries. A partitioning approach divides the power grid into several overlapping windows where only the middle of each window is analyzed. The boundaries of each partition overlap with the adjacent partitions. This method of overlapping windows has been shown to be effective in industrial power grids to accelerate the power grid analysis process [469]. Some redundancy is applied during the analysis process which significantly reduces the error from spatial locality. This partitioning approach is illustrated in Fig. 23.15 where a flip-chip power grid with several C4 connections is partitioned into four overlapping windows. Each window consists of an analysis partition and an overlapping boundary. The size of each partition and overlapping boundary is chosen sufficiently large to minimize the error in the analysis partition. A tradeoff therefore also exists between computational complexity and induced error in the size of the overlapping boundary. When the size of the overlapping boundary is sufficiently large, the effect of the adjacent power grid partition is minimized. Alternatively, the computational complexity of

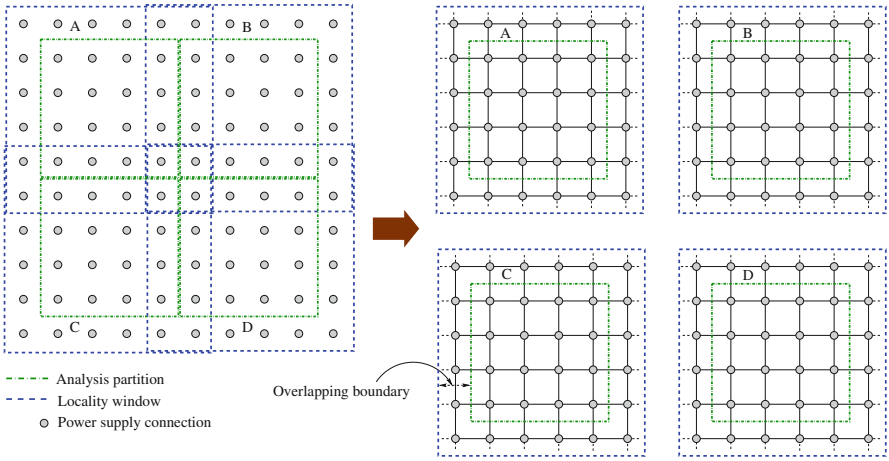


Fig. 23.15 Power grid divided into smaller partitions. Each partition consists of an analysis partition and overlapping boundary

the analysis process decreases when the size of the overlapping boundary is small. In this chapter, the size of each partition and the overlapping boundary is maintained larger than 100×100 and 20, respectively, making the approximation error less than 0.1 %. The partitioning approach also considers the locally uniform, globally non-uniform nature of the power grid. Each partition is treated as a uniform power grid. Different partitions can exhibit different impedance characteristics. Parallel computation of the *IR* drop analysis algorithm can also be implemented, further reducing the overall runtime of the analysis process.

23.3.4 Error Correction Windows

Several error reduction techniques can be used within this algorithm. One technique is the use of error correction windows, as illustrated in Fig. 23.16, where the supply connections, load circuits, error correction window, and analysis window are shown with, respectively, light dots, dark dots, light pink box, and green box. Since the voltage at each supply connection is known a priori, the induced error at a supply connection is the difference between the ideal supply voltage and the voltage determined from FAIR. This error primarily occurs when determining the current contribution of a power supply connection to the power grid. A correlation exists between the error at the supply connection node and the nodes within close proximity of the power supply. The error is generally maximum at the supply connection node and is lower with increasing distance from the supply connection. An error correction window for each supply connection is constructed based upon

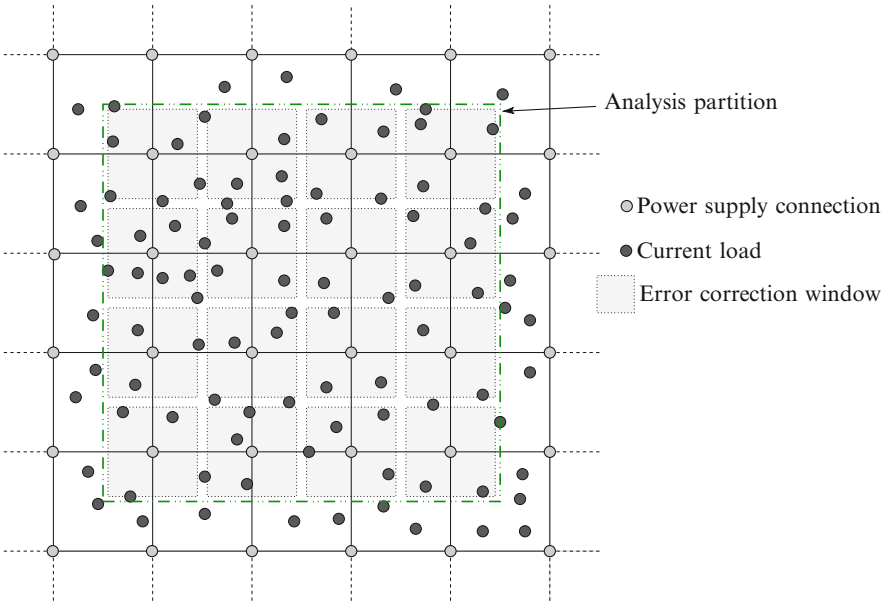


Fig. 23.16 Partition of a resistive flip-chip power grid with supply connections and load circuits denoted with, respectively, *light* and *dark dots*. The error correction windows are shown with small *light pink boxes* around each supply connection node

the error at the supply connection node. With the introduction of this error correction technique, the maximum error reduces to less than 0.3 % of the supply voltage, as described in Sect. 23.4.

23.4 Experimental Results

The validity of these algorithms to efficiently analyze a power grid for several scenarios is presented in this section. The algorithms are implemented using MATLAB and the computations are performed on a Unix workstation with a 3 GHz CPU and 10 GB of RAM. The accuracy of Algorithms I–III is compared with SPICE simulations. For simplicity, the resistance between two adjacent nodes in the power grid is assumed to be 1Ω and the voltage sources are assumed to be 1 V. The current loads are between 1 and 100 mA.

The validity of FAIR for one voltage supply and one current load is analyzed with a 1 V power supply connected at $N_{3,3}$ and the load sinking 100 mA at $N_{5,4}$. The maximum error is 1.44 mV, less than 0.2 % of the supply voltage. The error of the corresponding node voltages as compared to SPICE is listed in Table 23.2. The *light gray* box is the supply node and the *dark gray* box is the node connected to the current load.

Table 23.2 Error of Algorithm I as compared to SPICE. The voltage supply is connected at $N_{3,3}$ (*light gray*) and the load device is connected at $N_{5,4}$ (*dark gray*). The maximum error is less than 0.2 % of the supply voltage

	1	2	3	4	5	6	7	8
1	−0.12	−0.05	0.46	−0.27	−0.49	−0.26	−0.125	−0.15
2	−0.09	−0.55	0.79	0.152	−0.68	−0.37	−0.554	−0.14
3	0.33	0.62	0	1.13	−0.52	0.52	0	−0.26
4	−0.31	−0.83	0.21	−1.44	−0.31	−0.93	−0.64	−0.41
5	−0.25	−0.27	0.37	0.24	−1.10	0.24	−0.22	−0.38
6	−0.18	−0.04	0.18	−0.04	−0.77	−0.25	−0.18	−0.30
7	−0.13	−0.01	0	−0.23	−0.50	−0.36	−0.28	−0.30
8	−0.14	−0.04	−0.08	−0.27	−0.32	−0.34	−0.34	−0.34

Table 23.3 Error of Algorithm II as compared to SPICE. The voltage supply is connected at $N_{4,4}$ (*light gray*) and the load devices are connected at $N_{1,7}$, $N_{2,3}$, $N_{6,6}$, and $N_{2,7}$ (*dark gray*). The maximum error is less than 0.2 % of the supply voltage

	1	2	3	4	5	6	7	8
1	−0.17	−0.24	−0.06	0.07	−0.16	−0.38	−0.19	−0.30
2	0.01	−0.40	−0.06	0.32	−0.20	−0.14	−0.36	−0.03
3	−0.23	−0.25	−0.80	0.60	−0.64	−0.25	−0.18	−0.02
4	0.28	0.11	0.76	0	0.69	0.18	0.05	0.02
5	0.01	−0.40	−0.64	0.75	−0.50	−0.42	−0.06	−0.04
6	0	−0.49	−0.08	0.28	−0.39	−0.31	−0.36	−0.23
7	−0.40	−0.25	−0.32	0.12	−0.04	−0.45	−0.07	−0.13
8	−0.05	−0.35	1.11	0.08	−0.02	−0.29	−0.16	−0.17

Nodal voltage analysis of a power grid with one voltage supply and multiple current loads is evaluated when the voltage supply is connected to $N_{4,4}$ and four current loads are arbitrarily placed at $N_{1,7}$, $N_{2,3}$, $N_{6,6}$, and $N_{2,7}$. In this case, each load sinks 25 mA from the power grid. The error of Algorithm II as compared to SPICE is listed in Table 23.3. The maximum error of Algorithm II as compared to SPICE is 1.1 mV (less than 0.2 %).

The validity of Algorithm III is analyzed with three voltage supplies and a current load connected arbitrarily to the power grid. The current load sinks 100 mA current and the voltage supplies are 1 V. The maximum voltage drop is less than 100 mV. The error of Algorithm III as compared to SPICE is listed in Table 23.4. The maximum error is 1.41 mV which is smaller than 0.2 % of the supply voltage.

The complete algorithm, Algorithm IV, is validated for a larger power grid with multiple voltage supplies and multiple current loads arbitrarily placed within a 17×17 power grid. The results of Algorithm IV are compared with SPICE and the error is listed in Table 23.5. The current loads sink between 1 and 100 mA from the grid and the voltage supplies are 1 V. The maximum error is 4.03 mV which is less than 0.5 % of the supply voltage. When error correction is applied to Algorithm IV,

Table 23.4 Error of Algorithm III as compared to SPICE. Power supplies are connected at $N_{1,2}$, $N_{6,8}$, and $N_{8,1}$ (*light gray*) and current load is connected at $N_{5,4}$ (*dark gray*). The maximum error is 1.41 mV (less than 0.2 % of the power supply voltage)

	1	2	3	4	5	6	7	8
1	1.33	0.67	0.75	0.62	0.31	0.6	0.71	0
2	1.24	1.33	1.11	0.87	−0.07	0.54	0.23	0.63
3	1.21	0.49	0.83	1.41	−0.61	1.2	0.45	0.47
4	0.77	0.32	−0.09	−0.58	0.44	−0.63	−0.27	0.15
5	0.67	0.62	0.65	1.36	−0.62	1.42	0.62	0.35
6	0.74	0.69	0.7	0.62	−0.3	0.8	0.57	0.41
7	0.65	0.68	0.6	0.4	−0.15	0.78	0.27	0.42
8	0.68	0.7	0.6	0.68	0.71	0.34	0.87	0.72

the maximum error is reduced to 2.35 mV, which is less than 0.3 % of the supply voltage, as listed in Table 23.6. Note that the nodes are shown in italic font if error correction has been applied.

The computational complexity of the random walk method is *O(LMN)* [437], where N is the number of nodes without power supply connections, L is the number of steps in a single walk, and M is the number of walks to determine the voltage at a node. The random walk method is faster for flip chip power grids as compared to wire bonded power grids or power grids with a limited number of on-chip power supplies since M is significantly larger. However, the computational complexity of the random walk method can be decreased with hierarchical methods [437, 473], although the property of locality is sacrificed.

Alternatively, the computational complexity of FAIR is linear with the size of the power grid. Since no iterations are required (i.e., $L = 1$) and the voltage is determined with closed-form expressions (i.e., $M = 1$), the computational complexity is *O(N)*. The computational complexity does not depend on the type of power grid (e.g., the same computational complexity for flip chip, wire bonded power grids, and power grids with on-chip power supplies).

To compare the computational runtime of FAIR with existing power grid analysis techniques, five differently sized circuits with evenly distributed C4 bumps 25 nodes from each other are considered. The runtime of FAIR is compared with the random walk method in [474], as shown in Table 23.7. The partition size for all of the circuits when utilizing locality is larger than 100×100 to maintain an approximation error of less than 0.1 %. The random walk method is applied for 20,000 iterations on each circuit to accurately determine the node voltages. The number of iterations of the random walk method is chosen to maintain a maximum error of less than 10 mV as compared to the algorithm with 20,000 iterations. The error of this method is also less than 10 mV for each circuit. This method without utilizing locality is over 26 times faster than the random walk method for circuits smaller than five million nodes. FAIR, when utilizing locality, is over 60 times faster for power grids smaller than five million nodes. For circuit sizes greater than 25 million nodes (e.g.,

Table 23.5 Error of Algorithm IV as compared to SPICE. Power supplies are connected at the corners (*light gray*) and current loads are connected at various nodes (*dark gray*). The maximum error is 4.03 mV (less than 0.5 % of the power supply voltage). Error correction is not used in this example and the maximum error occurs at the supply connection

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	0	0.36	0.10	−0.25	−0.32	−0.43	−0.57	−0.53	−0.28	0.07	0.39	0.87	1.1	1.71	2.26	3.06	4.03
2	0.4	−0.52	−0.4	−0.29	−0.45	−0.67	−0.93	−1.15	−0.68	−0.18	0.22	0.65	1.08	1.43	1.76	1.87	3.3
3	0.18	−0.27	−0.39	−0.45	−0.58	−0.85	−1.31	−2.35	−1	−0.32	0.19	0.62	0.91	1.28	1.56	1.94	2.5
4	−0.12	−0.24	−0.36	−0.44	−0.49	−0.73	−1.09	−1.17	−0.64	−0.2	0.24	0.56	0.84	1.22	1.55	1.75	2.1
5	−0.2	−0.26	−0.31	−0.36	−0.42	−0.46	−0.93	−0.56	−0.25	0.01	0.28	0.55	0.89	1.19	1.42	1.65	1.83
6	−0.28	−0.3	−0.38	−0.38	−0.31	−0.07	−0.9	−0.01	−0.07	0.09	0.23	0.58	0.89	1.12	1.37	1.48	1.66
7	−0.33	−0.29	−0.31	−0.44	−0.56	−0.83	−0.27	−0.54	−0.17	0.14	0.2	0.69	0.93	1.12	1.25	1.44	1.61
8	−0.34	−0.3	−0.34	−0.22	−0.15	−0.11	−0.28	0.43	0.2	0.57	0.18	0.96	0.91	1.06	1.25	1.43	1.48
9	−0.36	−0.33	−0.35	−0.23	0.18	−0.4	0.03	0.11	−0.16	0.12	0.59	0.39	0.67	0.99	1.18	1.35	1.52
10	−0.46	−0.47	−0.4	−0.48	−0.54	−0.2	−0.46	0.15	−0.06	0.99	0.3	1.05	1	1.15	1.19	1.38	1.51
11	−0.44	−0.48	−0.36	−0.24	0.07	−0.62	0.05	−0.11	0.37	0.16	0.21	0.75	1.02	1.25	1.28	1.55	1.56
12	−0.48	−0.5	−0.35	−0.3	−0.14	−0.36	0.17	0.6	0.13	0.81	0.58	0.84	1.05	1.18	1.37	1.44	1.67
13	−0.55	−0.48	−0.48	−0.37	−0.27	−0.18	0.1	0.3	0.26	0.7	0.68	1.02	1.14	1.38	1.39	1.74	1.84
14	−0.6	−0.65	−0.64	−0.53	−0.24	−0.12	0.12	0.2	0.28	0.51	0.87	0.97	1.15	1.37	1.55	1.8	2.06
15	−0.7	−0.93	−0.77	−0.55	−0.25	−0.09	0.09	0.24	0.44	0.68	0.82	1.09	1.29	1.44	1.66	1.91	2.39
16	−0.95	−1.58	−0.94	−0.56	−0.32	−0.11	0.03	0.25	0.46	0.72	0.88	1.12	1.35	1.57	1.83	1.89	3.04
17	−2.49	−0.84	−0.46	−0.51	−0.16	−0.03	0.16	0.3	0.52	0.65	0.84	1.14	1.37	1.73	2.21	2.92	3.47

Table 23.6 Error of Algorithm IV with error correction windows as compared to SPICE. The nodes where error correction is applied is shown in *italic font*. The maximum error is 2.35 mV which is less than 0.3 % of the power supply. Power supply and current load locations are denoted as *light gray* and *dark gray boxes*, respectively

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	0	0.36	0.1	−0.25	−0.32	−0.43	−0.57	−0.53	−0.28	0.07	*−0.19*	*−0.28*	*−0.63*	*−0.59*	*−0.62*	*−0.39*	*0*
2	0.4	−0.52	−0.4	−0.29	−0.45	−0.67	−0.93	−1.15	−0.68	−0.18	*−0.36*	*−0.5*	*−0.65*	*−0.87*	*−1.12*	*−1.58*	*−0.15*
3	0.18	−0.27	−0.39	−0.45	−0.58	−0.85	−1.31	−2.35	−1	−0.32	*−0.39*	*−0.53*	*−0.82*	*−1.02*	*−1.32*	*−0.94*	*−0.38*
4	−0.12	−0.24	−0.36	−0.44	−0.49	−0.73	−1.09	−1.17	−0.64	−0.2	*−0.34*	*−0.59*	*−0.89*	*−1.08*	*−0.75*	*−0.55*	*−0.2*
5	−0.2	−0.26	−0.31	−0.36	−0.42	−0.46	−0.93	−0.56	−0.25	0.01	*−0.3*	*−0.6*	*−0.84*	*−0.54*	*−0.31*	*−0.08*	*0.1*
6	−0.28	−0.3	−0.38	−0.38	−0.31	−0.07	−0.9	−0.01	−0.07	0.09	*−0.35*	*−0.57*	*−0.26*	*−0.03*	*0.22*	*0.33*	*0.51*
7	−0.33	−0.29	−0.31	−0.44	−0.56	−0.83	−0.27	−0.54	−0.17	0.14	*−0.38*	*0.11*	*0.35*	*0.54*	*0.67*	*0.86*	*1.03*
8	−0.34	−0.3	−0.34	−0.22	−0.15	−0.11	−0.28	0.43	0.2	0.57	0.18	0.96	0.91	1.06	1.25	1.43	1.48
9	−0.36	−0.33	−0.35	−0.23	0.18	−0.4	0.03	0.11	−0.16	0.12	0.59	0.39	0.67	0.99	1.18	1.35	1.52
10	−0.46	−0.47	−0.4	−0.48	−0.54	−0.2	−0.46	0.15	−0.06	0.99	0.3	1.05	1	1.15	1.19	1.38	1.51
11	*−0.08*	*−0.12*	*0*	*0.12*	*0.43*	*−0.26*	*0.41*	−0.11	0.37	0.16	*−0.29*	*0.25*	*0.52*	*0.75*	*0.78*	*1.05*	*1.06*
12	*0.23*	*0.21*	*0.36*	*0.41*	*0.57*	*0.35*	*0.53*	0.6	0.13	0.81	*0.08*	*−0.15*	*0.06*	*0.19*	*0.38*	*0.45*	*0.68*
13	*0.52*	*0.59*	*0.59*	*0.7*	*0.8*	*0.53*	*0.46*	0.3	0.26	0.7	*0.18*	*0.03*	*−0.35*	*−0.11*	*−0.1*	*0.25*	*0.35*
14	*0.82*	*0.77*	*0.78*	*0.89*	*0.83*	*0.59*	*0.48*	0.2	0.28	0.51	*0.37*	*−0.02*	*−0.34*	*−0.61*	*−0.43*	*−0.18*	*0.08*
15	*1.08*	*0.85*	*1.01*	*0.87*	*0.82*	*0.62*	*0.45*	0.24	0.44	0.68	*0.32*	*0.1*	*−0.2*	*−0.54*	*−0.82*	*−0.57*	*−0.09*
16	*1.18*	*0.55*	*0.84*	*0.86*	*0.75*	*0.6*	*0.39*	0.25	0.46	0.72	*0.38*	*0.13*	*−0.14*	*−0.41*	*−0.65*	*−1.08*	*0.07*
17	*0*	*1.29*	*1.32*	*0.91*	*0.91*	*0.68*	*0.52*	0.3	0.52	0.65	*0.34*	*0.15*	*−0.12*	*−0.25*	*−0.27*	*−0.05*	*0*

Table 23.7 Runtime comparison between FAIR and random walk method

	# of nodes	Random walk [474] (min:s)	FAIR			
			No partitioning (min:s)	Speed enhancement	Locality and error correction (min:s)	Speed enhancement
Circuit I	250 K	4:22	0:10	26×	0:03	87×
Circuit II	1 M	15:08	0:32	28×	0:13	70×
Circuit III	4 M	59:46	2:19	26×	0:58	62×
Circuit IV	25 M	1156:14	17:13	67×	6:33	176×
Circuit V	49 M	3418:05	38:55	88×	13:09	260×

Circuit IV and Circuit V in Table 23.7), FAIR with locality is over 175 times faster than the random walk method. The runtime of the random walk method depends strongly upon the number of power supply connections. When the number of power supply connections decreases, the runtime of the random walk method dramatically increases. Alternatively, the runtime of FAIR is lower with fewer number of power supply connections.

23.5 Summary

Closed-form expressions and related Fast Algorithms for fast *IR* (FAIR) voltage drop analysis are provided in this chapter. The physical distance between circuit components and the principal of spatial locality are exploited within FAIR. The primary conclusions can be summarized as follows.

- Significant computational time is required for power grid analysis when the size of the grid is large. Efficient algorithms are therefore required to reduce the computational runtime of the power grid analysis process
- A novel algorithm, fast algorithms based on an effective resistance for *IR* drop analysis, is described for analyzing locally uniform, globally non-uniform power grids with non-uniform current loads and non-uniform voltage supplies
- The power grid impedance characteristics and the Euclidean distance between the circuit components are utilized to develop the closed-form expressions
- Local analyses of power distribution networks can be performed with FAIR
- The principle of spatial locality is exploited to improve the accuracy and runtime of FAIR. Parallel computation of the *IR* drop analysis algorithm can also be implemented
- A novel error correction technique exploiting the principle of spatial locality is described to improve the accuracy of FAIR
- FAIR is more computationally efficient than existing *IR* drop analysis techniques since no iterations are required

Chapter 24
Stability in Distributed Power Delivery Systems

Delivering high quality power to support power efficient systems is a fundamental requirement of all ICs. While the quality of the power supply can be efficiently addressed with a point-of-load power delivery system [187, 191, 475–477], the complexity of a dynamically controllable distributed POL power supply system is a significant design issue. Hundreds of on-chip power regulators need to be co-designed with billions of nonlinear current loads within a power domain, imposing a critical stability challenge on distributed power delivery systems. To cope with the design complexity of complex analog systems, modeling, optimization, and synthesis techniques are typically used [478]. To automate the design of a power delivery system, accurate methods to evaluate performance metrics (e.g., quality of transient response, stability, and power) are required.

With the increasing diversity of modern systems, dynamic voltage scaling and fine grain power management are becoming increasingly common. These modern heterogeneous systems are typically partitioned into a fine grain structure, where the power is individually delivered and dynamically managed within each domain. With dynamic voltage scaling, maintaining the stability of these distributed power delivery systems has become highly challenging.

Low dropout regulators suitable for on-chip integration have recently been described [303, 316, 318, 333, 364–366, 375, 379–386, 479–485], exhibiting fast load regulation, high power efficiency, as well as stability over a wide range of current loads and process, voltage, and temperature (PVT) variations. The LDO is therefore a key component in on-chip power management. A distributed system with multiple LDO regulators delivering power to a single grid may exhibit instability due to complex interactions among the LDO regulators, power distribution network, and current loads. The stability of these parallel connected voltage regulators is therefore a primary performance concern and requires accurate evaluation. To provide a stable distributed power delivery system, a stability analysis criterion is necessary.

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0_24

The stability of a single closed loop system is traditionally determined by the phase margin of the open loop response of a system. In systems with multiple dependent loops, the open loop approach is, however, impractical because no straightforward method exists to identify unstable loops [486]. A computer-aided design framework based on the passivity and gain of a power grid has recently been described for evaluating the stability of distributed power delivery systems with LDO regulators [332, 334]. While recognizing stability challenges is an important cornerstone to the distributed power grid design process, the accuracy and efficiency of the stability analysis process requires demonstration on practical power delivery systems. In this chapter, an alternative passivity-based stability criterion (PBSC) is described for use with existing CAD tools and design flows, and is not limited to LDO based power delivery systems. Based on this passivity-based criterion, accurate system level requirements for evaluating the exponential and marginal stability of distributed power delivery systems are provided. Automating the design process of a power delivery system based on this stability criterion is also demonstrated by a parametric circuit performance modeling technique [478].

A distributed power delivery system with six ultra-small LDO regulators based on this stability criterion has been fabricated in an advanced 28 nm CMOS technology, and exhibits stable voltage regulation. The system is the first successful silicon demonstration of stable parallel analog LDO regulators without off-chip compensation.

The rest of the chapter is organized as follows. The stability criterion is described in Sect. 24.1. A distributed power delivery system is evaluated in Sect. 24.2 based on the stability criterion. Simulation and test results exhibit strong correlation between the stability of a distributed power delivery system and the PBSC criterion. Automated design with the stability criterion is described in Sect. 24.3. The chapter is concluded in Sect. 24.4.

24.1 Passivity-Based Stability of Distributed Power Delivery Systems

Understanding the effects of the frequency domain parameters on the time domain characteristics provides significant insight into the analysis and transient behavior of complex systems [487–491]. Traditionally, the phase margin of the open loop transfer function can be used to determine the stability of a single LDO regulator. Similarly, a straightforward criterion is required for determining the stability of a distributed power delivery system.

A distributed power delivery system with more than two power supplies driving a single power grid is depicted in Fig. 24.1a. In this distributed system, the power supplies can be combined into a single power delivery system, yielding an equivalent single port network, as shown in Fig. 24.1b. Note that the output impedance of the equivalent single port network, Z_{TOT} in Fig. 24.1b, is the parallel

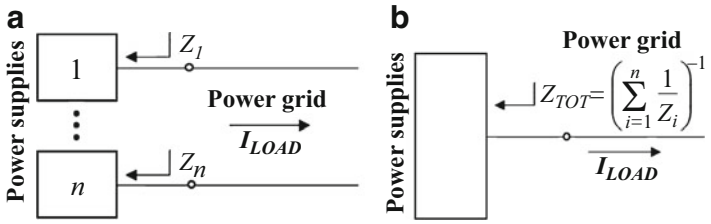


Fig. 24.1 Power delivery system (**a**) with $n \geq 2$ distributed power supplies, and (**b**) reduced single port network

combination of all of the output impedances $Z_i, i = 1, \ldots, N$ of the individual power supplies shown in Fig. 24.1a. The output impedance of a distributed power delivery system is, therefore, straightforward to evaluate based on the individual output impedance of the parallel connected components. Alternatively, there is no straightforward method to identify the single loop that causes instability in a system with multiple interacting feedback paths. The open loop transfer function, traditionally used to determine the stability of a lumped power delivery system, cannot be applied to a distributed power delivery system with multiple control loops [370]. A criterion for evaluating the stability of a multi-feedback path system composed of distributed power regulators is therefore needed.

Sufficient conditions for a stable distributed power delivery system are described in this section. These conditions are based on the observation, proven in [492], that a linear, time-invariant (LTI) system is stable when coupled to an arbitrary passive environment if and only if the driving point impedance is a passive system. Thus, a distributed power delivery system is stable if and only if the equivalent output impedance Z_{TOT} satisfies passivity requirements. The passivity of a linear time-invariant (LTI) system is described here in terms of frequency domain parameters.

An LTI system is passive if the system can only absorb energy, yielding, in mathematical terms [493],

$$\int_{-\infty}^{T} v(t)i(t)dt \geq 0, \forall\ T, \tag{24.1}$$

where $v(t)$ and $i(t)$ are, respectively, the voltage across the system and current flowing through the system. The total energy delivered to a passive system is determined from (24.1) based on the Parseval Theorem, exhibiting, for all positive currents,

$$\frac{1}{\pi}\int_{0}^{+\infty} Re[Z(j\omega)]|I(j\omega)|^2 d\omega \geq 0, \tag{24.2}$$

where $Z(s) = V(s)/I(s)$ is the system impedance, and $V(s)$ and $I(s)$ are, respectively, the phasor voltage and current of the system. The passivity condition based on (24.2), $\{Re[Z(\sigma + j\omega)] \geq 0, \forall \sigma > 0\}$, can be simplified based on [494] and specialized for a particular frequency range of interest S, yielding the following sufficient conditions for passivity of an LTI system: $Z(s)$ has no right half plane (RHP) poles, and the phase of $Z(s)$ is within the $(-90°, +90°)$ range $\forall s \in S$.

A distributed system is, therefore, exponentially stable (converges within an exponential envelope) if the impedance of the system satisfies these passivity requirements, marginally stable (oscillates with constant amplitude) if the voltage and current phasors are shifted by precisely 90°, and unstable otherwise. The phase of the output impedance is an efficient alternative to determine the stability of these distributed systems, since the traditional phase margin approach is not practical due to the multiple control loops.

24.2 Passivity Analysis of a Distributed Power Delivery System

In a distributed power delivery system, the total current load is shared among all of the power supplies. Voltage regulators in close proximity with the current load supply the greatest portion of the total current, which can be significantly higher than the average current supplied by a single regulator [191]. In addition, certain small signal parameters, such as the output resistance and output transconductance of individual regulators, are affected by the DC load current, changing the stability characteristics of each LDO regulator and the overall system. The stability of a distributed system is therefore a strong function of the local current shared among the distributed regulators.

To demonstrate the concept of a stable distributed power delivery system based on the passivity-based stability criterion, a power delivery system with six parallel connected LDO regulators is evaluated. A model of the power delivery system with six LDO regulators and a distributed power delivery network is shown in Fig. 24.2. Each power supply in the power delivery system is a standard LDO regulator [289] composed of an error amplifier (EA), output device (M_P), and compensation network $R_C C_C$, as depicted in Fig. 24.3. A three current mirror operational transconductance amplifier (OTA) topology [370] is used within each error amplifier.

The output impedance of parallel connected voltage regulators is a primary factor in determining the stability of a distributed power delivery system, and is a strong function of the poles and zeros of the individual LDO regulators. To maintain stability in a distributed power delivery system with n LDO regulators, the poles of the output impedance $Z_{OUT}^{TOT}(s)$ must be left plane poles and the phase of $Z_{OUT}^{TOT}(s)$ must be within the $(-90°, 90°)$ range $\forall\ s$.

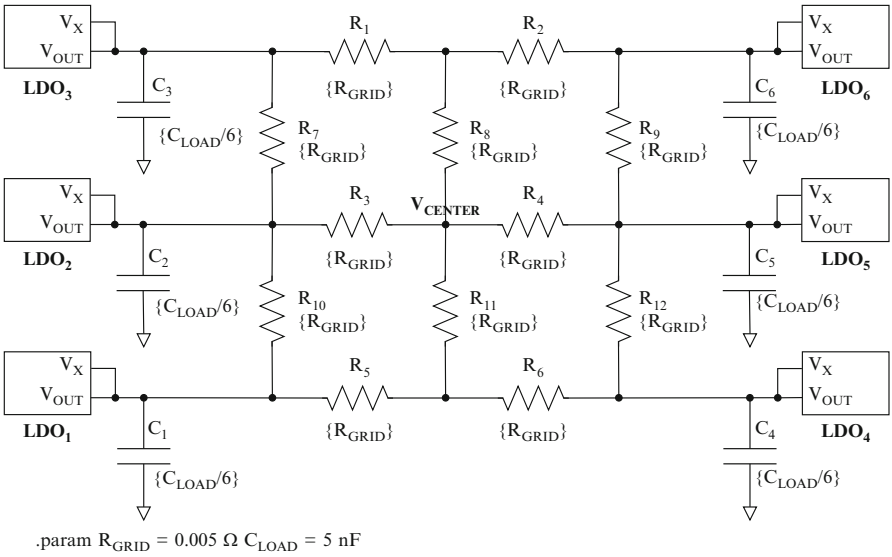


Fig. 24.2 Model of distributed LDO and power delivery system

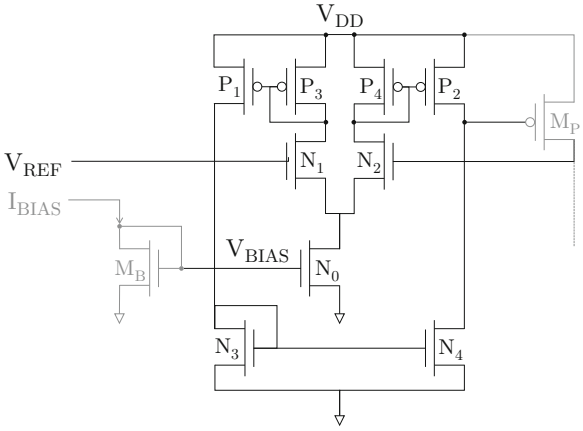


Fig. 24.3 Standard LDO topology [370]

The stability of the power delivery system is demonstrated on an example system assuming a total current load of 300 mA. Load sharing among the LDO regulators in the system exhibits a wide range of LDO currents (between 20 and 100 mA for an individual LDO regulator). The LDO in closest proximity with the current load supplies the largest portion (100 mA) of the total current requirements, which is higher by a factor of 2 than the average current load (52 mA) supplied by a

single LDO. Alternatively, remote LDOs supply significantly less current (down to 20 mA), only half of the average LDO load current. The output impedance of the system under this load sharing scenario is evaluated here for each of the LDO regulators and the combined distributed power delivery system.

The phase, gain, poles, and zeros within the range of interest are shown in Fig. 24.4, demonstrating a passive parallel combination of individually passive impedances. Note that the poles of the combined system output impedance are limited by the frequency range of the individual LDO poles. Thus, a distributed power delivery system with individually stable LDO regulators *under all feasible load currents* exhibits no right half plane poles (RHP). The stability of a multi-feedback system with individually stable power supplies is therefore limited by the phase of the combined output impedance of the system.

To demonstrate the effect of the phase of the output impedance on the stability of a distributed system, the transient response and phase of the output impedance $\angle Z_{OUT}$ of the distributed system with six LDO regulators are shown in Fig. 24.5. In agreement with the passivity-based stability criterion, the output response diverges (oscillates with increasing amplitude), and converges within an exponential envelope for, respectively, $|\angle Z_{OUT}| > 90°$ and $|\angle Z_{OUT}| < 90°$. Note that the system with $C_C = 0.5$ pF and $\max_{\forall f}\{\angle Z_{OUT}\} = 89°$ slowly converges to the steady-state solution, exhibiting an underdamped response inappropriate for voltage regulation in power delivery systems. Alternatively, a system with $C_C = 5$ pF and $\max_{\forall f}\{\angle Z_{OUT}\} = 70°$ exhibits an overdamped response with a significant stability margin. A strong correlation therefore exists between the phase shift of the output voltage and load current, and the effective stability margin of the system. Based on this observation, the phase margin of the output impedance for a distributed power delivery system is

$$PM(Z_{out}) = 90° - \max_{\forall f}\{\angle Z_{OUT}\}. \tag{24.3}$$

A distributed power delivery system is therefore unstable, stable, or marginally stable if the phase margin of the output impedance is, respectively, negative, positive, or zero. A safe phase margin of the output impedance should be determined based on specific design criteria to avoid excessively underdamped and overdamped voltage regulation systems.

A power delivery system with six LDO regulators has been designed and evaluated based on this passivity-based stability criterion. The system is fabricated in an advanced 28 nm CMOS technology. A die microphotograph of the LDO regulator is illustrated in Fig. 24.6. The area occupied by the LDO with all capacitors is $85 \times 42\,\mu$m. The measured transient response is illustrated in Fig. 24.7 for nominal input and output voltages of, respectively, 1.0 and 0.7 V, and a load current step (stepped from 52 to 788 mA in 5 ns). Based on these experimental results, the system of six parallel LDO regulators yields a stable response and voltage droop of 0.1 V.

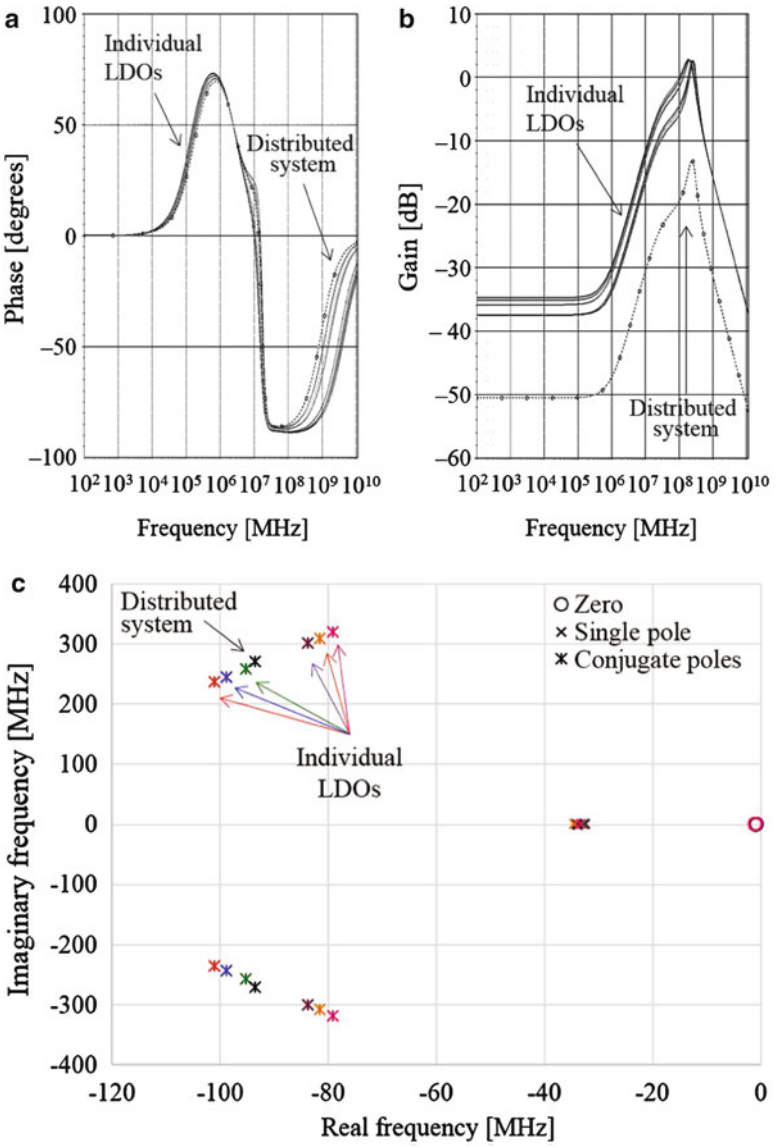


Fig. 24.4 Output impedance of individual LDO regulators loaded by different currents (between 20 and 100 mA) and the combined system output impedance, (**a**) phase $\angle Z_{OUT}$, (**b**) gain $|Z_{OUT}|$, and (**c**) poles and zeros

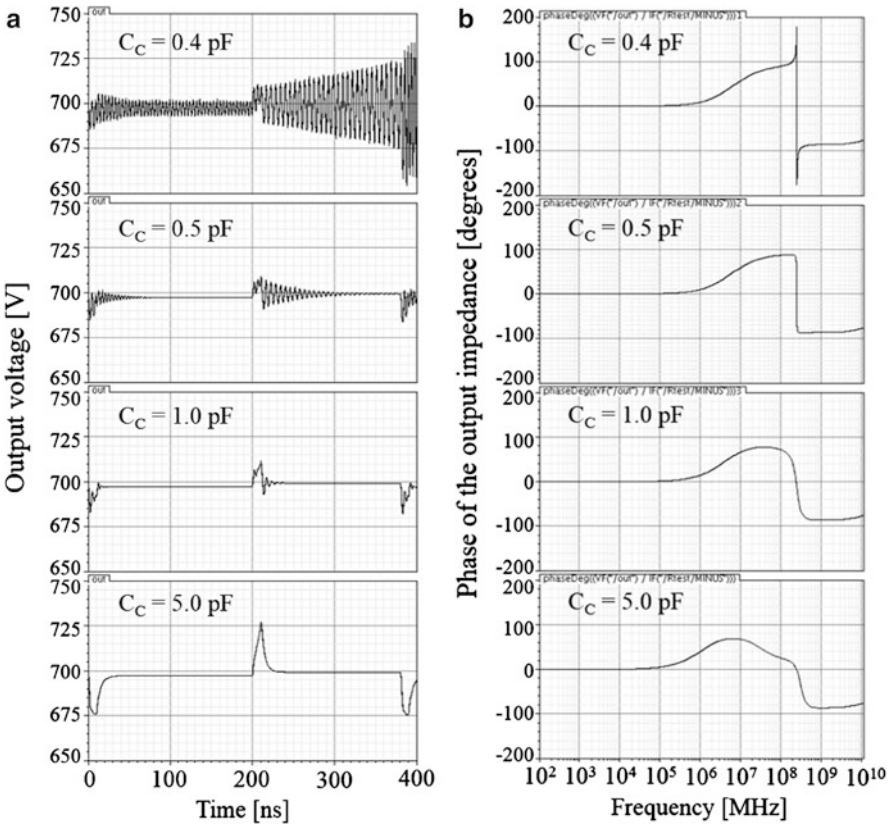


Fig. 24.5 Output response of a distributed power delivery system with different compensation capacitors ($C_C = 0.4$ pF, $C_C = 0.5$ pF, $C_C = 1$ pF, and $C_C = 5$ pF), illustrating the correlation between the (**a**) transient response, and (**b**) phase of the output impedance

24.3 Model of Parametric Circuit Performance

Existing automated techniques for designing analog circuits are based on numerical optimization and evaluation engines [478]. Parametric models characterize the performance of an analog circuit (e.g., gain, bandwidth (BW), slew rate (SR), or phase margin (PM)) based on certain circuit design variables (e.g., device sizes and voltage biases) [478]. The performance of an individual power supply is typically determined by a set of parameters, such as the DC gain, phase margin, DC offset, slew rate, and power. Alternatively, a distributed power delivery system should be evaluated based on both the performance of the individual power regulators and additional performance metrics characterizing the combined system, such as the *phase margin of the output impedance*. To reduce the design complexity of modern distributed power delivery systems, the passivity-based stability criterion should be

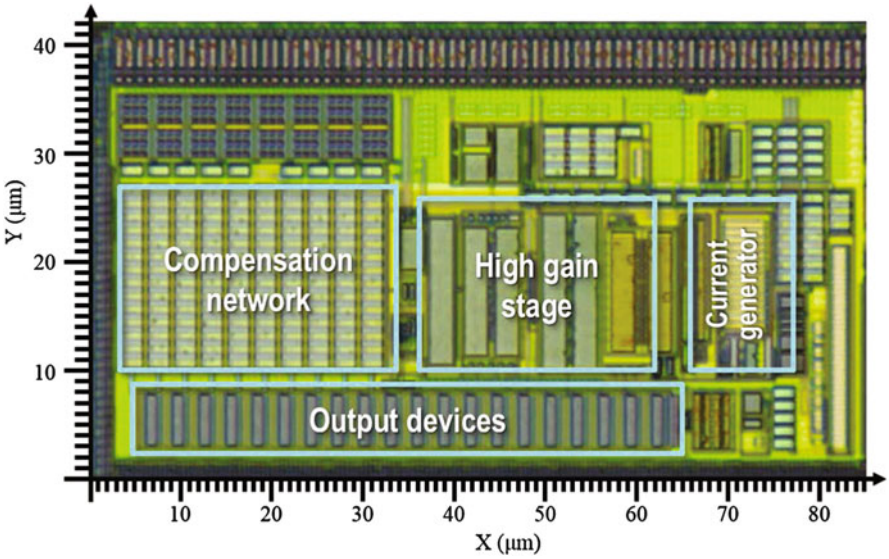


Fig. 24.6 Die microphotograph of an LDO regulator and current generating circuit

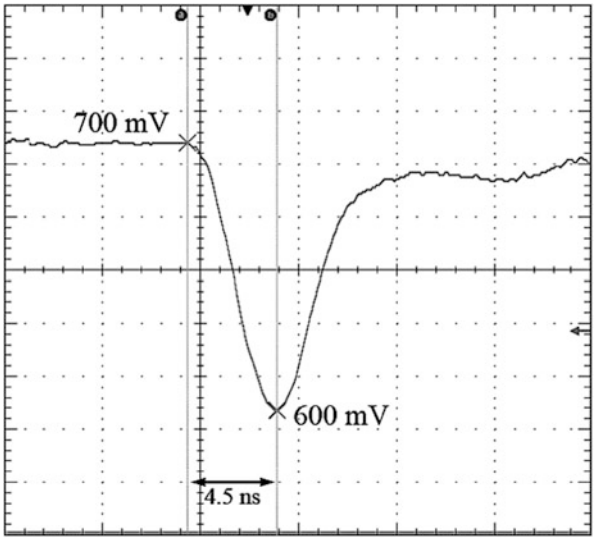


Fig. 24.7 Measured transient response for a load current step from 52 to 788 mA in 5 ns

integrated within existing automated design methodologies. An automated flow for designing a stable distributed power delivery system is shown in Fig. 24.8. The first stage of the flow is based on a standard parametric performance modeling technique [478]. During this stage, an LDO regulator is synthesized based on the specific LDO topology and design objectives. The output of the first stage is used during the second stage to determine the number and location of the parallel

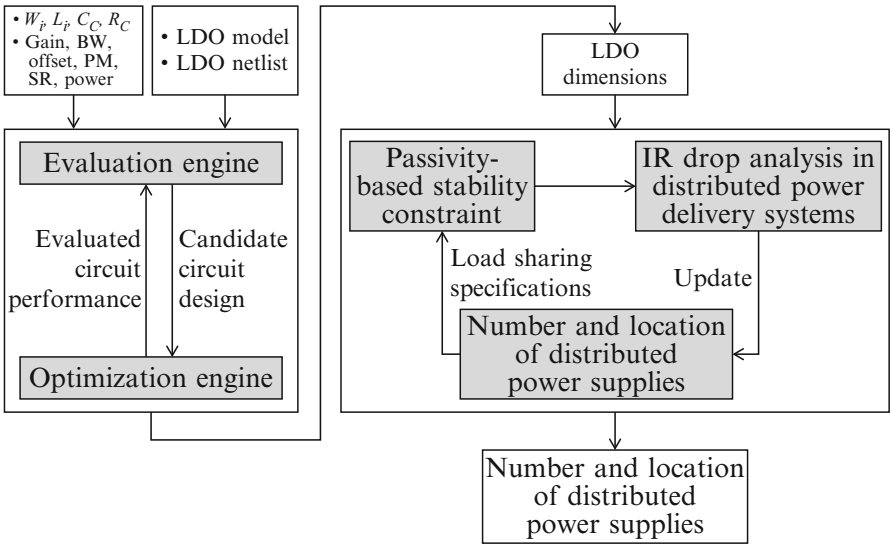


Fig. 24.8 Automated PBSC-based design flow for a distributed power delivery system

connected power supplies within a distributed power delivery system. During this second stage, a power delivery system, composed of distributed voltage regulators, is iteratively evaluated based on the passivity-based stability criterion and placement algorithms [191, 459]. During each iteration, the worst case load sharing scenario is determined for the specific power delivery system. The passivity-based stability of the distributed system is evaluated based on the individual current loads. If required, the number and location of the power supplies are updated. Finally, the number and location of the parallel connected power supplies that satisfies the quality of power and stability requirements of the distributed power delivery system are determined.

The operation of the second stage of the automated PBSC-based design flow is demonstrated based on the ISPD'11 placement benchmark suite of circuits [495]. The floorplan of the superblue5 (sb5), superblue10 (sb10), superblue12 (sb12), and superblue18 (sb18) circuits is illustrated in Fig. 24.9. Each of the circuits is composed of thousands of fine grain rectangular shapes. To reduce the complexity of the circuit evaluation process, the fine grain shapes are combined into larger rectangular nodes. Of the combined nodes, only the largest nodes are considered, exhibiting a reduced floorplan. The magnitude of the distributed current loads is proportional to the size of these nodes with a total load current of 1 A. The location of each of the current loads is in the center of the corresponding rectangular node. The number of fine grain shapes, large combined nodes, coverage of the reduced floorplan, and power grid data are listed in Table 24.1. Note that the nodes in the reduced floorplan occupy more than 85 % of the total active circuit area.

A constant voltage is ideally distributed to all of the current loads within a circuit. Practically, the quality of power is degraded in modern circuits due to parasitic

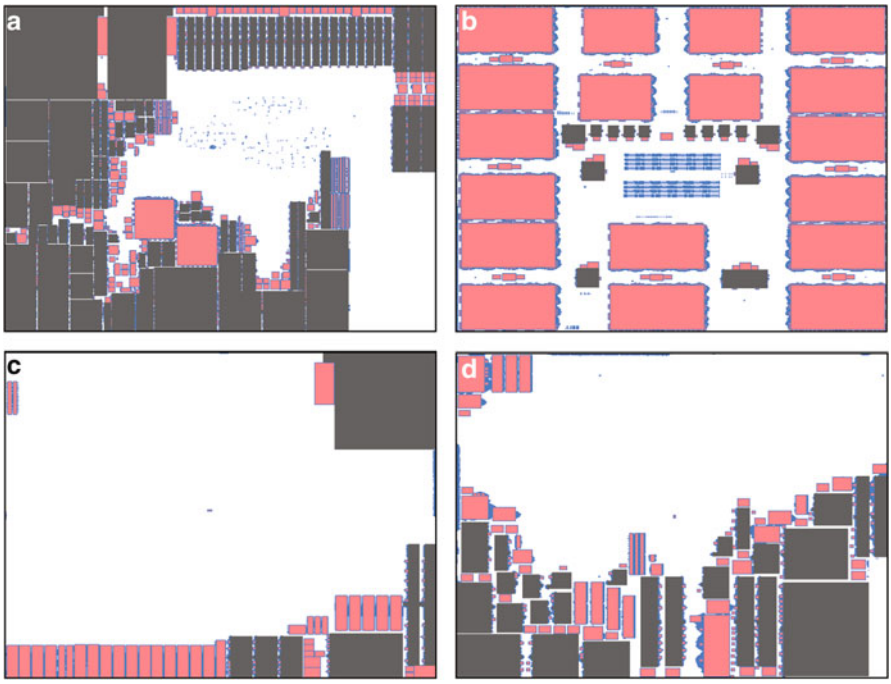


Fig. 24.9 Floorplan of ISPD'11 circuits [495] (**a**) superblue5, (**b**) superblue10, (**c**) superblue12, and (**d**) superblue18

Table 24.1 Properties of ISPD benchmark circuits

Circuit	Fine grain shapes	Large combined nodes	Coverage of reduced floorplan (%)	Power grid size	Nodes in power grid
sb5	29,736	129	85.0	774 × 713	551,862
sb10	2318	30	89.2	638 × 968	617,584
sb12	3578	15	98.4	444 × 518	229,992
sb18	6776	71	99.5	381 × 404	153,924

on-chip impedances. A voltage drop map of the superblue5 circuit without on-chip power supplies is shown in Fig. 24.10, yielding a maximum voltage drop of 23.4 %, assuming an off-chip voltage supply of 1 V. To address the quality of on-chip power, power delivery systems with a single on-chip power supply (case 1), three on-chip power supplies (case 2), and twelve on-chip power supplies (case 3) are considered. For each of the three cases, the *IR* drops of the distributed power delivery system is analyzed based on the *IR* drop algorithm for a power grid with multiple power supplies and current loads (see Chap. 23) [458]. The location of the power supplies in cases 1 and 2 is modeled as a mixed integer nonlinear programming problem [191], and optimized based on a general algebraic modeling system (GAMS) [496].

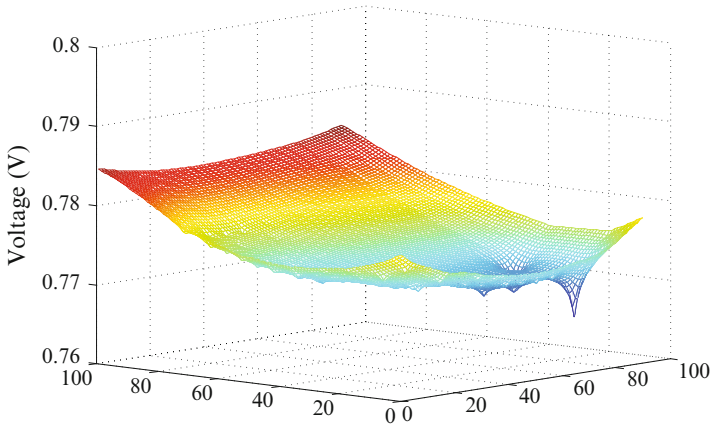


Fig. 24.10 Voltage drop map of superblue5 circuit

In case 3, the power supplies are uniformly distributed on-chip. The stability is evaluated for each of the three cases based on the passivity-based criterion. A map of the voltage drops and phase of the output impedance within superblue5 with a different number of on-chip power supplies is shown in Fig. 24.11. The maximum voltage drop is less with increasing number of power supplies, exhibiting a reduction in the maximum voltage drop of, respectively, 14.23 %, 20.29 %, and 22.29 % with a single, three, and twelve on-chip power supplies. Alternatively, the output current of the individual regulators changes with the number of power supplies, affecting the phase of the output impedance and stability characteristics of the distributed system. Based on the stability criterion, the superblue5 circuit is stable with a single power supply and three power supplies (the phase of the output impedance is within the $(-90°, 90°)$ range), and unstable with twelve power supplies (the minimum phase of the output impedance is $-95.1°$ which is less than $-90°$). While the distributed power delivery system with twelve power supplies exhibits a higher quality of power than systems with fewer power supplies, this system is shown to be unstable under an aggressive transient load response. Thus, a stable system with fewer power supplies is preferable to deliver power to the superblue5 circuit when considering both quality of power and stability challenges.

The second stage of the automated PBSC-based design flow, shown in Fig. 24.8, has been implemented in Matlab. Pseudo-code of the Matlab algorithm is summarized in Algorithm 24.1. A model of the LDO circuit is used to describe the small signal response of the on-chip power supplies, and evaluate the output impedance of the power supplies and overall power delivery system. The power delivery system for the ISPD'11 benchmark circuits, superblue5 (sb5), superblue10 (sb10), superblue12 (sb12), and superblue18 (sb18), has been evaluated based on this PBSC-based LDO placement algorithm. The maximum *IR* drop and stability results are listed in Table 24.2.

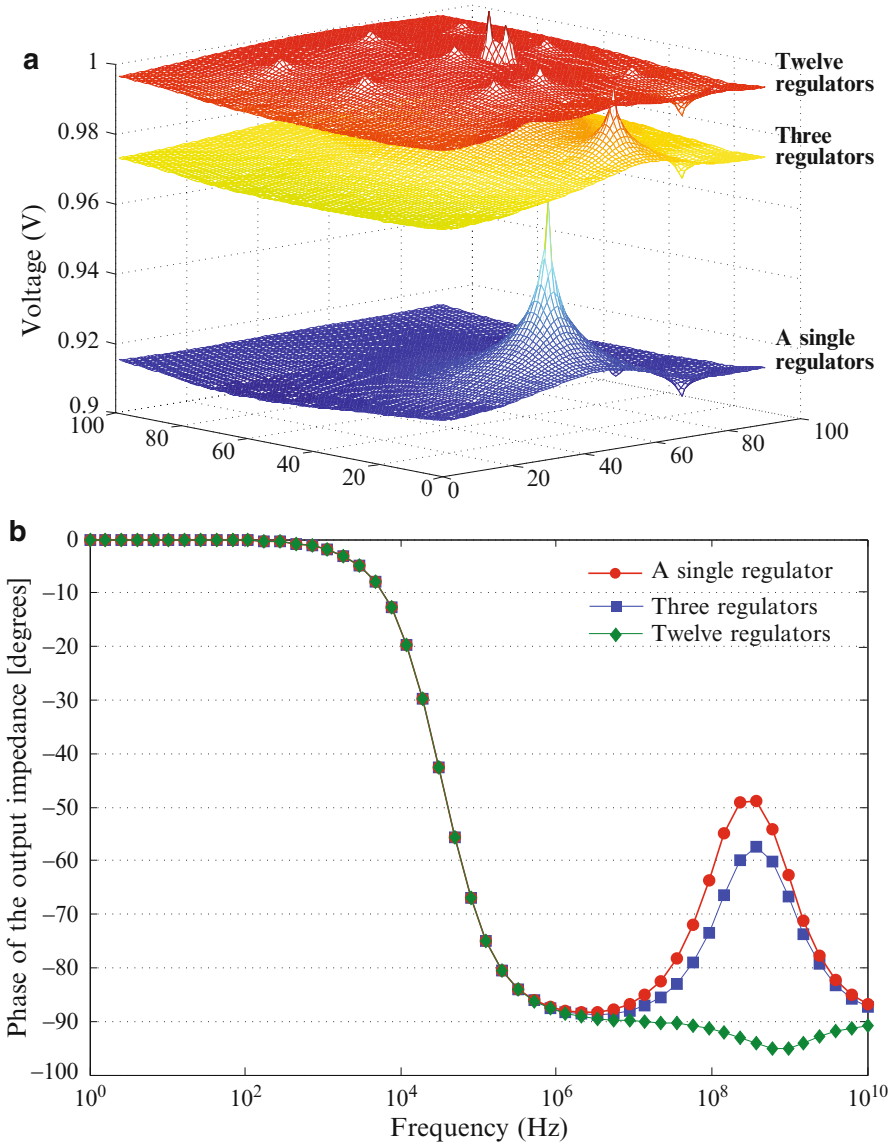


Fig. 24.11 Superblue5 circuit with a single, three, and twelve on-chip power supplies, (**a**) map of voltage drops, and (**b**) phase of the output impedance

Based on the evaluation of the benchmark circuits, the maximum voltage drop is significantly less with increasing number of on-chip power supplies. Alternatively, the stability of the distributed power delivery system is a function of the specific load distribution, and is affected by characteristics of the POL power delivery system. The automated PBSC-based design flow generates a distributed power delivery system that addresses both quality of power and stability requirements.

Algorithm 24.1 Automated PBSC-based design flow

LDOModel; % // A typical LDO model [497]
CircInfo; % // Supply voltages, load currents and locations
CircNodes; % // All nodes in the evaluated circuit
NumRegsList; % // List of numbers of LDOs to evaluate
PreferredNumRegs ← 0; % // Preferred number of LDOs
PreferredLocs ← N/A; % // Preferred location of LDOs
PreferredIRDrop ← 1; % // Maximum allowed *IR* drop

for all *NumRegs* ← *NumRegsList* **do**
  for all *RefNode* ← *CircuitNodes* **do**
    % // Find optimal LDO locations [496]
    OptLocs ← OPT_LOC(*CircInfo*, *NumRegs*);

    % // Analyze *IR* drop in a power grid [458]
    IRDrop(*RefNode*) ← CALC_IRDROP(*CircInfo*, . . .
                                OptLocs, . . .
                                RefNode);

    % // Calculate the output impedance of the LDOs
    for all *LDO* ← *LDOs* **do**
      ISupply ← Current delivered by the LDO;
      $Z_{OUT}^{LDO}(LDO)$ ← CALC_ZOUT(*LDOModel*, . . .
                                ISupply);
    end for

    % // Calculate the output impedance of the system
    $Z_{OUT}^{SYS} \leftarrow \left(\sum_{LDOs} \frac{1}{Z_{OUT}^{LDO}(LDO)} \right)^{-1}$;

    if $\max\{|\angle Z_{OUT}^{SYS}|\} < 90°$ **then**
      if $\max\{IRDrop\} < PreferredIRDrop$ **then**
        PreferredIRDrop ← $\max\{IRDrop\}$;
        PreferredNumRegs ← *NumRegs*;
        PreferredLocs ← *OptLocs*;
      end if
    end if
  end for
end for

24.4 Summary

Distributed on-chip power regulation and delivery are necessary for delivering high quality power to modern high performance integrated circuits. Significant load sharing and PVT variations however pose stability challenges on the co-design of these multiple on-chip voltage regulators. Thus, the design complexity of parallel voltage regulators driving the same power grid is traded off for high power quality.

To design a stable closed loop regulator, sufficient phase margin in the open loop transfer function is required. Phase margin is therefore a sufficient parameter for

Table 24.2 Maximum *IR* drop and stability in ISPD benchmark circuits

Circuit		sb5	sb10	sb12	sb18
No regulators	Maximum *IR* drop (%)	23.4	23.0	23.6	22.7
	Stability	N/A	N/A	N/A	N/A
A single regulator	Maximum *IR* drop (%)	9.17	10.8	10.7	10.7
	Stability	Stable	Stable	Stable	Stable
Three regulators	Maximum *IR* drop (%)	3.11	4.75	6.39	4.43
	Stability	Stable	Stable	Stable	Stable
Twelve regulators	Maximum *IR* drop (%)	1.11	2.43	4.88	1.54
	Stability	Unstable	Unstable	Unstable	Unstable

determining the stability of a single LDO. Evaluating the open loop characteristics is however not practical with parallel LDO regulators due to the multiple regulation loops. Evaluating the stability of a distributed power delivery system is therefore not possible with the traditional phase margin criterion.

- An alternative passivity-based stability criterion is described for evaluating the stability of parallel voltage regulators driving a single power grid
- Based on this criterion, a distributed power delivery system is stable if and only if the total output impedance of the parallel connected LDOs exhibits no right half plane poles and a phase between $-90°$ and $+90°$
- Similar to a single voltage regulator, the phase margin of the output impedance (the difference between the maximum phase and 90°) determines the stability of a distributed power delivery system
- A distributed system with six LDO regulators is used to demonstrate the application of the PBSC method to determine the stability of a distributed power delivery system
- Feasible load sharing variations are evaluated based on system specifications, and upper and lower limits for load sharing variations are described
- Each of the LDO regulators is designed with sufficient phase margin to deliver stable power while satisfying any load sharing limitations
- The system, fabricated in a 28 nm CMOS process, exhibits a stable response with excellent load regulation
- Integration of the stability criterion within an existing design automation flow is demonstrated on a set of benchmark circuits, yielding an efficient technique for the automated design of stable, distributed power delivery systems
- The passivity-based stability criterion is a simple and efficient method to evaluate the stability of distributed power delivery systems

Chapter 25
Power Optimization Based on Link Breaking Methodology

A change in voltage at the power node of a gate can significantly increase the delay of a logic gate [32, 498, 499], degrading the overall performance of a system [34]. Since different circuits are affected differently by a drop in the power supply voltage, the power distribution network should be designed to satisfy multiple constraints. The voltage level for those gates along the critical path can tolerate the least voltage degradation, whereas the gates along a noncritical path may satisfy speed constraints despite a higher voltage drop [289]. Circuits, such as a phase-locked loops and voltage controlled oscillators (VCOs), are highly sensitive to changes in the power supply voltage [500]. Alternatively, digital logic circuits can tolerate much higher variations in the power supply voltage. The voltage level of a power distribution network across an entire IC is typically maintained within 10 % degradation, while for a PLL, the voltage level should satisfy a maximum 2 % voltage degradation. To satisfy these constraints, the current supplied to the PLL is filtered by a DC-to-DC converter or a large on-chip decoupling capacitance placed near the PLL [501]. The decoupling capacitors and DC-to-DC converters however consume large area and can dissipate significant power [338].

Separate power networks can be designed to independently supply current to different parts of a circuit, thereby shielding different parts of an IC from each other. Separate power networks are widely used in mixed-signal circuits, where the current is supplied to the analog and digital circuits by different power networks [502]. For systems requiring the same voltage, this approach may, however, inefficiently utilize metal resources due to additional area and routing constraints [289]. I/O pads are also a limited resource, preventing the use of an excessive number of separate power networks [115]. In Fig. 25.1, a single and multiple separate power networks are illustrated. With a single network, as shown in Fig. 25.1a, the sensitive circuit (e.g., a PLL) and aggressor circuit (exemplified by a large digital logic circuit) share the same power network, lowering the network impedance. A sensitive circuit can, however, be highly affected by the noise generated from the aggressor circuit. With multiple power networks, as shown in Fig. 25.1b, one network can be dedicated

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0_25

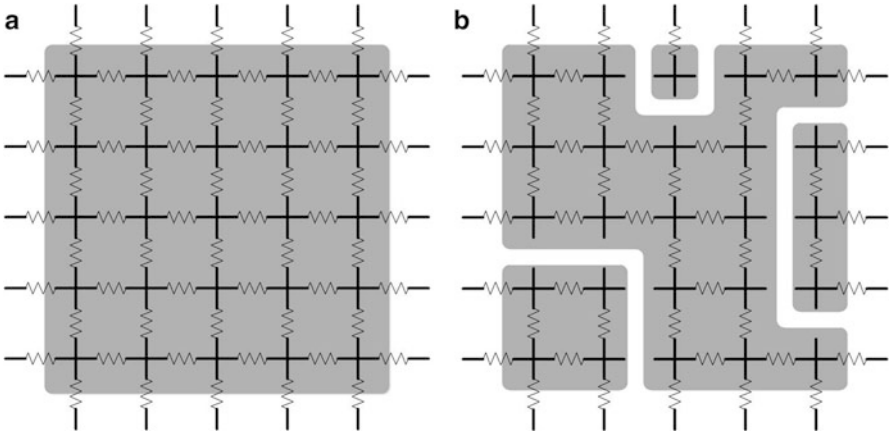


Fig. 25.1 Mesh structured power distribution network. (**a**) Single power distribution network focused on reducing the network impedance. (**b**) Multiple power distribution networks lower the noise at the expense of increasing the network impedance

to the aggressor circuit while another network can be dedicated to the sensitive circuits, minimizing noise coupling between the aggressor and sensitive circuits. This approach, however, results in an increase in the power network impedance and additional routability constraints. The methodology described in this chapter utilizes a single power network to provide a low network impedance and reduced routability constraints while disconnecting (or breaking) links within the on-chip power network between the aggressor and sensitive circuits, thereby reducing the noise coupling to the sensitive circuits.

This chapter is organized as follows. The primary design objective for reducing voltage variations is formulated in Sect. 25.1. An example where links within the on-chip mesh structured power distribution network are disconnected is described in Sect. 25.2. The sensitivity of the victim circuits to variations in the voltage within the power network is characterized in Sect. 25.3. In Sect. 25.4, the link breaking methodology is described. An algorithm for breaking links for a large number of aggressor and victim circuits connected to a common on-chip power distribution network is also described in this section. In Sect. 25.5, several design cases are evaluated. Degradation in the supply voltage and propagation delay before and after applying the link breaking methodology is summarized. Additional discussion related to enhancing the voltage levels within an on-chip power distribution network and the computational runtime of the algorithm is presented in Sect. 25.6. Finally, the conclusions are summarized in Sect. 25.7.

25.1 Reduction in Voltage Variations

The voltage drop ΔV_x at node x within a mesh structured power network, illustrated in Fig. 25.2, is a superposition of the voltage drop independently produced by each current source. Disconnecting a link on a mesh structured network increases the voltage drop at node x produced by current I_x. The voltage drop at node x produced by other currents $I_{j,j\neq x}$ is however reduced. If only a single node x is considered, the objective is to minimize the overall voltage drop ΔV_x.

Consider the case where circuits A and B are connected to a simple power distribution network, as illustrated in Fig. 25.3a. The current sunk by circuits A and B is, respectively, I_A and I_B. The impedance from the power supply to the circuit is, respectively, Z_A and Z_B. The impedance of the power network between circuits A and B is denoted as Z_{AB}.

The voltage drop on the power distribution network at node A (the location where circuit A is connected to the power network) due to the current sunk by circuit A is

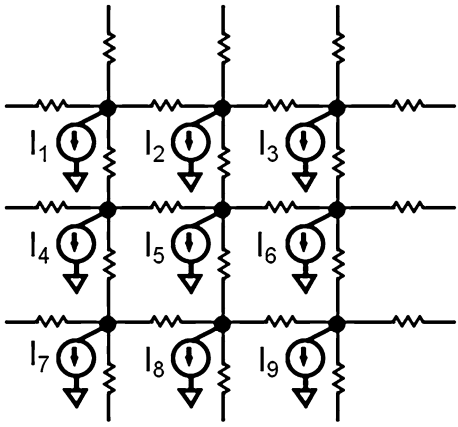


Fig. 25.2 Mesh structured power network with current sources

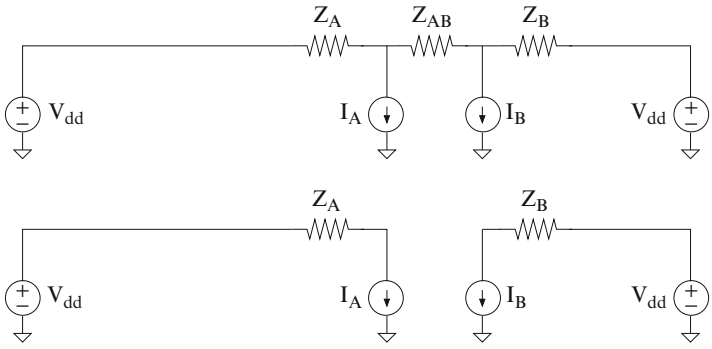


Fig. 25.3 Two circuits connected to a simple power distribution network, (**a**) common power network for both circuits, and (**b**) separate power networks for each circuit

$$\Delta V_A = I_A \cdot [Z_A \| (Z_{AB} + Z_B)]\,. \tag{25.1}$$

The voltage drop on the power distribution network at node A due to the current sunk by circuit B is treated as noise injected by circuit B at circuit A. This voltage drop is

$$\Delta V_A = I_B \cdot [(Z_A + Z_{AB}) \| Z_B] \cdot \frac{Z_A}{Z_A + Z_{AB}}. \tag{25.2}$$

The overall voltage drop at node A is the superposition of (25.1) and (25.2),

$$\begin{aligned} \Delta V_A =& I_A \cdot [Z_A \| (Z_{AB} + Z_B)] + \\ & I_B \cdot [(Z_A + Z_{AB}) \| Z_B] \cdot \frac{Z_A}{Z_A + Z_{AB}}. \end{aligned} \tag{25.3}$$

Similarly, the voltage drop at node B is

$$\begin{aligned} \Delta V_B =& I_B \cdot [Z_A \| (Z_{AB} + Z_B)] + \\ & I_A \cdot [(Z_A + Z_{AB}) \| Z_B] \cdot \frac{Z_B}{Z_B + Z_{AB}}. \end{aligned} \tag{25.4}$$

Assuming circuit B is an aggressor ($I_B \gg I_A$) and circuits A and B are located in close physical proximity ($Z_{AB} \ll Z_A$ and Z_B), the voltage drop at nodes A and B is dominated by the voltage drop ΔV_B. To protect circuit A from circuit B, link Z_{AB} should be disconnected, as illustrated in Fig. 25.3b, resulting in a voltage drop at nodes A and B, of respectively,

$$\Delta V_A = I_A Z_A, \tag{25.5}$$

and

$$\Delta V_A = I_B Z_B. \tag{25.6}$$

In this example, the objective is to determine if the link Z_{AB} needs to be broken. If the voltage drop at node A is lower when link Z_{AB} is disconnected as compared with the configuration where Z_{AB} is connected, link Z_{AB} should be broken. Note that by disconnecting link Z_{AB}, the voltage drop at node B also changes, requiring the voltage drop at node B to be evaluated and maintained below some limit.

Since every circuit is an aggressor and a victim, the problem formulation and objective needs to be generalized. Two parameters are therefore assigned to each circuit, one characterizing the aggressiveness and the second the sensitivity of a circuit. The aggressor parameter is directly related to the current sunk by a circuit. Simultaneously, every circuit exhibits a different sensitivity to variations in the power network voltage. For example, a PLL is highly sensitive to voltage variations

as compared to digital logic. Two circuits with a different critical path may also exhibit a different sensitivity to voltage variations: a slower critical path requires a smaller power drop, while a fast critical path can better tolerate a large voltage drop on the power network. A sensitivity factor is therefore assigned to each circuit connected to the power network. A more detailed discussion of the sensitivity factor is presented in Sect. 25.3.

In a system with multiple aggressors and victims, the objective is to minimize the effect of the voltage drop over the entire system. To improve the performance of an IC, the voltage drop is reduced in those circuits with high sensitivity at the expense of increasing the voltage drop in the less sensitive circuits.

Breaking a link between two circuits in a mesh structured power distribution network does not however completely isolate these circuits, rather resulting in an increase in the impedance between the two circuits. The larger impedance between the circuits lowers the noise coupling between the two nodes. Three specific nodes, the victim, aggressor, and power supply, within a mesh structured power distribution network, are illustrated in Fig. 25.4. The objective is to increase the network impedance between the victim and the aggressor nodes (Z_{AB}), reduce the influence of the aggressor on the victim node, while only minimally increasing the effective impedance between the aggressor and the power supply (Z_A).

The normalized effective resistance between the left and right nodes as a function of a specific disconnected link at a particular location (along the x-axis) is depicted in Fig. 25.5. A 20 × 20 node mesh structured network is illustrated in Fig. 25.6. The x-axis describes the location (or link number depicted in Fig. 25.6) of the disconnected link between two nodes. The largest increase in the effective resistance is achieved when breaking the link closest to either node. An 11 % increase in the resistance is caused by breaking a single link. This change confirms that breaking links within a mesh structured power distribution networks may result in a large change in the effective impedance; effectively shielding the victim from the aggressor.

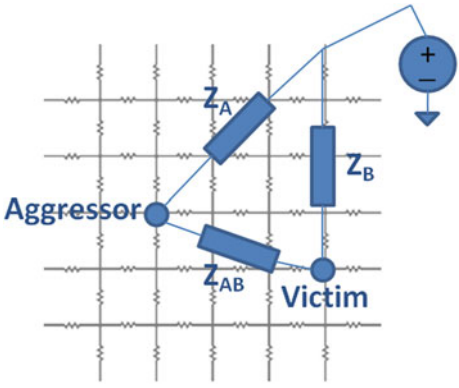


Fig. 25.4 Aggressor and victim circuits sharing a mesh structured power distribution network. The objective is to increase Z_{AB}, while insignificantly increasing Z_A, resulting in shielding the victim from an aggressor

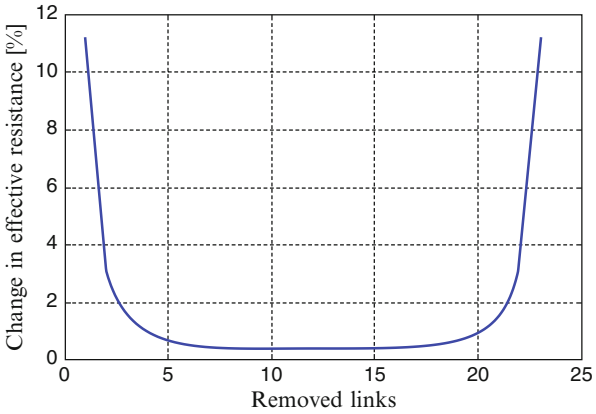


Fig. 25.5 A change in the effective resistance between the left and right nodes within a 20 × 20 mesh structured power distribution network (see Fig. 25.6) as a function of a specific location of a disconnected link between two nodes

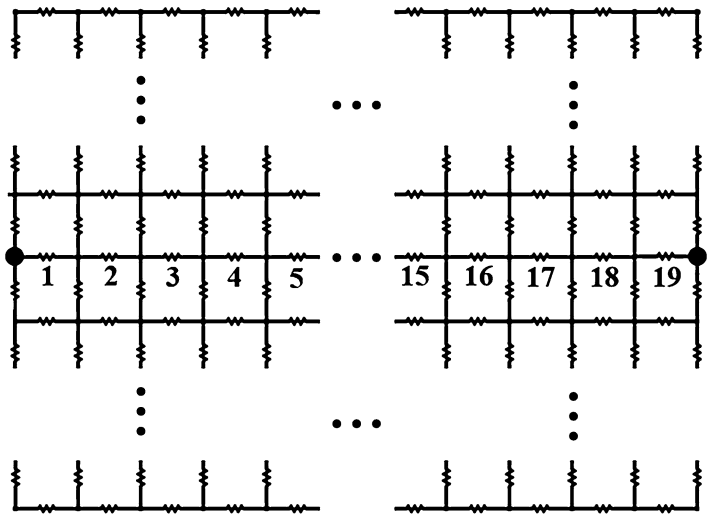


Fig. 25.6 20 × 20 node mesh structured network. The effective resistance is between the two bold nodes. The links are numbered based on the location along the horizontal path

25.2 Single Aggressor and Victim Example

A single aggressor and single victim example is provided in this section, intuitively illustrating the problem and solution. A 25 × 25 node mesh structured power network is illustrated in Fig. 25.7a. Two power pads are located at the top/right and bottom/left nodes. The aggressor, a large current sink, is connected in the left/center region of the mesh network, while the victim circuit is connected in the right/center

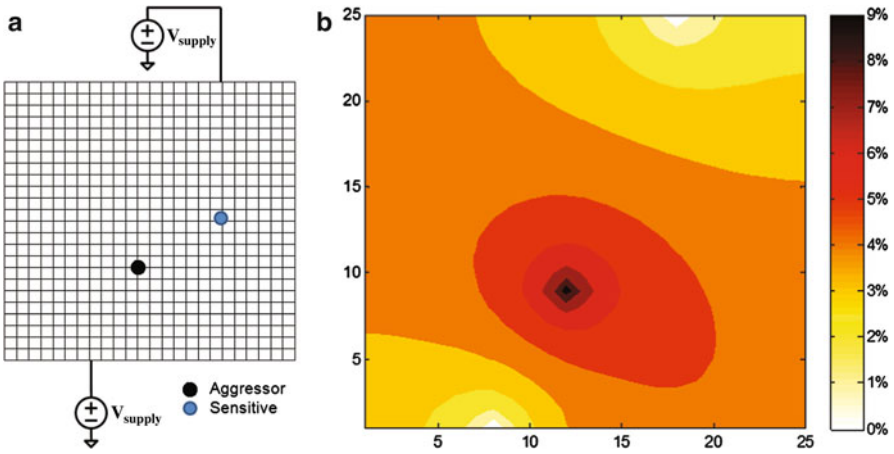


Fig. 25.7 Two circuits, an aggressor and a victim, are connected to a 25 × 25 node mesh structured power distribution network. (**a**) Schematic of the power network, and (**b**) map of voltage drops within the power distribution network before disconnecting any links

region. The current sunk from the power network by an aggressor is two orders greater than the current sunk by the victim circuit. A 1 V power supply voltage is assumed. The voltage drop is shown in Fig. 25.7b as a shade of color, where the darker shade represents a higher voltage drop. The aggressor and victim nodes exhibit, respectively, a 96 and 47 mV voltage drop.

As discussed previously, the design objective is to reduce the voltage drop at the victim node, while insignificantly increasing the voltage drop at the aggressor node. The disconnected links are therefore required to be far from the aggressor and close to the victim. The links are removed around the victim node, isolating the victim from the rest of the network while maintaining a single connection to the power supply (the lowest voltage drop). A large number of additional connections may also be provided based on reliability and current density constraints. The procedure is repeated until the target low voltage drop at the victim node is achieved. Note that the voltage drop at the aggressor node is simultaneously monitored while disconnecting links. The procedure is discontinued once the voltage drop at the aggressor node exceeds the target limit or the desired voltage drop at the victim node is achieved.

The voltage drop for the revised 25 × 25 node mesh structured network is illustrated in Fig. 25.8a. Nine sets of links have been disconnected, producing a 20 mV voltage drop at the victim node, while the voltage drop at the aggressor node has increased from 96 to 98 mV, as depicted in Fig. 25.8b. The improvement in the variation of the power voltage at the victim node is 135 %, while a voltage degradation of only 2.1 % is observed at the aggressor node. Since in practical applications each node within the network can be simultaneously both an aggressor

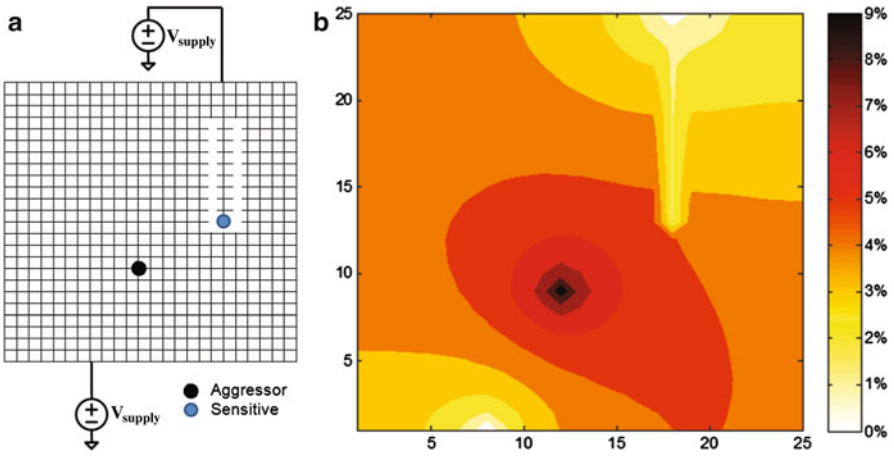


Fig. 25.8 Two circuits, an aggressor and a victim circuit, are connected to a 25 × 25 node mesh structured power distribution network. (**a**) Schematic of the power network, and (**b**) map of voltage drops within the power distribution network after disconnecting nine pairs of links. Note that the voltage drop for the victim circuit is significantly lower after disconnecting the links

and a victim, the methodology addresses this issue based on maximizing the overall performance of a circuit.

25.3 Sensitivity Factor

The sensitivity factor describes the relative importance of a change in voltage on the performance of a circuit. A method to describe the sensitivity factor is to investigate the sensitivity of the supplied voltage on the performance (for example, the propagation delay) of a particular circuit. The sensitivity factor is [503]

$$s = \left. \frac{\frac{\Delta delay}{delay(x)}}{\frac{\Delta V}{V(x)}} \right|_{x=V_{dd}} = \frac{\Delta delay}{\Delta V} \cdot \frac{V_{dd}}{delay_{min}}, \tag{25.7}$$

where $\Delta delay$ and $delay_{min}$ are, respectively, the change in the delay and the minimum delay. The minimum delay is achieved assuming a full V_{dd} at the power rail of the circuit. ΔV is the change in the supply voltage at the node supplied to the circuit. The sensitivity factor is dependent on the type of circuit.

Consider an example where two circuits are connected to the power network, as depicted in Fig. 25.9. With a power supply of 1 V (full V_{dd}) applied to $node_A$ and $node_B$, the propagation delay of the critical path within circuit A is 150 ps, while the propagation delay of the critical path within circuit B is 90 ps. Reducing the power level by 10 %, the delay of circuits A and B is, respectively, degraded to

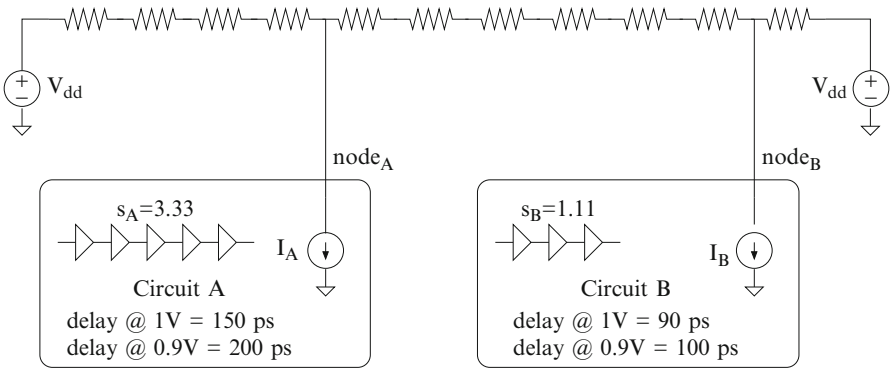


Fig. 25.9 Example of determining the sensitivity factor, where two circuits have different propagation delays

200 and 100 ps. The resulting sensitivity factor for circuits A and B is therefore, respectively,

$$S_A = \frac{200 - 150\,\text{ps}}{1 - 0.9} \frac{1}{200\,\text{ps}} = 3.33, \tag{25.8}$$

and

$$S_A = \frac{100 - 90\,\text{ps}}{1 - 0.9} \frac{1}{200\,\text{ps}} = 1.11. \tag{25.9}$$

25.4 Link Breaking Methodology

An algorithm for determining which links should be removed, thereby shielding the sensitive circuits, is described in this section. Since each circuit within a network can be characterized as both an aggressor and a victim, each node of interest is associated with a matrix composed of two parameters $[i, s]$. Parameter i is an aggressor related parameter, which is equal to the current sunk from the network. Parameter s is related to the victim, expressing the sensitivity of the circuit connected to the node. The objective is to enhance overall performance, such as minimize the average (25.10) or worst case (25.11) delay.

$$delay_{average} = \frac{1}{k} \sum_{j=1}^{k} \left(delay_j \right), \tag{25.10}$$

$$delay_{worst} = max \left(delay_1, delay_2, \ldots, delay_k \right), \tag{25.11}$$

LINK-BREAKING
1. Determine voltage drops at all k nodes
2. Calculate initial $delay_{initial}$ function based on (25.11)
3. Generate x randomly perturbed systems
4. Determine voltage drops at k nodes for x systems
5. Calculate $delay$ function based on (25.11) for x systems
6. For every x systems
7. Generate six different networks, where a link is broken at every direction
8. Determine new $delay$ values, maintaining network with lowest $delay$
9. Goto 7, if enchantment is achieved
10. Select system with lowest $delay$
11. If $delay_{initial} > delay$, $delay_{initial} \leftarrow delay$ and goto 3

Fig. 25.10 Pseudocode for link breaking algorithm

where

$$delay_j = delay_{\min -j} \cdot \left[\frac{s_j}{V_{dd}} \cdot \Delta V_j + 1 \right]. \tag{25.12}$$

ΔV_j is a change in the voltage at node j due to load currents $i_1, i_2, \ldots, i_k$ and the impedance of the mesh structured power distribution network. $delay_{min-j}$ is the minimum propagation delay of circuit j achieved by applying the maximum supply voltage V_{dd}. s_j is the sensitivity factor of circuit j.

Pseudocode of the LINK-BREAKING algorithm is provided in Fig. 25.10, with the objective of minimizing the worst case propagation delay. Other algorithms can be used which may yield enhanced computational efficiency or a global solution to the link breaking methodology. In line 1, the voltage drop at k nodes (all aggressor/sensitive nodes) is determined. Based on the voltage and sensitivity of the circuits, the initial value of the delay function $delay_{initial}$ is determined, as listed in line 2. The revised number of power networks x is generated, where each network is perturbed by removing a random link. In lines 4 and 5, the voltage drop and delay function are determined for each of the perturbed networks. A search for a local minimum is evaluated for each perturbed system in lines 6–9. The network with the lowest delay is selected in line 10. The process is repeated until the value of the delay function cannot be further reduced.

Since k nodes of interest are typically lower than the overall number of nodes in a system, a random walk procedure can be used to efficiently determine the voltages [504], trading off accuracy with runtime. The number of parallel random walk procedures is based on the target accuracy.

25.5 Case Studies

Five study cases are presented in this section. In each of the cases, the circuit is composed of nine blocks. In the first case study, the current sunk by every block is maintained equal. The sensitivity factor and critical delay of each block are however assumed different. For the following two cases, one block sinks significantly greater current, representing the case of a single dominant aggressor. In the final two cases, the current and delay of the nine blocks are varied, representing general circuits. The design objective is to minimize the worst case propagation delay; (25.11) is therefore used for all of the five case studies.

A mesh structured power distribution network with 20×20 number of nodes is considered. A block diagram of the circuit is schematically illustrated in Fig. 25.11. Four one volt power supplies are connected at the center of the four edges (left, right, top, and bottom). The maximum permitted degradation in supply voltage is 0.3 V.

For Case 1, a map illustrating the variation in voltage over the mesh structured power network is shown in Fig. 25.12a. A darker shaded color represents a lower voltage within the power network. After applying the link breaking methodology, 38 links are disconnected from 760 possible links within the power network. The voltage map after the link breaking methodology is shown in Fig. 25.12a. The

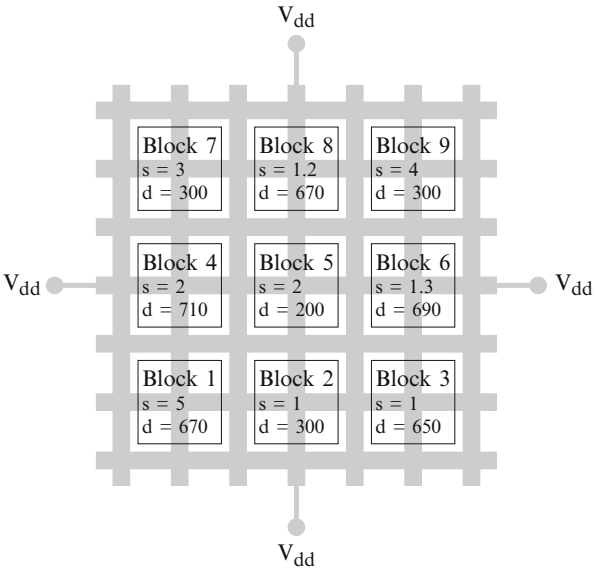
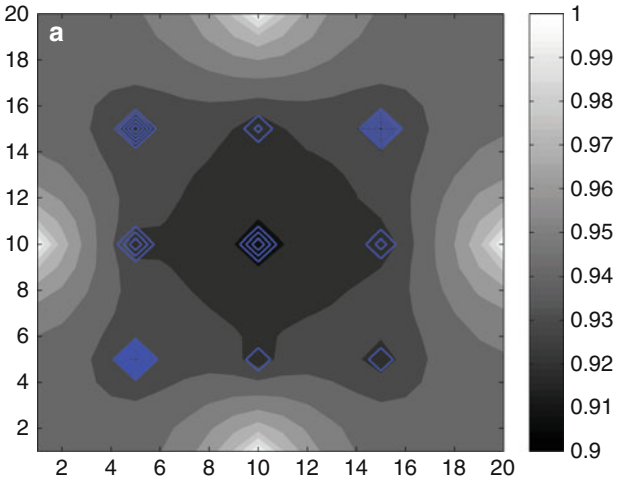
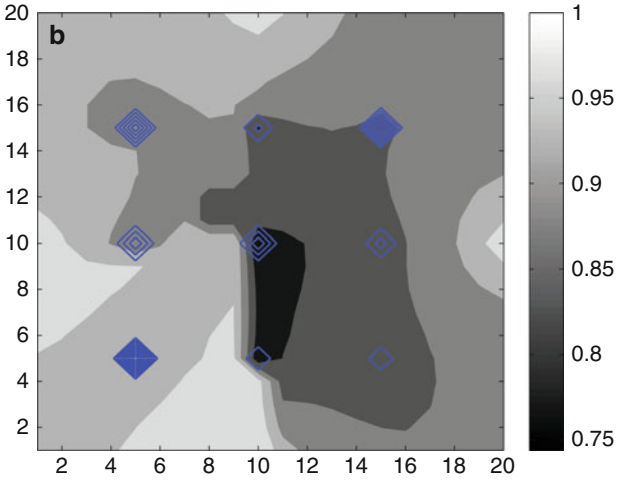


Fig. 25.11 Nine circuit blocks are connected to a mesh structured power distribution network. Four power supplies provide the current. The numbers within the blocks represent the sensitivity factor (s) and propagation delay in ps (d) when applying 1 V to the block. Note that the minimum propagation delay is achieved when applying a full power supply



Case 1: Supply voltage before applying the link breaking methodology.



Case 1: Supply voltage after applying the link breaking methodology.

Fig. 25.12 Supply voltage before and after the link breaking methodology for Case 1. The diamond shapes represent the location of the aggressor/victim circuit blocks. In this example, the current sunk by each of the nine blocks is assumed equal. The voltage drop is reduced in the more sensitive circuit blocks (resulting in a smaller delay), while increased in the less sensitive circuit blocks (resulting in a higher delay). (**a**) Case 1: Supply voltage before applying the link breaking methodology. (**b**) Case 1: Supply voltage after applying the link breaking methodology

resulting power network is illustrated in Fig. 25.13. Note that the power supply voltage is increased at the lower left corner due to the high sensitivity factor ($s = 5$)

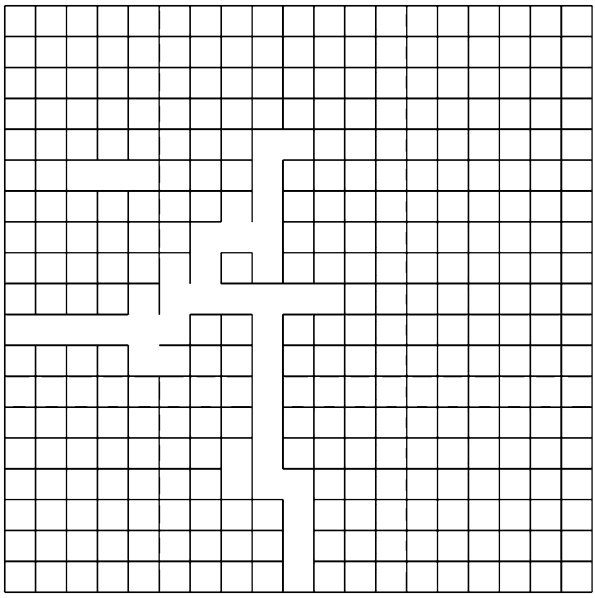


Fig. 25.13 Resulting power network after applying the link breaking methodology for Case 1

and propagation delay assigned to the block located in the lower left corner. The supply voltage is reduced at the other locations due to a lower sensitivity factor or small delay assigned to the block.

The voltage and propagation before and after application of the link breaking methodology for each block are listed in Table 25.1.

The sensitivity factor, current, minimum delay, and improvement or degradation in the voltage and delay are also listed. A close to 4 % improvement in the supply voltage, 95 % of the ideal power supply, is achieved for block 1. Note that the

Table 25.1 Case 1. Sensitivity factor, sunk current, minimum delay (achieved with 1 V at the power rail of the block), supply voltage, and propagation delay before and after the link breaking methodology for the nine circuit blocks. The improvement or degradation in the supply voltage, propagation delay, and maximum operating frequency are also listed. A 1 V power supply is used

Block number		1	2	3	4	5	6	7	8	9	f_{max}
Sensitivity factor (s)		5	1	1	2	2	1.3	3	1.2	4	N/A
Sunk current (A)		1	1	1	1	1	1	1	1	1	N/A
Delay (ps) @ $V_{dd} = 1$ V		670	300	650	710	200	690	300	300	300	N/A
Voltage	Before (mV)	914	914	913	914	900	910	913	910	912	N/A
	After (mV)	949	840	877	909	829	880	912	864	891	N/A
	Improvement (%)	3.8	−8.1	−3.9	−0.5	−7.9	−3.3	0.1	−5.1	−2.3	N/A
Delay	Before (ps)	1024	334	723	880	249	785	393	348	430	0.98 GHz
	After (ps)	861	372	760	892	274	816	387	357	440	1.16 GHz
	Improvement (%)	15.9	−11.4	−5.1	−1.4	−10.0	−3.9	1.5	−2.6	−2.3	15.9

maximum improvement in the supply voltage is 9 %, producing a supply voltage of 1 V. The improvement in voltage is achieved at the expense of a lower supply voltage at the other blocks. The performance of the overall circuit is increased since the worst case propagation delay at block 1 is reduced. Due to the higher supply voltage, the propagation delay at block 1 is lowered from 1 ns to 861 ps, permitting an increase in the maximum operating frequency.

For Cases 2, 3, 4, and 5, the current is different among the circuit blocks. The supply voltage map before and after application of the link breaking methodology, as well as the resulting power network, is illustrated in Fig. 25.14.

The current sunk for each of the cases, voltage before and after application of the methodology, sensitivity, propagation delay, and improvement in the supply voltage and propagation delay are listed in Table 25.2.

Case 2 (Fig. 25.14a, b) and Case 3 (Fig. 25.14d, e) illustrate those cases where the current sunk by a circuit (the aggressor) is significantly higher as compared to the other circuit blocks. The sensitivity factors and minimum delay are the same as in Case 1. The highest degradation in the supply voltage is within the aggressor circuit; however, the supply voltage is greater in those circuit blocks with a higher sensitivity and minimum delay, resulting in a reduction in the worst case delay and a higher maximum operating frequency. The increase in the supply voltage at block 1 is 5 %, achieving 97 % of the ideal power supply voltage and resulting in an improvement in the propagation delay of 16 %. Note that the improvement in the propagation delay is greater than the supply voltage due to the high sensitivity factor. After applying the link breaking methodology, blocks 1, 4, and 6 exhibit a similar worst case propagation delay, demonstrating the effectiveness of the methodology. In Case 3, the voltage at block 1 is increased by 2 %, achieving 96 % of the ideal power supply voltage and resulting in an improvement in the propagation delay of 5 %.

Case 4 (Fig. 25.14g, h) and Case 5 (Fig. 25.14j, k) represent cases where different current is sunk. After applying the link breaking methodology, the supply voltage at block 1 is increased by, respectively, 3 % and 5 %. The maximum operating frequency is enhanced by, respectively, 8 % and 17 %.

25.6 Discussion

The voltage drop within a power distribution network is evaluated for circuit blocks with different current levels and sensitivities. The minimum propagation delay ($delay_{min}$) is maintained the same. A 20 × 20 mesh structured power distribution network with two power supplies and two current sources (one aggressor and one victim) is considered. The voltage improvement at the victim and degradation at the aggressor are illustrated, respectively, in Fig. 25.15a, b. Note that by assigning a higher sensitivity to the victim circuit, the voltage drop on the power network at the victim is reduced. Simultaneously, the voltage drop at the aggressor is increased, while the aggressor is less sensitive to voltage variations. The tradeoff between

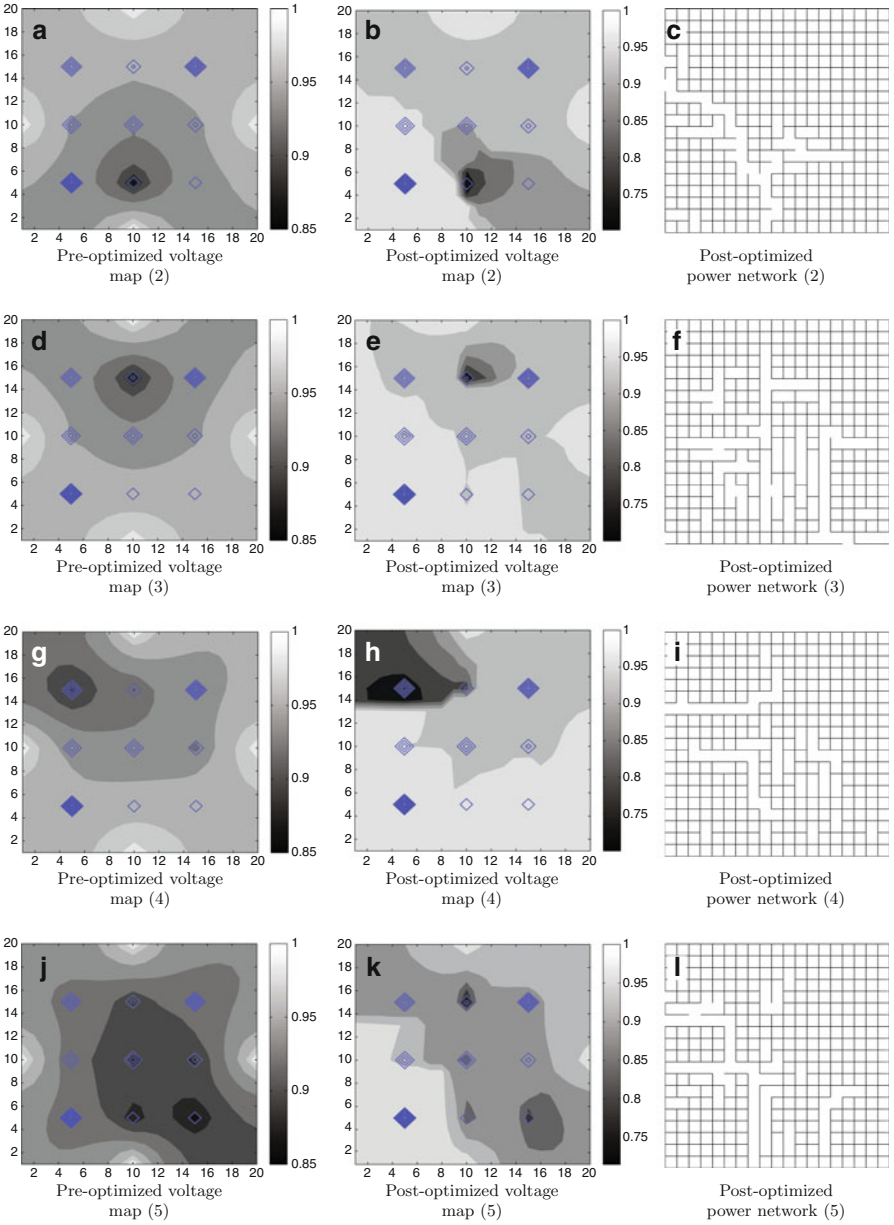


Fig. 25.14 Map of voltage variations before and after application of the link breaking methodology for Cases 2, 3, 4, and 5. The *diamond shapes* represent the location of the aggressor/victim circuit blocks. The resulting power network after the link breaking methodology is also illustrated. Cases 2 and 3 represent the cases where a single block sinks significantly higher current as compared to the other blocks. In Cases 4 and 5, the sunk current, sensitivity factor, and delay are different for different blocks, representing general design cases. (**a**) Pre-optimized voltage map (2).

reducing the voltage drop at the victim while increasing the voltage drop at the aggressor is an important aspect of the link breaking methodology.

The improvement and degradation of the voltage drop at, respectively, the victim and aggressor are depicted in Fig. 25.16 for different ratios of the current sunk by the victim and aggressor, assuming the two circuits have equal sensitivity. Note that a higher change in voltage is achieved at the victim when the current sunk by the aggressor is greater. This effect is due to the dominance of the aggressor on the victim circuit before applying the link breaking methodology. The link breaking methodology can therefore be used to reduce the voltage drop at the victim.

The computational runtime of the algorithm, depicted in Fig. 25.10, is evaluated for differently sized power distribution networks. The algorithm has been executed on a Linux eight-core with 8 GB RAM system. The runtime as a function of the number of nodes in the power network is depicted in Fig. 25.17. The runtime of the link breaking methodology can also be accelerated by utilizing multigrid-like techniques [505] and ignoring those current sources located farther from the target nodes. The number of aggressor and/or victim circuits is not a dominant factor affecting the runtime of the algorithm, as illustrated in Fig. 25.18. Initially, the runtime increases exponentially with the number of aggressor and victim circuits. With a further increase in the number of circuits, the computational runtime decreases due to the smaller number of links that can be disconnected. For those cases where only a small number of circuits are evaluated within a large power distribution network, the random walk method [504] can be used to estimate the voltage variations, significantly accelerating the link breaking methodology.

The worst case voltage drop (located at the aggressor) cannot be reduced by utilizing the link breaking methodology, since the methodology always increases the worst case impedance of the power network. However, the effect of the aggressor on other circuits with a higher sensitivity and propagation delay can be reduced, resulting in enhanced overall system performance.

25.7 Summary

The design of the power distribution network is an essential part of an IC design flow. The network is typically designed as a single network or multiple separate networks. The advantages of a single network are a reduced network impedance and fewer routability constraints, while multiple separate networks have the advantage of lower noise coupling. The link breaking methodology utilizes a single network,

Fig. 25.14 (**b**) Post-optimized voltage map (2). (**c**) Post-optimized power network (2). (**d**) Pre-optimized voltage map (3). (**e**) Post-optimized voltage map (3). (**f**) Post-optimized power network (3). (**g**) Pre-optimized voltage map (4). (**h**) Post-optimized voltage map (4). (**i**) Post-optimized power network (4). (**j**) Pre-optimized voltage map (5). (**k**) Post-optimized voltage map (5). (**l**) Post-optimized power network (5)

Table 25.2 Sensitivity factor, sunk current, minimum delay, supply voltage, and propagation delay before and after application of the link breaking methodology for the nine circuit blocks. The improvement or degradation in the supply voltage, propagation delay, and maximum operating frequency are also listed. Cases 2 and 3 represent the cases where a single block sinks significantly higher current as compared to the other blocks. In Cases 4 and 5, the sunk current, sensitivity factor, and delay are different for different blocks, representing general design cases

Block number		1	2	3	4	5	6	7	8	9	f_{max}
Sensitivity factor (s)		5	1	1	2	2	1.3	3	1.2	4	N/A
Delay (ps) @ $V_{dd} = 1$ V		670	300	650	710	200	690	300	300	300	N/A
Case 2 (see Fig. 25.14a, b)											
Sunk current (A)		1	10	1	1	1	1	1	1	1	N/A
Voltage	Before (mV)	924	850	924	936	920	933	943	940	942	N/A
	After (mV)	973	702	861	965	896	921	921	925	931	N/A
	Improvement (%)	5.3	−17.4	−6.8	3.0	−2.6	−1.3	−2.3	−1.6	-1.2	N/A
Delay	Before (ps)	988	369	748	856	248	803	376	343	395	1.01 GHz
	After (ps)	829	422	807	834	263	828	403	356	417	1.20 GHz
	Improvement (%)	16.0	−15.0	−7.8	2.6	−5.8	−3.2	−7.4	−3.7	-5.7	15.8
Case 3 (see Fig. 25.14d, e)											
Sunk current (A)		1	1	1	1	1	1	1	10	1	N/A
Voltage	Before (mV)	945	944	944	937	922	934	925	850	924	N/A
	After (mV)	966	934	924	947	936	906	939	700	909	N/A
	Improvement (%)	2.1	−1.1	−2.0	−1.0	1.6	−3.0	1.5	−17.6	−1.6	N/A
Delay	Before (ps)	904	336	727	847	245	794	390	375	413	1.11 GHz
	After (ps)	855	349	762	857	243	844	387	445	446	1.17 GHz
	Improvement (%)	5.4	−3.9	−4.8	−1.2	0.8	−6.3	0.8	−18.7	−8.0	5.4
Case 4 (see Fig. 25.14g, h)											
Sunk current (A)		5	1	1	2	2	1.3	3	12	4	N/A
Voltage	Before (mV)	939	949	949	930	924	914	850	899	927	N/A
	After (mV)	967	953	959	950	938	898	700	775	923	N/A
	Improvement (%)	3.0	0.4	1.0	2.1	1.5	−1.7	−17.6	−13.7	−0.4	N/A
Delay	Before (ps)	929	334	724	858	244	814	461	357	411	1.08 GHz
	After (ps)	851	332	712	853	243	851	621	415	428	1.17 GHz
	Improvement (%)	8.3	0.6	1.7	0.7	0.4	−4.5	−34.7	−16.2	−4.1	7.8
Case 5 (see Fig. 25.14j, k)											
Sunk current (A)		1	5	5	2	2	3	1.3	4	1.2	N/A
Voltage	Before (mV)	907	861	850	901	874	875	907	876	901	N/A
	After (mV)	958	825	781	928	838	864	852	716	890	N/A
	Improvement (%)	5.5	−4.3	−8.1	2.9	−4.1	−1.2	−6.1	−18.2	-1.3	N/A
Delay	Before (ps)	1050	366	800	910	268	860	411	369	448	0.95 GHz
	After (ps)	870	378	847	870	283	869	463	430	463	1.15 GHz
	Improvement (%)	17.1	−3.2	−5.8	4.5	−6.0	−1.1	−12.7	−16.5	-3.3	17.1

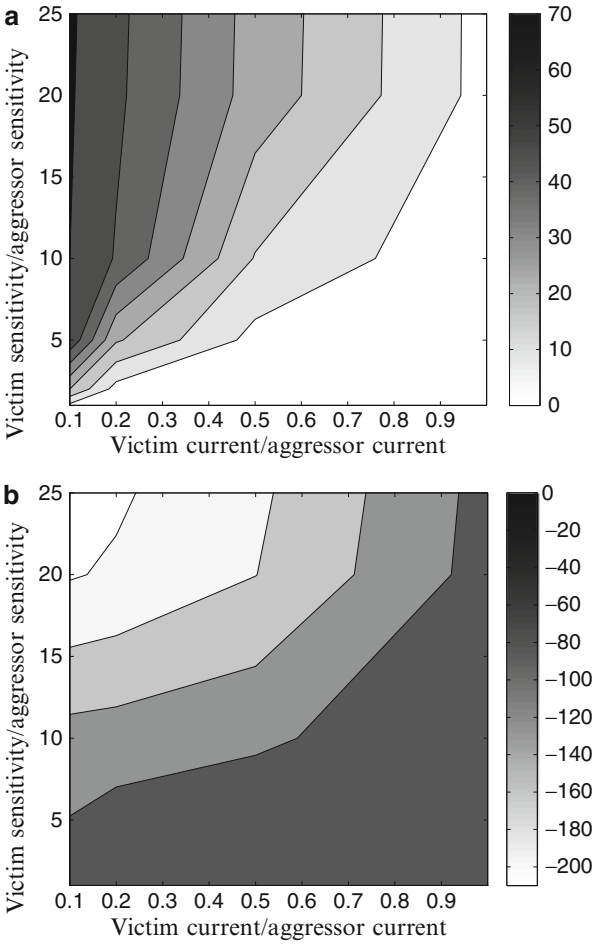


Fig. 25.15 Change in voltage drop at the (**a**) victim, and (**b**) aggressor circuit. The *darker shade* represents a greater reduction in the voltage drop at the victim and a lower increase in the voltage drop at the aggressor

disconnecting links between the aggressive and sensitive circuits; thereby isolating the victim from the aggressor. This approach reduces the noise, while maintaining a low network impedance.

- Sensitivity to changes in the supply voltage varies for different circuits
- Voltage variations at the more sensitive circuits need to be reduced at the expense of increased voltage variations at the less sensitive circuits
- A smaller voltage drop is also important in long critical paths as compared to shorter less critical logic paths

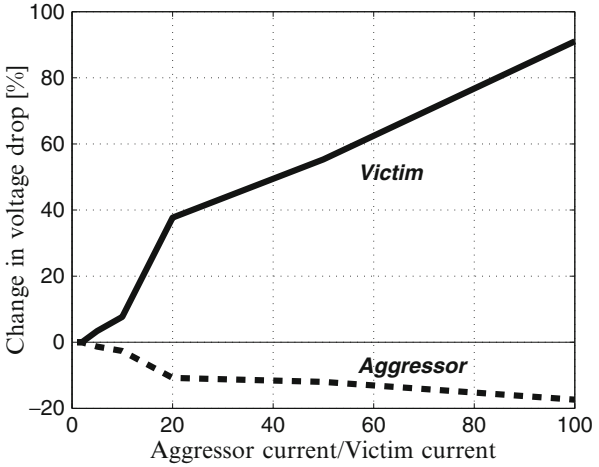


Fig. 25.16 Change in voltage within the power distribution network for the victim and aggressor circuits as a function of the ratio of the current sunk by the aggressor and victim circuits. The sensitivity factor is assumed equal for both circuits

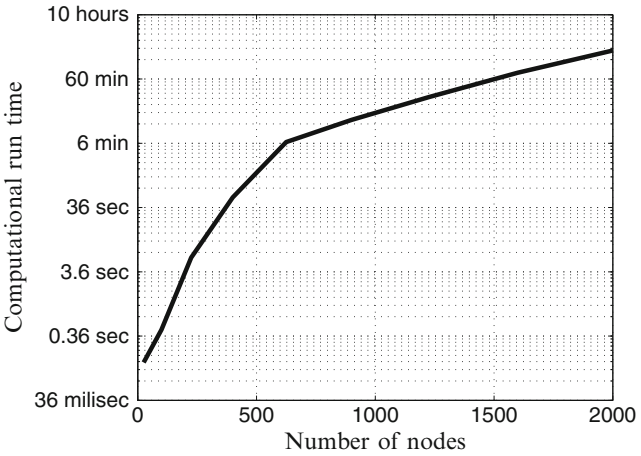


Fig. 25.17 Computational runtime of the link breaking methodology as a function of the number of nodes within the power distribution network

- The aggressiveness and sensitivity of circuits are considered during the link breaking process.
- The methodology is evaluated for two cases, targeting the reduction in the worst case propagation delay by increasing the supply voltage at blocks with a high propagation delay

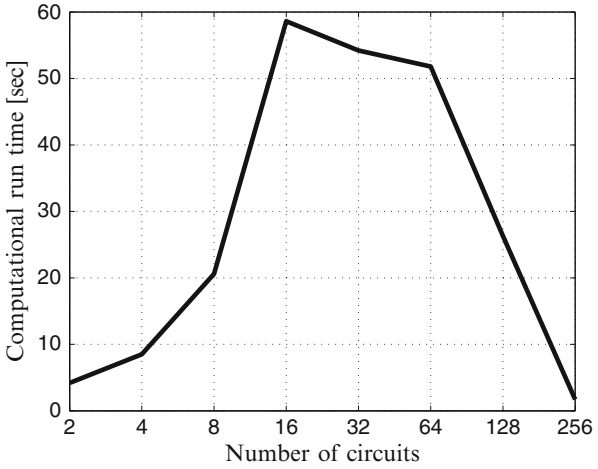


Fig. 25.18 Computational runtime of the link breaking methodology as a function of the number of victim and aggressor circuits. The runtime initially increases with a higher number of circuits. After reaching a peak, the runtime decreases due to the smaller number of links that can be removed

- An average enhancement of 4 % in power supply voltage at nodes with high sensitivity and high propagation delay is achieved, resulting in, on average, 96 % of the ideal power supply voltage at these nodes
- An average improvement of 11 % in the maximum operating frequency is achieved when utilizing the link breaking methodology

Chapter 26
Power Supply Clustering in Heterogeneous Systems

On-chip power integration is necessary for delivering high quality power to modern high performance circuits. The tradeoff between power efficiency and area for switching and linear power supplies is discussed in Chap. 16. To optimize the power efficiency of a system with existing power supplies, the power should be primarily converted with a few power efficient switching supplies, delivered to on-chip voltage clusters, and regulated with linear low dropout regulators within the individual power domains. This principle with multiple voltage clusters is illustrated in Fig. 26.1 by a heterogeneous power delivery system with multiple power domains, off-chip/in-package/on-chip SMPS power converters, and on-chip LDO power regulators.

Several schemes for heterogeneous power delivery [191, 310, 362, 476, 506, 507] that consider tens to hundreds of on-chip power regulators have recently been described. Optimizing the power delivery process in terms of the co-design of the on-chip voltage regulators, decoupling capacitors, and current loads is described in [191, 310, 362, 476, 507]. The co-design of hundreds to thousands of on-chip regulators with multiple switching converters is a new design objective. In energy efficient systems, the voltage and current are dynamically scaled within the individual power domains, affecting the voltage drop across the LDO regulators and the overall efficiency of the power delivery system. Optimal real-time clustering of the power supplies decreases the voltage drop within the LDO regulators, increasing overall power efficiency. Exhaustive approaches for clustering power supplies are computationally impractical in DVS systems with hundreds to thousands of power domains. Other existing approaches for on-chip power delivery are ad hoc in nature and not optimal. A computationally efficient methodology to co-design in run time switching converters and on-chip LDO regulators within a heterogeneous system is described in this chapter, achieving high quality power and efficiency within limited on-chip area. The power savings with this approach are evaluated with IBM power

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0_26

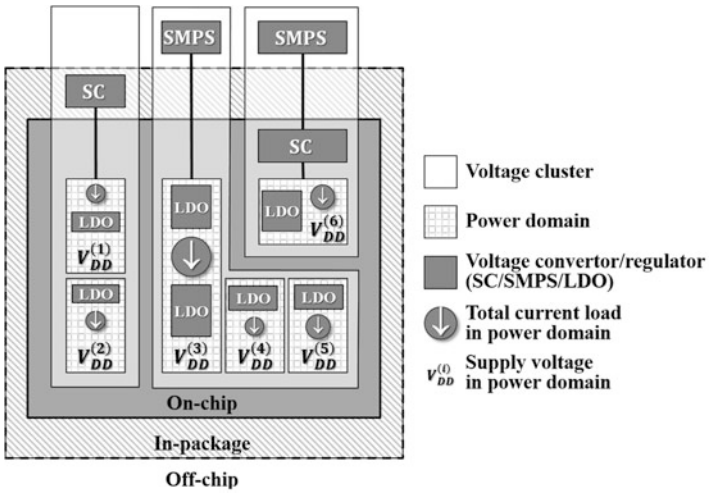


Fig. 26.1 Heterogeneous power delivery with multiple power domains

grid benchmark circuits, demonstrating up to 24 % increase in power efficiency with the voltage clusters [508, 509]. Significant speedup is exhibited with the recursive clustering algorithm, exhibiting polynomial computational complexity.

The rest of the chapter is organized as follows. Heterogeneous power delivery is described in Sect. 26.1 to both decrease the noise and increase the efficiency of the supplied power. The effects of the number of power supplies, on-chip power supply voltages, and clustering topology on the power efficiency of a dynamically controlled heterogeneous power delivery systems is described in Sect. 26.2. Computationally efficient algorithms for optimal and near-optimal clusters of power supplies are demonstrated in Sect. 26.3. Separation of power conversion and regulations in benchmark circuits is evaluated in Sect. 26.4. The chapter is concluded in Sect. 26.5.

26.1 Heterogeneous Power Delivery System

Both linear and switching power regulators are characterized by an undesirable power-area tradeoff, exhibiting either high power in compact linear regulators or large area in power efficient SMPS, as depicted in Fig. 26.2. Thus, the overhead of a power delivery system composed of only switching or linear regulators is significant. Several power delivery solutions exist that exhibit intermediate power losses and area as compared to either linear or traditional SMPS systems. For example, in a PSiP system, lower power losses as compared to a linear system, and smaller area as compared to a traditional off-chip SMPS system, are traded off

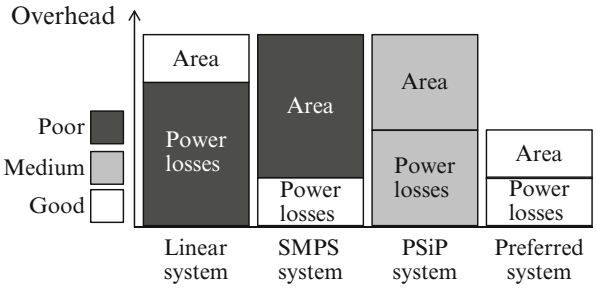


Fig. 26.2 Power and area overhead of a linear, SMPS, PSiP, and preferred power conversion system

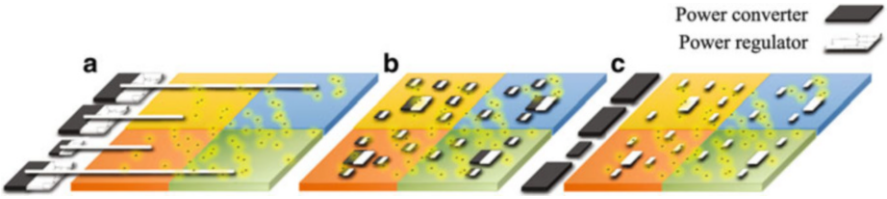


Fig. 26.3 Power delivery system with four voltage domains, utilizing (**a**) off-chip power supplies, (**b**) distributed POL power supplies, and (**c**) a heterogeneous system with off-chip converters and on-chip regulators

for greater design complexity. A desirable power delivery system minimizes power losses while satisfying on-chip area constraints, yielding both high power efficiency and small area, as depicted in Fig. 26.2.

To exploit the advantages of switching and linear converters, a heterogeneous power delivery system is considered that converts the power in off-chip switching power supplies and regulates the on-chip power with compact linear power supplies, minimizing LDO voltage drops and on-chip power losses. In a heterogeneous power delivery system, the area overhead is primarily constrained by the compact LDOs that regulate the on-chip power, while the power overhead is dictated by the power efficient switching converters. Power conversion is therefore decoupled from power regulation, lowering the power and area overhead of the overall power delivery system. A heterogeneous power delivery system moderates the drawbacks and exploits the advantages of the historically power efficient power supplies that both convert and regulate the power off-chip with more recent trends for area efficient distributed power supplies that both convert and regulate the power on-chip. Off-chip, on-chip distributed, and heterogeneous power delivery topologies are illustrated in Fig. 26.3.

Consider a heterogeneous power delivery system with L on-chip LDOs and S off-chip SMPSs that deliver power to N voltage domains $\{(V_{DD}^{(i)}, I_{DD}^{(i)})\}_{i=1}^{N}$ with an operating voltage $V_{DD}^{(i)}$ and current $I_{DD}^{(i)}$. To supply the required voltages, $V_{DD}^{(i)} \neq V_{DD}^{(j)}$ $\forall i \neq j$, the number of on-chip power supplies L should be equal to or greater

than the number of voltage domains $N \leq L$. Alternatively, each SMPS drives one or more LDOs, yielding the relation, $S \leq L$. The effects of the number of on-chip power regulators, off-chip power converters, and distribution of the on-chip power supplies in a heterogeneous power delivery system are described, respectively, in Sects. 26.1.1, 26.1.2, and 26.1.3.

26.1.1 Number of On-Chip Power Regulators

The area of an LDO is proportional to the current load (see (16.18)), and the power efficiency is primarily dictated by the current load and voltage drop V_{Drop} across the power transistor within the LDO (see (16.19)). Thus, a single LDO that provides a specific current and voltage to a load consumes approximately the same area and dissipates similar power as numerous LDOs providing the same total current and voltage to a load. Consider l_k on-chip distributed LDOs to maintain a regulated voltage V_{DD} and load current I_{DD} within a specific voltage domain (V_{DD}, I_{DD}). Let I_i $(i = 1, \ldots, l_k)$ be a local current load supplied by a single LDO within a domain, such that $\sum I_i = I_{DD}$. The LDO area A_i is assumed here to be linearly proportional to the supply current I_i within a specific current range (see (16.18)), $A_i = \alpha \cdot I_i$. The l_k LDOs form a distributed on-chip power regulation system with a total size, $A = \sum A_i = \alpha \cdot \sum I_i = \alpha \cdot I_{DD}$. Therefore, the total area of the distributed regulation system does not strongly depend on l_k, the number of LDOs. To maximize the power efficiency of a system, all of the LDOs operate at the minimum voltage drop V_{Drop}, exhibiting a total power loss $V_{Drop} \cdot \sum I_i = V_{Drop} \cdot I_{DD}$ which is independent of l_k. Alternatively, the distance between an LDO and a current load is reduced at higher values of l_k, decreasing the on-chip voltage drops and increasing the quality of the supplied power.

26.1.2 Number of Off-Chip Power Converters

Intuitively, the number of off-chip voltage levels increases with the larger number of off-chip converters, increasing the granularity of the voltage levels supplied to the on-chip regulators and lowering the voltage drop across the hundreds of ultra-small regulators distributed on-chip. To minimize the voltage drop across an on-chip linear regulator, each off-chip SMPS converter should drive a single on-chip LDO. In practice, however, the number of power converters that can be placed off-chip is limited. Thus, each off-chip SMPS supplies power to several on-chip LDOs within an SMPS cluster. As a result, the voltage drop across the on-chip regulators is greater, degrading the overall power efficiency of the system. The upper and lower bounds of the power efficiency of a heterogeneous system for a specific number of SMPS are described in this section.

Given N power domains $\{(V_{DD}^{(i)}, I_{DD}^{(i)})\}_{i=1}^{N}$ sorted by the supply voltages $V_{DD}^{(i)} < V_{DD}^{(j)}\ \forall i < j$, and l_k power supplies in the kth power domain, $L = \sum_{k=1}^{k=N} l_k$ linear power supplies should be distributed on-chip to deliver high quality power to the load circuitry. To explore the area-power efficiency tradeoff in a heterogeneous power delivery system, a single linear regulator is assumed capable of providing sufficient high quality current within a power domain, yielding $l_k = 1\ \forall k$ and $L = N$. The voltage supplied by an LDO to a power domain cannot be stepped up by an LDO. The output voltage of each SMPS is therefore higher than the voltage within the individual power domains, increasing the voltage drop across the LDOs within an SMPS cluster, degrading power efficiency.

An expression for determining the optimal LDO clustering within the SMPS clusters is presented below. Consider a system with S off-chip or in-package SMPS converters and L on-chip LDOs, delivering power to L power domains with N different supply voltages $\{(V_{DD}^{(i)}, I_{DD}^{(i)})\}_{i=1}^{N}$. Intuitively, the LDOs that regulate power domains with similar supply voltages should be assigned to the same voltage cluster. Thus, to explore the power efficiency of a heterogeneous power delivery system, $L = N \geq S$ is assumed. The ith SMPS supplies power to l_i LDOs ($\Sigma l_i = L = N$), forming the ith voltage cluster. The heterogeneous system is illustrated in Fig. 26.4 with off-chip converters. Note that the SMPS, LDO, and supply voltages are assumed to be ordered such that

$$V_{SMPS}^{(i)} < V_{SMPS}^{(j)} \quad \text{if } \{i < j\}, \tag{26.1}$$

$$V_{LDO}^{(i,m)} < V_{LDO}^{(j,n)} \quad \text{if } \{i < j\} \text{ or } \{i = j, m < n\}, \tag{26.2}$$

$$V_{DD}^{(i)} < V_{DD}^{(j)} \quad \text{if } \{i < j\}, \tag{26.3}$$

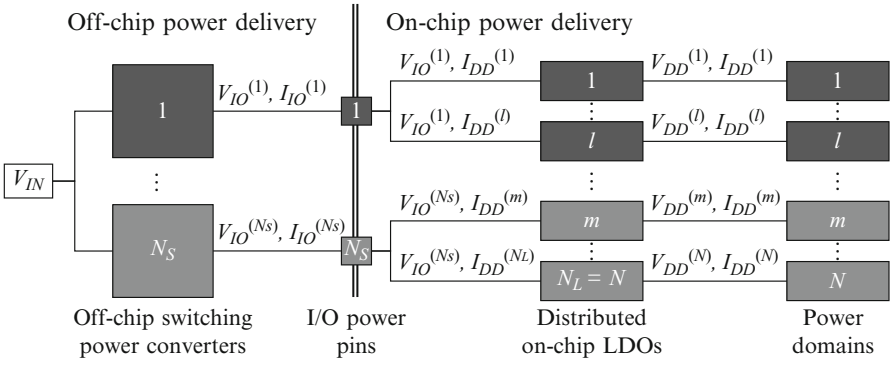


Fig. 26.4 Heterogeneous power delivery system with S off-chip switching converters, $L = N = \Sigma l_i$ on-chip linear regulators, and N on-chip power domains

where $V_{SMPS}^{(i)}$ is the output voltage of the SMPS in the ith cluster, $V_{LDO}^{(i,m)}$ is the output voltage of the mth LDO in the ith cluster, and $V_{DD}^{(j)}$ is the voltage supplied to the jth power domain.

To increase the power efficiency of a heterogeneous power delivery system, the voltage drops across the distributed on-chip LDOs should be reduced. The granularity of the converted voltage levels supplied on-chip increases with additional off-chip SMPS converters, reducing the power losses within the on-chip LDOs. At the limit, $S = L = N$ switching power converters are placed off-chip, providing voltages $\{V_{SMPS}^{(i)}\}_{i=1}^{N}$ at the I/O power pins. In the configuration with $S = L = N$, the on-chip LDOs operate with a minimum output voltage drop V_{Drop}, yielding

$$V_{SMPS}^{(i)} = V_{DD}^{(i)} + V_{Drop}, \quad i = 1, \ldots, N, \tag{26.4}$$

where V_{Drop} is the voltage dropout of the output transistor within the LDO. Assuming ideal power efficiency of the off-chip SMPS, the power efficiency of a system with the maximum number of SMPS converters ($S = L$) is

$$\varphi_{S=L=N} = \frac{P_{Load}}{P_{IN}} = \frac{\sum_{i=1}^{N} V_{DD}^{(i)} I_{DD}^{(i)}}{\sum_{i=1}^{S} V_{SMPS}^{(i)} I_{SMPS}^{(i)}} = \frac{\sum_{i=1}^{N} V_{DD}^{(i)} I_{DD}^{(i)}}{\sum_{i=1}^{N} \left(V_{DD}^{(i)} + V_{Drop} \right) I_{DD}^{(i)}}. \tag{26.5}$$

In this case, the power efficiency is only limited by the dropout voltage across the transistor, and exhibits a high power efficiency for low V_{Drop} devices.

Area and I/O power pin constraints exist, however, that limit the number of off-chip power supplies, degrading the overall power efficiency. Let S be the maximum number of off-chip switching power converters in a heterogeneous power delivery system. To minimize the voltage drop across the on-chip LDOs for the worst case power efficiency scenario where $S = 1$, the off-chip SMPS produces a voltage $V_{SMPS}^{(1)}$ that is higher than the maximum domain voltage by one dropout voltage V_{Drop},

$$V_{SMPS}^{(1)} = \max_{1 \le i \le N} \left\{ V_{DD}^{(i)} \right\} + V_{Drop}, \tag{26.6}$$

exhibiting a power efficiency,

$$\varphi_{S=1,L=N} = \frac{P_{Load}}{P_{IN}} = \frac{\sum_{i=1}^{N} V_{DD}^{(i)} I_{DD}^{(i)}}{V_{SMPS}^{(1)} I_{SMPS}^{(1)}} = \frac{\sum_{i=1}^{N} V_{DD}^{(i)} I_{DD}^{(i)}}{\max_{1 \le i \le N} \left\{ V_{DD}^{(i)} + V_{Drop} \right\} \sum_{i=1}^{N} I_{DD}^{(i)}}. \tag{26.7}$$

In a system with a single off-chip SMPS, the power loss within each domain, in addition to the output voltage drop V_{Drop}, is determined by the difference between the domain voltage and maximum voltage in the system. Those power domains

with lower voltages exhibit greater power losses, significantly degrading the power efficiency of a heterogeneous system. The upper and lower bounds of the power efficiency of a heterogeneous system for a specific number of switching power converters are given, respectively, by (26.7) and (26.5), yielding

$$\frac{\sum_{i=1}^{N} V_{DD}^{(i)} I_{DD}^{(i)}}{\max\limits_{1\le i\le N}\left\{V_{DD}^{(i)} + V_{Drop}\right\} \sum_{i=1}^{N} I_{DD}^{(i)}} \le \varphi \le \frac{\sum_{i=1}^{N} V_{DD}^{(i)} I_{DD}^{(i)}}{\sum_{i=1}^{N} \left(V_{DD}^{(i)} + V_{Drop}\right) I_{DD}^{(i)}}. \tag{26.8}$$

Thus, the power efficiency of a heterogeneous system is a strong function of the number of off-chip power converters.

26.1.3 Power Supply Clusters

In a practical heterogeneous power delivery system, the number of SMPS converters is smaller than the number of on-chip LDO regulators ($S < L$). Several options therefore exist to include the on-chip LDOs within SMPS clusters, affecting the power efficiency of the overall power delivery system. To illustrate the effects of the clustering topology on the power efficiency of a power delivery system, a heterogeneous system is considered with two switching converters and three linear regulators, supplying equal current I_{DD} to three power domains $\{V_{DD}^{(i)}\} = \{1.8\,\text{V}, 1.1\,\text{V}, 1.0\,\text{V}\}$. Assume $V_{Drop} = 0.1\,\text{V}$. The power supply clusterings $K_1 = \{1, 2\}$ and $K_2 = \{2, 1\}$ for the heterogeneous system with $S = 2$ and $L = N = 3$ are shown in Fig. 26.5. The voltage at the output of the switching converters is determined from (26.10), yielding a power efficiency, $\varphi(K_1) = P_{Load}/P_{IN} = 91\,\%$ and $\varphi(K_2) = P_{Load}/P_{IN} = 80\,\%$. Determining the optimal clustering of the on-chip power supplies is a primary challenge in heterogeneous power efficient systems.

The power efficiency of a general heterogeneous power delivery system, as illustrated in Fig. 26.4, is

$$\varphi = \frac{\sum_{i=1}^{N} V_{DD}^{(i)} I_{DD}^{(i)}}{\sum_{i=1}^{S} \left(V_{SMPS}^{(i)} \cdot \sum_{m=1}^{l_i} I_{LDO}^{(i,m)}\right)}. \tag{26.9}$$

The voltage supplied by an LDO to a power domain cannot be stepped up. The output voltage of each SMPS is therefore higher than the output voltage of the LDOs within an SMPS cluster, yielding $V_{SMPS}^{(i)} \ge V_{LDO}^{(i,m)},\ \forall 1 \le m \le l_i$. Alternatively, the power efficiency of a power delivery system increases with smaller LDO dropout voltages ($V_{SMPS}^{(i)} - V_{LDO}^{(i,m)},\ \forall 1 \le m \le l_i$). Thus, to minimize the power loss within the *i*th SMPS cluster, the output voltage of an SMPS is

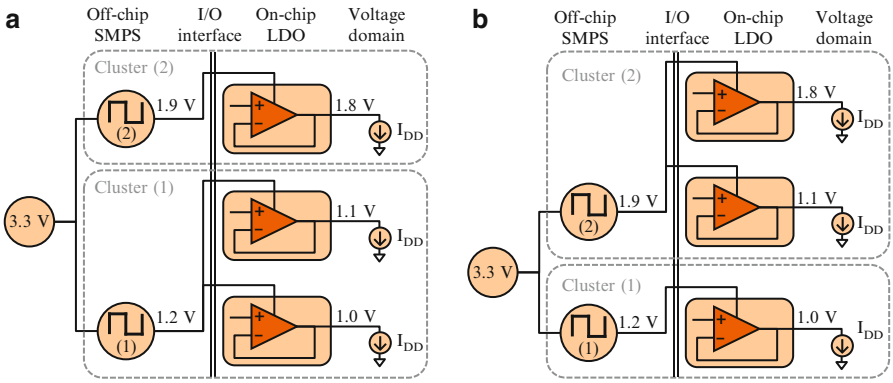


Fig. 26.5 Power supply clusterings for a heterogeneous power delivery system with $S = 2$ and $L = N = 3$, (**a**) $K_1 = \{1, 2\}$, and (**b**) $K_2 = \{2, 1\}$

$$V_{SMPS}^{(i)} = \max_{1 \le m \le l_i} V_{LDO}^{(i,m)} + V_{Drop}. \tag{26.10}$$

The preferred SMPS output voltage in (26.10) with power efficiency φ described by (26.9) yields the optimum power efficiency for a specific choice of clusters,

$$\varphi = \frac{\sum_{i=1}^{N} V_{DD}^{(i)} I_{DD}^{(i)}}{\sum_{i=1}^{S} \left(\max_{1 \le m \le l_i} V_{LDO}^{(i,m)} + V_{Drop} \right) \sum_{m=1}^{l_i} I_{LDO}^{(i,m)}}. \tag{26.11}$$

For each SMPS converter, the voltage drop across the driven LDOs increases with a wider range of voltages included within that SMPS cluster, increasing overall power dissipation. Intuitively, for any power supply cluster, adding a power domain with a specific voltage within a SMPS cluster that includes a similar voltage range results in a lower voltage drop and power loss than including the same power domain in a SMPS cluster with a significantly different range of voltages. Thus, the choice of power clustering directly affects the efficiency of the power delivery system. To minimize power losses in a heterogeneous power delivery system, a power distribution network with a higher φ is preferred. The optimal solution with minimum power losses can be obtained by comparing the power efficiency φ (see (26.14)) for all possible clusters $\{K_i\}$, and choosing the configuration with the maximum efficiency φ^{OPT},

$$\varphi^{OPT} = \max_{\{K_i\}} \{\varphi\} = \max_{\{K_i\}} \left\{ \frac{\sum_{i=1}^{N} V_{DD}^{(i)} I_{DD}^{(i)}}{\sum_{i=1}^{S} \left(\max_{1 \le m \le l_i} V_{LDO}^{(i,m)} + V_{Drop} \right) \sum_{m=1}^{l_i} I_{LDO}^{(i,m)}} \right\}. \tag{26.12}$$

The power efficiency of a heterogeneous system is also a strong function of the current distribution, which is not necessarily equally distributed to the individual power domains. Optimizing the power efficiency of a heterogeneous system based on the current distribution within the power domains requires additional assumptions or knowledge of the behavior and specifications of the currents. The purpose here is to provide a framework for an efficient system-wide power delivery methodology and specific rules for delivering power.

26.2 Dynamic Control in Heterogeneous Power Delivery Systems

DVS is a primary objective for efficiently managing the power budget in 100- and 1000-core ICs, further increasing the design complexity of the power delivery system. As the voltage supplied by an LDO to a power domain changes, the voltage dropout within the LDO varies. As a result, the power saved during low power operation is dissipated within the regulators. Varying the load currents affects the efficiency of the power supplies in a similar way. Thus, in a system with fixed power supply clusters, the energy efficiency of the power delivery system is not optimal. To avoid excessive dissipation of power, SMPS clusters should be dynamically reconfigured in every time slot Δt based on the temporarily required voltage and current levels within the individual power domains. A heterogeneous system for real-time power management in modern high performance integrated circuits is illustrated in Fig. 26.6.

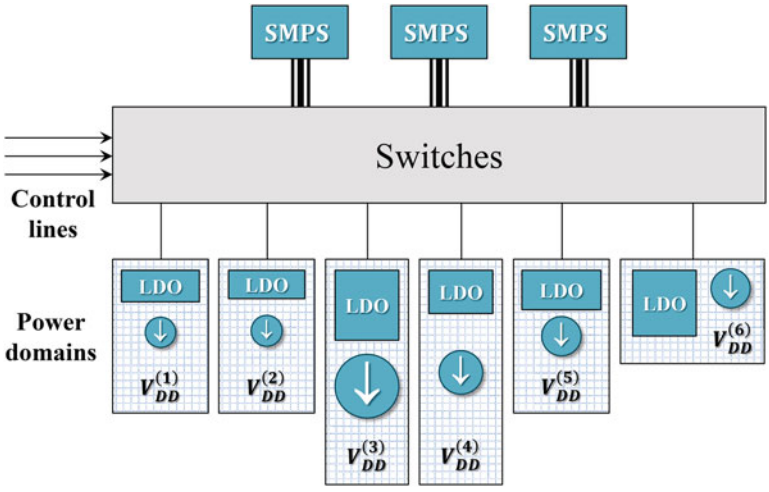


Fig. 26.6 Heterogeneous power delivery with multiple dynamically controllable power domains

The optimal power supply clusters need to be determined during each control time slot, decreasing the voltage dropout within the regulators and increasing the overall energy efficiency within shorter time slots. Alternatively, the duration of the control time slot Δt is inversely proportional to the power dissipated by an MOSFET switch in the ith power domain,

$$P_{SW}^{(i)} = f_{sw} \cdot t_{ave} \cdot V_{Off} \cdot I_{DD}^{(i)}, \tag{26.13}$$

where $f_{sw} = 1/\Delta t$ is the system switching frequency, t_{ave} is the MOSFET average on/off time, V_{Off} is the switch off voltage, and $I_{LDO}^{(i,m)} + I_{LDO,Q}^{(i,m)} \approx I_{LDO}^{(i,m)}$. Power losses of this system, therefore, increase with higher switching frequencies, and are considered here to determine the preferred duration of the control time slot $\Delta t = 1/f_{sw}$. In a power efficient system, power switching losses should not exceed the power savings due to dynamic control over the power delivery process. The optimal power efficiency in (26.9) with switching power losses is

$$\varphi = \frac{\sum_{i=1}^{N} V_{DD}^{(i)} I_{DD}^{(i)}}{\sum_{i=1}^{S} \left(V_{SMPS}^{(i)} + f_{sw} \cdot t_{ave} \cdot V_{Off} \right) \sum_{m=1}^{l_i} I_{LDO}^{(i,m)}}. \tag{26.14}$$

Both analog and digitally controlled LDO regulators with voltage drops as low as 0.15 V down to 0.05 V have recently been reported [358, 365, 386, 510], yielding $V_{Drop} \approx V_{Off}$. Thus, for a sufficiently long control time slot $\Delta t = 1/f_{sw} >> t_{ave}$, the power dissipated within the switches is significantly lower than the power dissipated within the LDO and can therefore be neglected. Modern MOSFET switches are capable of switching within tens of nanoseconds [511], exhibiting a practical target for the $\Delta t >> t_{ave}$ requirement in dynamically controlled heterogeneous systems. Alternatively, a real-time power delivery management system poses a significant computational challenge. Thus, a computationally efficient method to co-design the on-chip power supplies in modern high performance circuits is required.

26.3 Computationally Efficient Power Supply Clustering

The optimal clustering topology with minimum power losses can be obtained by exhaustively comparing the power efficiency φ (see (26.14)) for all possible clusterings, and choosing the configuration with the maximum efficiency. The number of possible clusterings, however, grows exponentially with S, producing a computationally infeasible solution. To efficiently determine the preferable power supply clusters, an alternative computationally efficient solution is required. Binary and linear near-optimal power supply clusterings with, respectively, $\mathcal{O}(S)$ and $\mathcal{O}(N)$ are described in Sect. 26.3.1. An optimal power supply clustering with dynamic programming with $\mathcal{O}(N^2 \cdot S)$ is described in Sect. 26.3.2.

26.3.1 Near-Optimal Power Supply Clustering

Intuitively, to reduce the voltage drop across the on-chip LDOs, LDOs that regulate the voltage domains with a small difference in voltage levels should be assembled into a voltage cluster driven by the same SMPS, minimizing the voltage range within each cluster. A binary power supply clustering, based on a greedy algorithm, identifies in each step the voltage cluster with the widest voltage range and distributes the LDOs into two separate clusters. Pseudo-code of the algorithm is provided in Algorithm 26.1. The algorithm produces a set of S SMPS voltage clusters *List_of_Clusters* with a binary clustering of power supplies. The third step is executed S times, yielding an algorithm that exhibits linear complexity $\mathcal{O}(S)$ with the number of switching converters. The primary weakness of the binary power supply clustering algorithm is the greedy nature of the algorithm. The number of voltage clusters S is only considered when the algorithm is terminated, reducing the power efficiency of the overall power delivery system. Consider a heterogeneous power delivery system with three switching converters and four LDO regulators that supply power to four voltage domains. The voltage and current levels within the voltage domains are (1 V, 1 A), (1.49 V, 1 A), (1.51 V, 1 A), and (2 V, 1 A). The optimal and binary power supply clusterings, SMPS output voltages, and power efficiency are listed in Table 26.1, exhibiting, respectively, 93 % and 87 % power efficiency for $V_{Drop} = 0.1$ V.

Algorithm 26.1 Algorithm for binary power supply clustering

1: **procedure** BINARY_CLUSTERING($\left\{V_{DD}^{(i)} \mid i = 1 \ldots N\right\}$)
2: $List\_of\_Clusters \leftarrow \{\}$;
3: $Next\_Cluster \leftarrow \left\{V_{DD}^{(i)} \mid i = 1 \ldots N\right\}$;
4: $(Low\_Cluster, High\_Cluster) \leftarrow$ DISTRIBUTE_A_CLUSTER($Next\_Cluster$);
5: $List\_of\_Clusters \leftarrow \{List\_of\_Clusters, Low\_Cluster, High\_Cluster\}$;
6: **if** (number of clusters in $List\_of\_Clusters < S$) **then**
7: **for all** ($Cluster$'s in $List\_of\_Clusters$) **do**
8: Find $Cluster$ such that (max $\{Cluster\}$ − min $\{Cluster\}$) is maximal;
9: **end for**
10: $Next\_Cluster \leftarrow Cluster$;
11: Return to line 4;
12: **end if**
13: **return** $List\_of\_Clusters$;
14: **end procedure**

15: **procedure** DISTRIBUTE_A_CLUSTER($Next\_Cluster$)
16: $V_{Mean} \leftarrow 1/2\,(\min\{Next\_Cluster\} + \max\{Next\_Cluster\})$;
17: $Low\_Cluster \leftarrow \left\{V_{DD}^{(i)} \in Next\_Cluster \mid V_{DD}^{(i)} \leq V_{Mean}\right\}$;
18: $High\_Cluster \leftarrow \left\{V_{DD}^{(i)} \in Next\_Cluster \mid V_{DD}^{(i)} > V_{Mean}\right\}$;
19: **return** $(Low\_Cluster, High\_Cluster)$;
20: **end procedure**

Table 26.1 Power supply clustering for a heterogeneous power delivery system with $S = 3$, $L = N = 4$, $V_{Drop} = 0.1$ V, and voltage domains (1 V, 1 A), (1.49 V, 1 A), (1.51 V, 1 A), and (2 V, 1 A)

Clustering type	Power supply clustering $K = \{l_1, l_2, l_3\}$	SMPS voltages (V) $\left\{V_{SMPS}^{(i)} \mid i = 1, 2, 3\right\}$	Efficiency (%)
Binary	$\{2, 1, 1\}$	$\{1.59, 1.61, 2.10\}$	87
Optimal	$\{1, 2, 1\}$	$\{1.10, 1.61, 2.10\}$	93

Algorithm 26.2 Algorithm for linear power supply clustering

1: **procedure** LINEAR_CLUSTERING(sorted supply voltages $\left\{V_{DD}^{(i)} \mid i = 1 \ldots N\right\}$)
2: *List_of_Clusters* $\leftarrow \{\}$;
3: *Cluster_Range* $\leftarrow \left(\max\{V_{DD}^{(i)} \mid i = 1 \ldots N\} - \min\{V_{DD}^{(i)} \mid i = 1 \ldots N\}\right)/S$;
4: **for each** $\left(V_{DD} \in \left\{V_{DD}^{(i)} \mid i = 1 \ldots N\right\}\right)$ **do**
5: $k \leftarrow \left\lfloor \left(V_{DD} - \min\left\{V_{DD}^{(i)} \mid i = 1 \ldots N\right\}\right)/Cluster\_Range \right\rfloor + 1$;
6: Add V_{DD} to the k^{th} cluster in *List_of_Clusters*;
7: **end for**
8: **if** (number of clusters in *List_of_Clusters* $< S$) **then**
9: **for all** (*Cluster*'s in *List_of_Clusters*) **do**
10: Find *Cluster* such that $(\max\{Cluster\} - \min\{Cluster\})$ is maximal;
11: **end for**
12: *Next_Cluster* $\leftarrow$ *Cluster*;
13: (*Low_Cluster*, *High_Cluster*) $\leftarrow$ DISTRIBUTE_A_CLUSTER(*Next_Cluster*);
14: *List_of_Clusters* $\leftarrow \{List\_of\_Clusters, Low\_Cluster, High\_Cluster\}$;
15: Return to line 8;
16: **end if**
17: **return** *List_of_Clusters*;
18: **end procedure**

Alternatively, a linear power supply clustering produces a topology by linearly distributing the LDOs within S voltage clusters, as described by the algorithm represented by the pseudo-code provided in Algorithm 26.2. If less than S SMPS voltage clusters are produced within steps 1 through 3 in the linear power supply clustering algorithm, the linearly generated clusters are distributed into additional clusters using a binary algorithm. This algorithm produces a set of S SMPS voltage clusters *List_of_Clusters* with a linear power supply clustering. In the worst case, the third and fourth steps are executed, respectively, N and S times, yielding an algorithmic complexity $\mathcal{O}(N)$ that is linear with the number of voltage domains.

To compare the power efficiency of the near-optimal and optimal power delivery networks, a heterogeneous power delivery system with a small number of voltage domains is initially considered due to the computational complexity of the exhaustive optimal algorithm. The exhaustive algorithm determines the most power efficient clustering by comparing the power efficiency of all possible clusterings. The efficiency of the optimal power network produced by the exhaustive algorithm is compared here to the power efficiency of the near-optimal clustering

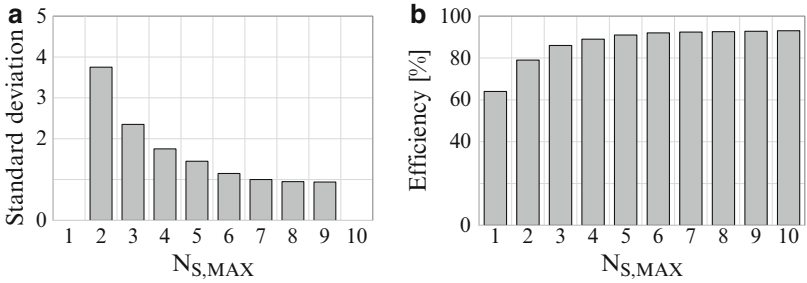


Fig. 26.7 Heterogeneous power delivery system (**a**) standard deviation, and (**b**) average efficiency using an exhaustive power supply clustering algorithm

algorithm. To estimate the power efficiency of the optimal power supply clusters, a heterogeneous power delivery system S_H with ten voltage domains ($N = 10$) and ten on-chip linear regulators ($L = 10$) is considered. The maximum number of off-chip switching converters is evaluated for one to ten converters ($1 \leq S \leq 10$). A voltage threshold of $V_{Drop} = 0.1\,V$, and domain voltages and currents of, respectively, 0.5–2 V and 0.5–3.5 A, are considered. Simulation results are sampled for 100 iterations. The power efficiency of a heterogeneous power delivery system with the power supply clusters, determined by an exhaustive analysis, is presented in Fig. 26.7a. A power efficiency above 80 % is demonstrated for $S \geq 2$, and a maximum 93 % power efficiency is achieved for $S = N$. Thus, the power efficiency of a heterogeneous power delivery system with an optimal power clustering exhibits a reasonable power efficiency of 80 %, using only two off-chip switching converters. The efficiency increases rapidly with additional off-chip converters. Based on the Monte Carlo integration technique [512], the average error in the efficiency is bounded by $\sigma_M/\sqrt{M}$, where σ_M is the standard deviation of a power efficiency sample and M is the number of samples. The standard deviation of the power efficiency is shown in Fig. 26.7b for $2 \leq S \leq 9$. Values of σ_M range from 3.7 for $S = 2$–0.9 for $S = 9$, bounding the power efficiency error for $M = 100$ by, respectively, 0.37–0.09 %. Power supply clustering for $S = 1$ and $S = N$ is explicit, yielding no error in the power efficiency. To evaluate the power efficiency of the near-optimal power supply clustering topologies described in Sect. 26.4, algorithms for binary and linear power supply clusterings have also been evaluated in Matlab. The same heterogeneous system S_H is considered for both linear and binary distributed power supplies. For a heterogeneous system with a single off-chip SMPS converter ($S = 1$) or maximum number of off-chip SMPS converters ($S = L$), the linear, binary, and optimal clustering of the on-chip LDO regulators is identical. For $S = 1$, all of the LDOs are driven by a single SMPS converter, while for $S = L$, each LDO is driven by a different SMPS converter. Thus, the power efficiency of a heterogeneous system with $S = 1$ or $S = L$ is optimal with either a linear or binary power supply clustering. Alternatively, for $S < L$, a linear and binary clustering of the power supplies may differ from the exhaustive optimal

solution, exhibiting a lower than optimal power efficiency. Due to the uniform nature of the linear approach, a linear clustering of the on-chip LDO regulators within the off-chip SMPS converters exhibits near optimal efficiency for power delivery systems with near uniformly distributed domain voltages. Alternatively, for a power delivery system with domain voltages that exhibit significant deviation from a uniform distribution, the power efficiency with the binary power supply clustering may be higher than with the linear clustering. This behavior is due to the greedy nature of the binary approach that iteratively identifies the on-chip power supply cluster with the lowest power efficiency and splits the cluster, increasing the overall efficiency of the system.

To demonstrate the power efficiency of the binary and linear clusterings, the reduction in efficiency in both the binary and linear power supply clusterings is simulated for two different power profiles, exhibiting a maximum 4 % drop in power efficiency. The optimal solution with zero reduction in power efficiency is demonstrated for both power profiles in Fig. 26.8 for $S = 1$ and $S = L$. In the first power profile, the voltage levels are assumed to be randomly distributed between 0.5 and 2 V, yielding an average power efficiency generated from over 100 iterations, as depicted in Fig. 26.8a. In this case, for $1 < S < L$, the exhaustive optimal solution produces a power supply that is uniformly distributed, while the linear power supply clustering yields a higher power efficiency. In the second power profile, the voltage levels are assumed to be normally distributed within each of the [0.5, 1.5), [1.5, 1.8), and [1.8, 2] ranges, prioritizing the mean value of the groups. Due to the non-uniform clustered nature of the voltage domain profile, for a heterogeneous system with three off-chip SMPS converters, intuitively, the on-chip LDO regulators should be non-uniformly distributed into three clusters covering the ranges [0.5, 1.5), [1.5, 1.8), and [1.8, 2]. In this case, a system with uniformly distributed clusters with voltage ranges [0.5, 1), [1, 1.5), [1.5, 2) is less power efficient. This heterogeneous system is therefore more suitable for a binary power supply clustering rather than a linear power supply clustering. The average power efficiency for the second power profile, generated from over 100 iterations, is depicted in Fig. 26.8b. In this case, specifically for $S = 3$, the optimal solution produces three non-uniform SMPS

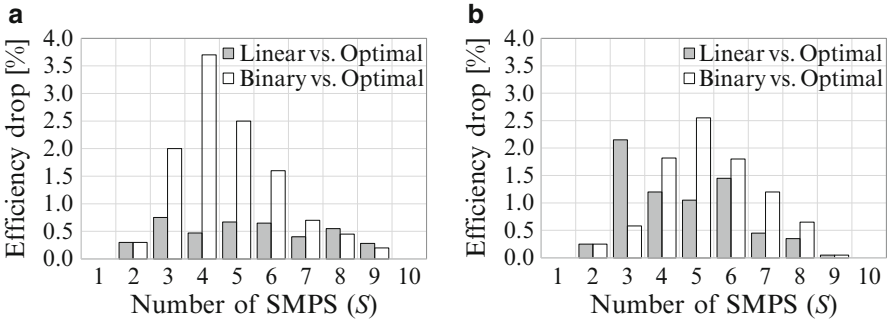


Fig. 26.8 Decrease in linear and binary power efficiency from the optimal power efficiency for (**a**) randomly distributed voltage levels, and (**b**) voltage levels grouped within three voltage ranges

clusters, covering the three ranges, [0.5, 1.5), [1.5, 1.8), and [1.8, 2]. The binary power supply clustering with $S = 3$ also produces three SMPS clusters with voltage ranges, [0.5, 1.25), [1.25, 1.625), and [1.625, 2], exhibiting a higher power efficiency than the efficiency produced by a linear power supply clustering. Based on a Monte Carlo integration technique, the error in estimating the drop in power efficiency, illustrated in Fig. 26.8, is smaller than 0.63 % for all values of S.

Due to the greedy nature of the binary power supply clustering algorithm, the binary clustering algorithm is better for those voltage domain levels grouped near specific voltage levels. Alternatively, the number of SMPS clusters S is only considered at the termination of the binary clustering algorithm, potentially reducing the effectiveness of the binary clustering algorithm in those systems with uniformly distributed voltage domains. As expected, for most values of S and power supply specifications, the drop in power efficiency for the linear power supply clustering algorithm is lower than with the binary clustering algorithm. However, the second power profile that forms three non-uniform voltage groups is better addressed by the binary power supply clustering algorithm, producing a more efficient heterogeneous power delivery system for $S = 3$. Thus, a heterogeneous power delivery system with a higher power efficiency is usually produced with a linear power supply clustering algorithm. However, for certain power profiles, a binary power supply clustering is preferable. To increase the power efficiency of a heterogeneous power delivery system, a combined hybrid approach should be employed. The power efficiency should be evaluated with both the binary and linear algorithms, and the configuration with the higher power efficiency should be employed. Considering the results depicted in Fig. 26.8 based on this combined hybrid approach, the drop in power efficiency from the optimal solution is reduced to 1.5 %, yielding a computationally efficient $\mathcal{O}(S + N)$ complexity, near-optimal, and high fidelity power supply clustering.

26.3.2 Power Supply Clustering with Dynamic Programming

Determining the optimal clusters of the on-chip power supplies is an important challenge in a heterogeneous power efficient system. An optimal power supply clustering algorithm with $\mathcal{O}(N^2 \cdot S)$ is described in this section. A recursive analytic expression is provided for power supply clustering with L LDO regulators and S SMPS in smaller power delivery systems ($l < L$ LDO regulators and $s = S - 1$ SMPS). Power supply clusters are determined recursively with dynamic programming. Recursive power supply clustering, similar to exhaustive power supply clustering, yields the optimal set of power supply clusters.

The key idea behind the algorithm is determining the number of voltage regulators within a high voltage SMPS cluster in $\mathcal{O}(N)$. Once the number of LDO regulators in a high voltage SMPS cluster is determined, the problem of power supply clustering is reformulated for the remaining LDO regulators and a smaller number of SMPS clusters. To exemplify this solution, consider a heterogeneous system with three switching converters ($S = 3$) and five linear regulators

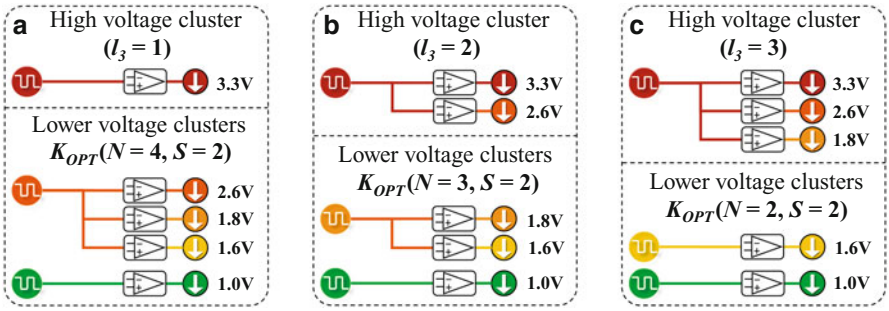


Fig. 26.9 A single step of the recursive power supply clustering algorithm for a heterogeneous power delivery system with $S = 3$ and $L = N = 5$, (**a**) a single LDO ($l_3 = 1$), (**b**) two LDO regulators ($l_3 = 2$), and (**c**) three LDO regulators ($l_3 = 3$) in a high voltage SMPS cluster

$(L = 5)$, supplying equal current I_{DD} to five power domains $(N = 5)$ $\{V_{DD}^{(i)}\} = \{3.3\text{ V}, 2.6\text{ V}, 1.8\text{ V}, 1.6\text{ V}, 1.0\text{ V}\}$. Dynamic programming is used to determine the optimum set of power supply clusters. The optimum clustering $K_{OPT}(5, 3) = \{l_1, l_2, l_3\}$, $\Sigma l_i = 5$ is recursively determined based on the number of LDO regulators in the high voltage cluster l_3, and lower order optimal supply clustering $K_{OPT}(4, 2)$, $K_{OPT}(3, 2)$, and $K_{OPT}(2, 2)$. A single recursive step is illustrated in Fig. 26.9, demonstrating three possible alternatives for clustering with one ($l_3 = 1$), two ($l_3 = 2$), and three ($l_3 = 3$) LDO regulators within the high voltage SMPS cluster. Given the lower order clustering $K_{OPT}(4, 2)$, $K_{OPT}(3, 2)$, and $K_{OPT}(2, 2)$, the optimum clustering $K_{OPT}(5, 3)$ is determined with linear computational complexity by comparing the power efficiencies $\varphi\{K_{OPT}(4, 2), 1\}$, $\varphi\{K_{OPT}(3, 2), 2\}$, and $\varphi\{K_{OPT}(2, 2), 3\}$, and choosing the clustering topology that minimizes the power losses.

For a general clustering algorithm, consider clustering L on-chip LDO regulators within S SMPS clusters to deliver power to L power domains with $N = L$ different supply voltages. The optimal clustering topology of a system with N different supply voltages and S SMPS $K_{OPT}(N, S) = \{l_i\}_{i=1}^{S}$, $\sum l_i = N$ is determined recursively by

$$K_{OPT}(N, S) = \{K_{OPT}(N - n_0, S - 1), l_S\}, \tag{26.15}$$

with the initial conditions,

$$K_{OPT}(N, 2) = \{N - l_S, l_S\}, \tag{26.16}$$

$$K_{OPT}(N, S = N) = \{1, 2, \ldots, N\}, \tag{26.17}$$

where $1 \leq l_S \leq (N - S)$ is the number of LDO regulators in the high voltage SMPS cluster. To maximize the overall power efficiency of the system, the number of LDO regulators in the last SMPS cluster is

$$\varphi(K_{OPT}(N, S)) = \max_{l_S} \varphi\left(\{K_{OPT}(N - l_S, S - 1), l_S\}\right). \tag{26.18}$$

Once the power supply clusters are recursively determined with dynamic programming, the maximum voltage level within each SMPS cluster determines the SMPS output and LDO input voltage based on (26.10). Pseudo-code of the algorithm for determining the LDO input voltages based on the clustering algorithm is shown in Algorithm 26.3.

Algorithm 26.3 Dynamic programming algorithm to determine LDO input voltages for power efficient clustering

procedure DP_CLUSTERING($\left\{V_{DD}^{(i)}, I_{DD}^{(i)} \mid i = 1 \ldots N\right\}$,$S$,$V_{Drop}$)
  if $S == 1$ **then**
    % There is only one cluster
    $V_{LDO}(:) \leftarrow \max V_{DD}(:) + V_{Drop}$;
  else if $S == N$ **then**
    % The number of clusters equals the number of voltage levels
    $V_{LDO}(:) \leftarrow V_{DD}(:) + V_{Drop}$;
  else
    % Initiate a matrix to store all the clusters for lower order systems
    $all\_V_{LDO}(:) \leftarrow zeros(S, L = N, N)$;
    % Update initial conditions based on (26.16) and (26.17)
    for $l = 1$ **to** N **do**
      $all\_V_{LDO}(1, l, 1 : l) \leftarrow ones(1, l) \cdot \left(V_{DD}(l) + V_{Drop}\right)$;
    end for
    for $s = 1$ **to** S **do**
      $all\_V_{LDO}(s, s, 1 : s) \leftarrow V_{DD}(1 : s) + V_{Drop}$;
    end for
    % Find all the lower order ($s \leq S$ and $l \leq L$) clusters
    for $(s = 2$ **to** $S)$ **do**
      for $(l = s + 1$ **to** $N)$ **do**
        $all\_V_{LDO,OPT}(:) \leftarrow zeros(1, N)$;
        $\varphi_{OPT} \leftarrow 0$;
        for $(l_s = 1$ **to** $l - s + 1)$ **do**
          % l_s and $(l - l_s)$ are the number of LDO regulators in, respectively,
          % the highest voltage cluster and the rest $(s - 1)$ clusters
          $V_{LDO,TMP}(1 : l - l_s) \leftarrow all\_V_{LDO}(s - 1, l - l_s, :)$;
          $V_{LDO,TMP}(1 - l_s + 1 : l) \leftarrow V_{DD}(l) + V_{Drop}$;
          $\varphi_{TMP} \leftarrow 100 \cdot \frac{\sum V_{DD}(1:l) \cdot I_{DD}(1:l)}{\sum V_{LDO,TMP}(1:l) \cdot I_{DD}(1:l)}$;
          if $(\varphi_{TMP} > \varphi_{OPT})$ **then**
            % Store the clustering with the highest efficiency
            $V_{LDO,OPT} \leftarrow V_{LDO,TMP}$;
            $\varphi_{OPT} \leftarrow \varphi_{TMP}$;
          end if
        end for
        % Update all_V_{LDO} with the optimal clusters
        $all\_V_{LDO}(s, l, :) \leftarrow V_{LDO,OPT}$;
      end for
    end for
    $V_{LDO}(:) \leftarrow all\_V_{LDO}(S, L = N, :)$;
  end if
  return V_{LDO};
end procedure

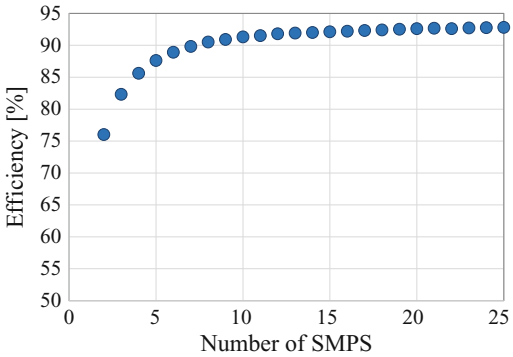


Fig. 26.10 Power efficiency of heterogeneous power delivery system based on a recursive power supply clustering algorithm

The LDO input voltages in a system with a single ($S = 1$) switching converter and the maximum number of SMPS ($S = N$) are determined, respectively, at lines 2–4 and 5–7. To determine the optimal clustering of a general system with ($1 < S < N$) switching converters, lines 8 through 42 are executed. The LDO input voltages for all of the systems with $s \leq S$ SMPS and $l \leq L$ LDO regulators are determined progressively and stored in matrix all_V_{LDO}. The matrix is allocated and initiated based on (26.16) and (26.17) at lines 9–17. The voltage levels at the LDO input voltages are determined in a loop (see lines 19–20) for systems with a progressively increasing number of power supplies. All of the high voltage cluster configurations with a different number of LDO regulators are determined at lines 26–28. The power efficiency of different configurations is compared at lines 29–33, determining the most power efficient system. The number of efficiency comparisons required to determine the optimal set of clusters $K_{OPT}(N, S)$ given all of the optimal clusterings of lower order $K_{OPT}(n < N, s < S)$ is $N - S$. The computational complexity to determine the most power efficient clusters with $N = L$ LDO regulators and S SMPS converters is therefore

$$\sum_{s=1}^{S}\left(\sum_{n=s}^{N}\mathcal{O}(n-s)\right) = \mathcal{O}(N^2 \cdot S), \quad N \geq S. \tag{26.19}$$

To estimate the power efficiency of the recursive power supply clustering algorithm, a heterogeneous power delivery system with 25 supply voltage levels ($N = 25$) and 25 on-chip linear regulators ($L = 25$) is considered. The number of off-chip switching converters is evaluated for 1–25 converters ($1 \leq S \leq 25$). A voltage drop of $V_{Drop} = 0.1$ V and 100 random profiles of domain voltages and currents of, respectively, 0.5–2 V and 0.5–3.5 A are considered. The maximum power efficiency with an average domain voltage and current of, respectively, 1.25 V and 2 A is evaluated based on (26.5), yielding 93 % power efficiency for $S = 25$. The power efficiency of a heterogeneous power delivery system with the power supply clusters determined by a recursive analysis is illustrated in Fig. 26.10. As expected, a maximum 93 % power efficiency is achieved for $S = N$. An average power

efficiency above 82 % is demonstrated for $S \geq 3$. Thus, the power efficiency of a heterogeneous power delivery system with an optimal voltage clustering exhibits a reasonable power efficiency, despite only three switching converters. The efficiency increases rapidly and saturates with additional SMPS converters. An excessive number of off-chip or in-package converters is therefore avoided with the power supply clustering algorithm, producing a power and area efficient system of power converters and regulators.

26.4 Demonstration of Co-design of Power Delivery System

The optimal exhaustive power delivery network, the hybrid near-optimal, and the recursive (based on dynamic programming) optimal solutions described in Sect. 26.3 are evaluated in Matlab. To determine the power efficiency with and without power separation, several test cases are evaluated based on IBM power grid benchmark circuits [192]. The test cases and simulation results with the linear, binary, and recursive clustering algorithms are described in Sect. 26.4.1. Power delivery with on-chip regulation in circuits with multiple power domains [476] is considered in Sect. 26.4.2.

26.4.1 Power Supply Clustering of IBM Power Grid Benchmark Circuits

Five test cases have been considered based on IBM power grid benchmark circuits [192] to evaluate the efficiency of the power separation principle in circuits with hundreds of power domains and tens of different supply voltages. The voltage map of a V_{DD} and V_{GND} $M6$ metal layer at normalized (x, y) locations is presented in Fig. 26.11a. The actual voltage drop in the metal layer $M6$ is $V_{DD} - V_{GND}$, as shown in Fig. 26.11b for the ibmpg1 benchmark circuit. Different circuits in ibmpg1 operate with different supply voltages, varying from 0.5 to 1.5 V. A total number of 66 voltage domains with voltage levels ranging from 0.5 to 1.8 V with a 0.02 voltage shift $\{V_{DD}^{(i)} = 0.5V + i \cdot 0.02V\}_{i=0}^{i=65}$ is considered in this analysis. Each of the benchmark circuits is partitioned into voltage domains based on the voltage maps, and the area of each domain is determined. The current within a benchmark circuit is assumed to be uniformly distributed. The current load of a domain is therefore assumed to be proportional to the area of the domain. The voltage within each domain is regulated by an LDO, ensuring that the total number of on-chip LDO regulators is the same as the number of voltage domains. To illustrate the process in which a test case is generated, specifications for several voltage domains in ibmpg1 are listed in Table 26.2. The total number of nodes processed in ibmpg1 is 30,027.

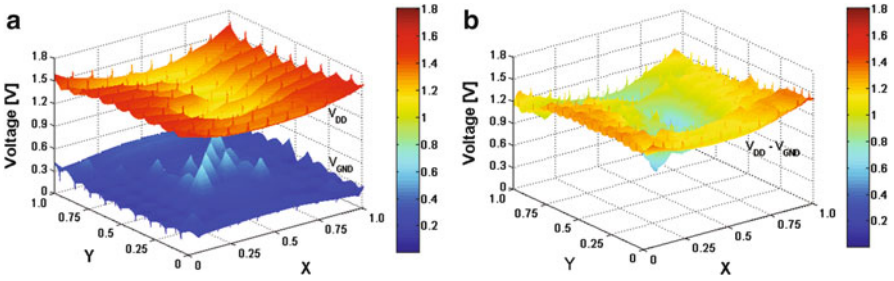


Fig. 26.11 Test case ibmpg1 with (**a**) V_{DD} and V_{GND} voltage map, and (**b**) $V_{DD} - V_{GND}$ voltage drop across the circuits in metal layer $M6$

Table 26.2 Power grid specifications for 0.8, 1.0, 1.2, 1.4, and 1.6 V domains based on ibmpg1 test case

Voltage domain id	15	25	35	45	55
Supply voltage (V)	0.8	1.0	1.2	1.4	1.6
Number of nodes	636	820	1297	19	0
Area (%)	2.12	2.73	4.320	0.062	0
Supply current (mA)	212	273	432	6.2	0

A total current of 10 A is assumed to be consumed by the circuits represented by ibmpg1.

Test cases for the IBM benchmark circuits, ibmpg2, ibmpg3, ibmpgnew1, and ibmpgnew2, are generated in a similar way. The power supply clustering algorithm is demonstrated in Matlab and applied to the test cases on a multi-core system with four Intel(R) Core(TM) i3-2120 CPU @ 3.30 GHz processors and 2.498 MB memory. A voltage drop of 0.1 V within an LDO is assumed. The power grid specifications and simulation results with and without power supply clustering are listed in Table 26.3.

Power clustering with the algorithms exhibits orders of magnitude shorter CPU time than the exhaustive approach. Those power grids with a range of LDO output voltages up to 0.2 V (ibmpgnew1 and ibmpgnew2) exhibit a high power efficiency of 93 % despite only two SMPS clusters. The power efficiency of these grids increases by 2.5 % as compared to a power delivery system without power separation. Increasing the power efficiency in those power grids with a large number of SMPS clusters (94 % with ten switching converters) requires excessive area. Alternatively, the ibmpg1 benchmark circuit exhibits a wider range of LDO output voltages, 0.5–1.5 V, and therefore, a low power efficiency of 68 % without power supply clustering. The effectiveness of power separation is shown to be significant in ibmpg1 with a 10.2 % and 21.1 % increase in power efficiency with, respectively, $S = 2$ and $S = 10$ SMPS clusters as compared to $S = 1$. Separation of power conversion and regulation is therefore particularly important in those systems with a wide range of on-chip supply voltages and voltage drops. To provide high

Table 26.3 Power efficiency in circuits with and without separation of power conversion and regulation. CPU time for hybrid (with near-optimal efficiency), dynamic programming (DP) based, and exhaustive clustering is provided

Benchmark		ibmpg1	ibmpg2	ibmpg3	ibmpgnew1	ibmpgnew2
Voltage domains/ LDOs	Number	49	21	15	11	11
	Voltage range (V)	(0.50,1.46)	(1.12,1.52)	(1.44,1.70)	(1.52,1.72)	(1.52,1.72)
Power efficiency with S voltage clusters (%)	Without power separation	68.1	77.2	88.6	90.3	90.4
	$S = 2$	78.3	87.0	91.9	92.9	92.9
	$S = 3$	82.4	89.2	92.8	93.5	93.5
	$S = 5$	86.5	91.0	93.6	94.1	94.1
	$S = 10$	89.2	92.1	94.1	94.3	94.3
CPU time (s) ($S = {}^1/_2N$)	Hybrid	0.023	0.018	0.022	0.022	0.02
	DP	0.299	0.038	0.012	0.011	0.011
	Exhaustive	>2000	>2000	>2000	5.537	5.477

quality power in dynamically scaled multi-voltage circuits, the efficiency of the power supply clustering algorithm is dynamically evaluated. The power supply clustering DP algorithm exhibits an order of magnitude smaller CPU run time as compared with the exhaustive method, while providing identical clusters. Additional clustering speedup is achieved with the hybrid approach with near-optimal power efficiency. With this approach, the switching converters and linear regulators can be co-designed in run time for power and area efficient management of the energy budget.

26.4.2 Power Supply Clustering and Existing Power Delivery Solutions

The separation principle is illustrated in circuits C_1 and C_2 with multiple power domains $V_{DD}^{(A)} = 1.4\,\text{V}$, $V_{DD}^{(B)} = 1.2\,\text{V}$, and $V_{DD}^{(C)} = 1.0\,\text{V}$, and the on-chip voltage conversion and regulation scheme used in [476]. A maximum voltage drop (V_{Drop}) of 5 % of the input voltage at the I/O interface (V_{SMPS}) is allowed at POLs within all of the power domains ($V_{Drop} \leq 0.05 V_{SMPS}$). The total area of the circuits, C_1 and C_2, is similar. The area of C_1 is dominated by the low power domain C, while the area of C_2 is dominated by the high performance power domain A. The current density J and normalized area per power domain are listed in Table 26.4 for both circuits. To support on-chip voltage conversion and regulation within the power domains A, B, and C in C_1 (C_2), a single input voltage $V_{SMPS} = 1.45\,\text{V}$ ($V_{SMPS} = 1.50\text{V}$) is used in the original configuration without power separation

Table 26.4 Parameter setup for circuits C_1 and C_2

Power domain	J (A/unit area)	Normalized area	
		Circuit C_1	Circuit C_2
A	2/3 V	1	6
B	1/2 V	2	2
C	1/3 V	6	1

Table 26.5 On-chip voltage regulation with and without separation of power conversion and regulation

Circuit	Without power separation [476]		With power separation (current work)	
	V_{SMPS}	φ (%)	$V_{SMPS}^{(A)}$, $V_{SMPS}^{(B)}$, $V_{SMPS}^{(C)}$	φ (%)
C_1	1.45 V	77.3	1.47 V, 1.26 V, 1.05 V	95.2
C_2	1.50 V	90.7	1.47 V, 1.26 V, 1.05 V	95.2

[476]. Note that the input voltage $V_{SMPS} = 1.45\,\text{V}$ ($V_{SMPS} = 1.50\,\text{V}$) is higher than the highest supply voltage $V_{DD}^{(A)} = 1.4\,\text{V}$ to maintain the required margin for on-chip voltage regulation with a reasonable number of 50 (60) on-chip LDO regulators in C_1 (C_2). Alternatively, the power can be converted separately off-chip or in-package for each power domain and regulated on-chip within each power domain. The input voltages and power efficiency for on-chip voltage regulation with and without separation of power conversion and regulation is listed in Table 26.5 for both circuits. In this configuration, three off-chip or in-package SMPS supply three different voltages, $V_{SMPS}^{(A)} = 1.47\,\text{V}$, $V_{SMPS}^{(B)} = 1.26\,\text{V}$, and $V_{SMPS}^{(C)} = 1.05\,\text{V}$ to, respectively, power domains A, B, and C, yielding a power efficiency as high as 95.2 %. To regulate the on-chip voltage with the required number of LDO regulators, the voltage at the output of each SMPS is designed with a margin of 5 % of the required on-chip supply voltage ($V_{DD}^{(i)} = 1.05 V_{DD}^{(i)}$). Without separating power conversion and regulation, the choice of off-chip supply voltage is determined by the supply voltage in the high performance power domain, exhibiting a higher voltage drop ($V_{SMPS} - V_{DD}^{(C)} = 0.45\,\text{V}$ in C_1) and a lower power efficiency (77.3 % in C_1) in low power circuits. Alternatively, an enhanced choice of $V_{SMPS}^{(i)}$ voltages is possible by separating power conversion and regulation, exhibiting a lower voltage drop ($V_{SMPS}^{(i)} - V_{DD}^{(i)} \leq 0.07\,\text{V}$ in C_1) and a higher power efficiency (95.2 % in C_1).

26.5 Summary

A heterogeneous power delivery system is described in this chapter. A computationally efficient method to co-design the on-chip power supplies in modern high performance circuits is described. The primary conclusions can be summarized as follows.

- To achieve a power efficient system, power should be primarily converted off-chip, in-package, and/or on-chip with power efficient switching supplies, and

regulated with ultra-small linear low dropout regulators at the point-of-load, exhibiting a heterogeneous power delivery system
- To avoid excessive usage of switching converters, the preferred number of power converters should be determined based on application-specific minimum power efficiency requirements
- Dynamically co-designing (clustering) tens of power converters with hundreds to thousands of on-chip regulators is a primary concern in an energy efficient power delivery system
- Algorithms to cluster a heterogeneous power supply system with polynomial computational complexity are described. An order of magnitude speedup is exhibited with the clustering algorithms as compared with an exhaustive clustering algorithm
- With clustering algorithms, switching converters and linear regulators can be co-designed in run time for power and area efficient management of the energy budget
- A clustering approach is evaluated based on partitioned IBM benchmark circuits
- Power grids with a narrow range of supply voltages (within a range of up to 0.2 V) exhibit a high power efficiency despite a small number of clusters (90.5 % without power supply clustering and 93 % with two clusters). Additional clusters do not significantly increase power efficiency in these power grids, while requiring excessive physical and design resources
- Power grids with a wider range of supply voltages, (within a range of at least 1 V) exhibit a low power efficiency of 68 % without power supply clustering. Power supply clusters are particularly effective in these power grids, increasing system power efficiency by 10.2 % and 21.1 % with, respectively, two and ten power supply clusters

Chapter 27
Conclusions

As the power delivery system transforms from a lumped network with a few off-chip converters into a heterogeneous, distributed, dynamically controlled system with many thousands of on-chip power components, the design, synthesis, and control objectives of the power delivery process need to be rethought at the system level. To cope with the complexity of system-wide power optimization, efficient techniques to model, analyze, and optimize a power delivery system are required.

The generation and distribution of high quality power to the load circuitry are two primary issues in the power delivery process. The interconnect network distributing power within a modern microprocessor typically contains many millions (to several billions) of nodes. The design of power distribution networks is typically performed at several design stages, iteratively increasing the accuracy of the current flow estimate while reducing the granularity of the power distribution network. Furthermore, the power distribution network is allocated over a large area with the entire network interacting, requiring near full scale simulation. Efficient and accurate analysis is therefore a key factor in the design of high performance power delivery systems. A number of synthesis and analysis algorithms are available to enhance the power delivery design process. Accuracy and computational efficiency is the primary tradeoff within these design algorithms and tools. A methodology is described based on a model of a two-port infinite mesh, providing both high accuracy and computational efficiency when analyzing large scale power networks.

While the quality of the power supply can be efficiently addressed with distributed on-chip power supplies, the stability of these parallel connected voltage regulators is also a primary concern. To maintain a stable power delivery system composed of multiple parallel connected regulators, a passivity-based stability criterion can be used. A distributed power delivery system is stable if the total output impedance of the parallel connected LDOs exhibits no right half plane poles and a phase between $-90°$ and $+90°$. This PBSC-based approach is evaluated on a fully integrated power delivery system with distributed on-chip low dropout regulators,

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0_27

fabricated in a 28 nm CMOS process. The experimental results of a distributed power delivery system satisfy this passivity-based criterion, yielding a stable system response.

An additional objective is the development of design methodologies to efficiently utilize the available resources, such as area, metal, and power. A novel link breaking methodology is discussed to optimize a mesh structured power distribution network. The resulting power distribution network is a combination of a single power distribution network to lower the network impedance, and multiple networks to reduce noise coupling among the circuits. Since the sensitivity to supply voltage variations within a power distribution network can vary among different circuits, this methodology reduces the voltage drop at the more sensitive circuits while penalizing the less sensitive circuits. This methodology is evaluated for multiple case studies, reducing voltage drops in the sensitive circuits (the critical paths).

The parasitic impedance of the interconnects, decoupling capacitances, load circuits, and on-chip power regulators are computationally expensive to simultaneously analyze. The distinctive properties of a power network have been exploited to develop closed-form expressions for the effective resistance between circuit components. This effective resistance model is based on the physical distance between circuit components within a two layer mesh where the horizontal and vertical unit resistances may be different. This effective resistance model is utilized in the development of a power grid analysis algorithm to compute the node voltage without requiring any iterations. This algorithm drastically improves computational efficiency since the iterative procedures commonly used to determine *IR* drop and *L di/dt* noise are no longer needed. The symmetric nature of the power and ground distribution networks and the principle of spatial locality are also exploited to further enhance the computational efficiency and accuracy of the analysis process.

Exhaustive ad hoc approaches for co-designing power supplies at different levels of hierarchy within a power delivery system exhibit significant design complexity and are computationally impractical in DVS/DVFS systems with hundreds to thousands of power domains. A computationally efficient methodology to co-design switching converters and on-chip linear regulators within a heterogeneous system is described, achieving high quality power and efficiency within limited on-chip area. Dynamically clustering a heterogeneous power delivery system is demonstrated on a suite of IBM benchmark circuits, exhibiting a computationally and power efficient alternative to existing ad hoc methodologies that employ either switching or linear on-chip power supplies.

Part VI
Noise in Power Distribution Networks

Several aspects of noise within power distribution networks are described in Part VI. The noise is primarily affected by the impedance of the power distribution network; focus is therefore placed on the impedance characteristics of the network. With increasing frequency, the inductive portion of the network impedance plays a significant role, requiring enhanced understanding of the inductive properties of power distribution networks. A variety of design tradeoffs is presented in Part VI. Noise reduction techniques are also described in this part. A summary of the chapters within this part is presented below.

An analysis of the inductive properties of power distribution grids is presented in Chap. 28. Three types of power grids are considered. The dependence of the grid inductance on grid type, grid line width, and grid dimensions is discussed. The concept of a sheet inductance to characterize the inductive properties of a power grid is described.

The variation of the inductance of power distribution grids with frequency is discussed in Chap. 29. The physical mechanisms underlying the dependence of inductance on frequency are discussed. The variation of the inductance of the power grids is interpreted in terms of these mechanisms. The variation of inductance with frequency in paired, interdigitated, and non-interdigitated grid types is compared. The dominant mechanisms for the variation of inductance with frequency are identified for each grid type. The dependence of the frequency variation characteristics on grid type and line width is reviewed.

Inductance/area/resistance tradeoffs in high performance power distribution grids are analyzed in Chap. 30. Two tradeoff scenarios are considered. In the first scenario, the area overhead of a power grid is maintained constant and the resistance versus inductance tradeoff is explored as the width of the grid lines is varied. In the second scenario, the total metal area occupied by the grid lines (and, consequently, the grid resistance) is maintained constant and the area versus inductance tradeoff is explored as the width of the grid lines is varied.

The effect of the on-chip interconnect inductance on the high frequency impedance characteristics of a power distribution system is discussed in Chap. 31. Scaling trends of the chip-package resonance is described. The propagation of the power

noise through an on-chip power distribution network is discussed. The significance of the inductance of on-chip power lines in nanometer circuits is demonstrated.

On-chip power noise reduction techniques in high performance ICs are the primary subject of Chap. 32. A design technique to lower ground bounce in noise sensitive circuits is described. An on-chip noise-free ground is added to divert ground noise from the sensitive nodes. An on-chip decoupling capacitor tuned in resonance with the parasitic inductance of the interconnects is shown to provide an additional low impedance ground path, reducing the power noise. The dependence of ground noise reduction mechanisms on various system parameters is also discussed.

A shielding technique to reduce crosstalk noise is presented in Chap. 33. The deleterious effects of power noise on the power lines utilized as shield lines are analytically evaluated. Design guidelines for inserting shield lines for different technology nodes and noise levels are provided based on practical power/ground noise models.

Chapter 28
Inductive Properties of On-Chip Power Distribution Grids

The inductive properties of power distribution grids are evaluated in this chapter. As discussed in Sect. 1.3, the inductance of power grids is an important factor in determining the impedance characteristics of a power distribution network operating at high frequencies.

This chapter is organized as follows. In Sect. 28.1, the inductive characteristics of a current loop formed by a power transmission circuit are established. The analysis approach is outlined in Sect. 28.2. The three types of grid structures considered in this study and analysis setup are described in Sect. 28.3. The dependence of the grid inductance characteristics on the line width is discussed in Sect. 28.4. The differences in the inductive properties among the three types of grids are reviewed in Sect. 28.5. The dependence of the grid inductance on the grid dimensions is described in Sect. 28.6. The chapter concludes with a summary.

28.1 Power Transmission Circuit

Consider a power transmission circuit as shown in Fig. 28.1a. The circuit consists of the forward current (power) path and the return current (ground) path forming a transmission current loop between the power supply at one end of the loop and a power consuming circuit at the other end. In a simple case, the forward and return paths each consist of a single conductor. In general, a "path" refers to a multi-conductor structure carrying current in a specific direction (to or from the load), which is the case in power distribution grids. The circuit dimensions are assumed to be sufficiently small for a lumped circuit approximation to be valid. The inductance of both terminating devices is assumed negligible as compared to the inductance of the transmission line. The inductive characteristics of the current loop are, therefore, determined by the inductive properties of the forward and return current paths.

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0_28

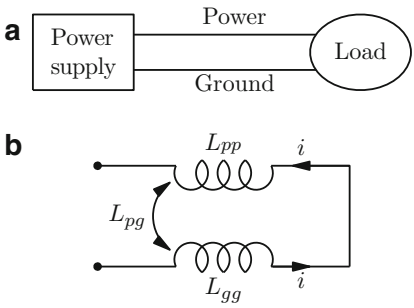


Fig. 28.1 A simple power transmission circuit; (**a**) block diagram, (**b**) equivalent inductive circuit

The power transmission loop consists of forward and return current paths. The equivalent inductive circuit is depicted in Fig. 28.1b. The partial inductance matrix for this circuit is

$$L_{ij} = \begin{bmatrix} L_{pp} & -L_{pg} \\ -L_{gp} & L_{gg} \end{bmatrix}, \tag{28.1}$$

where L_{pp} and L_{gg} are the partial self-inductance of the forward and return current paths, respectively, and L_{pg} is the absolute value of the partial mutual inductance between the paths. The loop inductance of the power transmission loop is

$$L_{\text{loop}} = L_{pp} + L_{gg} - 2L_{pg}. \tag{28.2}$$

The mutual coupling L_{pg} between the power and ground paths reduces the loop inductance. This behavior can be formulated more generally: *the greater the mutual coupling between antiparallel (flowing in opposite directions) currents, the smaller the loop inductance of a circuit*. The effect is particularly significant when the mutual inductance is comparable to the self-inductance of the current paths. This is the case when the line separation is comparable to the dimensions of the line cross section.

The effect of coupling on the net inductance is reversed when currents of the two inductive coupled paths flow in the same direction. For example, in order to reduce the partial inductance L_{11} of line segment 1, another line segment 2 with partial self-inductance L_{22} and coupling L_{12} is placed in parallel with segment 1. A schematic of the equivalent inductance is shown in Fig. 28.2. The resulting partial inductance of the current path is

$$L_{1\|2} = \frac{L_{11}L_{22} - L_{12}^2}{L_{11} + L_{22} - 2L_{12}}. \tag{28.3}$$

For the limiting case of no coupling, this expression simplifies to $\frac{L_{11}L_{22}}{L_{11}+L_{22}}$. In the opposite case of full coupling of segment 1, $L_{11} = L_{12}$ ($L_{11} \leq L_{22}$) and the total inductance becomes L_{11}. For two identical parallel elements, (28.3) simplifies to

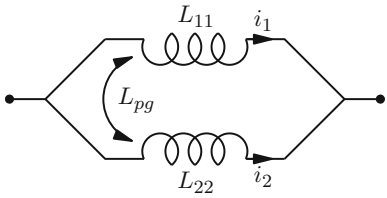


Fig. 28.2 Two parallel coupled inductors

$L_{\parallel} = (L_{\text{self}} + L_{\text{mutual}})/2$. In general, *the greater the mutual coupling between parallel (flowing in the same directions) currents, the greater the loop inductance of a circuit*. To present this concept in an on-chip perspective, consider a 1000 μm long line with a 1 × 3 μm cross section, and a partial self-inductance of 1.342 nH (at 1 GHz). Adding another identical line in parallel with the first line with a 17 μm separation (20 μm line pitch) results in a mutual line coupling of 0.725 nH and a net inductance of 1.033 nH (∼ 55 % higher than $L_{\text{self}}/2 = 0.671$ nH).

Inductive coupling among the conductors of the same circuit can, therefore, either increase or decrease the total inductance of the circuit. To minimize the circuit inductance, coupling of conductors carrying current in the same direction can be reduced by increasing the distance between the conductors. Coupling of conductors carrying current in opposite directions should be increased by physically placing the conductors closer to each other.

This optimization naturally occurs in grid structured power distribution networks with alternating power and ground lines. Three types of power distribution grid are analyzed to demonstrate this effect, as described in Sect. 28.3.

28.2 Simulation Setup

The inductance extraction program FastHenry [70] is used to explore the inductive properties of grid structures. FastHenry efficiently calculates the frequency dependent self- and mutual impedances, $R(\omega) + \omega L(\omega)$, in complex three-dimensional interconnect structures. A quasi-magnetostatic approximation is assumed, meaning the distributed capacitance of the line and any related displacement currents associated with the capacitance are ignored. The accelerated solution algorithm employed in the program provides approximately a 1 % worst case accuracy as compared with the direct solution of the system of linear equations characterizing the system [70].

A conductivity of 58 S/μm ≃ $(1.72\,\mu\Omega \cdot \text{cm})^{-1}$ is assumed for the interconnect material. The inductive portion of the impedance is relatively insensitive to the interconnect resistivity in the range of 1.7–2.5 μΩ · cm (typical for advanced processes with copper interconnect [97], [513, 514]). A conductivity of 40 S/μm $(2.5\,\mu\Omega \cdot \text{cm})^{-1}$ yields an inductance that is less than 4 % larger than the inductance obtained for a conductivity of 58 S/μm.

A line thickness of 1 μm is assumed for the interconnect structures. In the analysis, the lines are split into multiple filaments to account for skin and proximity effects, as discussed in Sect. 2.2. The number of filaments is chosen to be sufficiently large to achieve a 1 % accuracy of the computed values. Simulations are performed at three frequencies, 1, 10, and 100 GHz. Typical simulation run times for the structures are under 1 min on a Sun Blade 100 workstation.

28.3 Grid Types

To assess the dependence of the inductive properties on the power and ground lines, the coupling characteristics of three types of power/ground grid structures have been analyzed. In the grids of the first type, called *non-interdigitated grids*, the power lines fill one half of the grid and the ground lines fill the other half of the grid, as shown in Fig. 28.3a. In *interdigitated grids*, the power and ground lines are alternated and equidistantly spaced, as shown in Fig. 28.3b. The grids of the third type are a variation of the interdigitated grids. Similar to interdigitated grids, the power and ground lines are alternated, but rather than placed equidistantly, the lines are placed in equidistantly spaced pairs of adjacent power and ground lines, as shown in Fig. 28.3c. These grids are called *paired grids*. Interdigitated and paired grids are grids with alternating power and ground lines.

The number of power lines matches the number of ground lines in all of the grid structures. The number of power/ground line pairs is varied from one to ten. The grid lines are assumed to be 1 mm long and are placed on a 20 μm pitch. The specific line length is unimportant since at these high length to line pitch ratios the inductance scales nearly linearly with the line length, as discussed in Sect. 28.6.

An analysis of these structures has been performed for two line cross sections, 1×1 μm and 1×3 μm. For each of these structures, the following characteristics have been determined: the partial self-inductance of the power (forward current) and ground (return current) paths L_{pp} and L_{gg}, respectively, the power to ground path coupling L_{pg}, and the loop inductance L_{loop}. When determining the loop inductance, all of the ground lines at one end of the grid are short circuited to form a ground terminal, all of the power lines at the same end of the grid are short circuited to form a power terminal, and all of the lines at the other end of the grid are short circuited to complete the current loop. This configuration assumes that the current loop is completed on-chip. This assumption is valid for high frequency signals which are effectively terminated through the on-chip decoupling capacitance which provides a low impedance termination as compared to the inductive leads of the package. The on-chip inductance affects the signal integrity of the high frequency signals. If the current loop is completed on-chip, the current in the power lines is always antiparallel to the current in the ground lines.

The loop inductance of the three types of grid structures operating at 1 GHz is displayed in Fig. 28.4 as a function of the number of lines in the grid. The partial self- and mutual inductance of the power and ground current paths is shown in

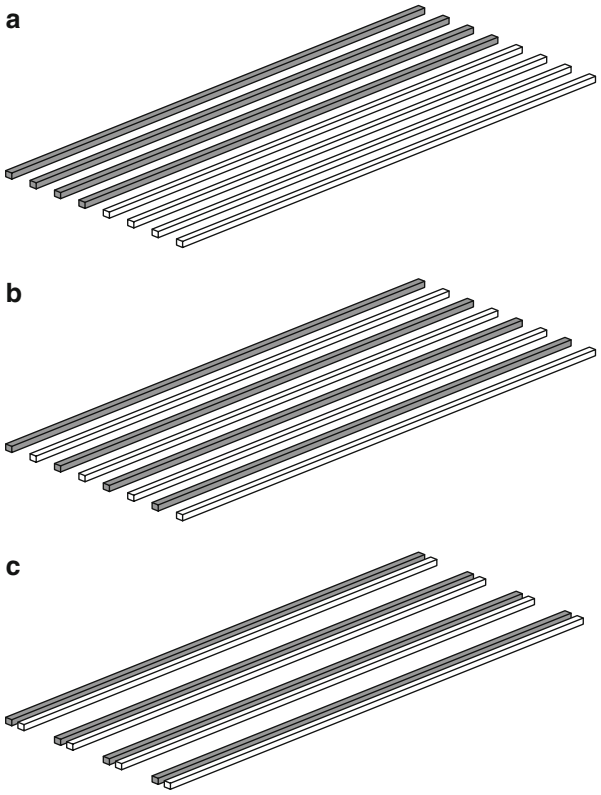


Fig. 28.3 Power/ground grid structures under investigation; (**a**) non-interdigitated grid, (**b**) grid with the power lines interdigitated with the ground lines, (**c**) paired grid, the power and ground lines are in close pairs. The power lines are *gray* colored, the ground lines are *white* colored

Fig. 28.5 for grid structures with $1 \times 1\,\mu m$ cross section lines and in Fig. 28.6 for grids with $1 \times 3\,\mu m$ cross section lines. The inductance data for 1 and 100 GHz are summarized in Table 28.1. The data depicted in Figs. 28.4, 28.5, and 28.6 are discussed in the following three sections.

The signal lines surrounding the power grids are omitted from the analysis. This omission is justified by the following considerations. First, current returning through signal lines rather than the power/ground network causes crosstalk noise on the lines. To minimize this undesirable effect, the circuits are designed in such a way that the majority of the return current flows through the power and ground lines. Second, the signal lines provide additional paths for the return current and only decrease the inductance of the power distribution network. The value of the grid inductance obtained in the absence of signal lines can therefore be considered as an upper bound.

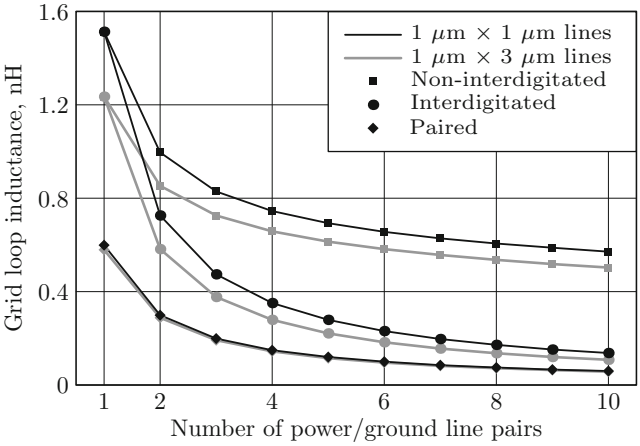


Fig. 28.4 Loop inductance of the power/ground grids as a function of the number of power/ground line pairs (at a 1 GHz signal frequency)

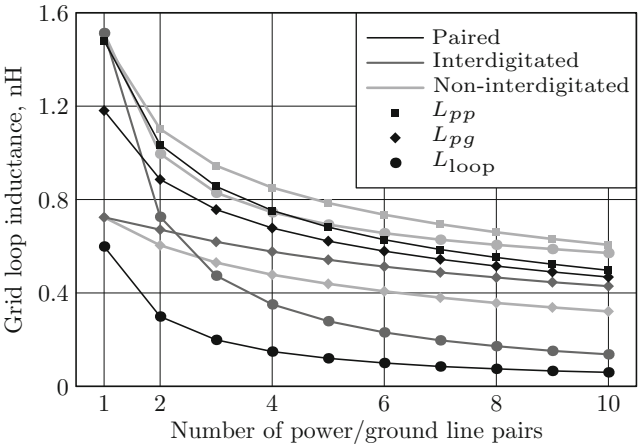


Fig. 28.5 Loop and partial inductance of the power/ground grids with 1 × 1 μm cross section lines (at a 1 GHz signal frequency)

28.4 Inductance Versus Line Width

The loop inductance of the grid depends relatively weakly on the line width. Grids with 1 × 3 μm cross section lines have a lower loop inductance than grids with 1 × 1 μm cross section lines. This decrease in inductance is dependent upon the grid type, as shown in Fig. 28.4. The largest decrease, approximately 21 %, is observed in interdigitated grids. In non-interdigitated grids, the inductance decreases by approximately 12 %. In paired grids, the decrease in inductance is limited to 3–4 %.

Table 28.1 Inductive characteristics of power/ground grids with a 1000 μm length and a 40 μm line pair pitch operating at 1 and 100 GHz

# of P/G pairs	Cross section (μm)	L_{pp}, L_{gg} (nH)			L_{pg} (nH)			L_{loop} (nH)		
		N/int	Int	Paired	N/int	Int	Paired	N/int	Int	Paired
1 GHz										
1	1 × 1	1.481	1.481	1.481	0.724	0.724	1.181	1.513	1.513	0.599
1	1 × 3	1.342	1.342	1.342	0.725	0.725	1.053	1.235	1.235	0.579
2	1 × 1	1.102	1.035	1.035	0.604	0.671	0.886	0.996	0.726	0.299
2	1 × 3	1.031	0.963	0.966	0.604	0.672	0.822	0.853	0.582	0.288
3	1 × 1	0.945	0.856	0.857	0.530	0.619	0.757	0.829	0.474	0.199
3	1 × 3	0.894	0.807	0.810	0.531	0.618	0.714	0.726	0.377	0.192
4	1 × 1	0.851	0.753	0.753	0.478	0.577	0.678	0.745	0.351	0.149
4	1 × 3	0.809	0.714	0.717	0.479	0.575	0.645	0.658	0.279	0.144
5	1 × 1	0.785	0.682	0.682	0.439	0.542	0.622	0.693	0.279	0.120
5	1 × 3	0.748	0.649	0.652	0.440	0.539	0.594	0.614	0.221	0.115
6	1 × 1	0.735	0.628	0.629	0.407	0.513	0.579	0.656	0.231	0.100
6	1 × 3	0.700	0.600	0.602	0.409	0.509	0.555	0.582	0.183	0.096
7	1 × 1	0.694	0.586	0.587	0.380	0.488	0.544	0.628	0.197	0.085
7	1 × 3	0.661	0.561	0.563	0.383	0.483	0.522	0.557	0.156	0.082
8	1 × 1	0.660	0.552	0.552	0.357	0.466	0.515	0.606	0.172	0.075
8	1 × 3	0.629	0.529	0.531	0.361	0.461	0.495	0.536	0.136	0.072
9	1 × 1	0.631	0.523	0.523	0.338	0.446	0.490	0.588	0.152	0.066
9	1 × 3	0.601	0.502	0.504	0.342	0.441	0.472	0.518	0.120	0.064
10	1 × 1	0.606	0.497	0.498	0.321	0.429	0.468	0.571	0.137	0.060
10	1 × 3	0.577	0.478	0.480	0.326	0.424	0.451	0.503	0.108	0.057
100 GHz										
1	1 × 1	1.468	1.468	1.457	0.724	0.724	1.181	1.486	1.486	0.551
1	1 × 3	1.315	1.315	1.291	0.725	0.725	1.062	1.180	1.180	0.457
2	1 × 1	1.088	1.022	1.023	0.604	0.670	0.886	0.968	0.703	0.275
2	1 × 3	1.010	0.945	0.940	0.605	0.670	0.826	0.810	0.548	0.228
3	1 × 1	0.928	0.845	0.848	0.533	0.616	0.756	0.789	0.458	0.183
3	1 × 3	0.873	0.793	0.792	0.534	0.615	0.716	0.678	0.355	0.152
4	1 × 1	0.830	0.741	0.745	0.483	0.571	0.676	0.695	0.340	0.138
4	1 × 3	0.788	0.702	0.702	0.484	0.570	0.645	0.607	0.263	0.114
5	1 × 1	0.762	0.671	0.674	0.444	0.536	0.619	0.634	0.270	0.110
5	1 × 3	0.727	0.639	0.640	0.446	0.535	0.594	0.560	0.208	0.091
6	1 × 1	0.710	0.618	0.621	0.414	0.506	0.575	0.591	0.224	0.092
6	1 × 3	0.680	0.591	0.592	0.416	0.505	0.554	0.527	0.173	0.076
7	1 × 1	0.668	0.576	0.579	0.389	0.480	0.540	0.559	0.191	0.079
7	1 × 3	0.641	0.553	0.554	0.391	0.479	0.522	0.501	0.147	0.065
8	1 × 1	0.633	0.542	0.545	0.367	0.458	0.510	0.533	0.167	0.069
8	1 × 3	0.609	0.522	0.523	0.369	0.457	0.494	0.480	0.128	0.057
9	1 × 1	0.604	0.513	0.516	0.348	0.439	0.485	0.511	0.148	0.061
9	1 × 3	0.582	0.495	0.496	0.351	0.438	0.471	0.463	0.114	0.050
10	1 × 1	0.578	0.488	0.491	0.332	0.422	0.463	0.493	0.133	0.055
10	1 × 3	0.558	0.472	0.473	0.334	0.421	0.450	0.448	0.102	0.045

N/int – non-interdigitated grids; *Int* – interdigitated grids; *Paired* – paired grids

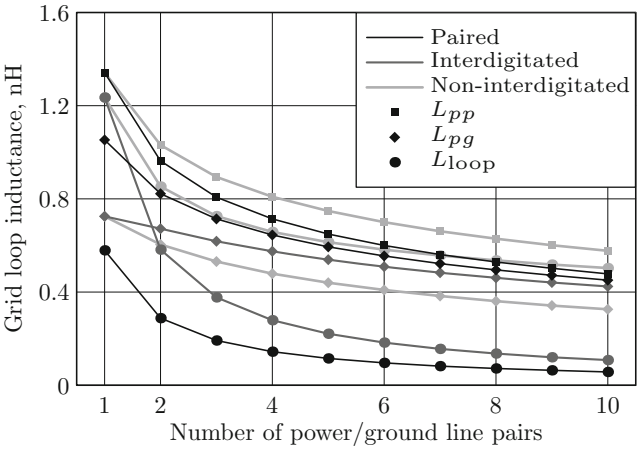


Fig. 28.6 Loop and partial inductance of the power/ground grids with 1 × 3 μm cross section lines (at a 1 GHz signal frequency)

This behavior can be explained in terms of the partial inductance, L_{pp} and L_{pg} [72, 73]. Due to the symmetry of the power and ground paths, $L_{gg} = L_{pp}$, the relation of the loop inductance to the partial inductance (28.2) simplifies to

$$L_{\text{loop}} = L_{pp} + L_{gg} - 2L_{pg} = 2(L_{pp} - L_{pg}). \tag{28.4}$$

According to (28.4), L_{loop} increases with larger L_{pp} and decreases with larger L_{pg}. That is, decreasing the self-inductance of the forward and return current paths forming the current loop decreases the loop inductance, while increasing the inductive coupling of the two paths decreases the loop inductance. The net change in the loop inductance depends, therefore, on the relative effect of increasing the line width on L_{pp} and L_{pg} in the structures of interest.

The self-inductance of a single line is a weak function of the line cross-sectional dimensions, see (3.1) [44]. This behavior is also true for the complex multi-conductor structures under investigation. Comparison of the data shown in Fig. 28.5 with the data shown in Fig. 28.6 demonstrates that changing the line cross section from 1 × 1 μm to 1 × 3 μm decreases L_{pp} by 4–6 % in all of the structures under consideration.

The dependence of inductive coupling L_{pg} on the line width, however, depends on the grid type. In non-interdigitated and interdigitated grids, the line spacing is much larger than the line width and the coupling L_{pg} changes insignificantly with the line width. Therefore, the loop inductance L_{loop} in non-interdigitated and interdigitated grids decreases with line width primarily due to the decrease in the self-inductance of the forward and return current paths, L_{pp} and L_{gg}.

In paired grids, the line width is comparable to the line-to-line separation and the dependence of L_{pg} on the line width is non-negligible: L_{pg} decreases by 4–6 %,

as quantified by comparison of the data shown in Fig. 28.5 with the data shown in Fig. 28.6. In paired grids, therefore, the grid inductance decreases more slowly with line width as compared with interdigitated and non-interdigitated grids, because a reduction in the self-inductance of the current paths L_{pp} is significantly offset by a decrease in the inductive coupling of the paths L_{pg}.

28.5 Dependence of Inductance on Grid Type

The grid inductance varies with the configuration of the grid. With the same number of power/ground lines, grids with alternating power and ground lines exhibit a lower inductance than non-interdigitated grids; this behavior is discussed in Sect. 28.5.1. The inductance of the paired grids is lower than the inductance of the interdigitated grids; this topic is discussed in Sect. 28.5.2.

28.5.1 Non-interdigitated Versus Interdigitated Grids

The difference in inductance between non-interdigitated and interdigitated grids increases with the number of lines, reaching an approximately 4.2 difference for ten power/ground line pairs for the case of a $1 \times 1\,\mu$m cross section line, as shown in Fig. 28.4 ($\sim$4.7 difference for the case of a $1 \times 3\,\mu$m cross section line). This difference is due to two factors.

First, in non-interdigitated grids the lines carrying current in the same direction (forming the forward or return current paths) are spread over half the width of the grid, while in interdigitated (and paired) grids both the forward and return paths are spread over the entire width of the grid. The smaller the separation between the lines, the greater the mutual inductive coupling between the lines and the partial self-inductance of the forward and return paths, L_{pp} and L_{gg}, as described in Sect. 28.1. This trend is confirmed by the data shown in Figs. 28.5 and 28.6, where interdigitated grids have a lower L_{pp} as compared to non-interdigitated grids.

Second, each line in the interdigitated structures is surrounded with lines carrying current in the opposite direction, creating strong coupling between the forward and return currents and increasing the partial mutual inductance L_{pg}. Alternatively, in the non-interdigitated arrays (see Fig. 28.3a), all of the lines (except for the two lines in the middle of the array) are surrounded with lines carrying current in the same direction. The power-to-ground inductive coupling L_{pg} is therefore lower in non-interdigitated grids, as shown in Figs. 28.5 and 28.6.

In summary, the interdigitated grids exhibit a lower partial self-inductance L_{pp} and a higher partial mutual inductance L_{pg} as compared to non-interdigitated grids [72, 73]. The interdigitated grids, therefore, have a lower loop inductance L_{loop} as described by (28.4).

28.5.2 Paired Versus Interdigitated Grids

The loop inductance of paired grids is 2.3 times lower than the inductance of the interdigitated grids for the case of a $1 \times 1\,\mu m$ cross section line and is 1.9 times lower for the case of a $1 \times 3\,\mu m$ cross section line, as shown in Fig. 28.4. The reason for this difference is described as follows in terms of the partial inductance. The structure of the forward (and return) current path in a paired grid is identical to the structure of the forward path in an interdigitated grid (only the relative position of the forward and return current paths differs). The partial self-inductance L_{pp} is therefore the same in paired and interdigitated grids; the two corresponding curves completely overlap in Figs. 28.5 and 28.6. The values of L_{pp} and L_{gg} for the two types of grids are equal within the accuracy of the analysis. In contrast, due to the immediate proximity of the forward and return current lines in the paired grids, the mutual coupling L_{pg} is higher as compared to the interdigitated grids, as shown in Figs. 28.5 and 28.6. Therefore, the difference in loop inductance between paired and interdigitated grids is due to the difference in the mutual inductance [72, 73].

Note that although the inductance of a power distribution network (i.e., the inductance of the power-ground current loop) in the case of paired power grids is lower as compared to interdigitated grids, the signal self-inductance (i.e., the inductance of the signal to the power/ground loop) as well as the inductive coupling of the signal lines is higher. The separation of the power/ground line pairs in paired grids is double the separation of the power and ground lines in interdigitated grids. The current loops formed between the signal lines and the power and ground lines are therefore larger in the case of paired grids. Interdigitated grids also provide enhanced capacitive shielding for the signal lines, as each power/ground line has the same number of signal neighbors as a power/ground line pair. Thus, a tradeoff exists between power integrity and signal integrity in the design of high speed power distribution networks. In many circuits, signal integrity is of primary concern, making interdigitated grids the preferred choice.

28.6 Dependence of Inductance on Grid Dimensions

The variation of grid inductance with grid dimensions, such as the grid length and width (assuming that the width and pitch of the grid lines is maintained constant) is considered in this section. The dependence of the grid loop inductance on the number of lines in the grid (i.e., the grid width) is discussed in Sect. 28.6.1. The dependence of the grid loop inductance on the length of the grid is discussed in Sect. 28.6.2. The concept of sheet inductance is described in Sect. 28.6.3. A technique for efficiently and accurately calculating the grid inductance is outlined in Sect. 28.6.4.

28.6.1 Dependence of Inductance on Grid Width

Apart from a lower loop inductance, paired and interdigitated grids have an additional desirable property as compared to non-interdigitated grids. The loop inductance of paired and interdigitated grids depends inversely linearly with the number of lines, as shown in Fig. 28.4. That is, for example, the inductance of a grid with ten power/ground line pairs is half of the inductance of a grid with five power/ground line pairs, all other factors being the same. For paired grids, this inversely linear dependence is exact (i.e., any deviation is well within the accuracy of the inductance extracted by FastHenry). For interdigitated grids, the inversely linear dependence is exact within the extraction accuracy at high numbers of power/ground line pairs. As the number of line pairs is reduced to two or three, the accuracy deteriorates to 5–8 % due to the "fringe" effect. The electrical environment of the lines at the edges of the grid, where a line has only one neighbor, is significantly different from the environment within the grid, where a line has two neighbors. The fringe effect is insignificant in paired grids because the electrical environment of a line is dominated by the pair neighbor, which is physically much closer as compared to other lines in the paired grid.

As discussed in Sect. 28.1, the inductance of conductors connected in parallel decreases slower than inversely linearly with the number of conductors if inductive coupling of the parallel conductors is present. The inversely linear decrease of inductance with the number of lines may seem to contradict the existence of significant inductive coupling among the lines in a grid. This effect can be explained by (28.4). While the partial self-inductance of the power and ground paths L_{pp} and L_{gg} indeed decreases slowly with the number of lines, so does the power to ground coupling L_{pg}, as shown in Figs. 28.5 and 28.6. The nonlinear behavior of L_{pp} and the nonlinear behavior of L_{pg} effectively cancel each other (see (28.4)), resulting in a loop inductance with an approximately inversely linear dependence on the number of lines [71, 73].

From a circuit analysis point of view this behavior can be explained as follows. Consider a paired grid. The coupling of a distant line to a power line in any power/ground pair is nearly the same as the coupling of the same distant line to a ground line in the same pair due to the close proximity of the power and ground lines within each power/ground pair. The coupling to the power line counteracts the coupling to the ground line. As a result, the two effects cancel. Applying the same argument in the opposite direction, the effect of coupling a specific line to a power line is canceled by the line coupling to the ground line immediately adjacent to the power line. Similar reasoning is applicable to interdigitated grids, however, due to the equidistant spacing between the lines, the degree of coupling cancellation is lower for those lines at the periphery of the grid. The lower degree of coupling cancellation is the cause of the aforementioned "fringe" effect.

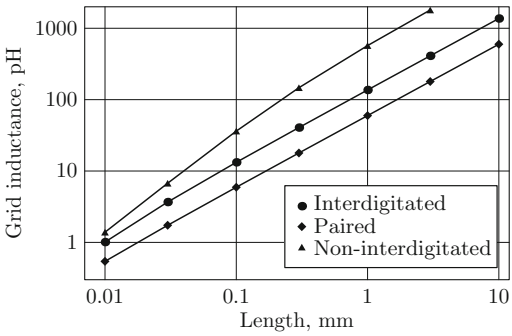


Fig. 28.7 Grid inductance versus grid length

28.6.2 Dependence of Inductance on Grid Length

The length of the grid structures described in Sect. 28.3 is varied to characterize the variation in the grid inductance with grid length. The results are shown in Fig. 28.7. The grid inductance varies virtually linearly over a wide range of grid length. This behavior is due to the cancellation of long distance inductive coupling, as described in Chap. 3. The linear dependence of inductance on length is analogous to the dependence of inductance on the number of lines [71]. Similar to the variation in the loop inductance, illustrated in Fig. 3.4, the variation in the grid inductance deviates from the linear behavior at grid lengths comparable to the separation between the forward and return current paths. In non-interdigitated grids, the effective separation between the forward and return currents is greatest. The range of the linear variation of the inductance with grid length is therefore limited as compared to paired and interdigitated grids, as shown in Fig. 28.7.

28.6.3 Sheet Inductance of Power Grids

As discussed in this section, the inductance of grids with alternating power and ground lines is linearly dependent upon the grid length and number of lines. Furthermore, the grid inductance of interdigitated grids is relatively constant with frequency, as described in Chap. 29. These properties of the grid inductance greatly simplify the procedure for evaluating the inductance of power distribution grids, permitting the efficient assessment of design tradeoffs.

The resistance of the grid increases linearly with grid length and decreases inversely linearly with grid width (i.e., the number of parallel lines). Therefore, the resistive properties of the grid can be conveniently described as a dimension independent grid sheet resistance $R_{\square}$, similar to the sheet resistance of an interconnect layer. The linear dependence of the grid inductance on the grid dimensions is similar to that of the grid resistance. As with resistance, it is convenient to express the inductance of a power grid as a dimension independent *grid sheet inductance*

$L_\square$ (i.e., Henrys per square), rather than to characterize the grid inductance for a particular grid with specific dimensions. Thus,

$$L_\square = L_{\text{grid}} \cdot \frac{PN}{l}, \tag{28.5}$$

where l is the grid length, P is the line (or line pair) pitch, and N is the number of lines (line pairs). This approach is analogous to the *plane* sheet inductance of two parallel power and ground planes (e.g., in a PCB stack), which depends only on the separation between the planes, not on the specific dimensions of the planes. Similarly, the grid sheet inductance reflects the overall structural characteristics of the grid (i.e., the line width and pitch) and is independent of the dimensions of a specific structure (i.e., the grid length and the number of lines in the grid). The sheet inductance is used as a dimension independent measure of the grid inductance in the discussion of power grid tradeoffs in Chap. 30.

28.6.4 Efficient Computation of Grid Inductance

The linear dependence of inductance on the grid length and width (i.e., the number of lines) has a convenient implication. The inductance of a large paired or interdigitated grid can be extrapolated with good accuracy from the inductance of a grid consisting of only a few power/ground pairs.

For an accurate extrapolation of the interdigitated grids, the line width and pitch of the original and extrapolated grids should be maintained the same. For example, for a grid consisting of $2N$ lines (i.e., N power-ground line pairs) of length l, width W, and pitch P, the inductance L_{2N} can be estimated as

$$L_{2N} \approx \frac{L_2}{N}, \tag{28.6}$$

where L_2 is the inductance of a loop formed by two lines with the same dimensions and pitch. The inductance of a two-line loop L_2 can be calculated using (3.4). The power grid is considered to consist of uncoupled two-line loops. The accuracy of the approximation represented by (28.6) is about 10 % for practical line geometries. Alternatively, the grid inductance L_{2N} can be approximated as

$$L_{2N} \approx L_4 \frac{2}{N}, \tag{28.7}$$

where L_4 is the inductance of a power grid consisting of four lines of the original dimensions and pitch. A grid consisting of four lines can be considered as two coupled two-line loops connected in parallel. The loop inductance of a four-line grid can be efficiently calculated using analytic expressions (28.3) and (3.4).

For practical geometries, a four-line approximation (see (28.7)) offers an accuracy within 5 % as compared to the two-line approximation (see (28.6)), as the coupling to non-neighbor lines is partially considered.

In paired grids, the effective inductive coupling among power/ground pairs is negligible and (28.6) is practically exact. The effective width of the current loop in paired grids is primarily determined by the separation between the power and ground lines within a pair. The spatial separation between pairs has (almost) no effect on the grid loop inductance. Expression (3.4) should be used with caution in estimating the inductance of adjacent power and ground lines. Expression (3.4) is accurate only for low frequencies and moderate W/T ratios; (3.4) does not consider proximity effects in paired grids at high frequencies.

Alternatively, the grid sheet inductance $L_{\square}$ can be determined based on (28.6) and (28.7). For example, using (3.4) to determine the loop inductance of a line pair L_{pair}, the grid sheet inductance of a paired grid becomes

$$L_{\square} = L_{\text{pair}} \frac{P}{l} \approx 0.4P \left(\ln \frac{S}{H+W} + \frac{3}{2} \right) \frac{\mu\text{H}}{\square}, \tag{28.8}$$

where P is the line pair pitch, and S is the separation between the line centers in the pair. The inductance of grids with the same line width and pitch is

$$L_{\text{grid}} = L_{\square} \cdot \frac{l}{PN}. \tag{28.9}$$

To summarize, the inductance of regular grids with alternating power and ground lines can be accurately estimated with analytic expressions.

28.7 Summary

The inductive properties of single layer regularly structured grids have been characterized. The primary results are summarized as follows.

- The inductance of grids with alternating power and ground lines varies linearly with grid length and width
- The inductance of grids with alternating power and ground lines varies relatively little with the cross-sectional dimensions of the lines and the signal frequency
- The inductance of grids with alternating power and ground lines can be conveniently expressed in a dimension independent form described as the sheet inductance
- The grid inductance can be analytically calculated based on the grid dimensions and the cross-sectional dimensions of the lines

Chapter 29
Variation of Grid Inductance with Frequency

The variation of inductance with frequency in high performance power distribution grids is discussed in this chapter. As discussed in Chap. 7, the on-chip inductance affects the integrity of the power supply in high speed circuits. The frequency of the currents flowing through the power distribution networks in high speed ICs varies from quasi-DC low frequencies to tens of gigahertz. Thus, understanding the variation of the power grid inductance with frequency is important in order to built a robust and efficient power delivery system.

This chapter is organized as follows. A procedure for analyzing the inductance as a function of frequency is described in Sect. 29.1. The variation of the power grid inductance with frequency is discussed in Sect. 29.2. The chapter concludes with a summary.

29.1 Analysis Approach

The variation of the grid inductance with frequency is evaluated for the three types of power/ground grids: non-interdigitated, interdigitated, and paired. These types of grid structures are described in Sect. 28.3. The grid structures are depicted in Fig. 29.1.

The analysis is analogous to the procedure described in Sect. 28.3. The inductance extraction program FastHenry [70] is used to explore the inductive properties of these interconnect structures. A conductivity of $58\,\mathrm{S}/\mu\mathrm{m} \simeq (1.72\,\mu\Omega \cdot \mathrm{cm})^{-1}$ is assumed for the interconnect material.

The inductance of grids with alternating power and ground lines, i.e., interdigitated and paired grids, behaves similarly to the grid resistance. With the width and pitch of the lines fixed, the inductance of these grid types increases linearly with the grid length and decreases inversely linearly with the number of lines, as discussed in Chap. 28. Consequently, the inductive and resistive properties of interdigitated

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0_29

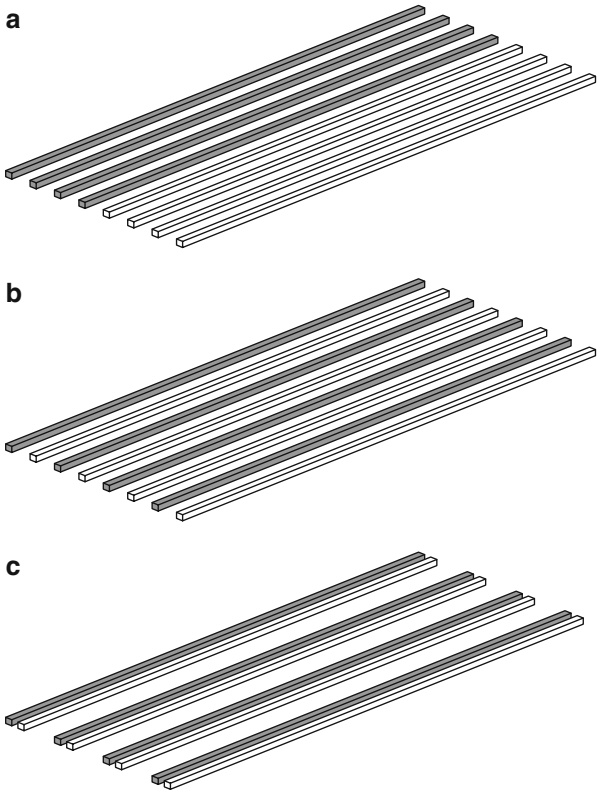


Fig. 29.1 Power/ground grid structures under investigation; (**a**) non-interdigitated grid, (**b**) interdigitated grid, the power lines are interdigitated with the ground lines, (**c**) paired grid, the power and ground lines are in close pairs. The power lines are *gray* colored, the ground lines are *white* colored

and paired grids with a specific line width and pitch can be conveniently expressed in terms of the sheet inductance $L_{\square}$, henrys per square, and the sheet resistance $R_{\square}$, ohms per square [125]. As with the sheet resistance, the sheet inductance is convenient since it is independent of a specific length and width of the grid; this quantity depends only on the pitch, width, and thickness of the grid lines. The impedance properties of interdigitated and paired grids can therefore be studied on structures with a limited number of lines. These results are readily scaled to larger structures, as described in Sect. 28.6.

The grid structures consist of ten lines, five power lines and five ground lines. The power and ground lines carry current in opposite directions, such that a grid forms a complete current loop. The grid lines are assumed to be 1 mm long and are placed on a 10 μm pitch (20 μm line pair pitch in paired grids). The specific grid length and the number of lines is not significant. As discussed in Sect. 28.6 and Chap. 3, the inductance scales linearly with the grid length and the number of lines, provided

the line length to line separation ratio is high and the number of lines exceeds eight to ten. An analysis of the aforementioned grid structures has been performed for line widths W of 1, 3, and 5 μm. The line thickness is 1 μm. The line separation within power-ground pairs in paired grids is 1 μm.

29.2 Discussion of Inductance Variation

The variation of grid inductance with frequency is presented and discussed in this section. Simple circuit models are discussed in Sect. 29.2.1 to provide insight into the variation of inductance with frequency. Based on this intuitive perspective, the data are analyzed and compared in Sect. 29.2.2.

29.2.1 Circuit Models

As discussed in Sect. 2.2, there are two primary mechanisms that produce a significant decrease in the on-chip interconnect inductance with frequency, the proximity effect and multi-path current redistribution. The phenomenon underlying these mechanisms is, however, the same. Where several parallel paths with significantly different electrical properties are available for current flow, the current is distributed among the paths so as to minimize the total impedance. As the frequency increases, the circuit inductance changes from the low frequency limit, determined by the ratio of the resistance of the parallel current paths, to the high frequency value, determined by the inductance ratio of the current paths. At high signal frequencies, the inductive reactance dominates the interconnect impedance; therefore, the path of minimum inductance carries the largest share of the current, minimizing the overall impedance (see Fig. 2.10). Note that parallel current paths can be formed either by several physically distinct lines, as in multi-path current redistribution, or by different paths within the same line, as in the proximity effect, as shown in Fig. 29.2. A thick line can be thought of as being composed of multiple thin lines bundled together in parallel. The proximity effect in such a thick line can be considered as a special case of current redistribution among multiple thin lines forming a thick line.

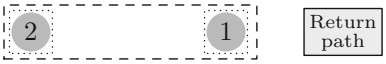


Fig. 29.2 A cross-sectional view of two parallel current paths (*gray circles*) sharing the same current return path (*gray rectangle*). The path closest to the return path, path 1, has a lower inductance than the other path, path 2. The parallel paths can be either two physically distinct lines, as shown by the *dotted line*, or two different paths within the same line, as shown by the *dashed line*

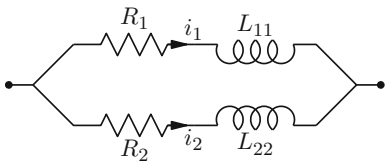


Fig. 29.3 A circuit model of two current paths with different inductive properties

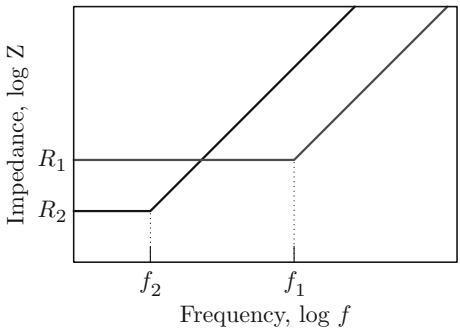


Fig. 29.4 Impedance magnitude versus frequency for two paths with dissimilar impedance characteristics

Consider a simple case with two current paths with different inductive properties. The impedance characteristics are represented by the circuit diagram shown in Fig. 29.3, where the inductive coupling between the two paths is neglected for simplicity. Assume that $L_1 < L_2$ and $R_1 > R_2$.

For the purpose of evaluating the variation of inductance with frequency, the electrical properties of the interconnect are characterized by the inductive time constant $\tau = L/R$. The impedance magnitude of these two paths is schematically shown in Fig. 29.4. The impedance of the first path is dominated by the inductive reactance above the frequency $f_1 = \frac{1}{2\pi}\frac{R_1}{L_1} = \frac{1}{2\pi\tau_1}$. The impedance of the second path is predominantly inductive above the frequency $f_2 = \frac{1}{2\pi}\frac{R_2}{L_2} = \frac{1}{2\pi\tau_2}$, $f_2 < f_1$. At low frequencies, i.e., from DC to the frequency f_1, the ratio of the two impedances is constant. The effective inductance at low frequencies is therefore also constant, determining the low frequency inductance limit. At high frequencies, i.e., frequencies exceeding f_2, the ratio of the impedances is also constant, determining the high frequency inductance limit, $\frac{L_1L_2}{L_1+L_2}$. At intermediate frequencies from f_1 to f_2, the impedance ratio changes, resulting in a variation of the overall inductance from the low frequency limit to the high frequency limit. The frequency range of inductance variation is therefore determined by the two time constants, τ_1 and τ_2. The magnitude of the inductance variation depends upon both the difference between the time constants τ_1 and τ_2 and on the inductance ratio L_1/L_2. Analogously, in the case of multiple parallel current paths, the frequency range and the magnitude of the variation in inductance is determined by the minimum and maximum time constants as well as the difference in inductance among the paths.

The decrease in inductance begins when the inductive reactance $j\omega L$ of the path with the lowest R/L ratio becomes comparable to the path resistance R, $R \sim j\omega L$. The inductance, therefore, begins to decrease at a lower frequency if the minimum R/L ratio of the current paths is lower.

Due to this behavior, the proximity effect becomes significant at higher frequencies than multi-path current redistribution. Significant proximity effects occur in conductors containing current paths with significantly different inductive characteristics. That is, the inductive coupling of one edge of the line to the "return" current (i.e., the current in the opposite direction) is substantially different from the inductive coupling of the other edge of the line to the same "return" current. In geometric terms, this characteristic means that the line width is larger than or comparable to the distance between the line and the return current. Consequently, the line with significant proximity effects is typically the immediate neighbor of the current return line. A narrower current loop is therefore formed with the current return path as compared to the other lines participating in the multi-path current redistribution. A smaller loop inductance L results in a higher R/L ratio. Referring to Fig. 2.10, current redistribution between paths one and two proceeds at frequencies lower than the onset frequency of the proximity effect in path one.

29.2.2 Analysis of Inductance Variation

The inductance of non-interdigitated grids versus signal frequency is shown in Fig. 29.5. At low frequencies, the forward and return currents are uniformly distributed among the lines. The two lines in the center of the grid form the smallest current loop while the lines at the periphery of the grid form wider current loops. The effective width of the current loop at low frequencies is relatively large, approx-

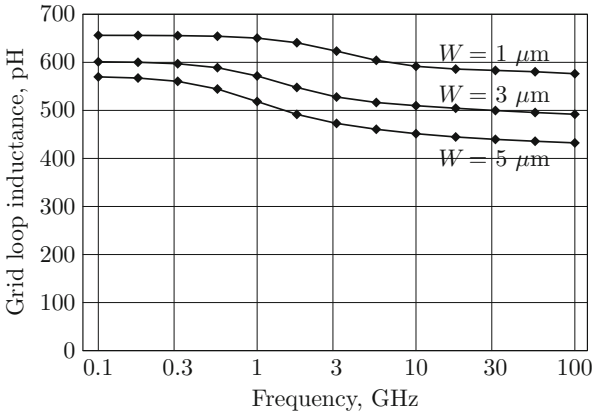


Fig. 29.5 Loop inductance of non-interdigitated grids versus signal frequency

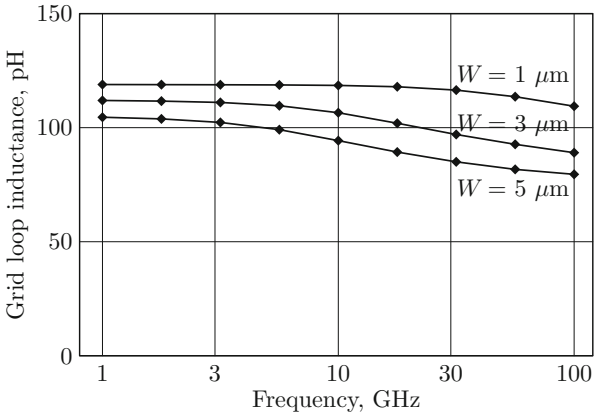


Fig. 29.6 Loop inductance of paired grids versus frequency

imately half of the grid width. Non-interdigitated grids, therefore, have a relatively large inductance L and a low R/L ratio as compared to the other two grid types, interdigitated and paired. Consequently, the onset of a decrease in inductance occurs at a comparatively lower frequency, as illustrated in Figs. 29.5, 29.6, and 29.7. As the signal frequency increases, the current redistributes toward the center of the grid to decrease the grid inductance. Since the width of the grid is much larger than the width of the grid line, the decrease in inductance is primarily due to multi-path current redistribution among the different lines while current redistribution within the line cross sections (the proximity effect) is a secondary effect. The low frequency inductance of a non-interdigitated grid increases with grid width. As the grid width (i.e., the number of lines) increases, the decrease in inductance with frequency becomes more significant and begins at a lower frequency [73].

In power grids with alternating power and ground lines (such as the interdigitated and paired grid structures illustrated in Fig. 29.1b, c, respectively), each line has the same resistance and self-inductance per length, and almost the same inductive coupling to the rest of the grid. As discussed in Chap. 30, long distance inductive coupling is cancelled out in grids with a periodic structure, such that the lines are inductively coupled only to the immediate neighbors, making inductive coupling effectively a local phenomenon. As a result, the distribution of the current among the lines at low frequencies (where the current flows through the path of lowest resistance) practically coincides with the current distribution at high frequencies (where the current flows through the path of lowest inductance). That is, the line resistance has a negligible effect on the current distribution within the grid, i.e., multi-path current redistribution is insignificant. Consequently, the decrease in inductance at high frequencies is caused primarily by the proximity effect which depends upon the line width, spacing, and material resistivity.

This situation is exemplified by paired grids, where multi-path current redistribution is insignificant and the proximity effect is more pronounced due to the

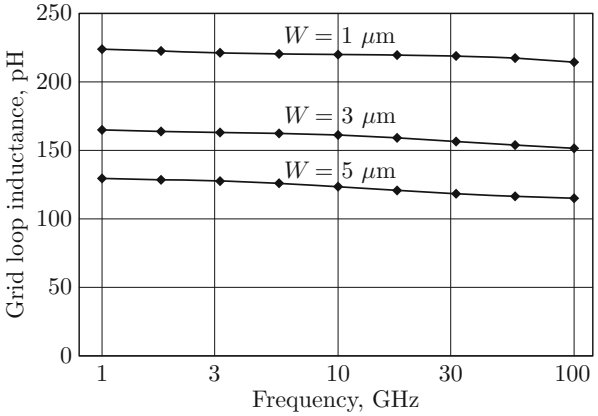


Fig. 29.7 Loop inductance of interdigitated grids versus frequency

small separation between adjacent power and ground lines. The loop inductance versus signal frequency for paired grids is shown in Fig. 29.6. The wider the line, the lower the frequency at which the onset of the proximity effect occurs and the larger the relative decrease in inductance [515, 516], as depicted in Fig. 29.6. Thus, the primary mechanism for a decrease in inductance in paired grids is the proximity effect.

The loop inductance versus signal frequency for interdigitated grids is shown in Fig. 29.7. As in paired grids, multi-path current redistribution is insignificant in interdigitated grids. However, the separation between grid lines is large as compared to the line width (unless the line width is comparable to the line pitch) and the proximity effect is, therefore, also insignificant [515, 516]. As shown in Fig. 29.7, the inductance of interdigitated grids is relatively constant with frequency, the decrease being limited to 10–12% of the low frequency inductance except for the case of very wide lines where the proximity effect becomes significant.

29.3 Summary

The variation of inductance with frequency in high performance power distribution grids is evaluated in this chapter. The variation of inductance with frequency in three types of power grids is analyzed in terms of the mechanisms of inductance variation, as discussed in Sect. 2.2. These results support the design of area efficient and robust power distribution grids in high speed integrated circuits. The chapter results are summarized as follows.

- The inductance of power distribution grids decreases with increasing signal frequency

- The decrease in the inductance of non-interdigitated grids is primarily due to multi-path redistribution of the forward and return currents
- Multi-path current redistribution is greatly minimized in interdigitated and paired grids due to the periodic structure of these grids
- The smaller the separation between the power and ground lines and the wider the lines, the more significant the proximity effects become and the greater the relative decrease in inductance with frequency
- The wider the grid lines, the lower the frequency at which the onset of the decrease in inductance occurs

Chapter 30
Inductance/Area/Resistance Tradeoffs

Tradeoffs among inductance, area, and resistance of power distribution grids are evaluated in this chapter. As discussed in Sect. 1.3, design objectives, such as low impedance (low inductance and resistance), small area, and low current densities (for improved reliability), are typically in conflict. It is therefore important to make a balanced compromise among these design goals based upon application-specific constraints. A quantitative model of the inductance/area/resistance tradeoff in high performance power distribution networks is therefore necessary to achieve an efficient power distribution network. Another important goal is to provide quantitative guidelines to these tradeoffs and to bring intuition to the design of high performance power distribution networks.

Two tradeoff scenarios are considered in this chapter. The inductance versus resistance tradeoff under a constant grid area constraint in high performance power distribution grids is analyzed in Sect. 30.1. The inductance versus area tradeoff under a constant grid resistance constraint is analyzed in Sect. 30.2. The chapter concludes with a summary.

30.1 Inductance vs. Resistance Tradeoff Under a Constant Grid Area Constraint

In the first tradeoff scenario, the fraction of the metal layer area dedicated to the power grid, called the grid area ratio and denoted as A, is assumed fixed, as shown in Fig. 30.1. The objective is to explore the tradeoff between grid inductance and resistance under the constraint of a constant area [125]. The area dedicated to the grid includes both the line width W and the minimum spacing S_0 necessary to isolate the power line from any neighboring lines; therefore, the grid area ratio can be expressed as $A = \frac{W+S_0}{P}$, where P is the line pitch.

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0_30

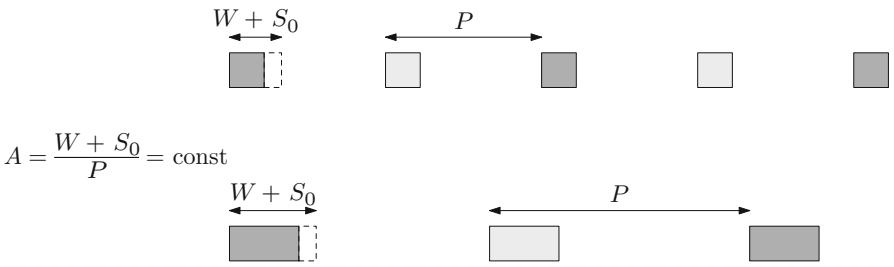


Fig. 30.1 Inductance versus resistance tradeoff scenario under a constant area constraint. As the line width varies, the grid area, including the minimum line spacing S_0, is maintained the same

The inductance of paired grids is virtually independent of the separation between the power/ground line pairs. The effective current loop area in paired grids is primarily determined by the line spacing within each power/ground pair, which is much smaller than the separation between the power/ground pairs [72]. Therefore, only paired grids with an area ratio of 0.2 (i.e., one fifth of the metal resources are allocated to the power and ground distribution) are considered here; the properties of paired grids with a different area ratio A (i.e., different P/G separation) can be linearly extrapolated. In contrast, the dependence of the inductance of the interdigitated grids on the grid line pitch is substantial, since the effective current loop area is strongly dependent on the line pitch. Interdigitated grids with area ratios of 0.2 and 0.33 are analyzed here.

To investigate inductance tradeoffs in power distribution grids, the dependence of the grid inductance on line width is evaluated using FastHenry. Paired and interdigitated grids consisting of ten P/G lines are evaluated. A line length of 1000 μm and a line thickness of 1 μm are assumed. The minimum spacing between the lines S_0 is 0.5 μm. The line width W is varied from 0.5 to 5 μm.

The grid inductance L_{grid} versus line width is shown in Fig. 30.2 for two signal frequencies: 1 GHz (the low frequency case) and 100 GHz (the high frequency case). The high frequency inductance is within 10 % of the low frequency inductance for interdigitated grids, as mentioned previously. The large change in inductance for paired grids is due to the proximity effect in closely spaced, relatively wide lines.

With increasing line width W, the grid line pitch P (and, consequently, the grid width) increases accordingly so as to maintain the desired grid area ratio $A = \frac{W+S_0}{P}$. Therefore, the inductance of a grid with a specific line width cannot be directly compared to the inductance of a grid with a different line width due to the difference in grid width. To perform a meaningful comparison of the grid inductance, the dimension specific data shown in Fig. 30.2 is converted to a dimension independent sheet inductance. The sheet inductance of a grid with a fixed area ratio A, $L_{\square}^{A}$, can be determined from L_{grid} through the following relationship,

$$L_{\square}^{A}(W) = L_{\text{grid}} \frac{NP}{l} = L_{\text{grid}} \frac{N}{l} \frac{W + S_0}{A}, \tag{30.1}$$

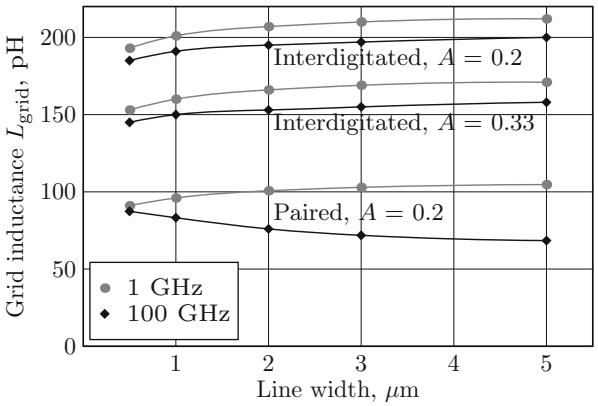


Fig. 30.2 The grid inductance versus line width under a constant grid area constraint for paired and interdigitated grids with ten P/G lines

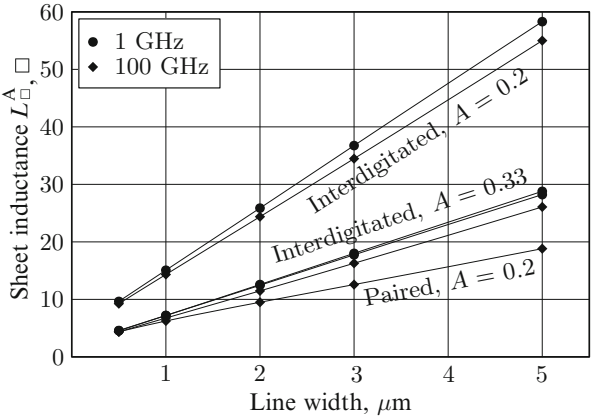


Fig. 30.3 The sheet inductance $L^A_{\square}$ versus line width under a constant grid area constraint

where N is the number of lines (line pairs), P is the line (line pair) pitch in an interdigitated (paired) grid, and l is the grid length. For each of the six L_{grid} data sets shown in Fig. 30.2, a correspondent $L^A_{\square}$ versus line width data set is plotted in Fig. 30.3. As illustrated in Fig. 30.3, the sheet inductance $L^A_{\square}$ increases with line width; this increase with line width can be approximated as a linear dependence with high accuracy.

The low frequency sheet resistance of a grid is $R_{\square} = \rho_{\square}\frac{P}{W}$. The grid resistance under a constant area ratio constraint, $A = \frac{W+S_0}{P} = \text{const}$, can be expressed as a function of only the line width W,

$$R^A_{\square} = \frac{\rho_{\square}}{A}\frac{W + S_0}{W}. \tag{30.2}$$

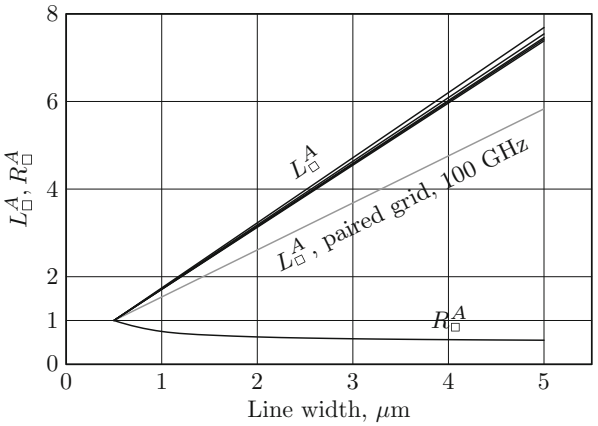


Fig. 30.4 Normalized sheet inductance and sheet resistance versus the width of the P/G line under a constant grid area constraint

This expression shows that as the line width W increases from the minimum line width $W_{\min} = S_0$ ($\frac{W+S_0}{W} = 2$) to a large width ($W \gg S_0$, $\frac{W+S_0}{W} \simeq 1$), the resistance decreases twofold. An intuitive explanation of this result is that at the minimum line width $W_{\min} = S_0$, only half of the grid area used for power routing is filled with metal lines (the other half is used for line spacing) while for large widths $W \gg S_0$, almost all of the grid area is metal.

In order to better observe the relative dependence of the grid sheet inductance and resistance on the line width, $L^A_\square$ and $R^A_\square$ are plotted in Fig. 30.4 normalized to the respective values at a minimum line width of 0.5 µm (such that $L^A_\square$ and $R^A_\square$ are equal to one normalized unit at 0.5 µm). As shown in Fig. 30.4, five out of six $L^A_\square$ lines have a similar slope. These lines depict the inductance of a paired grid at 1 GHz and the inductance of two interdigitated grids ($A = 0.2$ and $A = 0.33$) at 1 and 100 GHz. The line with a lower slope represents a paired grid at 100 GHz. This different behavior is due to pronounced proximity effects in closely placed wide lines with very high frequency signals.

The dependence of the grid sheet inductance on line width is virtually linear and can be accurately approximated by

$$L^A_\square(W) = L^A_\square(W_{\min}) \cdot \{1 + K \cdot (W - W_{\min})\}, \tag{30.3}$$

where $L^A_\square(W_{\min})$ is the sheet inductance of a grid with a minimum line width and K is the slope of the lines shown in Fig. 30.4. Note that while $L^A_\square(W_{\min})$ depends on the grid type and area ratio (as illustrated in Fig. 30.3), the coefficient K is virtually independent of these parameters (with the exception of the special case of paired grids at 100 GHz).

The grid inductance increases with line width, as shown in Fig. 30.4. The inductance increases eightfold (sixfold for the special case of a paired grid at

100 GHz) for a tenfold increase in line width [125]. The grid resistance decreases nonlinearly with line width. As mentioned previously, this decrease in resistance is limited to a factor of 2.

The inductance versus resistance tradeoff has an important implication in the case where at the minimum line width the peak power noise is determined by the resistive voltage drop IR, but at the maximum line width the inductive voltage drop $L\,dI/dt$ is dominant. As the line width decreases, the inductive $L\,dI/dt$ noise becomes smaller due to the lower grid inductance L while the resistive IR noise increases due to the greater grid resistance R, as shown in Fig. 30.4. Therefore, a minimum total power supply noise, $IR + L\,dI/dt$, exists at some target line width. The line width that produces the minimum noise depends upon the ratio and relative timing of the peak current demand I and the peak transient current demand $\frac{dI}{dt}$. The optimal line width is, therefore, application dependent. This tradeoff provides guidelines for choosing the width of the power grid lines that produces the minimum noise.

30.2 Inductance vs. Area Tradeoff Under a Constant Grid Resistance Constraint

In the second tradeoff scenario, the resistance of the power distribution grid is fixed (for example, by IR drop or electromigration constraints) [125], as shown in Fig. 30.5. The grid sheet resistance is

$$R_{\square} = \rho_{\square}\frac{P}{W} = \frac{\rho_{\square}}{M} = \text{const}, \tag{30.4}$$

where $\rho_{\square}$ is the sheet resistivity of the metal layer and $M = \frac{W}{P}$ is the fraction of the area filled with power grid metal, henceforth called the metal ratio of the grid. The constant resistance $R_{\square}$ infers a constant grid metal ratio M. The constraint of a constant grid resistance is similar to that of a constant grid area except that the line spacing is not considered as a part of the grid area. The objective is to explore

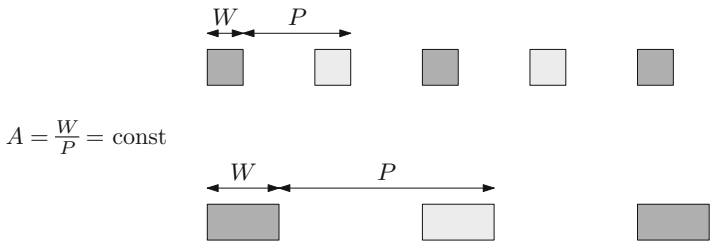


Fig. 30.5 Inductance versus area tradeoff scenario under a constant resistance constraint. As the line width varies, the metal area of the grid and, consequently, the grid resistance are maintained constant

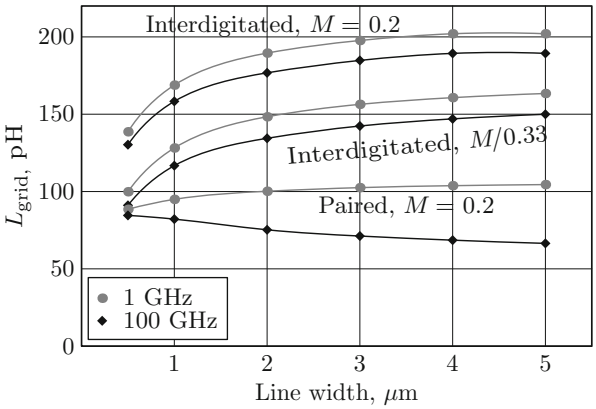


Fig. 30.6 The grid inductance versus line width under a constant grid resistance constraint for paired and interdigitated grids with ten P/G lines

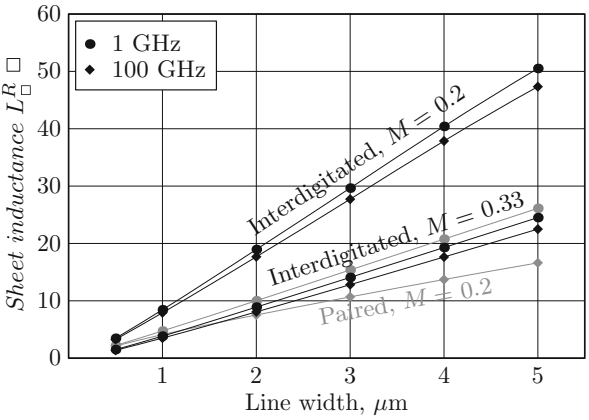


Fig. 30.7 The sheet inductance $L^R_\square$ versus line width under a constant grid resistance constraint

tradeoffs between the grid inductance and area under the constraint of a constant grid resistance. This analysis is conducted similarly to the analysis described in the previous section. The grid inductance L^R_{grid} versus line width is shown in Fig. 30.6. The corresponding sheet inductance $L^R_\square$ versus line width data set is plotted in Fig. 30.7. The normalized sheet inductance and grid area data, analogous to the data shown in Fig. 30.4, is depicted in Fig. 30.8.

As shown in Fig. 30.8, under a constant resistance constraint, the grid inductance increases linearly with line width. Unlike in the first scenario, the slope of the inductance increase with line width varies with the grid type and grid metal ratio. Paired grids have the lowest slope and interdigitated grids with a metal ratio of 0.33 have the highest slope. A slower increase in inductance with line width is preferable, as, under a target resistance constraint, the power network is either smaller and/or

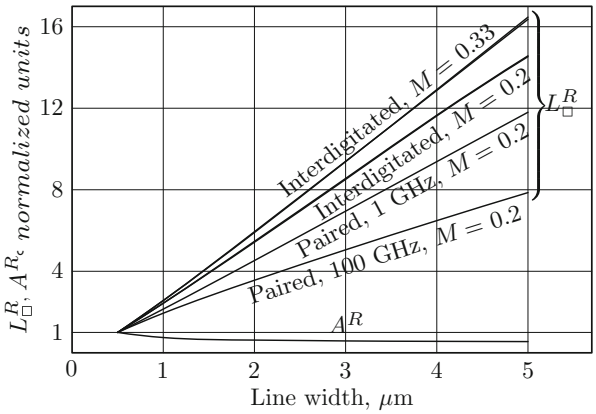


Fig. 30.8 Normalized sheet inductance $L^R_\square$ and grid area ratio A^R versus the width of the P/G line under a constant grid resistance (i.e., constant grid metal ratio M) constraint

less inductive. The slope of the inductance increase with line width is independent of frequency in interdigitated grids (the lines for 1 and 100 GHz coincide and are not discernible in the figure), while in paired grids the slope decreases significantly at high frequencies (100 GHz). The inductance increase varies from 8 to 16-fold, depending on grid type and grid resistance (i.e., grid metal ratio), for a tenfold increase in line width. A reduction in the grid area is limited by a factor of 2, similar to the decrease in resistance in the first tradeoff scenario.

30.3 Summary

Inductance/area/resistance tradeoffs in single layer power distribution grids are explored in this chapter. The primary conclusions can be summarized as follows.

- The grid inductance can be traded off against the grid resistance as the width of the grid lines is varied under a constant grid area constraint
- The grid inductance can be traded off against the grid area as the width of the grid lines is varied under a constant grid resistance constraint
- The grid inductance varies linearly with line width when either the grid resistance or the grid area is maintained constant
- The associated penalty in grid area (or resistance) is relatively small as long as the line width remains significantly greater than the minimum line spacing

Chapter 31
Noise Characteristics of On-Chip Power Networks with Decoupling Capacitors

The high frequency response of a power distribution system is the focus of this chapter. The impedance of the power distribution system at high frequencies is determined by the characteristics of the on-chip power distribution network. The impedance of a power system at a specific on-chip location is determined by the local resistive, inductive, and capacitive characteristics of the on-chip network. In this chapter, the impedance characteristics of both the on-chip power interconnect and the decoupling capacitors are combined to evaluate the noise characteristics of a power network. The inductance of an on-chip power distribution network is shown under specific conditions to be a significant design issue in high speed integrated circuits.

As discussed in Chap. 7, the inductance of the on-chip power and ground interconnect affects the impedance characteristics at relatively high frequencies; specifically, from the chip-package resonance to the highest frequencies of interest. The on-chip interconnect is a part of the current loop from the on-chip decoupling capacitors to the package decoupling capacitors. Typically, the inductance of this current loop is dominated by other parts of the loop—the bonding solder bumps, package conductors, and package decoupling capacitors. This situation is changing with technology scaling, as discussed in Sect. 31.1. The propagation of the power supply noise through the on-chip power distribution network is discussed in Sect. 31.2. The on-chip interconnect also provides a current path between the on-chip decoupling capacitors and the load. As the switching speed of the load increases, the inductance of the on-chip power lines can degrade the effectiveness of the on-chip capacitors, as discussed in Sect. 31.3. The chapter concludes with a summary.

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0_31

31.1 Scaling Effects in Chip-Package Resonance

The continuous improvement in the performance characteristics of integrated circuits is primarily due to decreasing feature sizes, as discussed in Chap. 1. Technology scaling, however, has highly unfavorable implications for the impedance characteristics of a power distribution system. The manner in which these scaling trends affect the impedance characteristics of a power distribution system are described in this section. Specifically, the impedance characteristics near the chip-package resonance is the topic of primary concern.

Ideal scaling theory is briefly reviewed in Sect. 5.1. The current density increases as S and the supply voltage decreases as $1/S$ in the ideal scaling scenario, as discussed in Chap. 5. To maintain the power noise margin at the same fraction of the power supply voltage, the impedance of the power distribution system will decrease as

$$Z_{\text{pds}} \propto \frac{V}{I} = \frac{1}{S^2 S_C^2}, \tag{31.1}$$

assuming that the circuit area increases by a factor S_C.

Power supply scaling is impeded in sub-100 nm technologies by the difficulties in reducing the transistor threshold voltages. A decrease in the power supply voltage therefore significantly deviates from the ideal scaling scenario [124]. Consider a scenario where the voltage levels are scaled by a factor S_V ($S_V < S$). The power supply V_{dd} decreases as $1/S_V$, while the transistor current I_{tr} scales as S/S_V^2. The current per circuit area I_{a} increases by a factor of $S^2 \cdot I_{\text{tr}} = S^3/S_V^2$. The impedance of the power supply system therefore decreases as

$$Z_{\text{pds}} \propto \frac{V}{I} \propto \frac{1}{S_V} \frac{1}{S_C^2 S^3/S_V^2} \propto \frac{S_V}{S} \frac{1}{S^2 S_C^2}. \tag{31.2}$$

This rate of decrease in the impedance is greater by a factor of S/S_V as compared to the ideal scaling scenario represented by (31.1).

The evolution of the impedance of a power distribution system in microprocessors is illustrated in Fig. 31.1. The rate of decrease in the impedance is approximately 2.7 times per technology generation (there are approximately four technology generations per decade). This rate is significantly greater than the dimension scaling factor $\sqrt{2}$, in good agreement with the scaling analysis characterized by (31.1) and (31.2). As described in Chap. 1, the rate of decrease in the target impedance has recently saturated (1.25 times per computer generation [286]). This decrease is due to the limited power dissipation capabilities of traditional air cooled packaging.

Consider the magnitude of the impedance at the frequency of the chip-package resonance, where the impedance is typically the greatest. The minimum impedance at the resonant frequency is the characteristic impedance of the tank circuit, $Z_0 = \sqrt{\frac{L}{C}}$, where C is the on-chip decoupling capacitance and L is the inductance

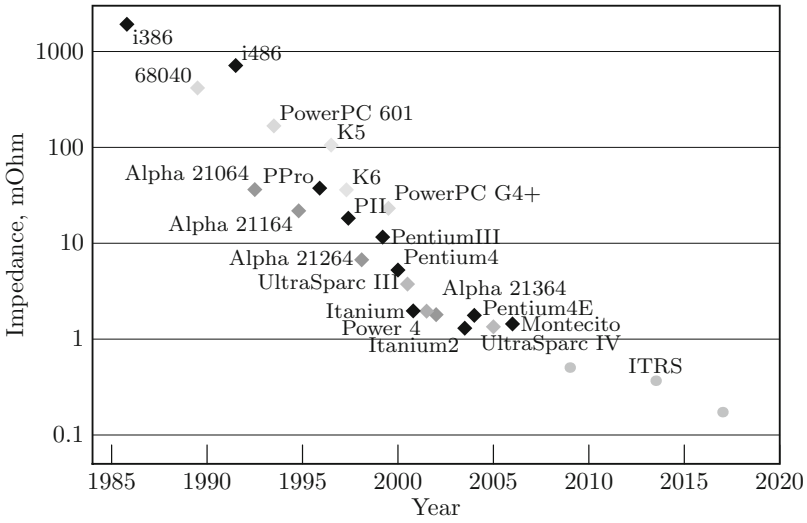


Fig. 31.1 Evolution of the impedance of a power distribution system in microprocessors. Several families of microprocessors and ITRS predictions [124] are shown in different shades of *gray*. The power supply noise margin is assumed to be 10 % of the power supply voltage

of the current loop from the on-chip load to the package decoupling capacitance, as discussed in Sect. 7.6. The decoupling capacitance per circuit area increases by a factor of S, and the overall capacitance of a circuit increases as $S_C^2 S$. The flip-chip contact density increases by a factor of S, assuming a contact pitch scaling factor of $\sqrt{S}$, as discussed in Chap. 5. Assuming a proportional decrease in the other components of the resonant inductance L, such as the inductance of the package conductors and the series inductance of the package capacitors, the overall loop inductance L decreases by a factor of S. Including an increase in the chip area as a factor of S_C^2, the inductance decreases by a factor of $1/S_C^2 S$. In this scenario, the resonant impedance only decreases by a factor of $\sqrt{\frac{1/S_C^2 S}{S_C^2 S}} = 1/S_C^2 S$, which is smaller than the requirements determined by (31.1) and (31.2). This reduction in impedance is therefore insufficient to satisfy a target noise margin. Further improvements in the circuit characteristics are necessary to approach the target specifications. The on-chip decoupling capacitance C is expensive to increase. When the available on-chip area is filled with decoupling capacitors, any additional decoupling capacitors increase die area and, consequently, the overall cost. Furthermore, in sub-100 nm technologies, the on-chip decoupling capacitors increase the static power consumption due to gate tunneling leakage current [517, 518].

The inductance of the current loop between the on-chip circuits and the package capacitors should be decreased to achieve the target impedance. The required decrease in inductance is particularly significant due to the square root dependence of the impedance on inductance. The inductance is reduced in advanced packaging

technologies through improvements in the structure of the package and package decoupling capacitors. Application of finely spaced metal layers and replacing the solder bump connections with denser microvia contacts have been described to achieve this objective [143].

Due to the aggressive reduction in the package inductance, the inductance of the off-chip portion of the current loop becomes comparable to the inductance of the on-chip power interconnect structures. This trend is in agreement with the general approach of using a system of hierarchical decoupling capacitors to achieve a low impedance power distribution system. The upper frequency limit of the low impedance characteristics of a power distribution system should be extended as the switching speed of the on-chip circuitry increases with technology scaling. Exclusively using on-chip capacitors to improve the high frequency impedance characteristics is a relatively expensive solution. Reducing the package inductance typically offers a more economical alternative by extending the frequency range of the package decoupling capacitors, thereby relaxing the requirements placed on the on-chip decoupling capacitors.

31.2 Propagation of Power Distribution Noise

The one-dimensional model described in Chap. 7 is inadequate to accurately describe the high frequency operation of a distributed circuit. As discussed in Chap. 8, the high frequency behavior of the on-chip power distribution network cannot be adequately described by a lumped model. A disturbance in the power supply voltage due to switching a local load propagates relatively slowly through the on-chip power distribution network. The power supply voltage is consequently non-uniform across the circuit die. These important effects are absent in a one-dimensional model. A two-dimensional model of the power distribution network is essential to accurately capture the high frequency impedance characteristics. The propagation of the power distribution noise through the on-chip power distribution network based on a simplified circuit model is discussed in this section.

The speed of noise propagation is an important characteristic of a power distribution network. The propagation speed of an undamped signal can be estimated using an idealized model of an on-chip power distribution network, where the decoupling capacitance is assumed to be uniformly distributed across a uniform power distribution grid. Assuming one-dimensional signal propagation, the power distribution grid is analogous to a capacitively loaded transmission line. The corresponding velocity of the signal propagation is

$$v = \frac{1}{\sqrt{L_{\square} C_{\mathrm{a}}}}, \tag{31.3}$$

where $L_{\square}$ is the sheet inductance of the power distribution grid (as described in Chap. 28) and C_{a} is the area density of the on-chip decoupling capacitance.

As a practical example, assume a global power distribution grid consists of two layers with mutually perpendicular lines. The width and pitch of the grid lines are 50 and 100 μm, respectively, similar to the characteristics of the upper layer of the two layer grid considered in Sect. 35.2. The corresponding sheet inductance is approximately $0.2 \frac{\text{nH}}{\square}$. A typical decoupling capacitance density C_a in high performance digital circuits manufactured in a 130 nm CMOS process is approximately $2\,\text{nF}^2/\text{mm}$. The velocity of the signal propagation based on these characteristics is approximately 1.6 mm/ns. This velocity is two orders of magnitude smaller than the speed of light in the circuit dielectric—approximately 150 mm/ns in silicon dioxide. The low velocity of the signal propagation is due to the high capacitance across the power and ground interconnect. This estimate is the upper bound on the signal velocity, as the resistance of the grid is neglected in (31.3). The resistance of the power lines further reduces the velocity of the signal. In overdamped power distribution networks, the signal propagation is determined by an *RC* rather than an *LC* time constant and approaches a diffusive *RC*-like signal behavior.

The relatively slow propagation of the power distribution noise has important circuit implications. From the perspective of a switching circuit, the low velocity noise propagation means that only the decoupling capacitance in the immediate proximity of the switching load is effective in limiting power supply variations at the load terminals. No charge sharing occurs during the switching transient between the load and the decoupling capacitors located farther than the propagation velocity times the switching time of the load, as described in Chap. 12. Alternatively, from the perspective of a quiescent circuit, the low propagation velocity means that the power supply level of the circuit is only affected by the switching loads that are located in close proximity to the circuit.

The idealized uniform model can also be used to estimate the inductive behavior of the on-chip power distribution grid. For one-dimensional signal propagation, the metric of inductive behavior for transmission lines described by (2.40) can be applied, yielding

$$\frac{t_r}{2\sqrt{L_\square C_\text{a}}} < l < \frac{2}{R_\square}\sqrt{\frac{L_\square}{C_\text{a}}}, \tag{31.4}$$

where $R_\square$ is the sheet resistance of the grid. Assuming a sheet resistance of $0.2\,\Omega/\square$ and a signal rise time t_r of the load current of 100 ps,

$$0.1\,\text{mm} < l < 3\,\text{mm}, \tag{31.5}$$

the power supply noise exhibits a significant inductive component only within a limited distance from the switching load. In other terms, the damping factor ζ of the current path within a power distribution grid,

$$\zeta = \frac{R_\square l}{2}\sqrt{\frac{C_\text{a}}{L_\square}}, \tag{31.6}$$

is smaller than unity if the path length is smaller than 3 mm. Within this distance from the load, the response of a power distribution network is underdamped and the power supply noise can exhibit significant ringing. At greater distances from the load, the propagation of the power supply noise approaches a diffusive *RC*-like behavior.

The inductance of a pair of wide global power and ground lines is comparable to that of an on-chip signal line. The capacitive load of the power lines, however, is approximately three orders of magnitude greater, while the resistance is ten to a hundred times lower. The range of length where the power interconnect exhibits inductive behavior, as indicated by (31.5), is similar to that of an on-chip signal line.

Note, however, that the characteristic impedance of a power-ground line pair is approximately two orders of magnitude lower than the characteristic impedance of an on-chip signal path. The magnitude of the power distribution noise induced by switching an on-chip signal line is therefore two orders of magnitude smaller than the swing of a signal line transition.

31.3 Local Inductive Behavior

The idealized model used in the preceding section provides a reasonable approximation of the noise propagation at a relatively large geometric scale, i.e., where the wavelength of the signal is significantly larger than the pitch of the power lines. At smaller scales (and, consequently, shorter propagation times), the discrete nature of both the power load and the decoupling capacitors may be significant under certain conditions. Particularly, the high frequency characteristics of the power distribution interconnect become crucial. These local effects are discussed in this section.

A low impedance power distribution system during and immediately after the switching of an on-chip load is maintained using on-chip decoupling capacitors. The on-chip decoupling capacitance limits the variation of the power supply until the package decoupling capacitors become effective. As discussed in Sect. 11.3, the intrinsic parasitic capacitance of the load circuit typically provides a small fraction of the required decoupling capacitance. The intrinsic capacitance is embedded in the circuit structure. Consequently, the impedance between the intrinsic capacitance and the switching load capacitance is small. The intentional decoupling capacitors augment the intrinsic capacitance of the circuit to reach the required level of capacitance. The intentional capacitance, however, is typically added at the final stages of the circuit design process and is often physically located at a significant distance from the switching load. As the switching time of the load decreases, the impedance of the power interconnect becomes increasingly important. The significance of the power line impedance on the efficacy of the decoupling capacitors is demonstrated in the following example.

Consider an integrated circuit manufactured in a sub-100 nm CMOS technology. A high power local circuit macro, 200 $\times$ 200 μm in size, switches a 20 pF load

capacitance C_{load} within a 100 ps time period t_r. Assuming a 1 V power supply, the maximum power current of the circuit is estimated as

$$I_{\max} \approx \frac{C_{load} V_{\mathrm{dd}}}{t_r/2} = \frac{20\,\mathrm{pF} \times 1\,\mathrm{V}}{100\,\mathrm{ps}/2} = 400\,\mathrm{mA}, \tag{31.7}$$

and the maximum current transient as

$$\left(\frac{dI}{dt}\right)_{\max} \approx \frac{I_{\max}}{t_r/2} = \frac{400\,\mathrm{mA}}{100\,\mathrm{ps}/2} = 8 \times 10^9\,\frac{\mathrm{A}}{\mathrm{s}}. \tag{31.8}$$

The decoupling capacitance embedded within the circuit is assumed to be insufficient, supplying only half of the required current. The rest of the current is supplied by the nearest on-chip decoupling capacitor. To limit the resistive and inductive voltage drops to below 100 mV, 10 % of the power supply in this 1 V system, the resistance and inductance of the current path between the load circuit and the decoupling capacitor should be smaller than, respectively,

$$R_{\max} = \frac{0.1 V_{\mathrm{dd}}}{0.5 I_{\max}} = \frac{0.1\,\mathrm{V}}{0.2\,\mathrm{A}} = 0.5\,\Omega \tag{31.9}$$

and

$$L_{\max} = \frac{0.1 V_{\mathrm{dd}}}{0.5\,(dI/dt)_{\max}} = \frac{0.1\,\mathrm{V}}{4 \times 10^9\,\mathrm{A/s}} = 25\,\mathrm{pH}. \tag{31.10}$$

These impedance specifications are demanding. Assume that the physical distance between the load and the capacitor is 100 µm. Consider a scenario where the load and capacitor are connected by two global power and ground lines that are 50 µm wide, 1 µm thick, and are placed on a 100 µm pitch, as illustrated in Fig. 31.2a. The resistance of the current path is approximately 0.08 Ω, well below the limit set by (31.9). The inductance of the path, however, is approximately 80 pH, exceeding the limit set by (31.10).

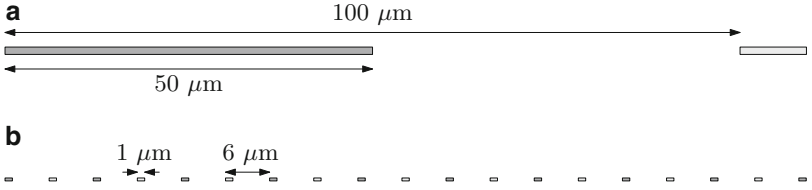


Fig. 31.2 Cross section of a current path connecting the load and decoupling capacitance. The power lines are shown in a *darker gray*, while the ground lines are shown in a *lighter gray*. The connection between the load and decoupling capacitance can be made using either (**a**) thick and wide global power lines or (**b**) finer local power lines. The dimensions are drawn to scale

Alternatively, if the load and decoupling capacitors are connected by fine lines of a local distribution network, for example, by 32 interdigitated power and ground lines with a $0.4 \times 1\,\mu$m cross section and a $6\,\mu$m line pitch, as illustrated in Fig. 31.2b, the inductance of the current path is reduced to 7 pH, well below the limit set by (31.10). The resistance in this case, however, is approximately $0.63\,\Omega$, exceeding the limit set by (31.9). The high current density in these fine lines is also likely to violate existing electromigration reliability constraints.

The target impedance characteristics are therefore more readily achieved if both wide and thick lines in the upper metal layers and fine lines in the lower metal layers are used. The superior impedance characteristics of such interconnect structures are described in Chap. 35. Alternatively, the width of the global power and ground lines in the upper metal layers should be greatly decreased. This approach will decrease the inductance of the grid, at the expense of a moderate increase in resistance, as discussed in Chap. 30.

Due to the limitations of the power interconnect, the on-chip decoupling capacitance should be placed in the immediate vicinity of the switching load in order to be effective. This requirement necessitates novel approaches to allocating the on-chip decoupling capacitance, as described in Chap. 12. A common design approach, where the bulk of the decoupling capacitance is placed among the circuit blocks after the initial design of the blocks has been completed, as shown in Fig. 11.18, does not permit placing the decoupling capacitors sufficiently close to the circuits far from the block boundary. As the feature size of the on-chip circuits decreases, the capacitance allocation process should be performed at a commensurately finer scale, as schematically illustrated in Fig. 31.3.

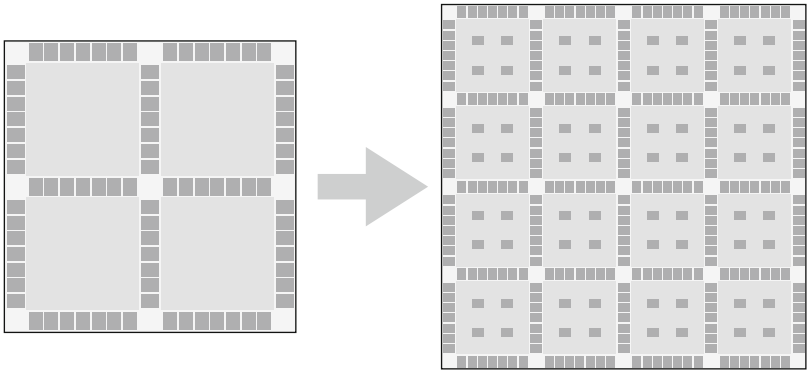


Fig. 31.3 The effect of circuit scaling on allocation of the on-chip decoupling capacitance. The circuit blocks are shown in *darker gray*, the decoupling capacitors are shown in *lighter gray*. As the circuit feature sizes decrease, the allocation of the on-chip decoupling capacitance should be performed at a commensurately finer scale. If the size of the circuit blocks does not decrease in proportion to the feature size, the decoupling capacitors are placed within the circuit blocks, as shown on the *right*

The high frequency characteristics of the on-chip power interconnect are particularly important when the power load is non-uniformly distributed. Those circuits with the greatest peak power consumption require a significant decoupling capacitance in close proximity while the surrounding lower power circuits tend to require a relatively low decoupling capacitance.

The power consumption is particularly non-uniform in high performance digital circuits with a highly irregular structure, such as microprocessors, which are comprised of both low and high power circuit blocks. More than half of the die area in state-of-the-art microprocessors is occupied by memory circuit structures, which are characterized by a low switching activity and, consequently, low power consumption. The power density of a microprocessor core can be an order of magnitude higher as compared to the memory arrays; the core and synchronization circuits dissipate the dominant share of the overall circuit power. The distribution of the power consumption within the high power blocks is also typically non-uniform. The peak power demand circuitry with high load capacitances and switching activities, such as the arithmetic units and bus drivers, can exceed severalfold the worst case requirements of the surrounding circuits.

Contemporary trends in circuit design exacerbate the uneven distribution of the dissipated power. Similar to microprocessors, system-on-chip circuits integrate diverse circuit structures and also tend to exhibit a highly non-uniform power distribution pattern. Since the power consumption of integrated circuits has become a primary design priority, as described in Chap. 1, aggressive power saving techniques have become mandatory. Clock and power gating have gained wider use in order to decrease dynamic and leakage power consumption, respectively, in idle circuit blocks [308, 519–531]. While the power consumed by the circuit blocks is greatly reduced in the gated mode, abrupt transients in the power current are induced when a circuit block transitions from a power saving mode to active operation or vice versa. Power gating presents particular challenges to the analysis and verification of power distribution networks. A significant share of the decoupling capacitance is often disconnected from the global network during a power-down mode. This change in the decoupling capacitance can potentially cause power integrity problems in the surrounding circuits.

31.4 Summary

The effect of the decoupling capacitance and the inductance of on-chip interconnect on the high frequency impedance characteristics of a power distribution system is discussed in this chapter. The primary conclusions are summarized as follows.

- The inductance of the on-chip interconnect becomes more significant as the inductance of the package conductors is reduced
- The power noise propagates through the on-chip power distribution network at a relatively low velocity

- The response of the on-chip power distribution network is underdamped in close proximity to the load
- The impedance of the current path between the on-chip load and the on-chip decoupling capacitors becomes a critical design parameter as the power supply and circuit switching times decrease
- Allocating the on-chip decoupling capacitance should be performed at a finer scale as the feature size of the on-chip circuits decreases

Chapter 32
Power Noise Reduction Techniques

Future generations of integrated circuit technologies are trending toward higher speeds and densities. The total capacitive load associated with the internal circuitry has been increasing for several generations of high complexity integrated circuits [149, 243]. As the operating frequencies increase, the average on-chip current required to charge and discharge these capacitances also has increased, while the switching time has decreased. As a result, a large change in the total on-chip current can occur within a brief period of time.

Due to the high slew rate of the currents flowing through the bonding wires, package pins, and on-chip interconnects, the ground and supply voltage can fluctuate (or bounce) due to the parasitic impedances associated with the package-to-chip and on-chip interconnects. These voltage fluctuations on the supply and ground rails, called ground bounce, ΔI noise, or simultaneous switching noise (SSN) [118], are larger since a significant number of the I/O drivers and internal logic circuitry switch close in time to the clock edges. SSN generates glitches on the ground and power supply wires, decreasing the effective current drive of the circuits, producing output signal distortion, thereby reducing the noise margins of a system. As a result, the performance and functionality of the system can be severely compromised.

In the past, research on SSN has concentrated on transient power noise caused by current flowing through the inductive bonding wires at the I/O buffers. SSN originating from the internal circuitry, however, has become an important issue in the design of nanoscale high performance ICs, such as systems-on-chip, mixed-signal circuits, and microprocessors. This increased importance is due to fast clock rates, large on-chip switching activities and currents, and increased on-chip inductance, all of which are increasingly common characteristics of nanoscale synchronous ICs.

Most of the work in this area falls into one of two categories: the first category includes analytic models that predict the behavior of the SSN, while the second category describes techniques to reduce ground bounce. A number of approaches have previously been described to analyze power and ground bounce and the effect

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0_32

of SSN on the performance of high complexity integrated circuits. Senthinathan et al. described an accurate technique for estimating the peak ground bounce noise by observing negative local feedback present in the current path of the driver [532]. This work suffers from the assumption that the switching currents of the output drivers are modeled as a triangular shape. In [533], Vaidyanath, Thoroddsen, and Prince relaxed this assumption by deriving an expression for the peak value of the ground bounce under the more realistic assumption that the ground bounce is a linear function of time during the output transition of the driver. Other research has considered short-channel effects in CMOS devices on the ground bounce waveform [534–536]. While most prior research has concentrated on the case where all of the drivers switch simultaneously, the authors in [535] consider the more realistic scenario when the drivers switch at different times. The idea of considering the effects of ground bounce on a tapered buffer has been presented in [537]. Tang and Friedman developed an analytic expression characterizing the on-chip SSN voltage based on a lumped *RLC* model characterizing the on-chip power supply rail rather than a single inductor to model a bonding wire [23]. In [538], Heydari and Pedram addressed ground bounce with no assumptions about the form of the switching current or noise voltage waveforms. The effect of ground bounce on the propagation delay and the optimum tapering factor of a multistage buffer is discussed. An analytic expression for the total propagation delay in the presence of ground bounce is also developed.

A number of techniques have been described to reduce SSN. In [539], a voltage controlled output buffer is described to control the slew rate. Ground bounce reduction is achieved by lowering the inductance in the power and ground paths by utilizing substrate conduction. An algorithm based on integer linear programming to skew the switching of the drivers to minimize ground bounce is presented in [540]. An architectural approach for reducing inductive noise caused by clock gating through gradual activation/deactivation units has been described in [541]. In [542], a routing method is described to distribute the ground bounce among the pads under a constraint of constant routing area. The total P/G noise of the system, however, is not reduced. Decoupling capacitors are often added to maintain the voltage on the P/G rails within specification, providing charge for the switching transients [538, 543]. Recently, several methods for reducing ground bounce have been suggested, such as bounce pre-generator circuits [544], supply current shaping, and clock frequency modulation [545].

Design techniques to reduce P/G noise in mixed-signal power distribution systems is the primary focus of this chapter. The efficiency of these techniques is based on the physical parameters of the system. This chapter is organized as follows. Ground noise reduction through the addition of a noise-free on-chip ground is described in Sect. 32.1. The efficiency of the technique as a function of the physical parameters of the system is evaluated in Sect. 32.2. Some specific conclusions are summarized in Sect. 32.3.

32.1 Ground Noise Reduction Through an Additional Low Noise On-Chip Ground

An equivalent circuit of an SoC-based power delivery system is shown in Fig. 32.1. Traditionally, noisy digital circuits share the power and ground supply with noise sensitive analog circuits (see Chap. 41). If a number of digital blocks switch simultaneously, the current I_D drawn from the power distribution network can be significant. This large current passes through the parasitic resistance R^p_{Gnd} and inductance L^p_{Gnd} of the package, producing voltage fluctuations on the ground

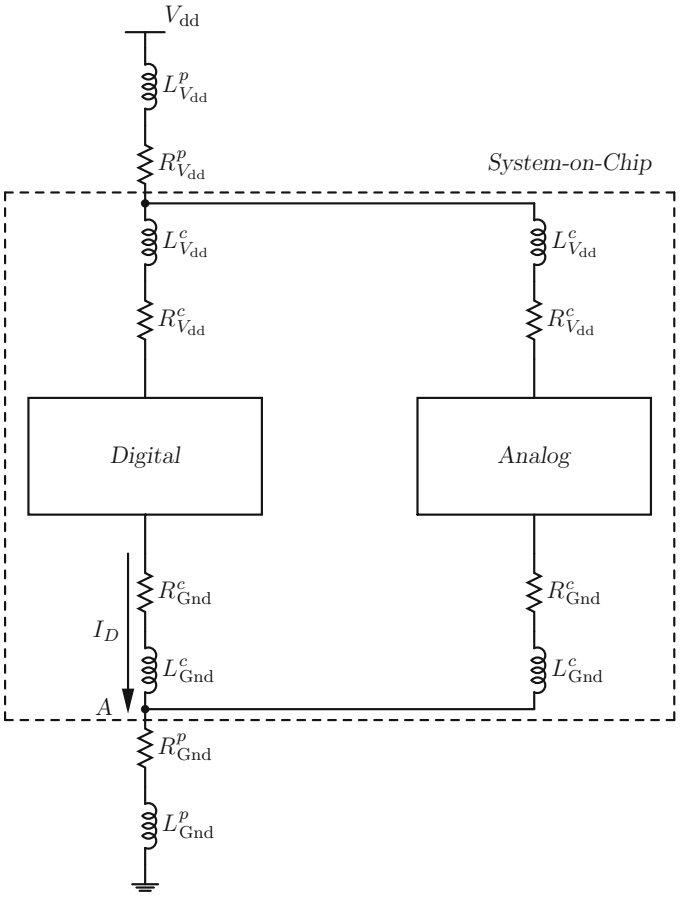


Fig. 32.1 An equivalent circuit for analyzing ground bounce in an SoC. The power distribution network is modeled as a series resistance and inductance. The superscripts p and c denote the parasitic impedance of, respectively, the package and on-chip power delivery systems. The subscript V_{dd} denotes the power supply voltage and the superscript G_{nd} denotes the ground

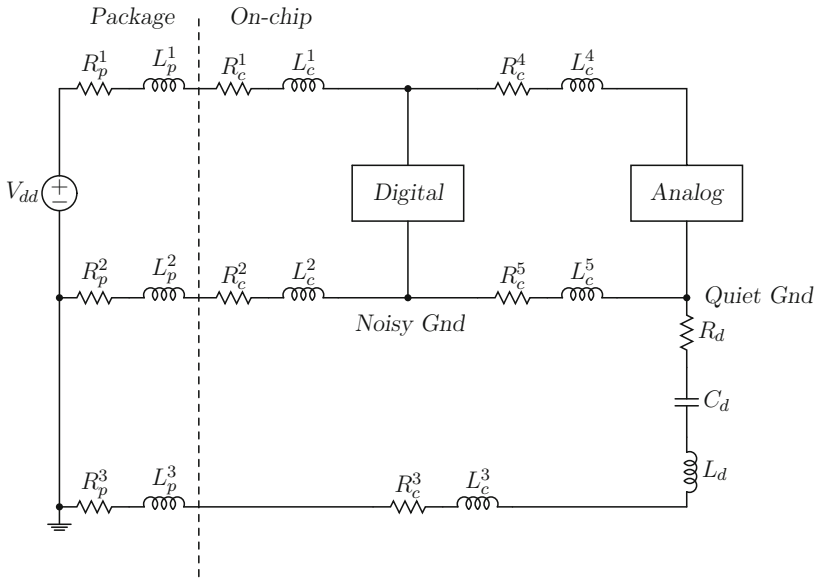


Fig. 32.2 Ground bounce reduction technique. The effective series resistance and effective series inductance of the decoupling capacitor are modeled by, respectively, R_d and L_d. R_c^5 and L_c^5 represent the physical separation between the noisy and noise sensitive blocks. The impedance of the additional on-chip ground is modeled by, respectively, R_c^3 and L_c^3

terminal (point A). As a result, ground bounce (or voltage fluctuations) appears at the ground terminal of the noise sensitive circuits.

To reduce voltage fluctuations at the ground terminal of the noise sensitive blocks, an on-chip low noise ground is added, as shown in Fig. 32.2. This approach utilizes a voltage divider formed by the impedance between the noisy ground terminal and the quiet ground terminal and the impedance of the path from the quiet ground terminal to the off-chip ground. The value of the capacitor is chosen to cancel the parasitic inductance of the additional low noise ground, i.e., the ESL of the capacitor L_d and the on-chip and package parasitic inductances of the dedicated low noise ground, respectively, L_c^3 and L_p^3. Alternatively, the capacitor is tuned in resonance with the parasitic inductances at a frequency that produces the greatest reduction in noise. The impedance of the additional ground path, therefore, behaves as a simple resistance.

The same technique can be used to reduce voltage fluctuations on the power supply. Based on the nature of the power supply noise, an additional ground path or power supply path can be provided. For instance, to ensure that the voltage does not drop below the power supply level, an on-chip path to the power supply is added. In the case of an overshoot, an additional ground path can be provided.

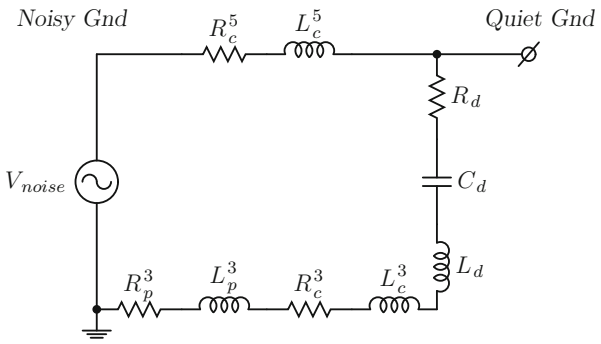


Fig. 32.3 Simplified circuit of the ground bounce reduction technique. The ground bounce due to simultaneously switching the digital circuits is modeled by a voltage source. The *Noisy Gnd* denotes an on-chip ground for the simultaneously switching digital circuits. The *Quiet Gnd* denotes a low noise ground for the noise sensitive circuits

32.2 Dependence of Ground Bounce Reduction on System Parameters

To determine the efficiency in reducing ground bounce, a simplified circuit model of the technique is used, as shown in Fig. 32.3. The ground bounce caused by simultaneously switching within the digital circuitry is modeled as a voltage source. A sinusoidal voltage source with an amplitude of 100 mV is used to determine the reduction in ground bounce at a single frequency. A triangular voltage source with an amplitude of 100 mV, rise time of 50 ps, and fall time of 200 ps is utilized to estimate the reduction in ground noise.

The dependence of the noise reduction technique on the physical separation between the noisy and noise sensitive circuits is presented in Sect. 32.2.1. The sensitivity of this technique to frequency and capacitance variations is discussed in Sect. 32.2.2. The dependence of ground noise on the impedance of an additional on-chip ground path is analyzed in Sect. 32.2.3.

32.2.1 Physical Separation Between Noisy and Noise Sensitive Circuits

To determine the dependence of the noise reduction technique on the physical separation between the noise source and noise receiver, the impedance of the ground path between the noisy and quiet terminals is modeled as a series *RL*, composed of the parasitic resistance and inductance per unit length. The peak voltage at the quiet ground is evaluated using SPICE where the distance between the digital and analog

Table 32.1 Ground bounce reduction as a function of the separation between the noisy and noise sensitive circuits

R_c^5	L_c^5	V_{quiet} (mV)		Noise reduction (%)	
(mΩ)	(fH)	Sinusoidal	Triangular	Sinusoidal	Triangular
13	7	90.81	97.11	9.2	2.9
26	14	82.99	94.68	17.0	5.3
39	21	76.30	92.63	23.7	7.4
52	28	70.54	90.55	29.5	9.5
65	35	65.53	89.36	34.5	10.6
78	42	61.16	88.06	38.8	11.9
91	49	57.33	86.93	42.7	13.1
104	56	53.94	85.93	46.1	14.1
117	63	50.91	85.05	49.1	15.0
130	70	48.23	84.28	51.8	15.7

$V_{noise} = 100\,\text{mV}, f = 1\,\text{GHz}, R_p^3 = 10\,\text{m}\Omega,$
$L_p^3 = 100\,\text{pH}, R_c^3 = 100\,\text{m}\Omega, L_c^3 = 100\,\text{fH}, R_d = 10\,\text{m}\Omega,$
$L_d = 10\,\text{fH}, C_d^{\text{Sin}} = 253\,\text{pF}, C_d^{\text{Triang}} = 63\,\text{pF}$

circuits is varied from one to ten unit lengths. The reduction in ground bounce as seen from the ground terminal of the noise sensitive circuit for sinusoidal and triangular noise sources is listed in Table 32.1.

Note that the reduction in ground noise increases linearly as the physical separation between the noisy and noise sensitive circuits becomes greater. A reduction in ground bounce of about 52 % for a single frequency noise source and about 16 % for a random noise source is achieved for a ground line (of ten unit lengths) between the digital and analog blocks. Enhanced results can be achieved if the impedance of the additional ground is much smaller than the impedance of the interconnect between the noisy and noise sensitive modules. From a circuits perspective, the digital and analog circuits should be placed sufficiently distant and the additional low noise ground should be composed of multiple parallel lines. Moreover, the additional ground should be placed close to the multiple ground pins.

Note that since this noise reduction technique utilizes a capacitor tuned in resonance with the parasitic inductance of an additional ground path, this approach is frequency dependent and produces the best results for a single frequency noise source. In the case of a random noise source, the frequency harmonic with the highest magnitude should be significantly reduced, thereby achieving the greatest reduction in noise. For example, the second harmonic is selected in the case of a triangular noise source.

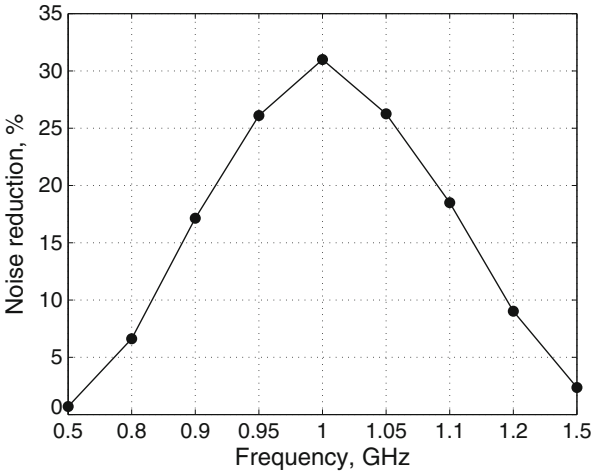


Fig. 32.4 Ground bounce reduction as a function of noise frequency. The reduction in noise drops linearly as the frequency varies from the target resonant frequency. The ground noise is modeled as a sinusoidal voltage source

32.2.2 Frequency and Capacitance Variations

To determine the sensitivity of the ground bounce reduction technique on frequency and capacitance variations, the frequency is varied by $\pm 50\,\%$ from the resonant frequency and the capacitor is varied by $\pm 10\,\%$ from the target value. The range of capacitance variation is chosen based on typical process variations for a CMOS technology. The efficiency of the reduction in ground bounce for a sinusoidal noise source versus frequency and capacitance variations is illustrated, respectively, in Figs. 32.4 and 32.5.

Note that the noise reduction drops linearly as the noise frequency varies from the target resonant frequency. The reduction in noise is slightly greater for higher frequencies. This phenomenon is due to the uncompensated parasitic inductance of the ground connecting the digital circuits to the analog circuits. As a result, at higher frequencies, the impedance of the ground path of a power delivery network increases, further reducing the noise. In general, the technique results in lower noise at higher frequencies. As illustrated in Fig. 32.5, the reduction in ground bounce is almost insensitive to capacitance variations. The efficiency of the technique drops by about $4\,\%$ as the capacitance is varied by $\pm 10\,\%$.

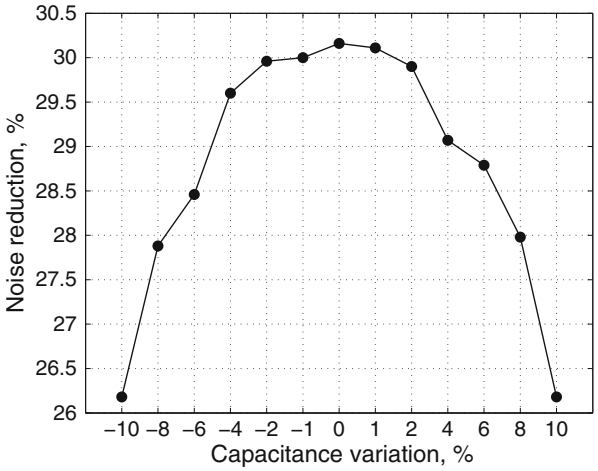


Fig. 32.5 Reduction in ground bounce as a function of capacitance variations. The reduction in ground bounce is almost insensitive to capacitance variations. The ground bounce is modeled as a sinusoidal voltage source

32.2.3 Impedance of an Additional Ground Path

As described in Sect. 32.1, the noise reduction technique utilizes a voltage divider formed by the ground of an on-chip power distribution system and an additional low noise ground. To increase the efficiency of the technique, the voltage transfer function of the voltage divider should be lowered, permitting a greater portion of the noise voltage to be diverted through the additional ground. As demonstrated in Sect. 32.2.1, placing the noisy and noise sensitive blocks farther from each other lowers the bounce at the ground terminal of the analog circuits. The ground noise can also be reduced by lowering the impedance of the low noise ground. The parasitic inductance of the additional ground is canceled by the capacitor tuned in resonance to the specific frequency. The impedance of the additional ground is therefore purely resistive at the resonant frequency. The reduction in noise for different values of the parasitic resistance of the low noise ground is listed in Table 32.2.

Note from Table 32.2 that by reducing the parasitic resistance of an on-chip low noise ground, the ground bounce can be significantly lowered. Noise reductions of about 68 % and 22 % are demonstrated for, respectively, sinusoidal and triangular noise sources. The results listed in Table 32.2 are determined for an average resistance and inductance of the on-chip power distribution ground of five unit lengths (see Table 32.1). Thus, the ground bounce can be further reduced if the analog and digital circuits are placed farther from each other. Even better results can be achieved if the parasitic resistance of the package pins R_p^3 and decoupling capacitor R_d are lowered. From a circuits perspective, the low noise on-chip ground

Table 32.2 Ground bounce reduction for different values of parasitic resistance of the on-chip low noise ground

R_c^3	V_{quiet} (mV)		Noise reduction (%)	
(mΩ)	Sinusoidal	Triangular	Sinusoidal	Triangular
100	60.54	87.88	39.5	12.1
80	56.52	86.57	43.5	13.4
60	51.67	84.98	48.3	15.0
40	45.79	83.03	54.2	17.0
20	38.59	80.60	61.4	19.4
10	34.37	79.15	65.6	20.9
5	32.08	78.37	67.9	21.6

$V_{noise} = 100\,\text{mV}, f = 1\,\text{GHz}, R_p^3 = 10\,\text{m}\Omega, L_p^3 = 100\,\text{pH}, L_c^3 = 100\,\text{fH}, R_c^5 = 80\,\text{m}\Omega, L_c^5 = 40\,\text{fH}, R_d = 10\,\text{m}\Omega, L_d = 10\,\text{fH}, C_d^{\text{Sin}} = 253\,\text{pF}, C_d^{\text{Triang}} = 63\,\text{pF}$

should be composed of many narrow lines connected in parallel to lower the parasitic resistance and inductance. A number of package pins should therefore be dedicated to the noise-free ground to lower the package resistance. A decoupling capacitor with a low ESR is also recommended.

32.3 Summary

Design techniques to reduce ground bounce in SoC and mixed-signal ICs are presented in this chapter and can be summarized as follows.

- A noise reduction technique with an additional on-chip ground is described to divert ground noise from the sensitive analog circuits
- The technique utilizes a decoupling capacitor tuned in resonance with the parasitic inductance of an additional low noise ground, making the technique frequency dependent
- The reduction in ground bounce, however, is almost independent of capacitance variations
- Noise reductions of 68 % and 22 % are demonstrated for, respectively, a single frequency and random ground noise
- The noise reduction efficiency can be further enhanced by simultaneously lowering the impedance of the additional noise-free ground and increasing the impedance of the ground path between the digital (noisy) and analog (noise sensitive) circuits

Chapter 33
Shielding Methodologies in the Presence of Power/Ground Noise

In highly scaled integrated circuits, crosstalk between adjacent interconnect has become a primary design issue. With aggressive technology scaling, the local interconnect has become more resistive and capacitive. The global interconnect has become more inductive. Capacitive and inductive coupling is therefore a significant design issue in global interconnects [546–548].

Shielding is widely used in integrated circuits to mitigate crosstalk between coupled lines. Two types of shielding methods have been developed, passive shielding [547–553] and active shielding [554–556]. In passive shielding, the power/ground lines are routed as shield lines between the critical interconnect to minimize the noise coupled from an aggressor to a victim line. Alternatively, active shielding [554–556] uses dedicated shield lines with switching signals rather than P/G lines. Although the performance of active shielding in reducing crosstalk noise voltage is superior to passive shielding, active shielding requires additional area and consumes greater power.

Power and ground networks are routed as shield lines in passive shielding to mitigate coupling noise. These P/G shield lines themselves can however be noisy. This noise, typically neglected in existing shielding methodologies, is due to inductive $L\ dI/dt$ noise and resistive IR voltage drops. With increasing device densities, the P/G noise voltage can be more than 20 % of the supply voltage [538, 557, 558]. Since the distance between the shield and victim lines is smaller than the distance between the aggressor and victim lines, the P/G noise on the shield line can produce more noise on the victim line than the crosstalk noise coupled from the aggressor to the victim. Hence, while a shield line reduces noise coupling from the aggressor interconnect, the shield line can also *increase* noise coupling due to P/G noise.

Although P/G noise has received significant attention in the design of robust power distribution networks [305, 538, 557, 558], existing works do not consider the deleterious effects of P/G noise on *shielding methodologies* [547–552, 554–556, 559]. P/G lines routed as shield lines have typically been treated as *ideal* ground or supply voltage connections, which do not accurately model the effects

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0_33

of noise on the shield line. Recently, noise on the P/G lines is mentioned in [553] without describing the effect of this noise on the victim line and related shielding methodologies. P/G noise on the shield lines is considered in this chapter to provide practical and more effective shielding methodologies.

An alternative method to reduce crosstalk is to increase the distance between the aggressor and victim lines without inserting a shield line. Tradeoffs between the two methods, shield insertion and physical spacing, are discussed in [549, 550] without considering P/G noise on the shield lines. P/G noise can however significantly affect the decision criteria between shielding and spacing, as discussed throughout this chapter. The primary objective here is to discuss the effects of P/G noise on shield lines within a passive shielding methodology. Comparisons between physical spacing and shield insertion techniques are provided. Boundary conditions are also identified to determine the efficacy regions of spacing and shield insertion. Once P/G noise is considered, spacing alone can be more useful than shield insertion under specific conditions, as described in this chapter. These results provide decision criteria in choosing between spacing or shielding in a noisy environment [560, 561].

The rest of the chapter is organized as follows. Background material is provided in Sect. 33.1. In Sect. 33.2, the effects of several technology and design parameters characterizing the interconnect and shield lines in terms of crosstalk noise on the victim line are discussed. In Sect. 33.3, a decision criterion for the critical interconnect length and width is provided to choose between shield insertion and physical spacing. The chapter is summarized in Sect. 33.4. Closed-form expressions for the interconnect resistance, capacitance, and inductance are provided in Appendix I.

33.1 Background

Background material is provided in this section for evaluating the effects of P/G noise on passive shielding methodologies. Specifically, an overview of crosstalk reduction techniques is provided in Sect. 33.1.1. An interconnect model and the design criterion used throughout this chapter are described in Sect. 33.1.2. The P/G noise model and the effects of this noise on crosstalk noise are described in Sect. 33.1.3.

33.1.1 Crosstalk Noise Reduction Techniques

Several techniques can be used to mitigate the effects of crosstalk noise in high complexity integrated circuits [546–552, 554–556, 559]. A brief overview of these techniques is provided in this section.

Increasing the physical distance between the aggressor and victim lines can reduce the coupling capacitance and mutual inductance between adjacent lines. The reduction in crosstalk capacitance is approximately inversely proportional with the

increase in spacing. The mutual inductance, however, is not significantly reduced with increasing distance since the mutual inductance is a long range phenomenon. To reduce the mutual inductance, additional return paths should be provided for the current to flow.

Inserting shield lines between the aggressor and victim lines reduces the capacitive and inductive coupling between adjacent blocks [547–552]. Shield insertion significantly reduces capacitive coupling between the aggressor and victim lines because capacitive coupling is a short range phenomenon and is significantly reduced in non-adjacent lines. Shield insertion moderately reduces the mutual inductance due to the current return path formed by the inserted shield line for both the aggressor and victim lines [552]. The difficulty in forcing the current return path complicates the inductive shielding process.

Active shielding is another shielding technique in which the shield line switches depending upon the switching pattern of the adjacent bus lines [554–556]. Capacitive (inductive) coupling is reduced with active shielding when the shield line is switched in the same (opposite) direction as the signal line [555]. The switching activity of the shield lines should therefore be tuned to the switching pattern which is different for *RC* dominated and *LC* dominated interconnect lines. The primary drawback of active shielding is increased power consumption and additional area of the logic circuitry controlling the active shield lines. Furthermore, process and environmental variations may unexpectedly affect the signal arrival times, degrading the efficiency of active shielding.

Sizing the buffer driving the aggressor and victim lines is another technique to reduce crosstalk noise [551]. The effective conductance of the driver increases with larger drivers. For the victim line, a larger driver can be used to maintain the victim line at a constant voltage by increasing the driver conductance. For the aggressor line, using a smaller driver decreases the crosstalk noise since the signal transition is slower due to the increased *RC* time constant, decreasing the induced noise on the victim line [551]. Proper sizing of the driver on the aggressor and victim lines can therefore produce lower crosstalk noise. This technique is however subject to delay constraints since a smaller driver increases the gate delay. Wire sizing can also be used to modify the line resistance, coupling capacitance, line-to-substrate capacitance, and self-inductance [562].

Repeater insertion is used to reduce the length of the long interconnect, decreasing the line resistance, and coupling capacitance and mutual inductance between lines [563]. Since the length of the switching portions of the adjacent lines decreases with additional inserted repeaters, the crosstalk noise on the victim line is reduced. The switching portions of the adjacent lines can be further reduced by interleaved repeater insertion [564]. Repeaters, however, consume power and area. Additionally, the jitter induced from each repeater can degrade the performance of certain sensitive signals such as the clock.

The primary focus of this chapter is to investigate passive shielding methodologies in the presence of P/G noise. Design guidelines are provided for choosing between spacing and shield insertion to enhance signal integrity under different conditions, as described in the following sections.

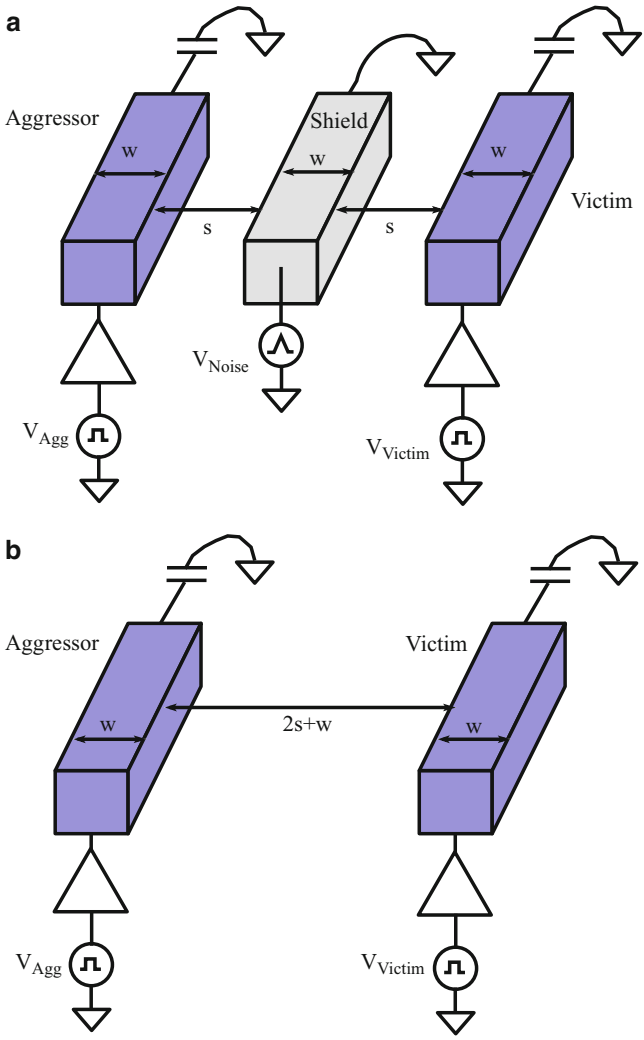


Fig. 33.1 Global interconnect model for (**a**) shield line between an aggressor and victim line, and (**b**) physical spacing between an aggressor and victim line. The aggressor and victim lines are modeled with a driver resistance at the near end and terminated with a load capacitance at the far end. P/G noise is modeled as a single voltage source at the near end of the shield line

33.1.2 Coupled Interconnect Model and Decision Criterion

A typical interconnect model with a shield line inserted between the aggressor and victim lines is depicted in Fig. 33.1a [549, 550]. The noise on the shield line is modeled as a single voltage source at the near end. The interconnect model used for physical spacing is depicted in Fig. 33.1b.

The objective is to compare the effect of inserting a shield line and physical spacing on the coupling noise at the far end of a victim line (sense node). The ratio K of the coupling noise at the sense node when only a shield line is present, $V_{sense\_with\_shielding}$, to the coupling noise when only physical spacing is used, $V_{sense\_with\_spacing}$, is the decision criterion used to determine the boundary conditions,

$$K = \frac{V_{sense\_with\_shielding}}{V_{sense\_with\_spacing}}. \tag{33.1}$$

If $K < 1$, inserting a shield line between the aggressor and victim lines is preferable because the crosstalk noise at the sense node is smaller with a shield than with additional spacing. Alternatively, if $K > 1$, increasing the spacing is a more effective technique. $K = 1$ is therefore treated as a design threshold. Spacing is more efficient above the threshold while shield insertion is more efficient below the threshold. Note that the area is maintained the same for both shield insertion and physical spacing to provide a fair comparison. The distance between the aggressor and victim lines is the same for both shield insertion and physical spacing, as depicted in Fig. 33.1. For instance, when the width of the shield line increases by Δw, the distance between the aggressor and victim lines increases by Δw/2 to maintain unaltered the distance between the shield line and the aggressor and victim lines. When comparing the effectiveness of shield insertion with physical spacing for a specific example, the distance between the aggressor and victim lines is increased by Δw to satisfy the same area constraints for both the shielding and spacing methods. Alternatively, when the distance between the aggressor and victim lines is increased using the spacing method, the distance between the shield line and the aggressor and victim lines is also increased with the shield insertion method to maintain the same area constraints.

To accurately evaluate the effects of inductive and capacitive coupling, the 2π *RLC* interconnect model [549] shown in Fig. 33.2 is used. The aggressor and victim

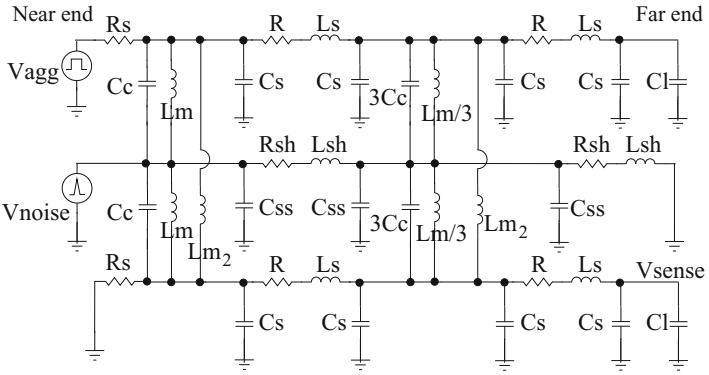


Fig. 33.2 2π *RLC* interconnect model with coupling capacitances and mutual inductances

Table 33.1 Interconnect parameters for 65 nm [405], 45 nm [567], and 32 nm [568] technology nodes

	W (μm)	S (μm)	T (μm)	H (μm)	ρ (10^{-8}) Ωm
65 nm	0.45	0.45	1.2	0.2	2.2
45 nm	0.40	0.40	0.72	0.2	2.2
32 nm	0.30	0.30	0.504	0.2	2.2

line parameters, R_s, R, C_s, C_c, C_l, and L_s, represent, respectively, the interconnect driver resistance, line resistance, line-to-substrate capacitance, coupling capacitance, load capacitance, and self-inductance. Additional parameters, R_{sh}, L_{sh}, C_{ss}, L_m, and L_{m2}, represent, respectively, the shield resistance, shield self-inductance, shield line-to-substrate capacitance, mutual inductance between the shield line and the aggressor and victim lines, and mutual inductance between the aggressor and victim lines. These circuit parameters have been extracted using the IBM Electromagnetic Field Solver Suite Tools (EIP) [565] for the 32, 45, and 65 nm technology nodes [405, 566–568] for the parameters listed in Table 33.1. The operating frequency is 1 GHz with 100 ps rise and fall transition times. The supply voltage is 1, 0.95, and 0.9 V for, respectively, the 65, 45, and 32 nm technology nodes.

33.1.3 Power/Ground Noise Model

P/G noise has become an important issue in the design of power distribution networks with technology scaling [23, 299, 538, 557, 558]. The effect of P/G noise on the uncertainty of the data signal delay, clock jitter, noise margin, and gate oxide reliability has been well studied [299]. The effect of noise coupling from the power and ground lines used for shielding sensitive data and clock lines, however, has not received significant attention. In this section, the detrimental effects of P/G noise on the shield insertion method is discussed.

To exemplify the detrimental effects of P/G noise on shield insertion, a representative noisy ground network is considered, as illustrated in Fig. 33.3. The power and ground networks are modeled as an inductive-resistive (*RL*) mesh structure. The active devices are modeled as current sources, and the corresponding current profile is modeled as a triangular waveform. Multiple ground connections and active devices are included to more accurately model the ground distribution network. 65 nm technology parameters are assumed.

Due to the resistive and inductive nature of the P/G distribution networks, *IR* and *L dI/dt* voltage drops degrade the signal integrity. The noise at a particular node strongly depends upon the distance among that node and the location of the ground connections and active devices. The maximum noise of the ground distribution network is maintained below 10 % of the supply voltage (i.e., the maximum ground

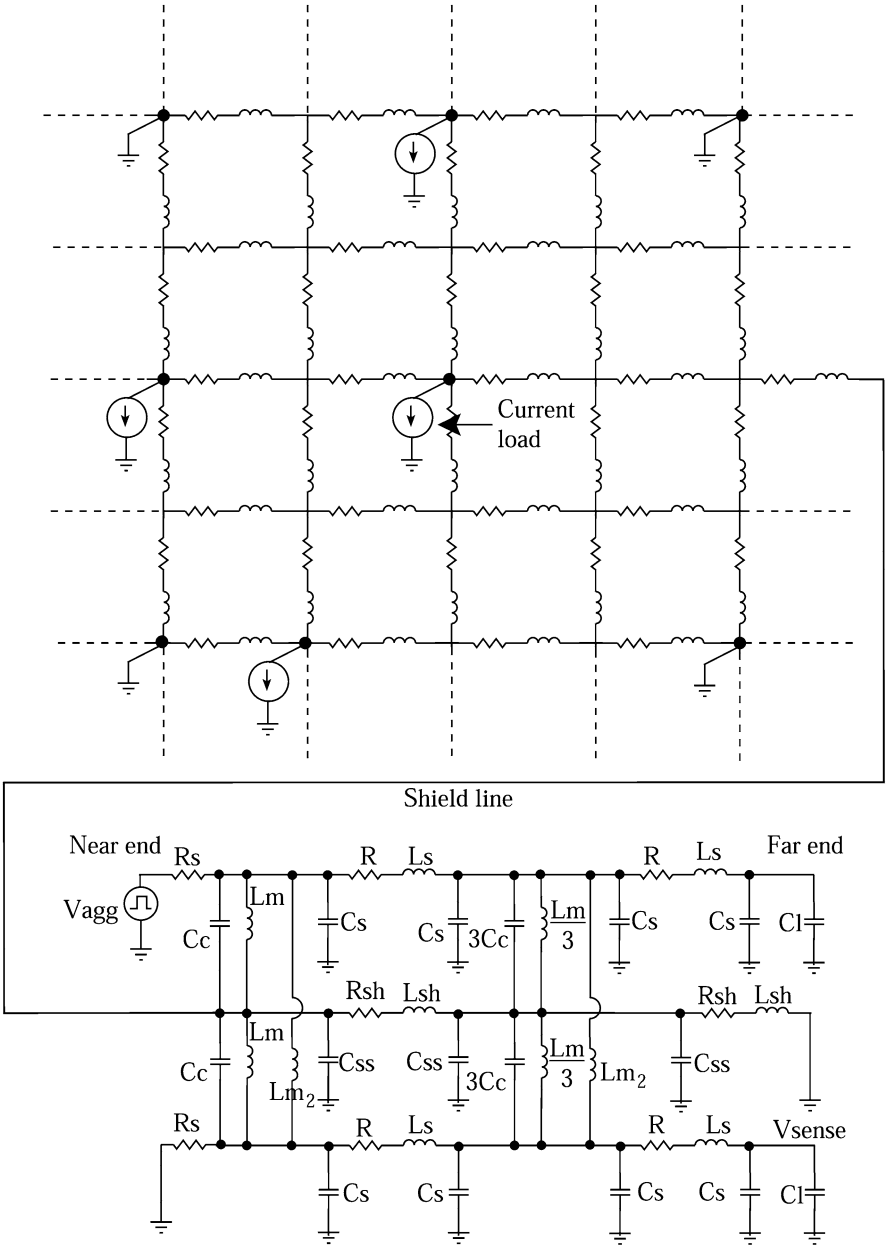


Fig. 33.3 The ground distribution network used as shield lines to evaluate the effect of P/G noise on crosstalk noise at the sense node for passive shielding. The ground distribution network consists of multiple ground connections, and the current loads are modeled as active devices connected to the ground network

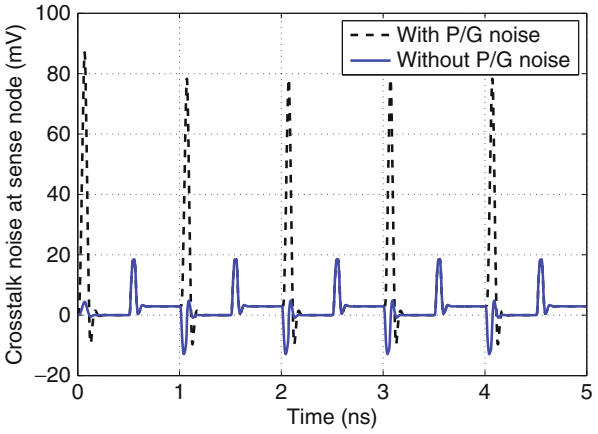


Fig. 33.4 Crosstalk noise at the sense node with a noisy shield line and a noise free shield line. Note that the crosstalk noise increases dramatically when P/G noise is present on the shield line

noise is less than 100 mV since, in this case, V_{DD} is 1 V). An arbitrary ground line is used as a shield. The crosstalk noise at the sense node is analyzed assuming a noisy and noise free shield line. The crosstalk noise is approximately five times larger when the shield line is noisy as compared to a noise free shield line, as illustrated in Fig. 33.4. Note that the detrimental effect of the P/G noise is significant for a system even when the ground noise is less than 10 % of the supply voltage. With continuous scaling of the supply voltage with each technology generation, the relative magnitude of the P/G noise to the supply voltage makes the victim lines increasingly sensitive to noise on the shield line.

33.2 Effects of Technology and Design Parameters on the Crosstalk Noise Voltage

Interconnect capacitance, inductance, and resistance increase with the length of the interconnect. The substrate and coupling capacitances increase and the self-inductance slightly decreases for wider interconnect. The coupling capacitance increases and the self-inductance slightly decreases for thicker interconnects. When the distance between adjacent interconnects increases, the coupling capacitance and mutual inductance decrease and the substrate capacitance increases. These trends are listed in Table 33.2.

The effects of technology scaling on the crosstalk noise voltage and the shield insertion process are discussed in Sect. 33.2.1. The effects of the interconnect line length and shield line width on the crosstalk noise are discussed, respectively, in Sects. 33.2.2 and 33.2.3. In Sect. 33.2.4, the effects of the ratio of the interconnect line resistance R_{line} to the interconnect driver resistance R_s on the coupling noise

Table 33.2 Effect of technology and design parameters on the resistance, capacitance, and inductance of the interconnect. Double arrows illustrate a significant change, single arrows illustrate a minor change, and ∼ illustrates no (or minimal) change

	$L \Uparrow$	$W \Uparrow$	$S \Uparrow$	$T \Uparrow$
R	⇑	⇓	∼	⇓
C_s	⇑	⇑	⇑	↑
C_c	⇑	⇑	⇓	⇑
L_s	⇑	⇓	∼	⇓
L_m	⇑	∼	⇓	∼

voltage are explored. The effects of the ratio of the line-to-substrate capacitance C_s to the coupling capacitance C_c on the coupling noise are discussed in Sect. 33.2.5. The effects of the interconnect self- and mutual inductance on crosstalk noise are reviewed in Sect. 33.2.6.

33.2.1 Effect of Technology Scaling on the Crosstalk Noise Voltage

The interconnect line parameters change with each technology generation, as listed in Table 33.1. In more advanced technologies, the interconnect is more resistive and the coupling between neighboring lines increases due to higher interconnect densities. A threefold challenge with technology scaling exists in terms of reducing crosstalk noise with shield insertion. First, the P/G network becomes more resistive due to interconnect scaling, increasing the *IR* voltage drop. The larger *IR* voltage drop increases the P/G noise on the shield line. Second, supply voltages scale with technology. P/G noise, however, does not scale significantly with technology, increasing the effects of P/G noise on circuit performance. Lastly, since the distance between adjacent interconnects also scales, the coupling capacitance and mutual inductance between the interconnect lines increase.

The crosstalk noise voltage is evaluated for different driver resistances, as illustrated in Fig. 33.5. When the P/G noise on the shield line is below 2–7 % of the supply voltage, a higher driver resistance is preferable to minimize the coupling noise at the sense node. When the P/G noise is greater than 2–7 % of the supply voltage, a lower driver resistance is preferable to minimize the crosstalk noise. Alternatively, when the P/G noise is greater than 7 % of the supply voltage, P/G noise is dominant whereas when the P/G noise is lower than 2 % of the supply voltage, the dominant noise source is the noise coupled from the aggressor line.

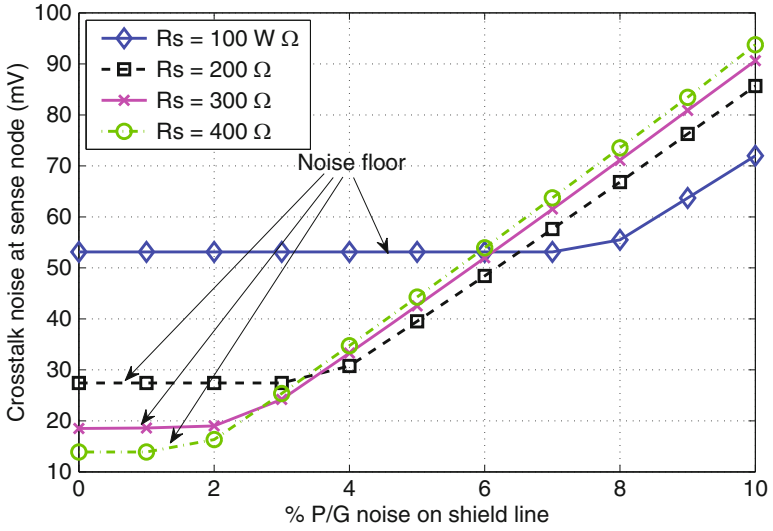


Fig. 33.5 Crosstalk noise at the sense node as the P/G noise is varied from 0 % to 10 % of the supply voltage for different driver resistances. Note that a noise floor exist for each driver resistance. This noise floor is due to the noise coupled from the aggressor line to the victim line when P/G noise is less than 7 % of the supply voltage with a small driver (i.e., driver resistance is 400 Ω) and less than 2 % with a large driver (i.e., driver resistance is 100 Ω)

The effect of the magnitude of the P/G noise on the crosstalk noise for different technology nodes is illustrated in Fig. 33.6. As expected, crosstalk noise is greater in more advanced technologies. Note that the noise floor when the P/G noise is below 3 % of the supply voltage is due to the noise coupled from the aggressor.

33.2.2 Effect of Line Length on Crosstalk Noise

The length of the global interconnect typically increases with technology scaling, causing greater signal noise [559, 563, 569]. The global interconnect can be longer than 4 mm [559, 563, 569]. Repeater insertion reduces the crosstalk noise and delay of the long interconnect. Inserting repeaters along the wide and thick global interconnects, however, can cause wire and via congestion as well as dissipate significant power [563]. The wire resistance, substrate capacitance, self-inductance of a wire, coupling capacitance, and mutual inductance between neighboring wires increase with longer line length.

For the interconnect model shown in Fig. 33.2, the coupling noise voltage at the sense node is compared to shield insertion and physical spacing for different interconnect lengths and driver resistances. These results are illustrated in Fig. 33.7, where $K = 1$ is the threshold (the same noise at the sense node occurs for both physical spacing and shield insertion).

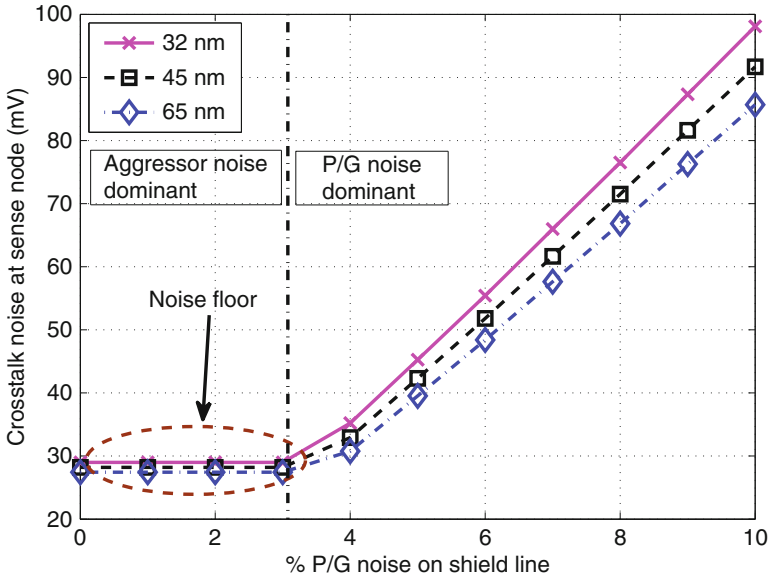


Fig. 33.6 Crosstalk noise at the sense node for several technology nodes when the P/G noise is varied from 0 % to 10 % of the supply voltage. The effects of P/G noise on the crosstalk noise increase with each technology generation. The noise floor is due to noise coupling from the aggressor to the victim. The P/G noise is dominant when the P/G noise is greater than 3 % of the supply voltage. Alternatively, the noise coupled from the aggressor is dominant when the P/G noise is less than 3 % of the supply voltage

At the 65 nm technology node, the peak value of K occurs at an interconnect length of 1.4 mm. K monotonically increases for interconnect lines shorter than 1.4 mm and monotonically decreases for interconnect lines longer than 1.4 mm. The crosstalk noise occurring at the sense node with physical spacing and shield insertion is shown, respectively, in Fig. 33.8a, b. The crosstalk noise at the sense node with physical spacing monotonically decreases with longer interconnect length. The crosstalk noise with shield insertion, however, exhibits a non-monotonic behavior since for a short interconnect line, the coupling capacitance and mutual inductance between adjacent lines dominate the line resistance. The crosstalk noise at the sense node, as shown in Fig. 33.8b, begins to decrease once the distance between the near and far end of the interconnect line is longer than the length where the effect of the line resistance dominates the effect of the coupling capacitance and mutual inductance (i.e., 1.4 mm for a 65 nm technology). Also note in Fig. 33.8 that inserting a shield line mitigates the effect of the driver resistance on the crosstalk noise, as discussed in Sect. 33.2.4. As a result, shield insertion is preferable for shorter lines and spacing is preferable for longer lines.

The effect of interconnect length is considered for different technology nodes. The critical interconnect length is determined for different driver resistances, as listed in Table 33.3. With each technology generation, the width and thickness of the

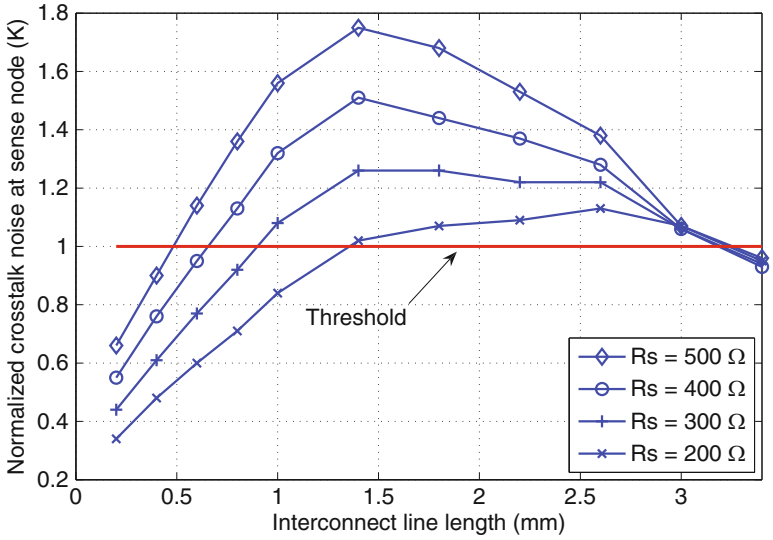


Fig. 33.7 Effect of interconnect length on crosstalk noise at the sense node for several driver sizes

interconnect scale with the minimum feature size. Since the line resistance increases with each technology generation, larger drivers (e.g., drivers with lower resistance) should be used to drive long victim lines. As listed in Table 33.3, shield insertion is more effective when both the aggressor and victim lines are driven by a large driver.

33.2.3 Effect of Shield Line Width on Crosstalk Noise

The effect of the cross-sectional area of the shield line on the coupling noise is discussed in this subsection. As the lines become more narrow and thin, the line resistance increases and the self-inductance decreases, making the lines more resistive. The coupling capacitance and mutual inductance between the shield line and the adjacent interconnect do not change significantly. To determine the effect of the cross-sectional area of the shield line on the crosstalk noise, the width of the shield line is evaluated for several driver resistances and interconnect lengths. A comparison of shield insertion and physical spacing is illustrated in Fig. 33.9 for a 1 mm long interconnect. Note that the distance between the aggressor and victim lines remains the same for both the physical spacing and shield insertion methods.

As the shield line width increases, shield insertion becomes less effective. Although increasing the width lowers the coupling from the aggressor to the sense node, P/G noise coupling to the sense node increases due to the lower resistance of the shield line and the higher mutual inductance. The P/G noise on the shield line propagates from the near end to the far end with less attenuation.

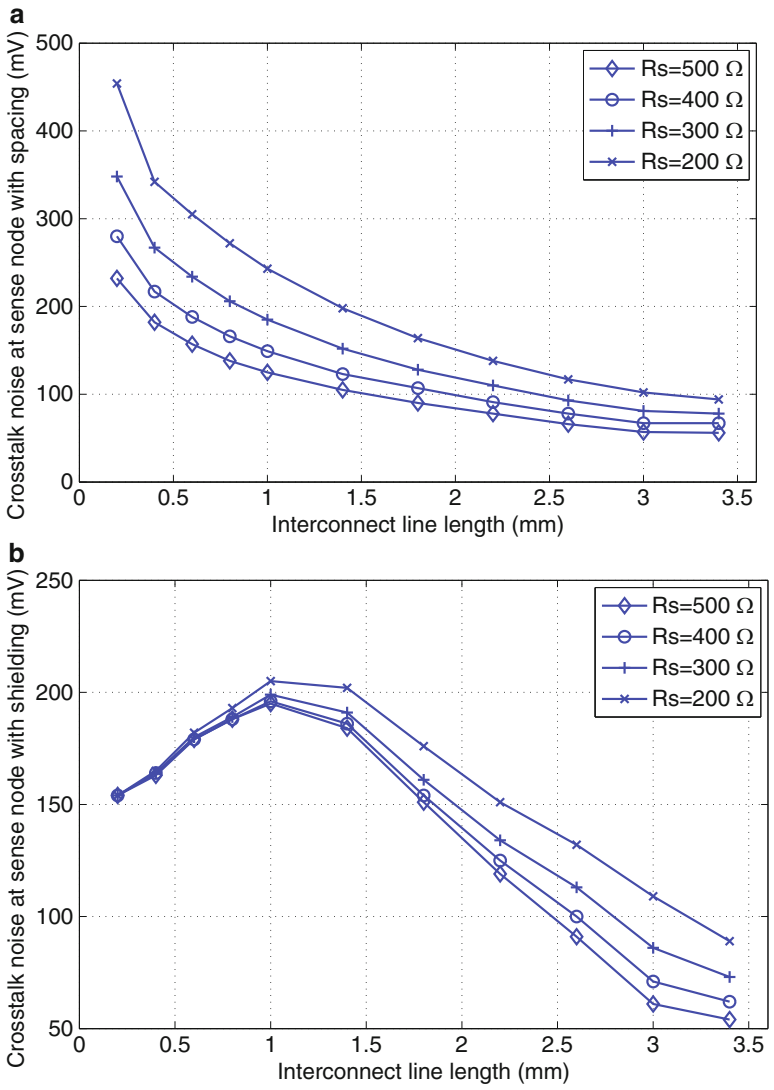


Fig. 33.8 Crosstalk noise occurring at the sense node for (**a**) physical spacing, and (**b**) shield insertion. Note that the behavior of the crosstalk noise with shield insertion is non-monotonic with increasing length

33.2.4 Effect of R_{line}/R_s on Crosstalk Noise

The driver resistance has a substantial effect on the behavior of global interconnects [570–572]. The driver resistance is less affected with technology scaling [114] because the oxide capacitance (C_{ox}) increases and the overdrive voltage ($V_{gs} - V_{th}$)

Table 33.3 Critical line length and driver resistance for several advanced technology nodes. Below the critical line length, shield insertion is preferable. Physical spacing is preferable for those interconnect lines longer than the critical line length

	65 nm				45 nm				32 nm			
Driver resistance (ohm)	100	200	300	400	100	200	300	400	100	200	300	400
Critical line length (mm)	1.4	1.3	1	0.8	1.3	1.1	0.9	0.8	1.3	0.8	0.2	0.1

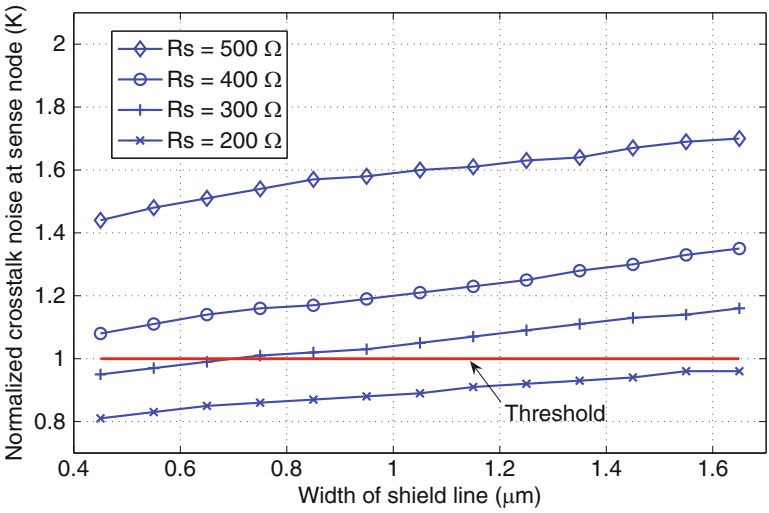


Fig. 33.9 Effect of shield line width on crosstalk noise for a 1 mm interconnect line. Note that signal integrity with shield insertion is degraded above the threshold $K = 1$

is lower with technology scaling. The line resistance, however, is a strong function of technology, increasing with each technology generation. The ratio of the line resistance to the driver resistance (R_{line}/R_s) therefore increases with each technology generation.

The effect of R_{line}/R_s on the crosstalk noise voltage is shown in Fig. 33.10 for several interconnect line lengths (for the 65 nm technology node). As mentioned previously, with increasing driver resistance, physical spacing becomes more efficient than shield insertion since coupling from the shield line is greater than coupling from the aggressor. The shield line exhibits no driver resistance so the P/G noise propagates to the sense node through the shield line whereas the aggressor noise voltage is attenuated by the large driver resistance at the near end of the aggressor. Alternatively, when the driver resistance is small, coupling from the aggressor dominates the P/G noise, making shield insertion preferable. Another observation is that the length of the interconnect significantly affects the speed, power, and area characteristics when choosing between spacing and shielding

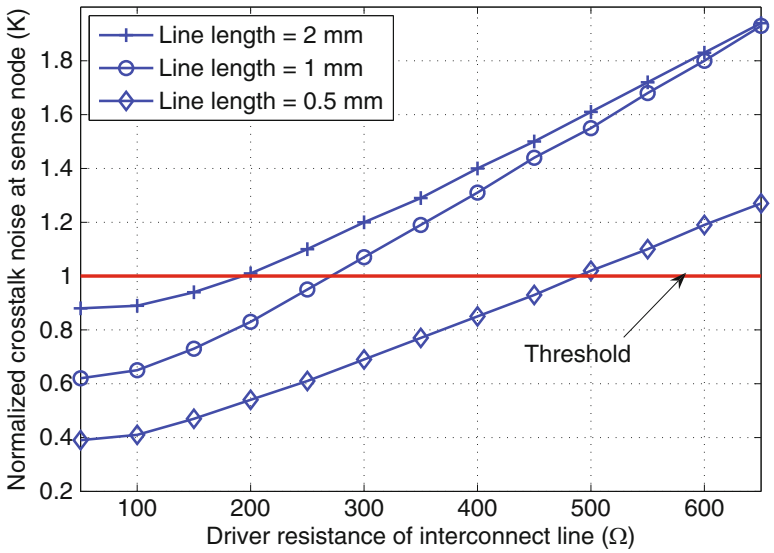


Fig. 33.10 Effect of R_{line}/R_s on the crosstalk noise voltage. The length of the interconnect line is 0.5 , 1 , and 2 mm

methodologies in a noisy environment. Spacing is preferable when the interconnect is longer whereas shielding is preferable for shorter interconnect lines, as shown in Fig. 33.10. Additionally, the R_{line}/R_s ratio increases in more advanced technologies. The crosstalk noise voltage is therefore more sensitive to P/G noise on the shield line. Either the driver resistance or the line width should be reduced in more advanced technologies.

33.2.5 Effect of the Ratio of Substrate Capacitance to Coupling Capacitance on Crosstalk Noise

The coupling capacitance between adjacent interconnect strongly depends upon the switching activity of the wires [573]. When the signals driving the adjacent lines switch in the same direction, the coupling capacitance is the same as the coupling capacitance between two adjacent quiet lines. When the signals driving the adjacent lines switch in the opposite direction, the coupling capacitance between the adjacent lines is two times the capacitance when only one of the adjacent lines is switching [573, 574].

The effect of the ratio of the line-to-substrate capacitance to the coupling capacitance has been evaluated for active and passive shielding structures [556], but without considering P/G noise on the shield lines. The effect of this ratio on the crosstalk noise at the sense node for different driver resistances is depicted in Figs. 33.11 and 33.12 for, respectively, interconnect line lengths of 0.5 and 1 mm. When the

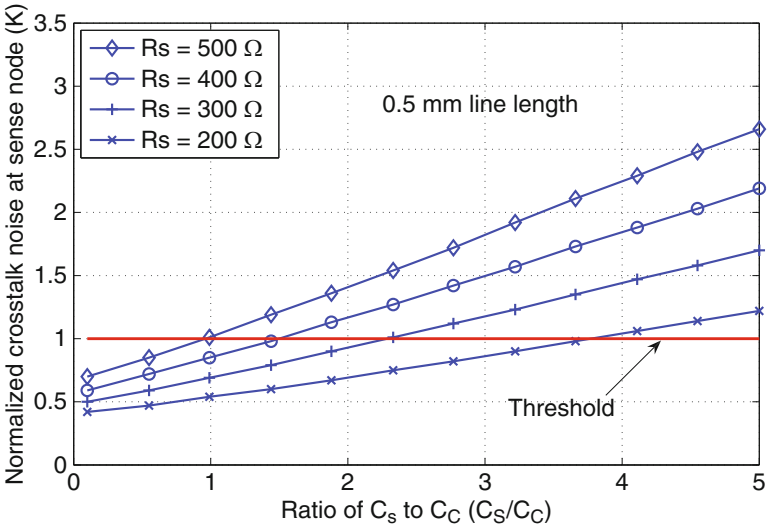


Fig. 33.11 Ratio of substrate capacitance to coupling capacitance versus normalized crosstalk noise when a P/G line is routed as a shield line. The interconnect length is 0.5 mm

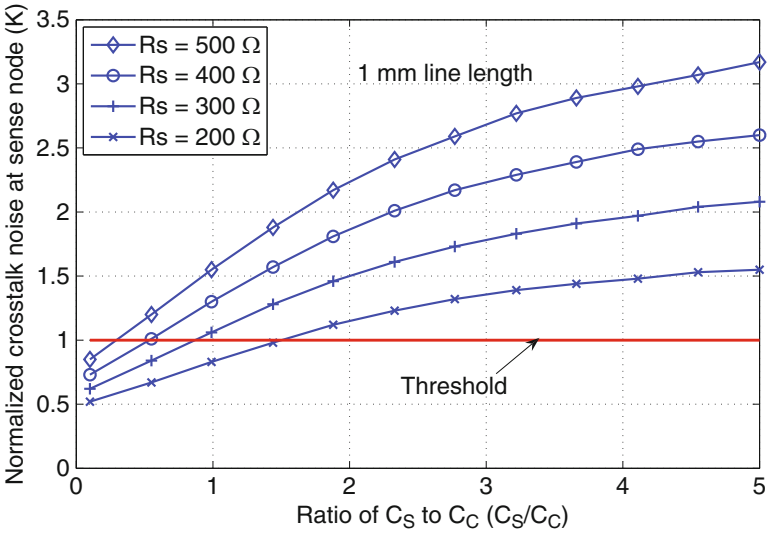


Fig. 33.12 Ratio of substrate capacitance to coupling capacitance versus normalized crosstalk noise when a P/G line is routed as a shield line. The interconnect length is 1 mm

coupling capacitance is greater than the line-to-substrate capacitance, shield insertion is more effective than additional spacing. As the line-to-substrate capacitance becomes greater than the coupling capacitance, physical spacing becomes more

efficient than shield insertion. For example, when R_s is equal to 300 Ω, spacing is preferable when C_s/C_c is greater than 2.3 for a 0.5 mm long line whereas for a 1 mm long line, spacing is preferable when C_s/C_c is greater than 0.9. The C_s/C_c ratio decreases with technology scaling, making shield insertion more effective than spacing in reducing crosstalk noise.

33.2.6 *Effect of Self- and Mutual Inductance on Crosstalk Noise*

The self- and mutual interconnect inductance strongly depends upon the technology and design parameters, as listed in Table 33.2. The effect of changes in the width, thickness, and spacing between the interconnects differs significantly for self- and mutual inductance. The self-inductance is constant for a range of mutual inductance between $0.5L_s$ to $1.2L_s$ for different driver resistances. When the ratio of L_m/L_s increases, spacing is more effective in reducing crosstalk noise, as depicted in Fig. 33.13. The crosstalk noise voltage generated at the sense node increases for both physical spacing and shield insertion when the L_m/L_s ratio increases. The increase in crosstalk noise voltage with shield insertion is however relatively high as compared to the increase in crosstalk noise voltage with physical spacing. The reason is that the noise coupled from the shield line is physically closer to the victim line than the noise coupled from the aggressor line. The relative effect of the change in the mutual inductance is therefore higher in shield insertion than physical spacing. This result is in good agreement with the results described in Sect. 33.2.3.

33.2.7 *Effect of Distance Between Aggressor and Victim Lines on Crosstalk Noise*

The crosstalk noise at the sense node is inversely proportional to the distance between the aggressor and victim lines since the coupling capacitance and mutual inductance decreases with increasing separation between lines. In this section, the effectiveness of shield insertion in a noisy environment is discussed. L_m decreases with greater separation between adjacent wires, lowering the L_m/L_s ratio. Alternatively, the C_s/C_c ratio increases with higher separation. Shield insertion is more efficient with a smaller L_m/L_s ratio. Conversely, additional spacing is preferable with a higher C_s/C_c ratio. The ratio $V_{sense\_with\_shielding}/V_{sense\_with\_spacing}$, denoted as K, therefore does not change significantly with increasing separation between the aggressor and victim lines. The distance between the aggressor and victim lines is varied from 0.8 to 2 μm, where the ratio of the crosstalk noise generated at the sense node with both shield insertion and spacing is shown in Fig. 33.14. Note that when comparing the effectiveness of shield insertion to physical spacing, the separation between the aggressor and victim lines is the same for both techniques.

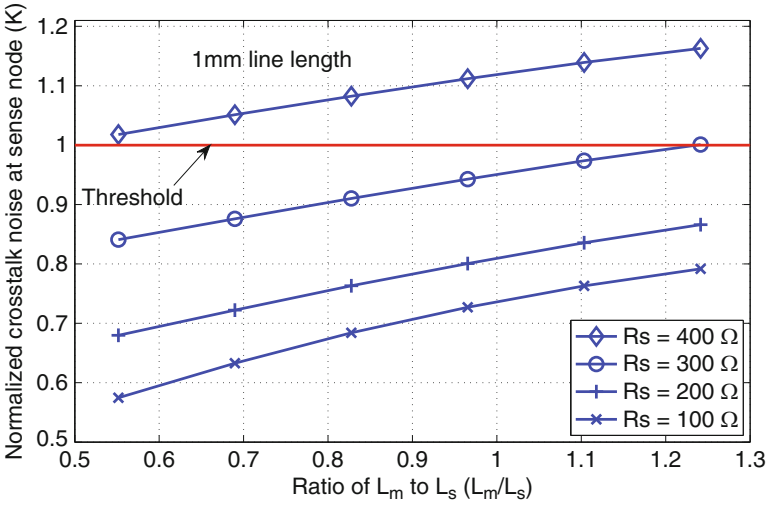


Fig. 33.13 Ratio of self-inductance to mutual inductance versus normalized crosstalk noise when a P/G line is routed as a shield line. The interconnect length is 1 mm

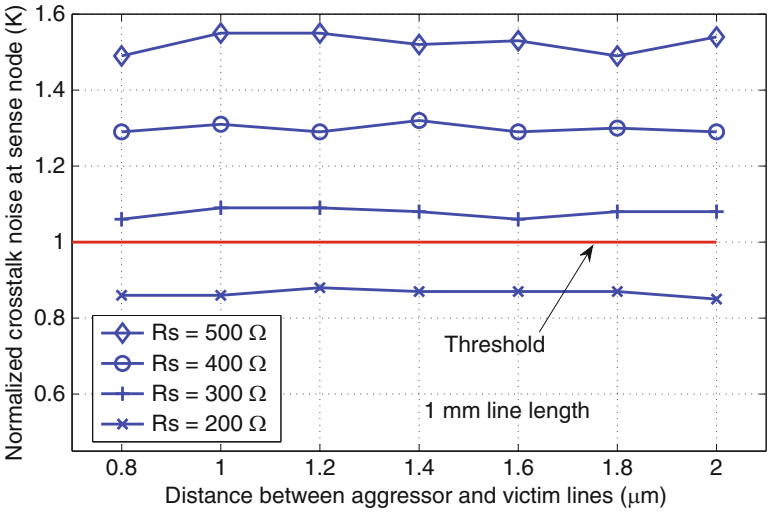


Fig. 33.14 Normalized crosstalk noise when a P/G line is routed as a shield line where the distance between the aggressor and victim line is varied from 0.8 to 2 μm. The interconnect length is 1 mm

33.3 Shield Insertion or Physical Spacing in a Noisy Environment

The decision criterion to choose between shield insertion and physical spacing in a noisy environment is summarized in this section. Shield insertion and physical spacing between adjacent interconnect are evaluated for several interconnect lengths and shield widths. Shield insertion is shown to be more efficient for shorter and narrower lines while additional space is preferable for longer and thicker lines. The effect of the driver resistance of the victim and aggressor lines on the crosstalk noise is also evaluated. Shielding is preferable for smaller driver resistance, and physical spacing is preferable for higher driver resistance. The ratio of the substrate capacitance to the coupling capacitance is explored in terms of mitigating coupling noise. Shield insertion is preferable for those lines with higher coupling capacitance than the line-to-substrate capacitance. Furthermore, when the mutual inductance between adjacent lines becomes higher than the self-inductance of the line, physical spacing becomes more efficient as compared to shield insertion in a noisy environment. A summary of the decision criteria is listed in Table 33.4 for different technology nodes.

A practical design example is analyzed that exemplifies the importance of P/G noise on the shield line when choosing between shield insertion and spacing. The circuit is shown in Fig. 33.3. Four different scenarios is considered: (1) a noise-free shield line, (2) a shield line with 40 mV peak noise, (3) a shield line with 100 mV peak noise, and (4) no shield line (physical spacing). The distance between the aggressor and victim lines is the same for shield insertion and physical spacing. The results are illustrated in Figs. 33.15 and 33.16 for, respectively, 0.5 and 1 mm interconnect lengths. For both cases, the crosstalk noise is greatest with a shield line with 100 mV P/G noise. The decision criteria, however, change when the P/G noise is 40 mV. For a 0.5 mm line length, the maximum noise with a shield line is

Table 33.4 Decision criterion for the critical interconnect length (width), $R_s = 300\,\Omega$. Shield insertion is preferable when the interconnect length (width) is smaller than the critical length (width). Spacing is preferable when the interconnect length (width) is greater than the critical length (width)

	Technology node	Shielding	Critical dimension	Spacing
Length[a]	65 nm	✓ <	1 mm	<✓
	45 nm	✓ <	0.9 mm	<✓
	32 nm	✓ <	0.2 mm	<✓
Width[b]	65 nm	✓ <	0.7 μm	<✓
	45 nm	✓ <	0.9 μm	<✓
	32 nm	✓ <	1.2 μm	<✓

[a]Width is maintained at 1 μm
[b]Length is maintained at 1 mm

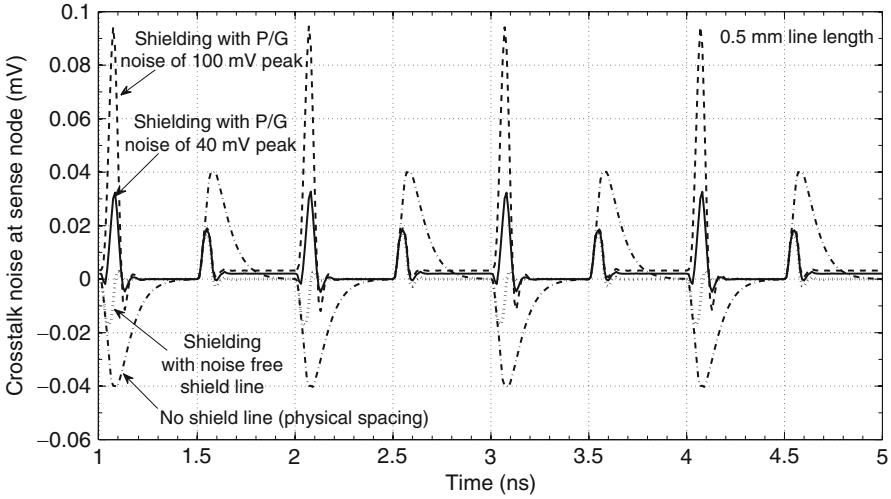


Fig. 33.15 Crosstalk noise at the sense node with an inserted shield line with different noise profiles (noise free, 40 , and 100 mV P/G noise on the shield line), and without a shield line (physical spacing). The interconnect length is 0.5 mm

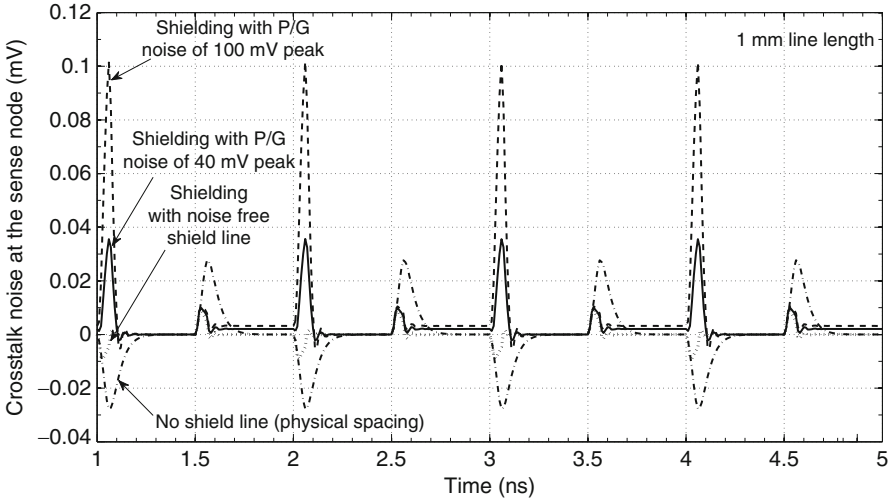


Fig. 33.16 Crosstalk noise at the sense node with an inserted shield line with different noise profiles (noise free, 40 , and 100 mV P/G noise on the shield line), and without a shield line (physical spacing). The interconnect length is 1 mm

greater than the noise without a shield line. The maximum noise with a 1 mm line is however greater without a shield line as compared to a shield line with 40 mV P/G noise. Additionally, when no P/G noise is present on the shield line, shield insertion is the preferred design method to mitigate crosstalk noise.

Two of the most important parameters to consider when choosing between shield insertion and physical spacing is the interconnect line length and the size of the transistors driving the aggressor and victim lines. For short interconnect lines, shield insertion is preferable while physical spacing is preferable for longer lines. This decision, however, also strongly depends upon the output resistance of the driver transistors and the width of the interconnect lines, as explained in Sect. 33.2. When R_s becomes smaller (i.e., a stronger driver strength), shield insertion is more efficient in reducing crosstalk noise.

33.4 Summary

With technology scaling, P/G noise has become a significant design issue. The P/G network has become more resistive, increasing the noise within the P/G distribution network. Additionally, with supply voltage scaling, the noise of the P/G network is more significant. Several solutions exist to mitigate this noise. Shield insertion and physical spacing of adjacent interconnect are described in this chapter. The efficiency of these noise mitigation methodologies is evaluated in the presence of P/G noise for different interconnect and shield parameters. The effects of the victim/aggressor driver resistance and line-to-substrate capacitance on crosstalk noise are evaluated. The decision criteria is illustrated based on a practical design example. The primary conclusions can be summarized as follows.

- Crosstalk noise is greater in more advanced technologies
- P/G noise on the shield line reduces the efficiency of shielding because this noise also couples to the victim lines
- P/G noise is the dominant source of crosstalk noise when the noise is greater than 7 % of the supply voltage
- Coupling from the aggressor to the victim is the dominant source of noise when P/G noise is less than 2 % of the supply voltage
- The length of the interconnect line and size of the transistors driving the lines are primary factors that determine the preferable technique for noise mitigation
- Shield insertion is preferable for shorter and narrower lines, and smaller driver resistance
- Physical spacing is preferable for longer and thicker lines and higher driver resistance
- When the line capacitance is dominated by coupling capacitance, shielding is more effective than additional spacing. Alternatively, physical spacing is preferable when line-to-substrate capacitance is greater than the coupling capacitance
- When the inductance of the adjacent lines is dominated by mutual inductance, physical spacing is more efficient as compared to shield insertion in a noisy environment

- Mitigation of crosstalk noise within a noise-free, low noise, and high noise environment is evaluated with and without shielding
- Shield insertion is less effective in a noisy environment
- Shield insertion efficiently mitigates crosstalk noise in distantly spaced adjacent lines under low P/G noise

Chapter 34
Conclusions

Noise is a fundamental issues in power supply networks, greatly degrading the performance of integrated circuits. These noise issues are discussed in Part VI. The impedance of a network is directly related to the noise in the power networks. *IR* and $L\,dI/dt$ voltage drops are produced, respectively, by the resistive and inductive portion of the network impedance. Determining the grid inductance, however, is a complicated task since crosstalk coupling between every line within a network needs to be evaluated.

In a commonly used interdigitated structure, the inductance can be treated as a local phenomenon, permitting the grid inductance to be efficiently and accurately estimated. A variety of tradeoffs among the inductance, resistance, and grid area is reviewed in this part. With increasing frequency, the inductive portion of the network impedance becomes more important than the resistive portion. Efficient and accurate models of the network inductance are therefore necessary when estimating the noise and applying noise reduction techniques.

A number of noise reduction techniques are reviewed. Physical separation of the power networks, local shielding, and decoupling capacitors are important techniques for reducing noise in modern ICs. The noise characteristics of the on-chip power supply networks are strongly affected by the location of the decoupling capacitors.

Noise evaluation, resistive and inductive impedance estimation, and noise reduction techniques are the primary foci of this part. These models and methodologies are evaluated, exhibiting good accuracy and computational efficiency.

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0_34

Part VII
Multi-layer Power Distribution Networks

Power distribution networks are typically allocated across a number of metal layers to enhance the performance characteristics of the network. These networks are reviewed in Part VII, with an emphasis on providing design intuition. The effects of multi-layer power distribution networks and tradeoffs among different properties of these networks are described in this part.

The impedance characteristics of on-chip multi-layer power distribution grids are described in Chap. 35. A circuit model of a multi-layer power distribution grid is reviewed. Analytic expressions describing the variation in the resistance and inductance of multi-layer grids with frequency are described. An intuitive explanation of the electrical behavior of power grids is offered. The results are supported with a case study.

Due to the large number of interconnect in interdigitated power and ground networks, excessive time is required to determine the inductance from electromagnetic simulation tools. In Chap. 36, a closed-form expression is described to accurately estimate the effective inductance of a single layer within an interdigitated power and ground distribution network. This expression is compared with previous models and FastHenry, exhibiting accurate and computationally efficient results. The inductance of a single layer within an interdigitated power and ground distribution network is bounded for any number of lines. The error of this expression decreases rapidly with increasing number of pairs within the network. The upper bound for the error of the closed-form model is also provided.

Two methods for optimizing a multi-layer interdigitated power and ground network are presented in Chap. 37. Based on the resistive and inductive (both self- and mutual) impedance, a closed-form expression for determining the optimal power and ground wire width that produces the minimum impedance for a single metal layer is described. Electromigration is also considered, permitting the appropriate number of metal layers to be determined. A tradeoff between the network impedance and current density is discussed. The optimal width as a function of metal layer is determined for different frequencies, suggesting important trends for interdigitated power and ground networks.

The global networks within conventional integrated circuits consists of three major types: power, ground, and clock distribution networks. These three networks consume most of the metal resources in the highest metal layers. The signals traversing the power and clock distribution networks are fundamentally different in terms of signal frequency and current flow. Combining the power and clock network into a multi-layer, globally integrated network is therefore possible. In Chap. 38, this general concept of a globally integrated power and clock (GIPAC) network is reviewed. The circuitry supporting this GIPAC system is also discussed.

Chapter 35
Impedance Characteristics of Multi-layer Grids

The power distribution network spans many layers of interconnect with disparate electrical properties. The impedance characteristics of multi-layer power distribution grids and the relevant design implications are the subject of this chapter.

Decoupling capacitors are an effective technique to reduce the effect of the inductance on power distribution networks operating at high frequencies. The efficacy of decoupling capacitors depends on the impedance of the conductors connecting the capacitors to the power load and source. The optimal allocation of the on-chip decoupling capacitance depends on the impedance characteristics of the interconnect. Robust and area efficient design of multi-layer power distribution grids therefore requires a thorough understanding of the impedance properties of the power distributing interconnect structures.

Power distribution networks in high performance digital ICs are commonly structured as a multi-layer grid, as shown in Fig. 35.1. The inductive properties of single layer power grids have been described in Chap. 28. In grid layers with alternating power and ground lines, long distance inductive coupling is greatly diminished due to cancellation, turning inductive coupling in single layer power grids into, effectively, a local phenomenon. The grid inductance, therefore, behaves similarly to the grid resistance: increases linearly with grid length and decreases inversely linearly with grid width (i.e., the number of lines in the grid). The electrical properties of power distribution grids can therefore be conveniently expressed by a dimension-independent sheet resistance $R_{\square}$ and sheet inductance $L_{\square}$ [125]. The inductance of the power grid layers can be efficiently estimated using simple models comprised of a few interconnect lines.

Area/inductance/resistance tradeoffs in power distribution grids have also been evaluated in Chap. 30. The sheet inductance of power distribution grids is shown to increase linearly with line width under two different tradeoff scenarios. Under the constraint of constant grid area, a tradeoff exists between the grid inductance and resistance. Under the constraint of a constant grid resistance, the grid inductance can be traded off against grid area.

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0_35

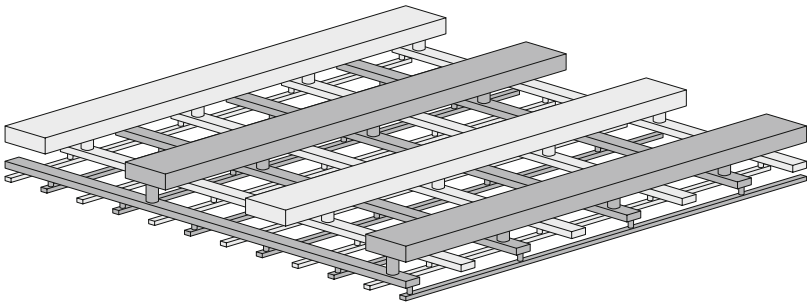


Fig. 35.1 A multi-layer power distribution grid. The ground lines are *light gray*, the power lines are *dark gray*

The variation of inductance with frequency in single layer power grids has been characterized in Chap. 29. This variation is relatively moderate, typically less than 10 % of the low frequency inductance. An exception from this behavior is power grids with closely spaced power and ground lines where the inductance variation with frequency is greater due to significant proximity effects.

Power distribution grids in modern integrated circuits typically consist of many grid layers, spanning an entire stack of interconnect layers. The objective of the present investigation is to characterize the electrical properties of these multi-layer grids, advancing the existing work beyond individual grid layers.

This chapter is organized as follows. The impedance characteristics of multi-layer power distribution grids are discussed in Sect. 35.1. A case study of a two layer power grid is presented in Sect. 35.2. The design implications of the impedance properties of a multi-layer grid are discussed in Sect. 35.3. The chapter concludes with a summary.

35.1 Electrical Properties of Multi-layer Grids

A circuit model of multi-layer power distribution grids is developed in this section. The impedance characteristics of multi-layer grids are determined based on this model. The impedance characteristics of individual layers of multi-layer power distribution grids are discussed in Sect. 35.1.1. The variation with frequency of the impedance characteristics of several grid layers forming a multi-layer grid is analyzed in Sect. 35.1.2.

35.1.1 Impedance Characteristics of Individual Grid Layers

The power and ground lines within each layer of a multi-layer power distribution grid are orthogonal to the lines in the adjacent layers. Orthogonal lines have zero mutual partial inductance as there is no magnetic linkage [45]. Orthogonal

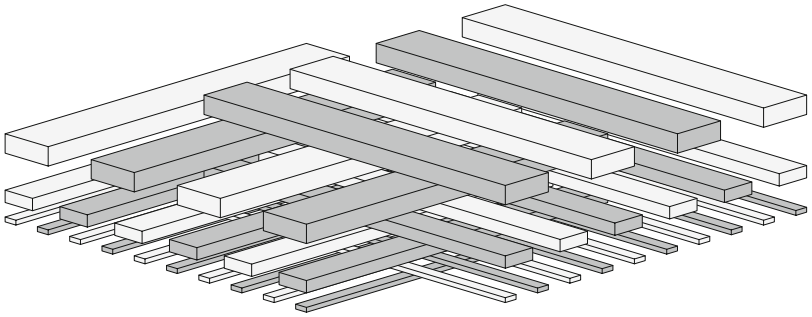


Fig. 35.2 A multi-layer grid consists of two stacks of layers. The lines in each stack are parallel to each other. The layers in one stack determine the resistive and inductive characteristics of the multi-layer grid in the direction of the lines in that stack, while the layers in the other stack determine the impedance characteristics in the orthogonal direction

grid layers can therefore be evaluated independently. A multi-layer grid can be considered to consist of two stacks of layers, with all of the lines in each stack parallel to each other, as shown in Fig. 35.2. Grid lines in one stack are orthogonal to the lines in the other stack. Grid layers in each stack only affect the grid inductance in the direction of the lines in the stack. This behavior is analogous to the properties of the grid resistance. The problem of characterizing a multi-layer grid is thereby reduced to determining the impedance characteristics of a stack of several individual grid layers with lines in the same direction.

The power and ground lines in power distribution grids are connected to the lines in the adjacent layers through vias. The vias (or clusters of vias) are distributed along a power line at a pitch equal to the power line pitch in the adjacent layer. The line pitch is much larger than the via length. The inductance and resistance of the two close parallel power lines in different metal layers are therefore much larger than the resistance and inductance of the connecting vias. The effect of vias on the resistance and inductance of a power grid is therefore negligible. This property is a direct consequence of the characteristic that the distance of the lateral current distribution in power grids (hundreds or thousands of micrometers) is much larger than the distance of the vertical current distribution (several micrometers). The power current is distributed among the metal layers over a distance comparable to a line pitch. The power and ground lines are effectively connected in parallel.

Each layer of a typical multi-layer power distribution grid has significantly different electrical properties. Lines in the upper layers tend to be thick and wide, forming a low resistance global power distribution grid. Lines in the lower layers tend to be thinner, narrower, and have a smaller pitch. The lower the metal layer, the smaller the metal thickness, width, and pitch. The upper grid layers therefore have a relatively high inductance and low resistance, whereas the lower layers have a relatively low inductance and high resistance [73, 125]. The lower the layer, the higher the resistance and the lower the inductance. In those circuits employing flip-chip packaging with a high density area array of I/O contacts, the

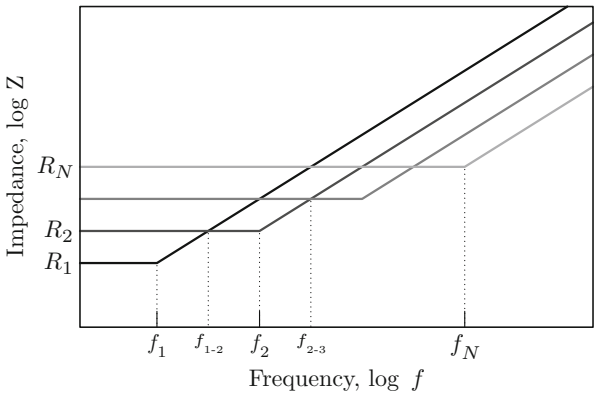


Fig. 35.3 Impedance of the individual grid layers comprising a multi-layer grid

interconnect layers in the package are tightly coupled to the on-chip interconnect, effectively extending the on-chip interconnect hierarchy. The difference in the electrical properties across the interconnect hierarchy is particularly significant in nanoscale circuits. While the cross-sectional dimensions of local on-chip lines are measured in tens of nanometers, the dimensions of package lines are of the order of tens of micrometers. The three orders of magnitude difference in dimensions translates to six orders of magnitude difference in resistance (proportional to the cross-sectional area of the grid lines) and to three orders of magnitude difference in inductance (proportional to the grid line density).

The variation with frequency of the impedance of each layer in a grid stack comprised of N grid layers is schematically shown in Fig. 35.3. The layers are numbered from 1 (the uppermost layer) to N (the lowest layer). The grid layer resistance increases with layer number, $R_1 < R_2 < \ldots < R_N$, and the inductance decreases with layer number $L_1 > L_2 > \ldots > L_N$. At low frequencies, the uppermost layer has the lowest impedance as the layer with the lowest resistance. This layer, however, has the highest inductance and, consequently, the lowest transition frequency $f_1 = \frac{1}{2\pi}\frac{R_1}{L_1}$, as compared to the other layers (see Fig. 35.3). The transition frequency is the frequency at which the impedance of the grid layer changes in character from resistive to inductive. At this frequency, the inductive impedance of a grid layer is equal to the resistive impedance (neglecting skin and proximity effects), i.e., $R_1 = \omega L_1$. The grid impedance increases linearly with frequency above f_1. The lowest grid layer has the highest resistance and the lowest inductance; therefore, this layer has the highest transition frequency f_N. As the inductance of the upper layers is higher than the lower layers, the impedance of an upper layer exceeds the impedance of any lower layer above a certain frequency. For example, the impedance of the first layer $R_1 + \omega L_1 \approx \omega L_1$ equals the magnitude of the second layer impedance $R_2 + \omega L_2 \approx R_2$ and exceeds the impedance of the second layer above frequency $f_{1-2} = \frac{1}{2\pi}\frac{R_2}{L_1}$, as shown in Fig. 35.3. Similarly, the impedance of layer k exceeds the impedance of layer l, $k < l$, at $f_{k-l} = \frac{1}{2\pi}\frac{R_l}{L_k}$.

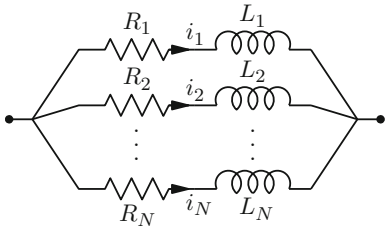


Fig. 35.4 Equivalent circuit of a stack of N grid layers

35.1.2 Impedance Characteristics of Multi-layer Grids

An entire stack of grid layers cannot be accurately described by a single *RL* circuit due to the aforementioned differences among the electrical properties of the individual grid layers. A stack of multiple grid layers can, however, be modeled by several parallel *RL* branches, each branch characterizing the electrical properties of one of the comprising grid layers, as shown in Fig. 35.4.

Due to the difference in the electrical properties of the individual layers, the magnitude of the current in each grid layer varies significantly with frequency. At low frequencies, the low resistance uppermost layer is the path of lowest impedance, as shown in Fig. 35.3. The uppermost layer has the greatest effect on the low frequency resistance and inductance of the grid stack, as the largest share of the overall current flows through this layer. As the frequency increases to $f_{1\text{-}2} = \frac{1}{2\pi}\frac{R_2}{L_1}$ and higher, the impedance of the uppermost layer ωL_1 exceeds the impedance of the second uppermost layer R_2, as shown in Fig. 35.3. The second uppermost layer, therefore, carries the largest share of the overall current and most affects the inductance and resistance within this frequency range. As the frequency exceeds $f_{2\text{-}3} = \frac{1}{2\pi}\frac{R_3}{L_2}$, the next layer in the stack becomes the path of least impedance and so on. The process continues until at very high frequencies the lowest layer carries most of the overall current.

As the frequencies increase, the majority of the overall current is progressively transferred from the layers of low resistance and high inductance to the layers of high resistance and low inductance. The overall grid inductance, therefore, decreases with frequency and the overall grid resistance increases with frequency. A qualitative plot of the variation of the grid inductance and resistance with frequency is shown in Fig. 35.5. At low frequency, all of the layers exhibit a purely resistive behavior and the current is partitioned among the layers according to the resistance of each layer. The share i_k of the overall current flowing through layer k is

$$i_k = \frac{I_k}{\sum_{n=1}^{N} I_n} = \frac{\prod_{n\neq k} R_n}{\sum_{m=1}^{N} \prod_{n\neq m} R_n}. \tag{35.1}$$

Note that $i_1 > i_2 > \ldots > i_N$ as $R_1 < R_2 < \ldots < R_N$. The resistance of a multi-layer grid R_0^{LF} at low frequency is therefore determined by the parallel connection of all of the individual layer resistances,

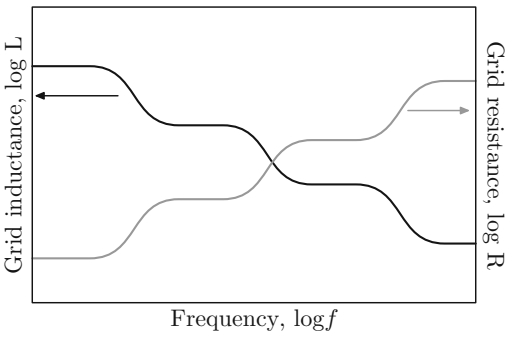


Fig. 35.5 Variation of the grid inductance and resistance of a multi-layer stack with frequency. As the signal frequency increases, the current flow shifts to the high resistance, low inductance layers, decreasing the inductance and increasing the resistance of the grid

$$R_0^{LF} = R_1 \| R_2 \| \ldots \| R_N = \frac{\prod_{n=1}^{N} R_n}{\sum_{m=1}^{N} \prod_{n \neq m} R_n}. \tag{35.2}$$

The low frequency inductance of a multi-layer grid L_0^{LF} is, however,

$$L_0^{LF} = L_1 i_1^2 + L_2 i_2^2 + \ldots + L_N i_N^2 \approx L_1, \tag{35.3}$$

due to $L_1 > L_k$ and $i_1 > i_k$ for any $k \neq 1$.

At very high frequencies, the resistance and inductance exchange roles. All of the grid layers exhibit a purely inductive behavior and the current is partitioned among the layers according to the inductance of each layer. The share of the overall current flowing through layer n is

$$i_k = \frac{I_k}{\sum_{n=1}^{N} I_n} = \frac{\prod_{n \neq k} L_n}{\sum_{m=1}^{N} \prod_{n \neq m} L_n}. \tag{35.4}$$

The relation among the currents of each layer is reversed as compared to the low frequency case: $i_1 < i_2 < \ldots < i_N$. The inductance of a multi-layer grid at high frequency L_0^{HF} is determined by the parallel connection of the individual layer inductances,

$$L_0^{HF} = L_1 \| L_2 \| \ldots \| L_N = \frac{\prod_{n=1}^{N} L_n}{\sum_{m=1}^{N} \prod_{n \neq m} L_n}. \tag{35.5}$$

The high frequency resistance of a multi-layer grid R_0^{HF} is

$$R_0^{HF} = R_1 i_1^2 + R_2 i_2^2 + \ldots + R_N i_N^2 \sim R_N, \tag{35.6}$$

due to $R_N > R_k$ and $i_N > i_k$ for any $k \neq N$.

The grid resistance and inductance vary with frequency between these limiting low and high frequency cases. If the difference in the electrical properties of

the layers is sufficiently high, the variation of the grid inductance and resistance with frequency has a staircase-like shape, as shown in Fig. 35.5. As the frequency increases, the grid layers consecutively serve as the primary current path, dominating the overall grid impedance within a specific frequency range [131].

35.2 Case Study of a Two Layer Grid

The electrical properties of a two layer grid are evaluated in this section to quantitatively illustrate the concepts described in Sect. 35.1. The grid parameters are described in Fig. 35.6.

The analysis approach used to determine the electrical characteristics of a grid structure is described in Sect. 35.2.1. Magnetic coupling between grid layers is discussed in Sect. 35.2.2. The inductive characteristics of a two layer grid are discussed in Sect. 35.2.3. The resistive characteristics of a two layer grid are discussed in Sect. 35.2.3. The impedance characteristics of a two layer grid are summarized in Sect. 35.2.5.

35.2.1 Simulation Setup

The inductance extraction program FastHenry [70] is used to explore the inductive properties of grid structures. FastHenry efficiently calculates the frequency dependent impedance $R(\omega) + \omega L(\omega)$ of complex three-dimensional interconnect structures under a quasi-magnetostatic approximation. In the analysis, the lines are split into multiple filaments to account for skin and proximity effects, as discussed in Sect. 2.2. A conductivity of $58\,\mathrm{S}/\mu\mathrm{m} \simeq (1.72\,\mu\Omega \cdot \mathrm{cm})^{-1}$ is used in the analysis where an advanced process with copper interconnect is assumed [514].

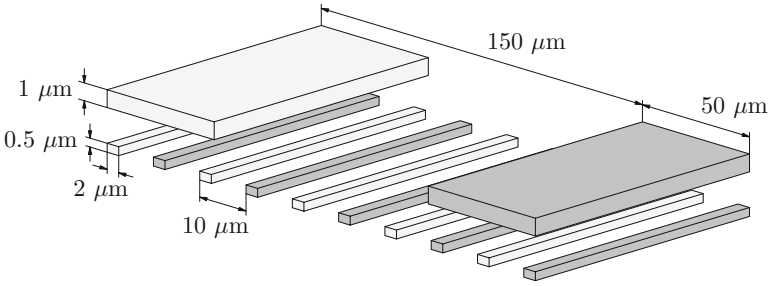


Fig. 35.6 General view of a two layer grid. The ground lines are *white colored*, the power lines are *gray colored*

When determining the loop inductance, all of the ground lines at one end of the grid are short circuited to form a ground terminal and all of the power lines at the same end of the grid are short circuited to form a power terminal. All of the lines at the other end of the grid are short circuited to complete the current loop. This configuration assumes that the power current loop is completed on-chip. This assumption is valid for high frequency signals which are effectively terminated through the on-chip decoupling capacitance which acts as a low impedance termination as compared to the inductive off-chip leads of the package. If the current loop is completed on-chip, the current in the power lines and the current in the ground lines always flow in opposite directions.

35.2.2 Inductive Coupling Between Grid Layers

An equivalent circuit diagram of a two layer power distribution grid is shown in Fig. 35.7. The partial mutual inductance between the lines in the two grid layers is significant as compared to the partial self-inductance of the lines. Therefore, the two grid layers are, in general, magnetically coupled, as indicated in Fig. 35.7. It can be shown, however, that for practical geometries, magnetic coupling is significant only in interdigitated grids under specific conditions.

The specific conditions are that the line pitch in both layers is the same and the separation between the two layers is smaller than the line pitch. The two layers with the same line pitch are spatially correlated, i.e., the relative position of the lines in the two layers is repeated throughout the structure. The net inductance of such grids depends upon the mutual alignment of the two grid layers. For example, consider a two layer grid with each layer consisting of ten interdigitated power and ground lines with a $1 \times 1\,\mu\text{m}$ cross section on an $8\,\mu\text{m}$ pitch, as in the cross section shown in Fig. 35.8. The separation between the layers is $4\,\mu\text{m}$. The variation of inductance of this two layer grid structure as a function of the physical offset between the two layers is shown in Fig. 35.9.

At 1 GHz, each of the layers has a loop inductance of 206 pH. The inductance of two identical parallel coupled inductors is

$$L_{1\|2} = \frac{L_{11}L_{22} - L_{12}^2}{L_{11} + L_{22} - 2L_{12}} = \frac{L_{11} + L_{12}}{2}\,. \tag{35.7}$$

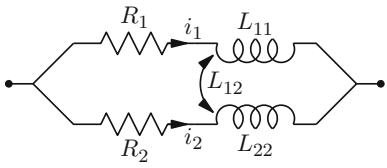


Fig. 35.7 An equivalent circuit diagram of a two layer grid

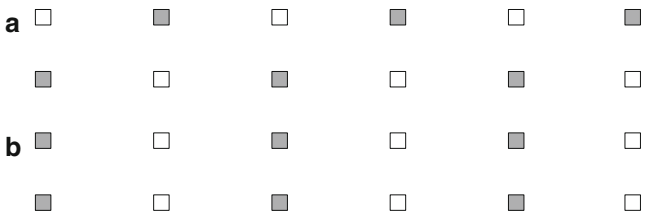


Fig. 35.8 Alignment of two layers with the same line pitch in a two layer grid resulting in the minimum and maximum grid inductance. The ground lines are *white colored*, the power lines are *gray colored*; (**a**) configuration with the minimum grid inductance: ground lines of one layer are aligned with the power lines of the other layer, (**b**) configuration with the maximum grid inductance: the ground lines of both layers are aligned with each other

In the case of zero coupling between the two grid layers, the net inductance of the two layer grid is approximately 206 pH/2 = 103 pH since the two grids are in parallel. If the ground lines in the top layer are placed immediately over the power lines of the bottom layer, as shown in Fig. 35.8a, a close return path for the power lines is provided as compared to the neighboring ground lines of the bottom layer. The magnetic coupling between the two grid layers is negative in this case, resulting in a net inductance of 81 pH for the two layer grid (which is lower than the uncoupled case of 103 pH), in agreement with (35.7). A two layer interdigitated grid effectively becomes a paired grid, as shown in Fig. 35.8a. (Rather than equidistant line spacing, in paired grids the lines are placed in close power-ground line pairs [72].) If, alternatively, the ground lines of the top layer are aligned with the ground lines of the bottom layer, as shown in Fig. 35.8b, the magnetic coupling between the two layers is positive and the net inductance of the two layer grid is 124 pH, higher than the uncoupled case, also in agreement with (35.7). As the offset between the grid layers changes between these two limits, the total inductance varies from a minimum of 81 pH to a maximum of 124 pH, passing a point where the effective coupling between the two layers is zero and the total inductance is 103 pH. The inductance at a 100 GHz signal frequency closely tracks this behavior at low frequencies.

If the layer separation is greater than the line pitch in either of the two layers, the net coupling from the lines in one layer to the lines in the other layer is insignificant. Coupling to the power lines is nearly cancelled by the coupling to the ground lines, carrying current in the opposite direction. This coupling cancellation is analogous to the cancellation of the long distance coupling within the same grid layer [72]. This cancellation also explains why two grid layers are effectively uncoupled if one of the layers is a paired grid. The power to ground line separation in a paired grid is smaller than the separation between two metalization layers with the grid lines in the same direction.

It is possible to demonstrate that in the case where the line pitch is not matched, as shown in Fig. 35.10, the layer coupling is effectively cancelled, and the grid inductance is independent of the layer alignment, as shown in Fig. 35.9. Metalization

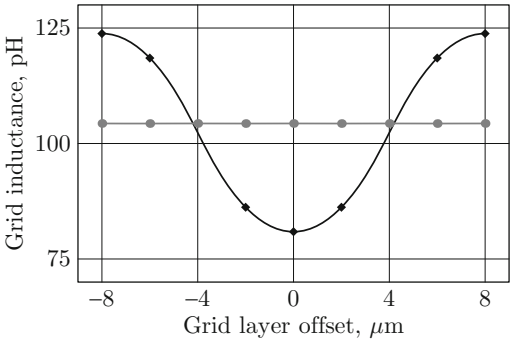


Fig. 35.9 Inductance of a two layer grid versus the physical offset between the two layers. The inductance of the grid with matched line pitch of the layers (*black line*) depends on the layer offset. (The low inductance alignment shown in Fig. 35.8a is chosen as the zero offset.) The inductance of the grid is constant where the line height, width, and pitch of the lower layer are twice as small as compared to the upper layer (the *gray line*)

Fig. 35.10 The cross section of a two layer grid with the line pitch of the upper layer a fractional multiple (5/4 in the case shown) of the line pitch in the bottom layer. Both effects illustrated in Fig. 35.8 occur at different locations (*circled*). The ground lines are *white colored*, the power lines are *gray colored*

layers in integrated circuits typically are of different thickness, line width, and line spacing. Therefore, unless intentionally designed otherwise, different grid layers typically have different line pitch and can be considered uncoupled, as has been implicitly assumed in Sect. 35.1.

35.2.3 Inductive Characteristics of a Two Layer Grid

The variation of the sheet inductance with signal frequency in the two layer grid is shown in Fig. 35.11. Note that the inductance of the individual grid layers, also shown in Fig. 35.11, is virtually constant with frequency [72, 515]. The sheet inductance of the upper layer L_1 is 268 pH/□ at 1 MHz (247 pH/□ at 100 GHz). The sheet inductance of the bottom layer L_2 is 19.6 pH/□ at 1 MHz (19 pH/□ at 100 GHz). The inductance of the bottom grid layer is approximately fifteen times lower than the inductance of the upper grid layer. This difference in inductance is primarily due to the difference in the line density of the layers. The line density of the bottom layer is fifteen times higher, as determined by the line pitch of the layers ($150/10 = 15$). The inductance of a single line is relatively insensitive to the

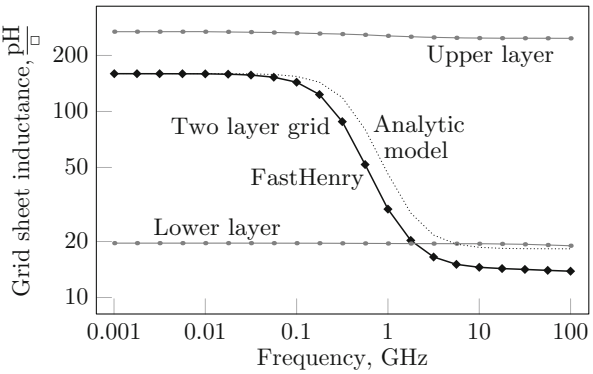


Fig. 35.11 Inductance of a two layer grid versus signal frequency. Both FastHenry data (*solid line*) and the analytic model data (*dotted line*) are shown. The individual inductance of the two comprising grid layers is shown for comparison (FastHenry data)

aspect ratio of the line cross section. The inductance of the two layer grid, however, varies significantly with signal frequency due to current redistribution, as discussed in Sect. 35.1.

The inductive characteristics of a two layer grid can also be analytically determined based on the simple model shown in Fig. 35.7. Assuming $L_{12} = 0$ as discussed in Sect. 35.2.2, the loop inductance of a two layer grid is

$$L_0 = \frac{L_1(R_2^2 + \omega^2 L_1 L_2) + L_2(R_1^2 + \omega^2 L_1 L_2)}{(R_1 + R_2)^2 + \omega^2 (L_1 + L_2)^2} \,. \tag{35.8}$$

At high frequencies, where the resistance of the grid layers has no influence on the current distribution between the layers, the grid inductance described by (35.8) asymptotically approaches the inductance of two ideal parallel inductors,

$$L_0^{HF} = \frac{L_1 L_2}{L_1 + L_2} \,, \tag{35.9}$$

in agreement with (35.5). At low frequencies, the grid inductance described by (35.8) approaches the low frequency limit of the grid inductance,

$$L_0^{LF} = L_1 \left(\frac{R_2}{R_1 + R_2} \right)^2 + L_2 \left(\frac{R_1}{R_1 + R_2} \right)^2 = 160\,\text{pH}/\square, \tag{35.10}$$

in agreement with (35.3).

The variation of the grid inductance with frequency according to the analytic model described by (35.8) is also illustrated in Fig. 35.11 by the dotted line. The analytic model satisfactorily describes the variation of grid inductance with

frequency. The discrepancy between the analytic and FastHenry data at high frequencies is due to proximity effects which are not captured by the model shown in Fig. 35.7.

35.2.4 Resistive Characteristics of a Two Layer Grid

The resistance of the two individual grid layers R_1 and R_2 and the resistance of the combined two layer grid R_0 are shown in Fig. 35.12. The resistance of the individual grid layers remains constant up to high frequencies. The resistance of the upper layer begins to moderately increase from approximately 0.5 GHz due to significant proximity effects in very wide lines. The resistance of both layers sharply increases above approximately 20 GHz due to significant skin effect. Note that the resistance of the grid comprised of the two layers exhibits significantly greater variation with frequency than either individual layer.

Similar to the grid inductance, the resistive characteristics of a two layer grid can be analytically determined from the properties of the comprising grid layers,

$$R_0 = \frac{R_1(R_1R_2 + \omega^2L_2^2) + R_2(R_1R_2 + \omega^2L_1^2)}{(R_1 + R_2)^2 + \omega^2(L_1 + L_2)^2} \,. \tag{35.11}$$

The grid resistance versus frequency data based on the analytic model described by (35.11) is shown by the dotted line in Fig. 35.12. The analytic solution describes well the general character of the resistance variation with frequency. At low frequencies, the resistance of the two layer grid approaches the parallel resistance of two grid layers,

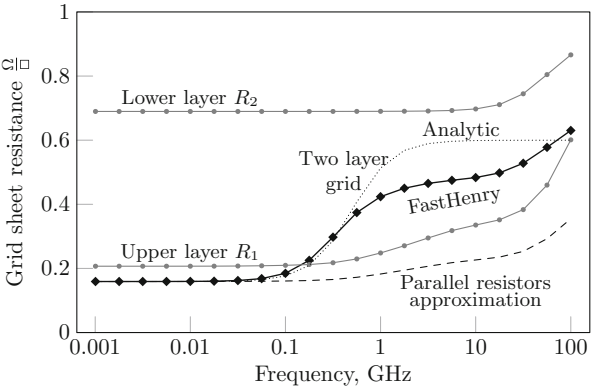


Fig. 35.12 Resistance of a two layer grid versus signal frequency. The individual resistance of the two comprising grid layers and the parallel resistance of the individual layer resistances are shown for comparison

$$R_0^{LF} = R_1 \| R_2 = \frac{R_1 R_2}{R_1 + R_2} = 0.16\,\Omega/\square, \tag{35.12}$$

in agreement with (35.2). The high frequency grid resistance asymptotically approaches

$$R_0^{HF} = R_1 \left(\frac{L_2}{L_1 + L_2} \right)^2 + R_2 \left(\frac{L_1}{L_1 + L_2} \right)^2 = 0.6\,\Omega/\square, \tag{35.13}$$

in agreement with (35.6). This analytically calculated high frequency resistance overestimates the FastHenry extracted resistance of $0.48\,\Omega/\square$ (at 10 GHz). The discrepancy is due to pronounced proximity and skin effects at high frequencies.

35.2.5 *Variation of Impedance with Frequency in a Two Layer Grid*

Having determined the variation with frequency of the resistance and inductance in the previous sections, it is possible to characterize the frequency dependent impedance characteristics of a two layer grid. The magnitude of the impedance calculated from the analytic models (35.8) and (35.11) is shown in Fig. 35.13 by the dotted line. Low frequency values of the individual layer inductance, L_1 and L_2, and resistance, R_1 and R_2, are used in the analytic model. The impedance magnitude based on FastHenry extracted data is shown by the solid line. The extracted impedance of the individual grid layers is also shown for comparison.

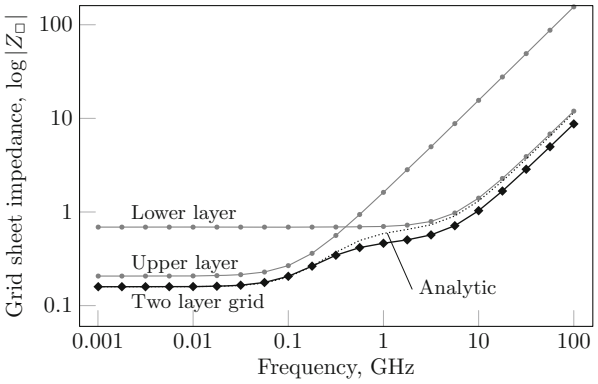


Fig. 35.13 Impedance magnitude of a two layer grid versus signal frequency. Both the extracted (*solid line*) and analytic (*dotted line*) data are shown. The impedance of the two comprising grid layers is also shown

Note that the impedance characteristics of the individual layers shown in Fig. 35.13 bear close resemblance to the schematic graph shown in Fig. 35.3.

As discussed in Sect. 35.1, the low resistance upper grid dominates the impedance characteristics at low frequencies, while the low inductance lower grid determines the impedance characteristics at high frequencies [131]. The analytic model satisfactorily describes the frequency dependent impedance characteristics. The discrepancy between the analytic and extracted data at high frequencies is due to overestimation of the high frequency inductance by the analytic model, as shown in Fig. 35.11.

35.3 Design Implications

The variation with frequency of the electrical properties of a multi-layer grid has several design implications. Modeling the resistance of a multi-layer grid as a parallel connection of individual layer resistances underestimates the high frequency resistance of the grid. The parallel resistance model, therefore, underestimates the resistive *IR* voltage drops during fast current transients. Representing a multi-layer grid inductance by the individual layer inductances connected in parallel is accurate only at very high frequencies. At lower frequencies, this model underestimates the grid inductance. Relatively low inductance and high resistance at high frequencies increase the damping factor of the power distribution grid (proportional to $R/\sqrt{L}$), thereby preventing resonant oscillations in power distribution networks at high frequencies. Conversely, resonant oscillations are more likely at lower frequencies, where the inductance is relatively high and the resistance is low.

Multi-layer grids with different grid layer impedance characteristics are well suited to distribute power in high speed integrated circuits. At low frequencies, where the grid impedance is dominated by the resistance, most of the current flows through the less resistive upper grid layers, decreasing the grid impedance. At high frequencies, where the grid impedance is dominated by the inductance, most of the current flows through the low inductance lower layers. Over the entire frequency range, the current flow changes so as to minimize the impedance of the grid. These properties of multi-layer power distribution grids support the design of power distribution networks with low impedance across a wide frequency range, necessary in high performance nanoscale integrated circuits.

The inductive properties of the interconnect changes the metal allocation strategy for global power distribution grids. In circuits based on resistance-only models, all of the metal area for the global power distribution is allocated in the upper layers with the lowest line resistance. The power interconnect in the lower metal layers connects the circuits to the global power grid and typically do not form continuous power grids. In multi-layer grids, however, significant metal resources are required to form continuous grids in the lower metal layers. This difference is a direct consequence of the inductive behavior of interconnect at high frequencies. A significant fraction of the high density lower metal layers should be used to lower

the high frequency impedance of the power grid. In this manner, the frequency range of the grid impedance is extended to match the increased switching speeds of scaled transistors.

Redistribution of the grid current toward the lower layers at high frequencies increases the current density in the power and ground lines in the lower grid layers, degrading the electromigration reliability of the power distribution grid. The significance of these effects will increase as the frequency of the current delivered through the on-chip power distribution grid increases with higher operating speeds. An analysis of these effects is therefore necessary to ensure the integrity of high speed nanoscale integrated circuits.

35.4 Summary

The electrical characteristics of multi-layer power distribution grids are evaluated in this chapter. The primary results are summarized as follows.

- The upper metal layers comprised of thicker and wider lines have low resistance and high inductance; the lower metal layers comprised of thinner and narrower lines have relatively high resistance and low inductance
- Inductive coupling between grid layers is shown to be insignificant in typical power distribution grids
- Due to this difference in electrical properties, the impedance characteristics of multi-layer grids vary significantly with frequency
- The current distribution among the grid layers changes with frequency, minimizing the overall impedance of the power grid
- As signal frequencies increase, the majority of the current flow shifts from the lower resistance upper layers to the lower inductance lower layers
- The inductance of a multi-layer grid decreases with frequency, while the resistance increases with frequency
- An analytic model describing the electrical properties of a multi-layer grid based on the inductive and resistive properties of the comprising grid layers is described
- A dense and continuous power distribution grid in the lower metal layers is essential to reduce the impedance of a power distribution grid operating at high frequencies

Chapter 36
Inductance Model of Interdigitated Power and Ground Networks

With high operating frequencies and scaled geometries, the power and ground distribution network requires greater design optimization to effectively provide higher current flow with minimal voltage variations. Low supply voltages and high currents in ICs place stringent constraints on the P/G distribution networks. Higher frequencies and smaller transistors produce shorter transition times, such that $L\,(di/dt)$ voltage drops can exceed IR voltage drops. All of these factors require the inductance to be considered in the design of on-chip P/G distribution networks. To optimize these large scale P/G distribution networks, the inductance needs to be accurately and efficiently determined.

An interdigitated P/G distribution network structure, depicted in Fig. 36.1, where a few wide lines are replaced by a large number of narrow lines, is often used to reduce the inductance effect [150, 151]. The advantages of an interdigitated structure are increased routing flexibility and reduced inductance effects. The current flow of the power and ground lines within a layer is assumed to flow in opposite directions, thereby reducing the loop inductance of the network [73].

The inductance is smaller with a large number of interdigitated pairs [73, 150, 151]. The complexity to estimate the inductance however increases with a large number of interdigitated pairs, since a larger number of mutual inductances needs to be calculated. The effective inductance of a practical interdigitated P/G distribution network is therefore a complex task.

This chapter is organized as follows. In Sect. 36.1, an estimate of the inductance of a four-pair interdigitated structure is provided. Based on the self- and mutual inductance, a closed-form expression for an interdigitated P/G distribution network is determined in Sect. 36.2. The accuracy of this expression and a comparison to other models are provided in Sect. 36.3. The upper bound of the error is also provided. The chapter is summarized in Sect. 36.4.

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0_36

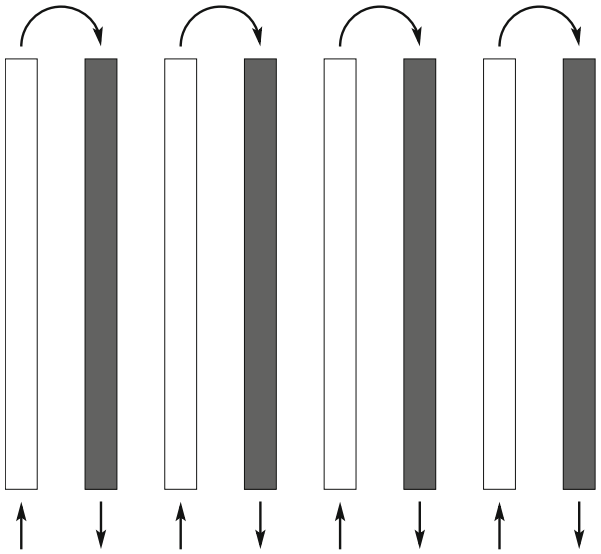


Fig. 36.1 A single metal layer of an interdigitated P/G distribution structure. The *darker* and *lighter lines* represent, respectively, the power and ground lines

36.1 Basic Four-Pair Structure

The loop inductance of two parallel wires with opposite current flow is

$$L_{loop} = L_{11} + L_{22} - 2M_{12}, \tag{36.1}$$

where L_{11}, L_{22}, and M_{12} are, respectively, the self-inductance of the power and ground lines, and the mutual inductance between these two wires.

The process of estimating the inductance becomes problematic with a large number of wires. To calculate the loop inductance, the mutual inductance terms among all of the wires need to be individually determined, a computationally expensive process. A closed-form expression characterizing this inductance would therefore be useful.

The inductance of a single layer within an interdigitated P/G distribution network structure with four pairs (eight wires), shown in Fig. 36.2, is

$$\frac{1}{L_{eff}} = \frac{1}{L_1} + \frac{1}{L_2} + \frac{1}{L_3} + \frac{1}{L_4}, \tag{36.2}$$

where L_1, L_2, L_3, and L_4 are, respectively, the inductance of the first, second, third, and fourth pair of a single layer within a P/G distribution network. The inductance of the first pair is

$$L_1 = L_{1_p} + L_{1_g} + \sum_{i=2}^{4} \left(M_{1_p i_p} + M_{1_g i_g}\right) - \sum_{i=1}^{4} \left(M_{1_p i_g} + M_{1_g i_p}\right), \tag{36.3}$$

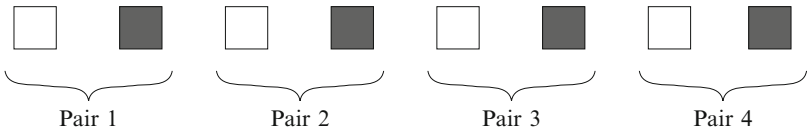


Fig. 36.2 Four pairs of a single layer within an interdigitated P/G distribution network

where subscripts p and g represent, respectively, power and ground. In this case, the overall inductance requires sixteen terms to determine each pair. For n pairs of P/G distribution networks, $2 \cdot 2n = 4n$ terms are required to characterize each pair, making complexity $O(n)$ for a single pair. For n pairs, the complexity in estimating the inductance of a single layer within a P/G network is $O(n^2)$.

36.2 P/G Network with Large Number of Interdigitated Pairs

The definition of the inductance between two loops, i and j, for a uniform current density is presented by the Neumann equation,

$$L_{ij} \equiv \frac{\mu_0 \mu_r}{4\pi} \oint_{C_i} \oint_{C_j} \frac{ds_i ds_j}{|R_{ij}|}, \tag{36.4}$$

where μ_0, μ_r, and R_{ij} are, respectively, the vacuum and relative permeability, and the distance between two loops. From [46], the mutual inductance between a pair of two rectangular conductors is

$$M = \frac{\mu_0 l}{2\pi} \left[ln \left(\frac{l}{d} + \sqrt{1 + \frac{l^2}{d^2}} \right) - \sqrt{1 + \frac{d^2}{l^2}} + \frac{d}{l} \right], \tag{36.5}$$

where l and d are, respectively, the length of the wire and pitch of two wires. If $l \gg d$, an approximate expression based on a Taylor series expansion is [575]

$$M = \frac{\mu_0 l}{2\pi} \left[ln \left(\frac{2l}{d} \right) - 1 + \frac{d}{l} \right]. \tag{36.6}$$

The self-inductance is derived in a similar way. For those cases where the length is larger than the width [48],

$$L_s = \frac{\mu_0 l}{2\pi} \left[ln \left(\frac{2l}{w+t} \right) + \frac{1}{2} + \frac{k(w+t)}{l} \right], \tag{36.7}$$

where w, t, and k are, respectively, the wire width, wire thickness, and fitting parameter ($k \approx 0.22$) for smaller length wires. In P/G distribution networks where $l \gg d$ and $l \gg w + t$, the last term characterizing the edge effect of the self- and mutual inductance can be neglected, simplifying (36.6) and (36.7) to, respectively,

$$M = \frac{\mu_0 l}{2\pi}\left[ln\left(\frac{2l}{d}\right) - 1\right], \tag{36.8}$$

$$L_s = \frac{\mu_0 l}{2\pi}\left[ln\left(\frac{2l}{w+t}\right) + \frac{1}{2}\right]. \tag{36.9}$$

The mutual component of the inductance within an interdigitated P/G distribution network decreases with increasing distance between the wires and can be treated as a local effect, according to [73]. In this case, the effective inductance of each pair is the sum of the self-inductances and a single mutual inductance between the two wires in the pair. This approach supports fast estimation of the effective inductance of a P/G distribution network; however, suffers in accuracy since the mutual inductance terms between all other parallel wires are neglected. Enhanced accuracy in estimating the mutual inductance terms is required.

The effective inductance of an arbitrary pair of power and ground lines m within an interdigitated P/G distribution network is presented in Fig. 36.3, and is

$$L_m = 2L_{m_s} - 2M_{m_p m_g} + \sum_{\substack{i=1 \\ i \neq m}}^{n} \left(M_{m_p i_p} - M_{m_p i_g} - M_{m_g i_p} + M_{m_g i_g}\right). \tag{36.10}$$

The terms $M_{m_p i_p} = M_{m_g i_g}$ are equal for any i in (36.10) since the distance between the power lines of pair m and i and the ground lines of pair m and i is the same. In addition, (36.10) can be rewritten as a function of distance $d = w + s$, where s is the spacing.

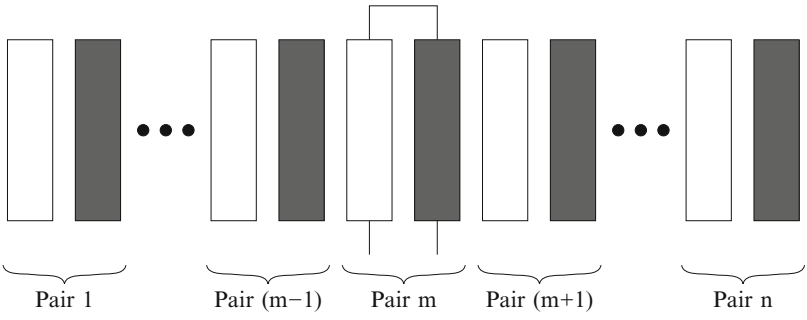


Fig. 36.3 n pairs of an interdigitated P/G distribution network. The focus of (36.10) is on the effective inductance of pair m

$$L_m = 2L_{m_s} - 2M(d) + \sum_{\substack{i=1 \\ i \neq m \\ k=|m-i|}}^{n} \left[2M\left(2dk\right) - M\left(2dk - d\right) - M\left(2dk + d\right)\right], \tag{36.11}$$

where

$$M(x) = \frac{\mu_0 l}{2\pi}\left[ln\left(\frac{2l}{x}\right) - 1\right]. \tag{36.12}$$

Equation (36.11) consists of three terms: the self-inductance of two wires, the mutual inductance between these two wires, and the sum of the mutual inductances between all of the other wires. The third term is neglected in [73]. Substituting (36.12) into (36.11), the summation term is

$$\sum M = \sum_{\substack{i=1 \\ i \neq m \\ k=|m-i|}}^{n} \frac{\mu_0 l}{2\pi}\left[2ln\left(\frac{2l}{2dk}\right) - ln\left(\frac{2l}{2dk - d}\right) - ln\left(\frac{2l}{2dk + d}\right)\right]. \tag{36.13}$$

The sum of the logarithmic terms is the product of a single logarithm, permitting (36.13) to be expressed as

$$\begin{aligned}\sum M &= \sum_{\substack{i=1 \\ i \neq m \\ k=|m-i|}}^{n} \frac{\mu_0 l}{2\pi} ln\left(\frac{2l}{2dk}\frac{2l}{2dk}\frac{2dk - d}{2l}\frac{2dk + d}{2l}\right) \\ &= \sum_{\substack{i=1 \\ i \neq m \\ k=|m-i|}}^{n} \frac{\mu_0 l}{2\pi} ln\left(\frac{2dk - d}{2dk}\frac{2dk + d}{2dk}\right) \\ &= \frac{\mu_0 l}{2\pi} \sum_{\substack{i=1 \\ i \neq m \\ k=|m-i|}}^{n} ln\left(\frac{2k - 1}{2k}\frac{2k + 1}{2k}\right).\end{aligned} \tag{36.14}$$

P/G distribution networks typically consist of a large number of interdigitated pairs and, as shown in Fig. 36.4, the terms of (36.14) quickly decline in magnitude to zero. The number of pairs on the left (and right) is therefore assumed to be infinite, permitting (36.14) to be formulated as

$$\sum M = \frac{\mu_0 l}{2\pi} lim_{n \to \infty}\left[2\sum_{k=1}^{n} ln\left(\frac{2k - 1}{2k}\frac{2k + 1}{2k}\right)\right]. \tag{36.15}$$

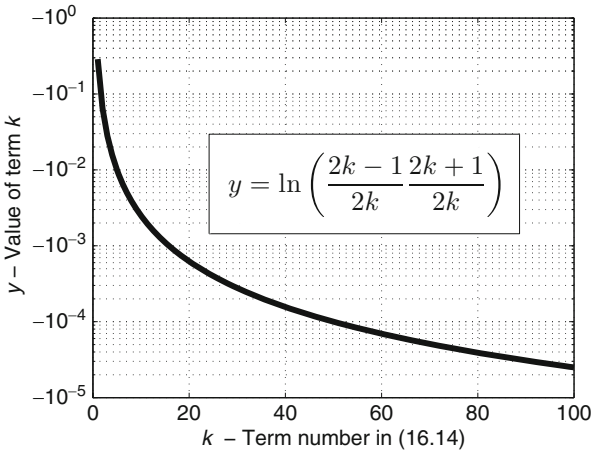


Fig. 36.4 Terms of (36.14). The values quickly decline in magnitude to zero

The factor of two originates from the two sides of the target pair. The infinite sum of (36.15) is presented as an infinite product,

$$\begin{aligned}\sum M &= 2\frac{\mu_0 l}{2\pi} ln\left[lim_{n\to\infty}\prod_{k=1}^{n}\left(\frac{2k-1}{2k}\frac{2k+1}{2k}\right)\right] \\ &= 2\frac{\mu_0 l}{2\pi} ln\left[lim_{n\to\infty}\prod_{k=1}^{n}\left(1-\frac{1}{(2k)^2}\right)\right].\end{aligned} \tag{36.16}$$

The limit of the product can be solved using the Wallis formula [576],

$$\frac{sin(x)}{x} = \prod_{n=1}^{\infty}\left(1-\frac{x^2}{\pi^2 n^2}\right), \tag{36.17}$$

at $x = \pi/2$, leading to the equality,

$$lim_{n\to\infty}\prod_{k=1}^{n}\left(1-\frac{1}{(2k)^2}\right) = \frac{2}{\pi}. \tag{36.18}$$

Based on (36.18), (36.10) may be presented in closed-form,

$$\begin{aligned}L_m &= 2\frac{\mu_0 l}{2\pi}\left[ln\left(\frac{2l}{w+t}\right)+\frac{1}{2}-ln\left(\frac{2l}{d}\right)+1+ln\left(\frac{2}{\pi}\right)\right] \\ &= 2\frac{\mu_0 l}{2\pi}\left[ln\left(\frac{d}{w+t}\right)+\frac{3}{2}+ln\left(\frac{2}{\pi}\right)\right].\end{aligned} \tag{36.19}$$

To estimate the overall inductance of a structure with N power and ground line pairs, the inductance of each pair is assumed to be equal. The mutual inductance between all of the other P/G pairs converges to a constant, making the inductance independent of the number of P/G pairs. The error is greatest in those cases where the number of pairs is smallest; however, in these cases, the effective inductance can be determined quickly with no approximation due to the small number of pairs. For those cases where the number of pairs is sufficiently large (eight pairs produce less than 10 % error), the effective inductance is

$$L_{eff} = \frac{2}{N}\frac{\mu_0 l}{2\pi}\left[ln\left(\frac{d}{w+t}\right) + \frac{3}{2} + ln\left(\frac{2}{\pi}\right)\right]. \tag{36.20}$$

Note than the effective inductance is described for a single layer within an interdigitated P/G network, where it is assumed that no metal layers are above or below the structure. In practical cases, the existence of different interconnect structures above or below the structure may reduce the accuracy of the model. For structures with large spacing, an interconnect structure below the target structure reduces the accuracy of the estimated inductance [577]. Interdigitated P/G networks, however, are designed with small spacing to exploit the available metal resources; therefore, the accuracy of the effective inductance model is maintained.

Additionally, since the current is assumed to flow throughout the entire interdigitated structure, the inductance determined in (36.20) represents the worst case effective inductance. Assuming that current is uniformly distributed throughout the interdigitated structure, the worst case effective inductance produces the largest voltage drop over the power and ground distribution network.

36.3 Comparison and Discussion

Three different models are compared in this section. The Grover model describes the inductance of each pair based on (36.10), where every mutual component is individually calculated [575]. While the individual inductance of each pair is determined, the effective inductance of a single layer within an interdigitated P/G network structure is estimated assuming the individual inductive lines are in parallel. Hence, the Grover model refers to the evaluation of every mutual term among all of the wires in a system. In [73], the effective inductance is determined based on an approximation, where the inductance is treated as a local effect, and the mutual inductance between other pairs is neglected. This model is called the Mezhiba model. The model, represented by (36.20), determines the effective inductance assuming the number of P/G pairs is infinite and named here the Jakushokas model. Since the magnitude of the mutual terms quickly declines to zero as a function of distance, this assumption is highly accurate.

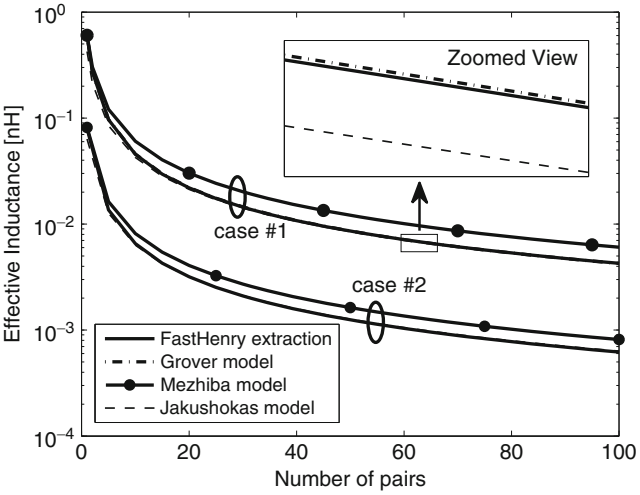


Fig. 36.5 Comparison of FastHenry, Grover, Mezhiba, and Jakushokas models for two different design cases

A comparison among FastHenry [70], a multipole 3-D inductance extraction program, the Grover, Mezhiba, and Jakushokas models, is summarized in this section. In addition, the complexity and accuracy of the Grover, Mezhiba, and Jakushokas models are compared. Two different structures of an interdigitated P/G distribution network are evaluated. For both structures, the width and spacing are maintained constant, $w = 1\ \mu\text{m}$ and $s = 1\ \mu\text{m}$; however, the length and thickness are different, $l_1 = 1\,\text{mm}$, $t_1 = 0.975\ \mu\text{m}$, and $l_2 = 100\ \mu\text{m}$, $t_2 = 0.17\ \mu\text{m}$. The thickness is based on a 65 nm CMOS technology [566] for the top (M8) and bottom (M1) metal layers. Both cases represent a single layer within a P/G distribution network. In Fig. 36.5, two structures are extracted using FastHenry, and compared to the Grover, Mezhiba, and Jakushokas models. The Grover and Jakushokas models exhibit enhanced accuracy as compared to the Mezhiba model. In Figs. 36.6 and 36.7, respectively, the accuracy and complexity are evaluated. The accuracy is evaluated by comparing the results with FastHenry. The complexity of FastHenry is, however, not evaluated since the required number of terms for the simulator is excessively large as compared to the analytic model.

The complexity and error of the Grover, Mezhiba, and Jakushokas models relative to FastHenry are evaluated. The Grover model considers all of the mutual terms and exhibits the lowest error (less than 1 % error); however, the complexity of the Grover model drastically increases for a large number of pairs. The complexity of the Mezhiba and Jakushokas models is independent of the number of power and ground pairs. The error of the Jakushokas model decreases with a larger number of P/G pairs, while the error of the Mezhiba model increases. The highest error ($\sim$30 %) of the Jakushokas model occurs with the fewest number of pairs, while the error of the Mezhiba model is highest with the greatest number of P/G pairs.

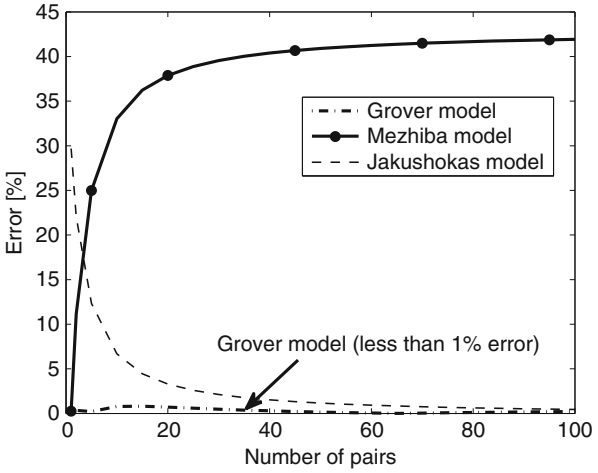


Fig. 36.6 Error comparison for the Grover, Mezhiba, and Jakushokas models. All of the models are compared to FastHenry

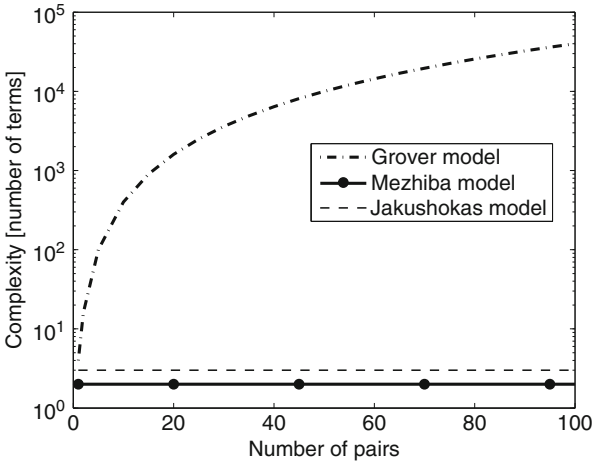


Fig. 36.7 Comparison of complexity of the Grover, Mezhiba, and Jakushokas models

Hence, the error of the Jakushokas model can be reduced using the Grover model, which is only computationally efficient for a few P/G pairs. Assuming the number of power and ground pairs is infinite, the effect of the mutual inductance terms is greater; therefore, the Jakushokas model underestimates the effective inductance. The mutual inductance of only a single pair is considered by the Mezhiba model, overestimating the inductance. The boundary conditions of the effective inductance are determined from the Mezhiba and Jakushokas models. These conditions permit the effective inductance of a single layer within an interdigitated P/G distribution

network structure to be determined for any number of power and ground line pairs. The boundary conditions for the effective inductance of an interdigitated P/G distribution network structure are therefore determined by the Jakushokas and Mezhiba models,

$$\frac{2}{N}\frac{\mu_0 l}{2\pi}\left[ln\left(\frac{d}{w+t}\right)+\frac{3}{2}+ln\left(\frac{2}{\pi}\right)\right] \leq L_{eff}(x) \leq \frac{2}{N}\frac{\mu_0 l}{2\pi}\left[ln\left(\frac{d}{w+t}\right)+\frac{3}{2}\right], \tag{36.21}$$

where x represents any number of pairs within a single layer of an interdigitated P/G distribution network. An expression is derived for the error between the Jakushokas and Grover models. The normalized error is

$$error = \left|\frac{L_{\text{Grover}} - L_{\text{Jakushokas}}}{L_{\text{Grover}}}\right| = \left|1 - \frac{L_{\text{Jakushokas}}}{L_{\text{Grover}}}\right|. \tag{36.22}$$

Since the Grover model cannot be expressed by a single equation, only the worst case error is determined. An assumption in the Jakushokas model is that the number of interdigitated pairs is infinite; therefore, the error is highest when only a single pair ($N = 1$) is present, expressing $error_{N=n} \leq error_{N=1}$. For this case, the inductance based on the Grover model is

$$L_{\text{Grover}(N=1)} = 2L_s - 2M = \frac{2\mu_0 l}{2\pi}\left[ln\left(\frac{d}{w+t}\right)+\frac{3}{2}\right]. \tag{36.23}$$

The inductance based on the Jakushokas model for $N = 1$ is

$$L_{\text{Jakushokas}(N=1)} = \frac{2\mu_0 l}{2\pi}\left[ln\left(\frac{d}{w+t}\right)+\frac{3}{2}+ln\left(\frac{2}{\pi}\right)\right]. \tag{36.24}$$

Substituting (36.23) and (36.24) into (36.22), the error bound is

$$ErrorBound_{N\geq 1} = \frac{ln\left(\frac{\pi}{2}\right)}{ln\left(\frac{d}{w+t}\right)+\frac{3}{2}}. \tag{36.25}$$

Based on the parameters of width, spacing, and thickness provided earlier in this chapter, the *ErrorBound* is less than 0.3 or 30 %. The error of the Jakushokas model drastically decreases with higher number of pairs, as shown in Fig. 36.6. Similarly, the *ErrorBound* can be expressed for those cases where $N \geq 2$,

$$ErrorBound_{N\geq 2} = \frac{ln\left(\frac{\sqrt{3}}{2}\frac{\pi}{2}\right)}{ln\left(\frac{d}{w+t}\right)+\frac{3}{2}+ln\left(\frac{\sqrt{3}}{2}\right)}. \tag{36.26}$$

The *ErrorBound* is less than 0.23 or 23 % for those cases where $N \geq 2$ with the aforementioned parameters of width, spacing, and thickness.

36.4 Summary

A closed-form expression is described to accurately estimate the inductance (self and mutual) of an interdigitated power and ground network. The primary results are summarized as follows.

- Estimating the inductance of a complex network is a complicated task since every mutual inductance element within a system must be considered
- An interdigitated power and ground distribution network reduces the effective inductance as compared to other P/G distribution network structures
- The closed-form Jakushokas model determines the inductance of an interdigitated P/G network, exhibiting good accuracy
- The error of the Jakushokas model is lower with higher number of interdigitated pairs, leading to less than 10 % error for those cases with more than ten pairs (typical interdigitated P/G networks are composed of 100's of interdigitated pairs)
- The Jakushokas model is compared with other models, demonstrating high accuracy and computational efficiency

Chapter 37
Multi-layer Interdigitated Power Networks

An interdigitated P/G distribution network structure is the most common structure in high complexity integrated circuits. Typically, a few wide lines are replaced by a large number of narrow lines to reduce the effects of inductance [150, 151]. Different P/G structures have been compared in [73], where the interdigitated structure is shown to achieve the greatest reduction in inductance.

An interdigitated P/G distribution structure is typically located on several metal layers. Each layer consists of interdigitated power and ground wires, where the direction of the wires is perpendicular to the direction of the wires in the previous layer, as depicted in Fig. 37.1. With advancements in technology, additional metal layers are provided [24], permitting the dedication of several metal layers to the P/G network. Due to electromigration, the maximum current is limited; therefore, a larger number of metal layers passes higher current to the microelectronic system while not surpassing any electromigration constraints.

The need for efficient P/G networks has been recognized, and several algorithms and techniques to optimize the P/G distribution network have been reported [116, 578]. A routing tool for standard cell circuits to efficiently supply and distribute power has been described in [579]. A typical high complexity IC however includes a variety of circuits, therefore, routing the supply network within a standard cell design flow can produce an ineffective network. To overcome this issue, several algorithms based on different optimization strategies have been developed [580, 581]; however, only the package inductance is considered in [581], neglecting the on-chip inductance. An algorithm based on partitioning the power/ground network into smaller sections is described in [456], where *IR* voltage drops are considered. With more advanced packaging techniques (such as flip-chip), the on-chip inductive noise ($L\ dI/dt$) is also important [114, 129]. To consider on-chip inductance in power/ground networks, a technique to simplify the mesh model of an *RLC* power/ground network is described [582], assuming the loads are treated as identical current sources. The significance of the on-chip inductance within paired and interdigitated power/ground network structures is described in [583], where the

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0_37

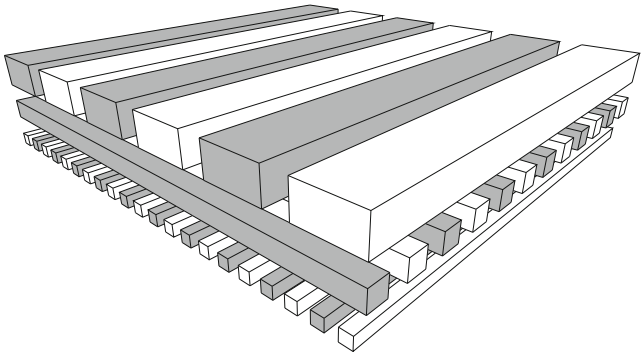


Fig. 37.1 Global interdigitated P/G distribution structure. The *darker* and *lighter lines* represent, respectively, the power and ground lines

inductance is treated as a local effect. In [412], the inductance model considers the mutual inductance between close and distant power/ground wires in interdigitated structures. Based on this model, a closed-form expression characterizing an interdigitated P/G network structure is described, permitting the optimal width of a power/ground network that minimizes the network impedance to be determined. Based on the optimum width of the power/ground lines, a methodology is described in this chapter to minimize the impedance under current density constraints for a multi-layer metal system.

This chapter is organized as follows. A closed-form expression describing the minimum impedance for a single metal layer is presented in Sect. 37.1. In Sect. 37.2, several methods to lower the current density across multiple metal layers are described. Two different approaches are suggested. The tradeoff between the impedance of a P/G network and the current density is presented in Sect. 37.3. This chapter is summarized in Sect. 37.4.

37.1 Single Metal Layer Characteristics

A single layer of a network is depicted in Fig. 37.2, which consists of N number of parallel power and ground wire pairs. The effective inductance of a single metal layer is

$$\frac{1}{L_{eff}} = \frac{1}{L_1} + \frac{1}{L_2} + \frac{1}{L_3} + \ldots + \frac{1}{L_N}, \tag{37.1}$$

where L_1, L_2, L_3, and L_N are, respectively, the effective inductance of the first, second, third, and Nth pair of an interdigitated P/G distribution network. Assuming the current flows in opposite directions in power and ground wires, the effective

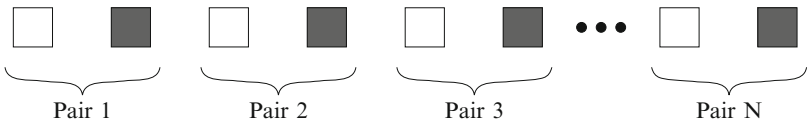


Fig. 37.2 *N* pairs of a single layer within an interdigitated P/G distribution structure

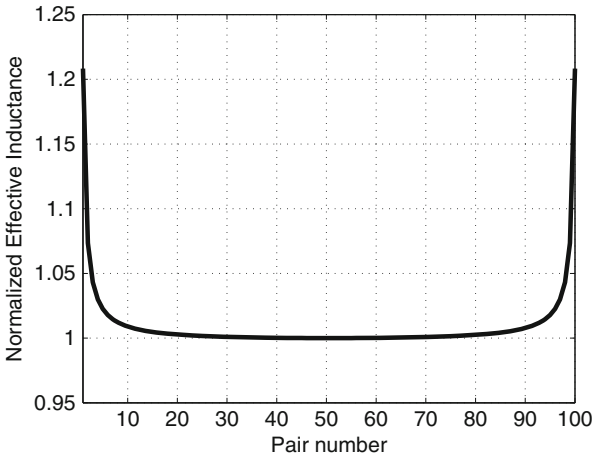


Fig. 37.3 Normalized effective inductance for each pair in a 100 pair interdigitated P/G distribution network

inductance of every pair can be determined based on [575]. In Fig. 37.3, the effective inductance normalized to the lowest effective inductance in a structure is depicted for each pair in a 100-pair interdigitated P/G distribution network.

The difference in inductance is small among all of the pairs, excluding those pairs closest to the boundary. The effect of the boundary is neglected, assuming each of the inductances is equal, permitting the effective inductance of a single layer within an interdigitated P/G distribution network structure to be determined [412]. A similar assumption is considered in [73], neglecting the mutual terms between the pairs, effectively treating the inductance as a local phenomenon. By not neglecting distant mutual effects, the effective inductance can be estimated with higher precision.

A derivation of the effective inductance expression is presented in Chap. 36 (also in [412]), based on the Wallis formula [576], resulting in

$$L_{eff} = \frac{\mu_0 l}{N\pi}\left[ln\left(\frac{w+s}{w+t}\right) + \frac{3}{2} + ln\left(\frac{2}{\pi}\right)\right], \tag{37.2}$$

where N, μ_0, l, w, t, and s are, respectively, the number of power and ground pairs, permeability of the vacuum, length, width, and thickness of a single power or ground wire, and the spacing between the power and ground wires. The mutual inductance

should be considered between all pairs, permitting the accuracy of the effective inductance to be improved by up to 30 %. The accuracy and complexity comparison of (37.2) and other models is provided in [412].

The area A allocated for a P/G network is typically constant,

$$A = l\left[N\left(w + s + w + s\right)\right] = 2lN\left(w + s\right), \tag{37.3}$$

where the first $w + s$ term is the width and space of the power line, and the second $w + s$ term is for the ground line. Substituting (37.3) into (37.2), the effective inductance is

$$L_{eff} = \frac{2l^2\mu_0\left(w + s\right)}{A\pi}\left[ln\left(\frac{w + s}{w + t}\right) + \frac{3}{2} + ln\left(\frac{2}{\pi}\right)\right]. \tag{37.4}$$

Note that (37.4) considers both the self- and mutual inductance, where the mutual inductance is between an infinite number of pairs. This expression requires low computational time while providing high accuracy when estimating the inductance of an interdigitated power and ground distribution network.

The optimal width for minimizing impedance is discussed in the rest of the section. The optimal line width for a single interdigitated metal layer is determined in Sect. 37.1.1. The accuracy of the optimal line width and related issues are discussed in Sect. 37.1.2.

37.1.1 Optimal Width for Minimum Impedance

An interdigitated P/G distribution network is typically allocated over an entire upper metal layer, where the network is designed for lowest impedance. The resistance of a single power and ground pair is

$$R = \rho\frac{2l}{tw}, \tag{37.5}$$

where ρ is the metal resistivity. N pairs of power and ground wires are in parallel; the effective resistance of the network is therefore

$$R_{eff} = \frac{1}{N}\rho\frac{2l}{tw}. \tag{37.6}$$

Substituting (37.3) into (37.6), the effective resistance for constant area is

$$R_{eff} = \frac{4l^2\rho\left(w + s\right)}{Atw}. \tag{37.7}$$

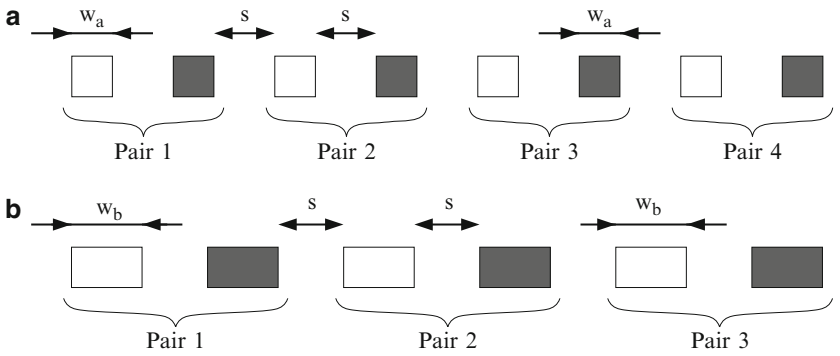


Fig. 37.4 Two different interdigitated power/ground networks are presented for the same physical area. The spacing s is maintained constant. The width w is different for each network, where the width w_a of network (**a**) is thinner than the width w_b of network (**b**). Increasing the width requires fewer interdigitated power/ground line pairs since the area is maintained constant

Under a constant area constraint, two interdigitated networks with different line widths, shown in Fig. 37.4, produce different network impedances. For constant area, according to (37.7), wider power and ground wires reduce the effective resistance. With multiple thin lines, a large area is consumed by the line-to-line spacing, increasing the effective resistance of the network. The inductance under a constant area constraint has the opposite effect since the mutual inductance is dominant in an interdigitated P/G distribution structure. A greater number of lines increases the mutual inductance, reducing the effective inductance, as described by (37.4).

The value of the effective impedance as a function of width (or number of pairs) is

$$Z_{eff}(w) = R_{eff}(w) + j2\pi f L_{eff}(w), \tag{37.8}$$

where f is the target frequency. At 5 GHz, (37.4) and (37.7) are compared to FastHenry, as shown in Fig. 37.5. A constant area of 1×1 mm at the top metal layer for a 65 nm CMOS technology [566] is assumed.

Since the effect of the resistive and inductive impedance behaves inversely with increasing width, the objective is to minimize the overall impedance at a specific frequency. The absolute value of the effective impedance as a function of line width is

$$|Z_{eff}(w)| = \sqrt{R_{eff}^2(w) + 4\pi^2 f^2 L_{eff}^2(w)}, \tag{37.9}$$

as illustrated in Fig. 37.6.

Since the effective inductance in (37.4) is a transcendental function of width, a closed-form analytic solution cannot be determined for the wire width that minimizes the impedance. A closed-form expression can, however, be determined

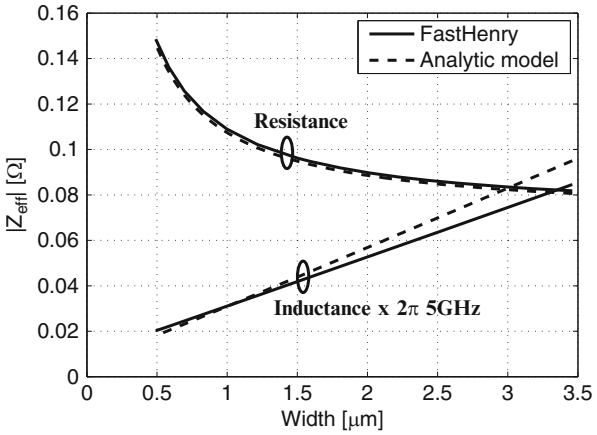


Fig. 37.5 Effective resistance and inductance at 5 GHz as a function of width for a single layer within an interdigitated P/G distribution network. The overall area is maintained constant

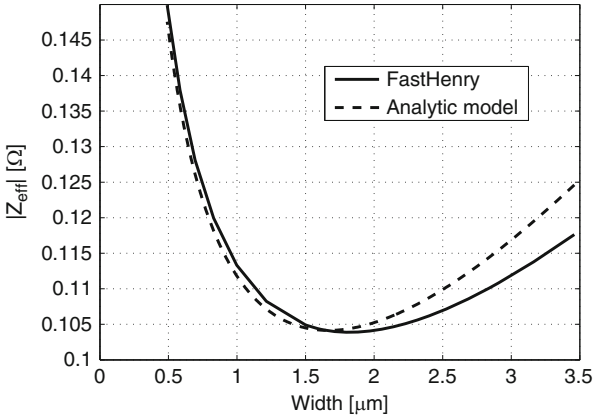


Fig. 37.6 Magnitude of the impedance of (37.9) and FastHenry

for the special case where the line-to-line spacing is equal to the thickness of the metal in the effective inductance equation, resulting in

$$w_{opt}^{(0)} \approx \sqrt[3]{0.91 \frac{s\rho^2}{\mu_o^2 t^2 f^2}}. \tag{37.10}$$

A detailed derivation of (37.10) is presented in the Appendix A. A numerical solution based on n iterations of the Newton–Raphson method is used to determine the optimal width for all other spacings,

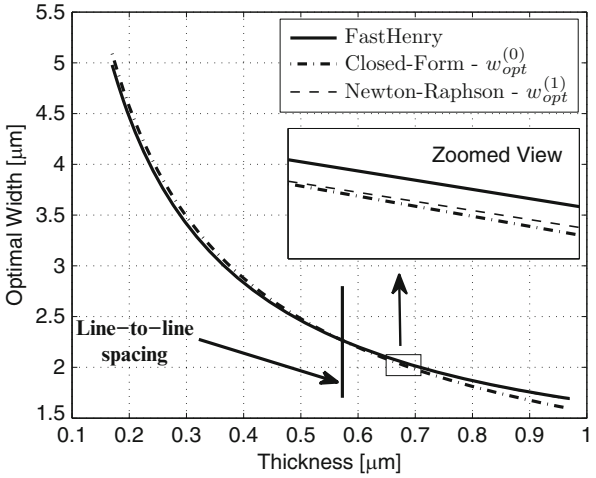


Fig. 37.7 Closed-form $w_{opt}^{(0)}$ and $w_{opt}^{(1)}$ based on the first iteration of the Newton-Raphson method as compared with FastHenry for different thicknesses

$$w_{opt}^{(n)} = w_{opt}^{(n-1)} - \frac{F\prime(w_{opt}^{(n-1)})}{F\prime\prime(w_{opt}^{(n-1)})}, \tag{37.11}$$

where $F \equiv |Z_{eff}(w)|$ and $w_{opt}^{(n-1)}$ is the $(n-1)$st estimate of the optimal wire width. The initial estimate is based on (37.10). The number of iterations can be increased to enhance the accuracy of the optimal width.

Considering the resistance and inductance (both self- and mutual) of a network, (37.10) combined with (37.11) can be used to determine the optimal line width of an interdigitated power/ground network. The optimal line width produces the minimum impedance network at a target frequency.

37.1.2 Optimal Width Characteristics

A comparison among FastHenry, (37.10), and $w_{opt}^{(1)}$ based on the first iteration of the Newton–Raphson method is shown in Fig. 37.7 for several different metal thicknesses. The spacing is chosen as the midpoint between the thinnest and thickest metal layers for a 65 nm CMOS technology. A 5 GHz frequency is assumed. The error between FastHenry and (37.10) reaches 6 %, while the error between FastHenry and $w_{opt}^{(1)}$ is below 1 %. For those cases where the target accuracy is below the error of the initial estimate, the closed-form expression of $w_{opt}^{(0)}$ is computationally efficient in determining the P/G line width. If higher accuracy is required, the interdigitated P/G wire width can be determined according to (37.11).

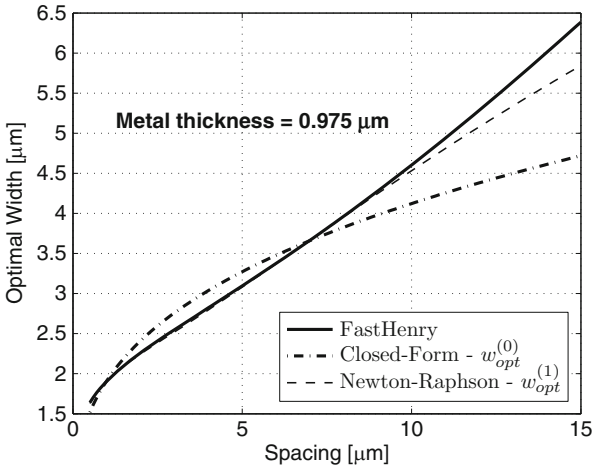


Fig. 37.8 Closed-form $w_{opt}^{(0)}$ and $w_{opt}^{(1)}$ based on the first iteration of the Newton-Raphson method as compared with FastHenry for different spacings

A comparison among FastHenry, (37.10), and $w_{opt}^{(1)}$ based on the first iteration of the Newton–Raphson method is shown in Fig. 37.8 for different spacings. The spacing ranges from 0.54 μm (the lowest permitted spacing) to 15 μm. When the spacing is equal to the thickness, $w_{opt}^{(0)}$ and $w_{opt}^{(1)}$ are equal since the logarithmic term of (37.4) is zero. At spacings below 7 μm, (37.10) exhibits an error below 9 % as compared to FastHenry. For spacings up to 15 μm, the error between FastHenry and (37.10) reaches 26 %, while the error with only the first iteration of the Newton–Raphson method is below 9 %. For spacings greater than 15 μm, additional iterations of the Newton–Raphson method are necessary.

The different number of iterations to determine w_{opt} are evaluated in Fig. 37.9. The error is relative to FastHenry. Note that (37.10) assumes the spacing and thickness are equal. The closed-form expression is therefore only accurate for small spacings. For wider spacings, the Newton–Raphson method is preferred to accurately determine the optimal width. A larger number of iterations is needed for wider spacings since the error decreases significantly for large number of iterations. Spacings up to 100 μm are evaluated, suggesting that four iterations are sufficient to determine the optimal width within 10 % accuracy as compared to FastHenry.

Since the width is optimized for a target frequency, the effect on the frequency range of interest (from DC to the target frequency) is important. In the following discussion, a target frequency of 5 GHz is assumed. In Fig. 37.10, the impedance of the network as a function of frequency is depicted. Three values of the width are evaluated—the optimal width, a width four times greater than the optimal width, and a width four times smaller than the optimal width. Note that the area is maintained constant. An increasing width corresponds to a fewer number of interdigitated pairs within the P/G network (a thinner line corresponds to a higher number of interdigitated pairs).

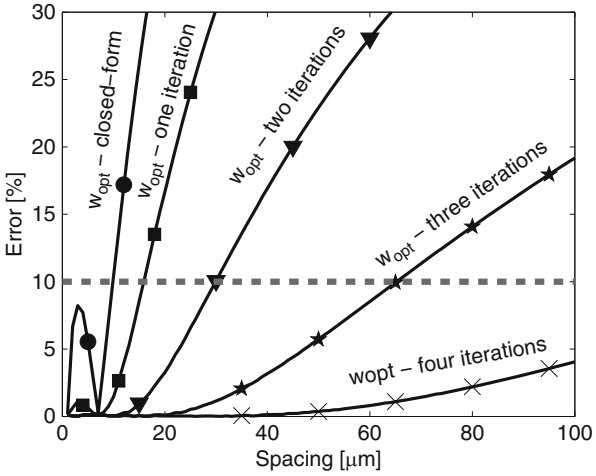


Fig. 37.9 Error of w_{opt} is evaluated for several spacings using closed-form and one to four iterations of the Newton–Raphson method. The error is relative to FastHenry

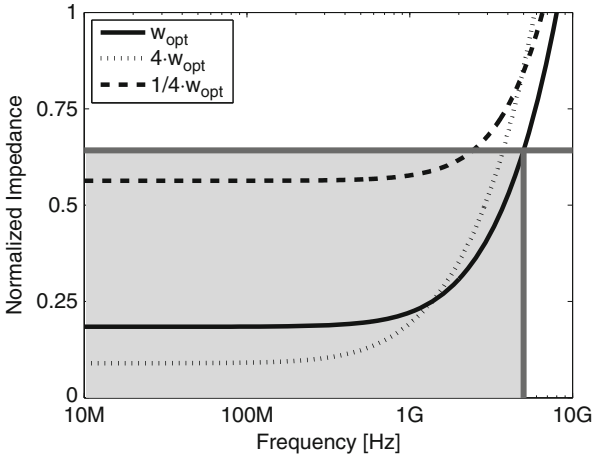


Fig. 37.10 Impedance of a single metal layer for interdigitated power distribution network over the frequency range of interest. Three different P/G network line widths are depicted. The impedance is minimum at the target frequency with the optimum width

As illustrated in Fig. 37.10, the minimum impedance at 5 GHz is achieved using the optimal width, while a lower and higher line width increases, respectively, the resistive and inductive component of the overall impedance. At low frequencies, the P/G network with wider lines produces a lower impedance, although more than the required impedance at the target frequency. In the example shown in Fig. 37.10,

the network impedance with wider lines is below the target impedance only below 3.7 GHz. As depicted in Fig. 37.10, the P/G network with a smaller width satisfies the impedance requirements only up to 2.5 GHz.

37.2 Multi-layer Optimization

Multi-layer systems can be approximated by the network shown in Fig. 37.11, where the resistance and inductance is, respectively, the effective resistance and inductance of a single layer within a P/G distribution network [73]. This model treats the system as worse case since all of the current is assumed to flow through the entire layer. Electromigration is considered when optimizing a multi-layer system.

The current density CD of an arbitrary layer m is

$$CD_m = \frac{|i_m|}{CroSec_m}, \tag{37.12}$$

where i_m and $CroSec_m$ are, respectively, the current and cross section of layer m. The skin effect is considered in determining the cross-section of the layer,

$$CroSec_m = \begin{cases} N_m w_m t_m & 2\delta > w \quad (37.13) \\ N_m w_m t_m & 2\delta > t \quad (37.14) \\ 2\delta\,[N_m(w_m + t_m) - 2\delta] & \text{otherwise}, \quad (37.15) \end{cases}$$

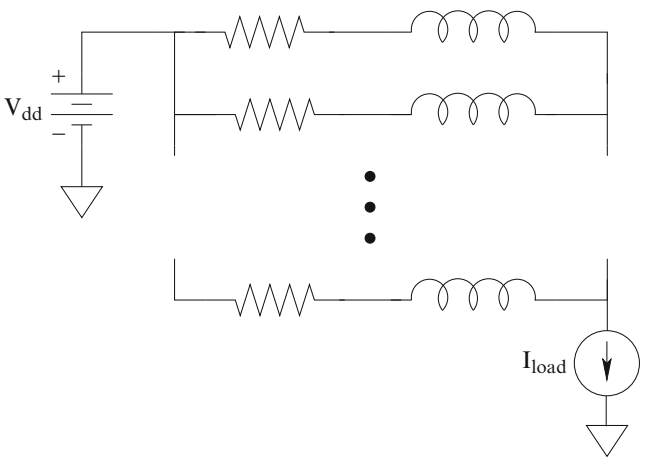


Fig. 37.11 Multi-layer P/G distribution network model. Each resistance and inductance represent, respectively, the effective resistance and inductance of a single layer within a P/G network

where δ is the skin depth. The skin depth is defined as

$$\delta \equiv \sqrt{\frac{1}{\pi f \mu_0 \sigma}}, \tag{37.16}$$

where σ is the conductivity of the material.

Allocating additional metal layers for the power and ground distribution network distributes the overall current among a larger number of metal layers. The current density of a particular metal layer is therefore lower.

Two different approaches are considered for optimizing a multi-layer P/G network. In the first approach, the current density per layer is maintained equal for all of the layers, while providing a low P/G network impedance. The second approach minimizes the impedance, while considering electromigration. A tradeoff exists between the current density and the impedance of a P/G distribution network. A lower impedance reduces the voltage drop, providing a higher noise margin.

37.2.1 First Approach: Equal Current Density

The first optimization approach for an interdigitated P/G distribution network structure is discussed in this section. The limiting current density is the highest current density among the layers. In this approach, the current density among the layers is maintained equal, minimizing the limiting current density of a P/G network. A lower limiting current density enhances the reliability of a multi-layer system. The current flowing through an arbitrary layer m is

$$|i_m| = \frac{V_{drop}}{|Z_m|}, \tag{37.17}$$

where V_{drop} and Z_m are, respectively, the voltage across the entire P/G distribution network and the impedance of the mth layer. Substituting (37.17) into (37.12), the current density of the mth metal layer is

$$CD_m = \frac{V_{drop}}{|Z_m|} \frac{1}{CroSec_m}. \tag{37.18}$$

Two layers, m and n, provide the same current density when

$$|Z_m|\, CroSec_m = |Z_n|\, CroSec_n. \tag{37.19}$$

While the width of a single metal layer is optimized for minimum impedance, the width of the remaining metal layers is chosen to maintain equal current density, as described by (37.19). Pseudo-code for determining the width of the individual metal layers within a multi-layer system based on maintaining equal current density among the metal layers is provided in Appendix B.

Based on the this methodology, an eight layer P/G distribution network is described for a 65 nm CMOS technology [584], where all of the metal layers are available for the P/G distribution, although in practical cases some metal layers are used for the signals, clock network, and shield lines. Based on a 65 nm CMOS technology, a width of 1.66 μm is initially determined from (37.11) for the top (eighth) metal layer to minimize the impedance of a single metal layer. The width of the additional metal layers is based on maintaining equal current density according to (37.19). The current density per multiple metal layers is depicted in Fig. 37.12. Increasing the width of the lower metal layer affects the current density in the lower layer as well as the upper metal layer. Increasing the width of the lower metal layer changes the impedance of the lower metal layer (decreasing the resistance

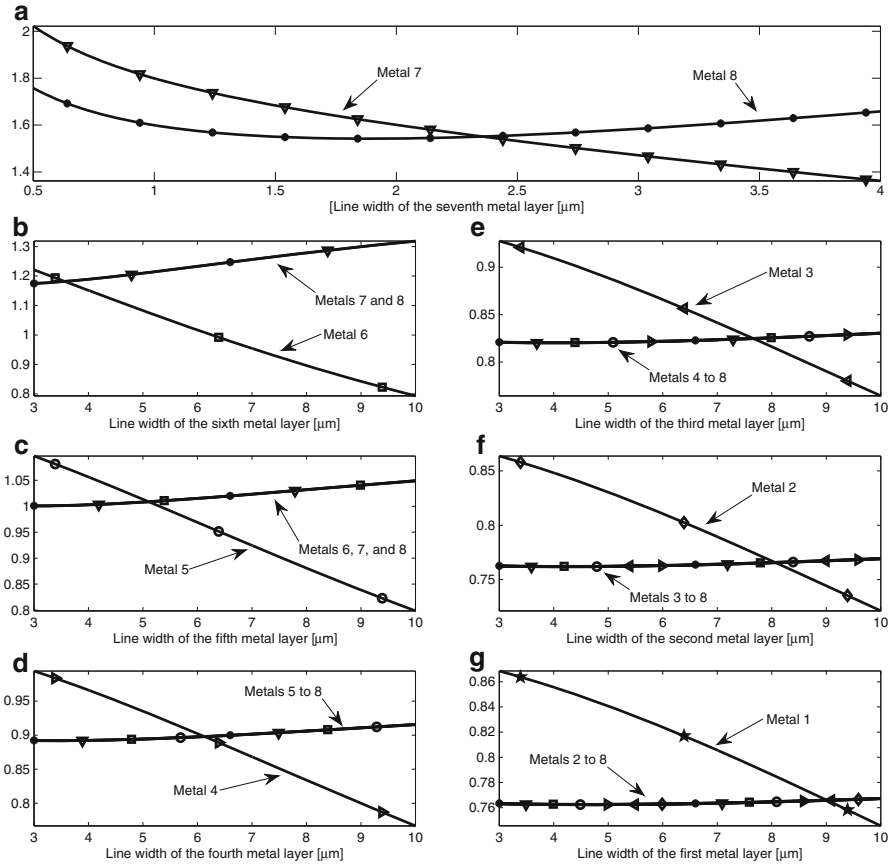


Fig. 37.12 Current density for multiple metal layers; (**a**) seventh and eighth, (**b**) sixth, seventh, and eighth, (**c**) fifth, sixth, seventh, and eighth, (**d**) fourth to eighth, (**e**) third to eighth, (**f**) second to eighth, (**g**) first to eighth. The width is determined at the intersection of the current density of the multiple metal layers. The *y-axis* for each figure is the current density in units of mA/μm^2, while the total current is assumed to be 1 A

Table 37.1 Spacing, thickness, width, and number of interdigitated pairs for an eight metal layer system. The eighth metal layer is the top metal layer. Since the lines are wider, the number of interdigitated pairs is lower for a constant area

Metal layer	Thickness (μm)	Spacing (μm)	Width (μm)	Number of pairs
8	0.975	0.540	1.66	227
7	0.650	0.360	2.36	183
6	0.430	0.240	3.56	131
5	0.300	0.165	5.11	94
4	0.250	0.140	6.13	79
3	0.200	0.110	7.67	64
2	0.190	0.105	8.07	61
1	0.170	0.105	9.02	54

and increasing the inductance). The current is distributed among the different metal layers based on the relative impedance of the metal layers, resulting in larger current in the metal layer with lower impedance while changing the current density in all of the metal layers.

The width is determined at the intersection of the current density of multiple metal layers. The intersection is the width where equal current density among the multiple metal layers is maintained, lowering the limiting current density. As inferred in Fig. 37.12, this intersection occurs at a greater width for the lower metal layers, since the metal layers are thinner. This structure is therefore called the *pyramid* structure. The spacing, thickness, width, and number of interdigitated pairs for each metal layer in an eight layer P/G system is summarized in Table 37.1.

The normalized current density is shown in Fig. 37.13 for an eight metal layer P/G network based on the equal current density methodology. While the maximum current density for a specific technology, physical area, and current is known, the required number of metal layers for an interdigitated P/G distribution network can be determined, as illustrated in Fig. 37.13.

Two additional P/G network structures are compared with the *pyramid* structure. These three structures are illustrated in Fig. 37.14. The width of the individual metal layers for the *pyramid* structure is listed in Table 37.1. This structure is shown in Fig. 37.14a. Note in the *pyramid* structure, the power and ground lines in the lower metal layers are wider. In conventional metal systems, the power and ground lines are wider at the higher metal layers, as illustrated in Fig. 37.14b. For this structure, the width of the metal layers is the opposite of the *pyramid* structure, and is therefore called the *inverted pyramid* (standard) structure. In Fig. 37.14c, the width of each metal layer is maintained constant at 5.5 μm; therefore, this structure is referred to as the *equal width* structure. The width, number of interdigitated pairs, effective impedance, and limiting current density for these three structures are listed in Table 37.2. For the current density evaluation, the metal layers are extracted individually using FastHenry.

In the *pyramid* structure, the current density is maintained equal among the layers, lowering the limiting current density. Since the thickness decreases with lower metal layers, the lines are wider, maintaining a constant current density.

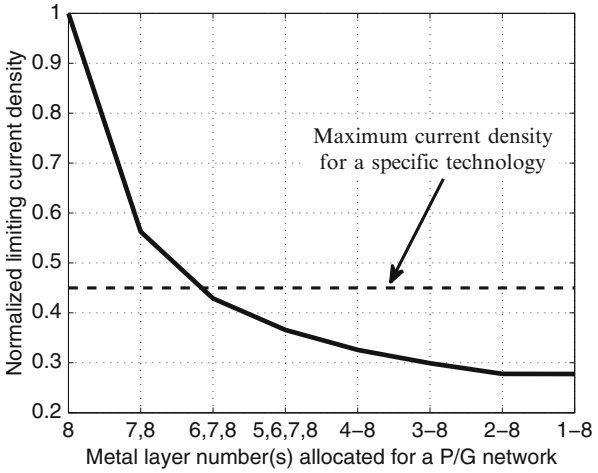


Fig. 37.13 Normalized limiting current density for different metal layers. The *x-axis* represents the metal layer number(s) allocated for a P/G network. The current density is highest when allocating only a single metal layer (layer number eight) for a P/G network. The current density is reduced as additional metal layers are added

Table 37.2 Three structures are compared for equal current density. The thickness, spacing, width, and number of interdigitated pairs per metal layer for each structure are listed

			Pyramid structure		*Inverted pyramid* structure		*Equal width* structure	
Metal layer	Thickness (μm)	Spacing (μm)	Width (μm)	Number of pairs	Width (μm)	Number of pairs	Width (μm)	Number of pairs
8	0.975	0.540	1.7	227	9.0	52	5.5	82
7	0.650	0.360	2.4	183	8.1	59	5.5	85
6	0.430	0.240	3.6	131	7.7	63	5.5	87
5	0.300	0.165	5.1	94	6.1	79	5.5	88
4	0.250	0.140	6.1	79	5.1	95	5.5	88
3	0.200	0.110	7.7	64	3.6	136	5.5	89
2	0.190	0.105	8.1	61	2.4	202	5.5	89
1	0.170	0.105	9.0	54	1.7	280	5.5	89
Effective impedance (mΩ)			30.6		46.0		38.2	
Limiting current density ($mA/\mu m^2$)			0.766		1.400		1.044	

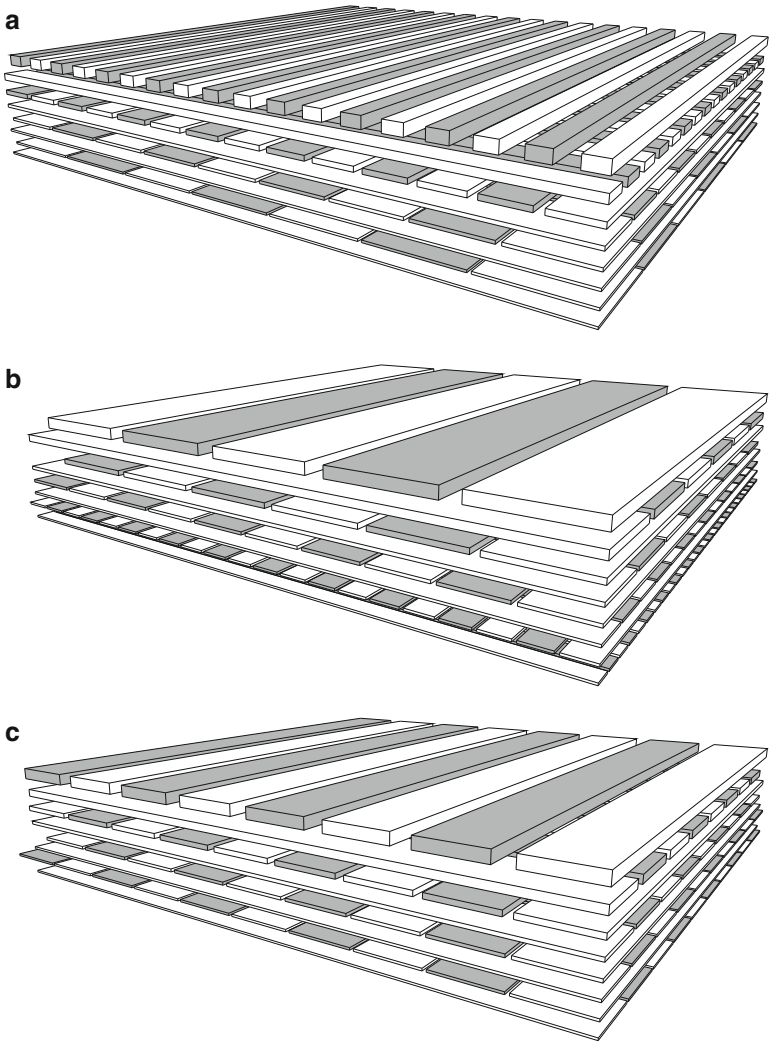


Fig. 37.14 Three P/G structures; (**a**) *pyramid* structure—the width decreases with higher metal layers, (**b**) *inverted pyramid* (standard) structure—the width increases with higher metal layers, (**c**) *equal width* structure—the width is maintained equal among all of the metal layers

In the *inverted pyramid* structure, the higher metal layers are wider, permitting greater routing flexibility; the reliability of the metal, however, decreases since the limiting current density is 82 % higher than the *pyramid* structure. In the *inverted pyramid* structure, most of the current flows in the higher metal layers increasing the effective impedance of the overall system. The impedance is 50 % higher than in the *pyramid* structure. The *equal width* structure exhibits a higher effective impedance and current density of 24 % and 36 %, respectively, as compared to the *pyramid*

structure. This trend is consistent with the change in importance of the inductance as compared to the resistance at higher frequencies.

37.2.2 Second Approach: Minimum Impedance

The focus of the second optimization approach is to minimize the impedance of the overall P/G distribution network. Assuming the metal layers are in parallel while optimizing each layer for minimum impedance, the lowest impedance of the overall system is achieved. The number of required metal layers is based on the current density constraint. Pseudo-code for this optimization algorithm is presented in Appendix C. The impedance of each of the eight metal layers is illustrated in Fig. 37.15.

Three different interdigitated P/G distribution structures, illustrated in Fig. 37.14, are compared. The structures are referred to by the same names as in the previous section, however, the widths are determined based on the minimum impedance algorithm rather than the equal current density algorithm. The width of the power and ground lines for the *pyramid* structure is based on the algorithm presented in Appendix C. In the *inverted pyramid* (standard) structure, the width increases with higher metal layers. The width of the metal layers is the opposite of the *pyramid* structure. The width of all eight metal layers is maintained constant at 2.4 μm for the *equal width* structure. In Table 37.3, the thickness, spacing, width, and number of interdigitated pairs are listed for each structure. The effective impedance and limiting current density for each structure are also summarized in Table 37.3.

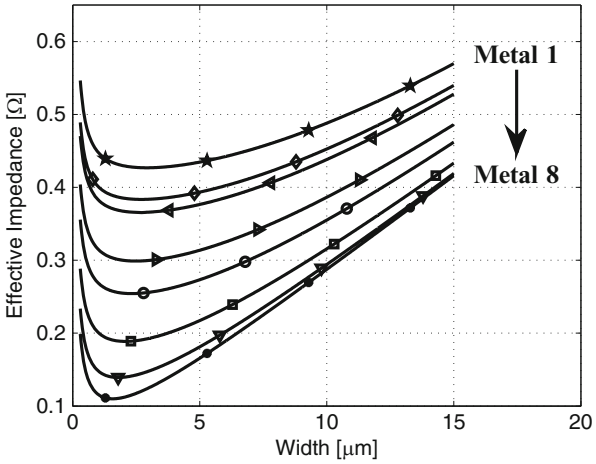


Fig. 37.15 Effective impedance as a function of width for each of eight metal layers within an interdigitated P/G distribution network. The overall area of each metal layer is maintained constant

Table 37.3 Three structures are compared for minimum impedance. The thickness, spacing, width, and number of interdigitated pairs per metal layer for each structure are listed

Metal layer	Thickness (μm)	Spacing (μm)	*Pyramid* structure		*Inverted pyramid* structure		*Equal width* structure	
			Width (μm)	Number of pairs	Width (μm)	Number of pairs	Width (μm)	Number of pairs
8	0.975	0.540	1.7	227	2.9	144	2.4	170
7	0.650	0.360	1.9	225	2.7	162	2.4	181
6	0.430	0.240	2.1	212	2.7	172	2.4	189
5	0.300	0.165	2.3	199	2.5	187	2.4	194
4	0.250	0.140	2.5	189	2.3	201	2.4	196
3	0.200	0.110	2.7	180	2.1	225	2.4	199
2	0.190	0.105	2.7	177	1.9	254	2.4	199
1	0.170	0.105	2.9	165	1.7	283	2.4	199
Effective impedance (mΩ)			29.5		31.5		30.5	
Limiting current density ($mA/\mu m^2$)			0.843		0.909		0.875	

The lowest effective impedance is achieved in the *pyramid* structure. The effective impedance is 6 % and 3 % higher for the *inverted pyramid* and *equal width* structures, respectively, as compared to the *pyramid* structure. The limiting current density in the *pyramid* structure is enhanced, respectively, by 8 % and 4 % as compared to the *inverted pyramid* and *equal width* structures. Hence, the effective impedance achieved by the *pyramid* structure is lower. This improvement is due to the relative importance of the inductance as compared to the resistance in high frequency systems.

37.3 Discussion

The following discussion is divided into four sections: a comparison between the two aforementioned design approaches (Sect. 37.3.1), a discussion of routability and the grid area ratio (Sect. 37.3.2), an estimate of the optimal power/ground line width for different frequencies and technologies (Sect. 37.3.3), and an investigation of the critical frequency in the design of multi-layer power/ground networks (Sect. 37.3.4).

37.3.1 Comparison

Evaluating both approaches, a tradeoff is observed between the impedance (or voltage drop) and the limiting current density of a P/G distribution network. When

Table 37.4 Comparison between two optimization approaches for a one, two, three, and eight metal layer system

Number of metal layers	First approach		Second approach	
	Z_{eff} (mΩ)	Limiting current density (mA/μm^2)	Z_{eff} (mΩ)	Limiting current density (mA/μm^2)
1	105.1	2.71	105.1	2.71
2	59.5	1.54	59.4	1.60
3	45.6	1.18	45.2	1.25
8	30.6	0.77	29.5	0.84

focusing only on the current density, the optimal solution suggests the P/G network should be as wide as possible; however, at high frequencies, the effective impedance increases significantly due to the higher inductance. Both approaches consider the effective impedance and current density, while the current density is the focus of the first approach, and the effective impedance is the focus of the second approach. A comparison between both approaches for a one, two, three, and eight metal layer system is listed in Table 37.4.

As observed from Table 37.4, the effective impedance is lowest using the second approach, while the limiting current density is lowest if the first approach is used. A tradeoff between the limiting current density and effective impedance is noted. A difference of less than 10 % between the two approaches for the impedance and current density is demonstrated. However, when additional constraints (such as routability) are considered, the optimal width may not be available for that particular layer of metal within the power/ground distribution network. In this situation, the optimization process is focused on minimizing the impedance or current density, resulting in a greater difference between the two approaches. These two approaches are presented here to satisfy both optimization flows.

37.3.2 Routability

To develop the methodology, these examples assume all of the metal layers and physical area can be used for the power/ground network. For a practical on-chip power and ground distribution network, routability, cost, and other issues should also be considered. Routability is an important issue primarily affecting the lower metal layers. Global power/ground networks tend to utilize the higher metal layers. To consider routability, a metric, the grid area ratio, is introduced as the ratio of the metal resources occupied by the power/ground network to the total metal area [299, 583],

$$A \equiv \frac{w + s_0}{p}, \tag{37.20}$$

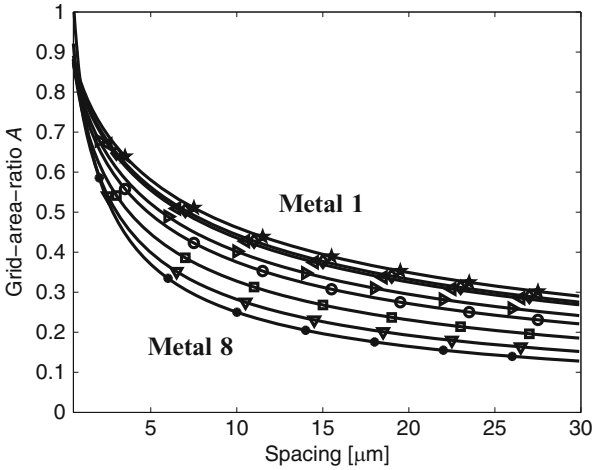


Fig. 37.16 Grid area ratio as a function of spacing between the lines for different metal layers. The line width is based on (37.11) to minimize the network impedance

where w, s_o, and p are, respectively, the line width, minimal spacing between power/ground lines, and the line pitch. The line pitch is the width and spacing between the power and ground lines. If the spacing between the power and ground lines is minimum, the grid area ratio is one. The grid area ratio is depicted in Fig. 37.16 for different spacings. As anticipated, increasing the distance between the lines reduces the grid area ratio. As illustrated in Fig. 37.16, the grid area ratio is higher for the lower metal layers, resulting in reduced routability for the lower metal layers as compared to the higher metal layers where routability is better.

In Fig. 37.17, several P/G networks with different line widths between the top and bottom metal layers are evaluated. A 5 GHz target frequency and 10 μm line-to-line spacing between the power and ground lines is chosen. Four interdigitated metal layers are allocated for the power network. In Fig. 37.17, the x-axis is w_{top}/w_{bottom}, permitting a comparison between the impedance and grid area ratio for several *pyramid*, *equal width*, and *inverted pyramid* structures. The vertical line at $w_{top}/w_{bottom} = 1$ represents the *equal width* structure. The region to the left of the *equal width* structure represents *pyramid* structures with increasing width at the bottom metal layers and decreasing width at the top layers. The region to the right represents *inverted pyramid* structures with decreasing width at the bottom metal layers and increasing width at the top layers. The lowest impedance among these structures is the *pyramid* structure with a line width based on (37.11). The grid area ratio however is lower in the *inverted pyramid* structure, indicating a tradeoff between the impedance and routability. The primary disadvantage of the *pyramid* structure is therefore a higher grid area ratio (lower routability) as compared to the conventional *inverted pyramid* structure. A power network to the right of the minimum impedance *pyramid* structure may therefore be a reasonable compromise to provide effective routability while tolerating a reasonable increase in network impedance.

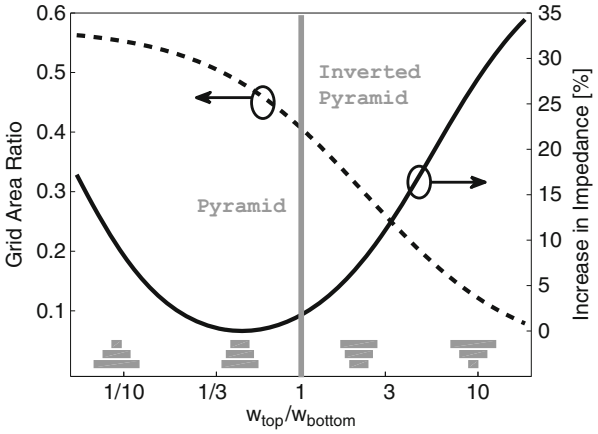


Fig. 37.17 Grid area ratio and increase in impedance for several interdigitated P/G structures. Four metal layers are allocated for the power network. The vertical line represents the *equal width* structure. The left region is for *pyramid* structures, while the right region is for *inverted pyramid* structures. The minimum impedance is achieved by the *pyramid* structure, where the line width is based on (37.11)

37.3.3 Fidelity

The required number of metal layers for the specified power levels is depicted in Fig. 37.18. The technology parameters are chosen based on a 65 nm CMOS technology with an area of 1×1 mm. The results are evaluated at three different frequencies, indicating that an additional metal layer is required at higher frequencies.

The optimal width as a function of the number of metal layers at 3 and 10 GHz is illustrated in Fig. 37.19 for a 65 nm, 45 nm [585], and 32 nm [586] CMOS technology. The optimal width is determined based on the Newton-Raphson method as described by (37.11). At higher frequencies, the optimal width is thinner since the inductive impedance is greater. The optimal width increases with thinner, less inductive metal layers to satisfy the minimum impedance constraint. With technology scaling, metal thicknesses typically decrease, requiring wider lines to compensate for the increase in resistivity.

The effect of frequency on the design of an interdigitated P/G distribution network is significant. At lower frequencies, where the resistive impedance is dominant, wide wires are used to reduce the impedance, while maintaining a constant cross-section to satisfy equal current density. At higher frequencies, the inductive impedance is dominant, suggesting that the power/ground lines should be less wide.

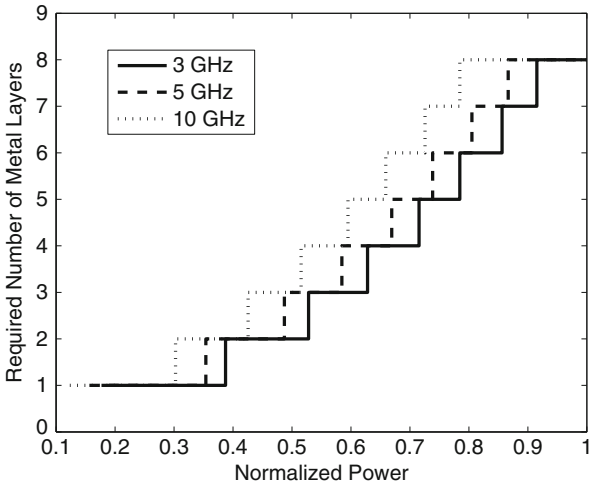


Fig. 37.18 Required number of metal layers for a P/G network as a function of normalized power evaluated at three different frequencies

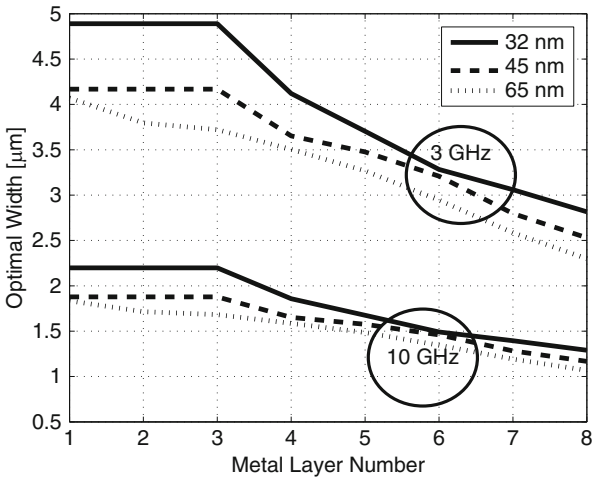


Fig. 37.19 Optimal width to minimize the effective impedance of each metal layer based on a 65, 45, and 32 nm CMOS technology for two different frequencies

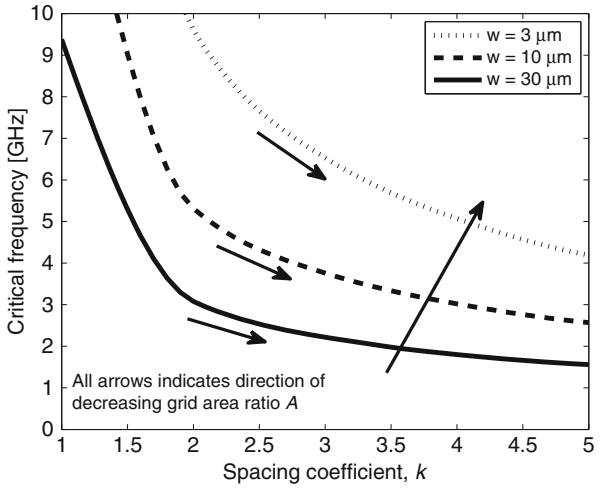


Fig. 37.20 Critical frequency at which the impedance of the higher metal layer is equal to the impedance of the lower metal layer. The width of the power/ground lines is maintained equal for both metal layers. The distance between the lines is the minimum spacing of each layer multiplied by a spacing coefficient k

37.3.4 Critical Frequency

The relationship between the two highest metal layers is evaluated, permitting the concept of a critical frequency to be defined. The critical frequency is described here as the frequency at which the impedance of two metal layers is equal. Assuming the width of both metal layers is the same, the critical frequency is determined for a variety of spacings, as depicted in Fig. 37.20. The arrows indicate the direction of decreasing grid area ratio A (increasing routability). The critical frequency is evaluated for three different line widths.

The critical frequency for different grid area ratios is illustrated in Fig. 37.21. A line width of 10 µm is assumed. The area above the target frequency is the region where the higher metal layer is dominant (the impedance of the higher metal layer is greater), while the area below the target frequency is the region where the lower metal layer is dominant (the impedance of the lower layer is greater). From Fig. 37.21, the higher metal layer is the dominant metal layer in terms of the impedance for signal frequencies above 3 GHz (for high routability) and 16 GHz (for low routability), assuming a 10 µm line width for both metal layers.

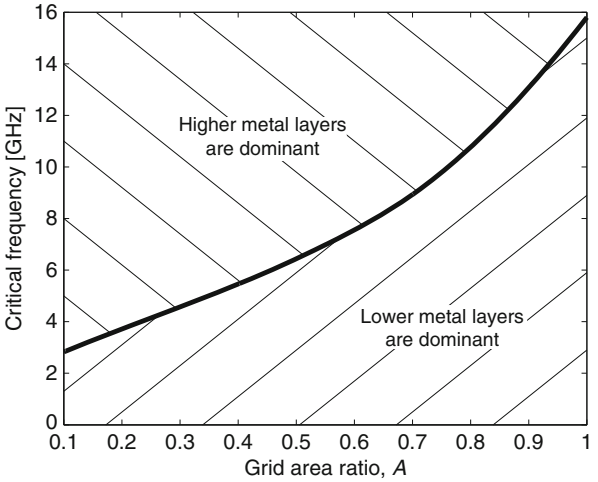


Fig. 37.21 Critical frequency as a function of grid area ratio. A line width of 10 μm is assumed for all power/ground lines. The area above the line indicates the region where the higher metal layer is dominant, while the region below the line indicates the region where the lower metal layer is dominant

37.4 Summary

A multi-layer interdigitated power and ground network is evaluated. The primary results are summarized as follows.

- A single interdigitated metal layer is explored under a constant area constraint, determining the optimal width of the power and ground lines that minimize the network impedance
- A closed-form expression for the optimum width, providing comparable accuracy to FastHenry, is described. The optimum width is determined for each metal layer based on 65, 45, and 32 nm CMOS technologies
- A *pyramid* shaped power/ground network structure is described for a multi-layer metal system
- Two approaches for designing multi-layer interdigitated distribution networks are presented
- With the first approach, the impedance of each metal layer is minimized, providing the lowest effective impedance of the overall power network
- With the second approach, the current density of each metal layer is maintained equal, providing the highest reliability of the overall power network
- Both multi-layer design approaches are compared among the *pyramid*, *inverted pyramid* (standard), and *equal width* P/G structures, where the lowest effective impedance and highest limiting current are demonstrated for the *pyramid* structure

- The frequency at which the higher metal layers is more dominant than the lower metal layers is determined under different routability constraints
- Several *pyramid*, *equal width*, and *inverted pyramid* structures are compared in terms of the impedance and grid area ratio, indicating a tradeoff between the impedance and routability of the network

Chapter 38
Globally Integrated Power and Clock Distribution Networks

Further increases in the density and performance of ICs require more complicated global interconnects, such as power, ground, and clock networks. On-chip metal resources are limited [299], further constraining the design of the global interconnect networks. These global networks require a significant portion of the overall metal resources [24].

The major networks, power, ground, and clock, consume most of the higher level metal resources. Each network is carefully designed to provide optimal circuit performance. These networks are typically designed independently of each other (power/ground network and clock network) since each network has different physical and electrical characteristics and constraints. These on-chip resources are highly utilized since each network is typically routed to every individual block, circuit, or gate within an IC.

For those circuits where the clock signal is generated on-chip, the clock and power signals may be combined onto a single layer to eliminate the on-chip global clock distribution network. This integrated network is named here as a *globally integrated power and clock* (GIPAC) distribution network. The power and clock signals are later separated from the GIPAC network into two different local networks using passive filters.

Signal splitting is fairly common in electrical and communications systems. Different signals are modulated, simultaneously transferring these signals over a single medium, and later demodulated [587]. A typical home phone system carries power and voice signals over the same network. Filters inside the device demultiplex the signals, routing the signals accordingly. The internet infrastructure also uses phone lines, sharing resources with the global phone network. Broadband communication over the power lines [588] further supports this approach of sharing a common global network.

The two signals, power and clock, are fundamentally different in nature. A primary difference between the power and clock signals is the operating frequency. While the clock signal exhibits a high frequency component, the power signal is

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0_38

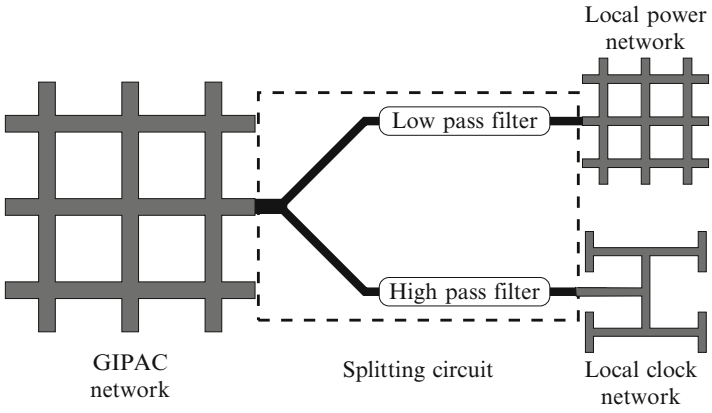


Fig. 38.1 Globally integrated power and clock distribution network. Low and high pass filters can be used to separate the GIPAC signal into local power and local clock signals

Table 38.1 Characteristics of power and clock signals

	Power signal	Clock signal
Frequency	Very low	High
Current	Very high	Low
Load	Very low resistance	Highly capacitive

ideally DC. These two signals can therefore be separated with high and low pass filters, as illustrated in Fig. 38.1. Additionally, the power signal carries high current, while high current is not necessary for the clock signal. Different characteristics of these two signals, listed in Table 38.1, distinguish the design of the high and low pass filters (the splitting circuit).

Previous research on power and clock networks has typically considered a single network at a time. Focusing on the power networks, the network impedance is a primary issue [136, 299]. For clock networks, the focus is typically on power, skew, and jitter [38]. In [589], the clock and power distribution networks are simultaneously considered to enhance immunity to power supply noise. However, no research on utilizing the same global network to distribute both power and clock has been described in the literature. This novel concept of combining these global networks is the focus of this chapter.

The chapter is organized as follows. The high level topology and related issues are discussed in Sect. 38.1. In Sect. 38.2, the strategy and related circuits for separating the power and clock signals are described. The circuit is evaluated in Sect. 38.3, and summarized in Sect. 38.4.

38.1 High Level Topology

The GIPAC structure is shown here to efficiently distribute power and clock within an SoC. In Fig. 38.2, the GIPAC network is depicted within an SoC, where multiple on-chip domains are characterized as a local network and the entire IC-based system as a global network. In this case, the GIPAC network distributes the integrated power and clock network over the entire circuit, while localized systems produce separate local power and clock networks. The GIPAC structure lowers the requirement for metal resources, supporting higher integration and functionality.

Noise is a primary issue in the design of a GIPAC network. The noise originates from three primary sources, as illustrated in Fig. 38.3:

1. Noise from the GIPAC into the power network
2. Noise from the GIPAC into the clock network
3. Noise from the power network into the clock network

Since the GIPAC network combines the clock and power signals, a fraction of the clock signal propagates through the low pass filter as noise into the power network. Two strategies exist for reducing these noise sources, a more effective low pass filter or a higher frequency clock signal. The disadvantage in a more effective low pass filter is increased area. The disadvantage of providing a higher clock frequency is the requirement for higher speed, thereby dissipating greater power.

Noise within the GIPAC also affects the clock network; however, only high frequency noise is significant since a high pass filter eliminates low frequency noise. The remaining noise, however, produces jitter in the clock signal.

The third noise path, noise propagation from the power network into the clock network, also needs to be considered. The on-chip circuitry powered by the power network switches at the same frequency as the clock signal, injecting high frequency

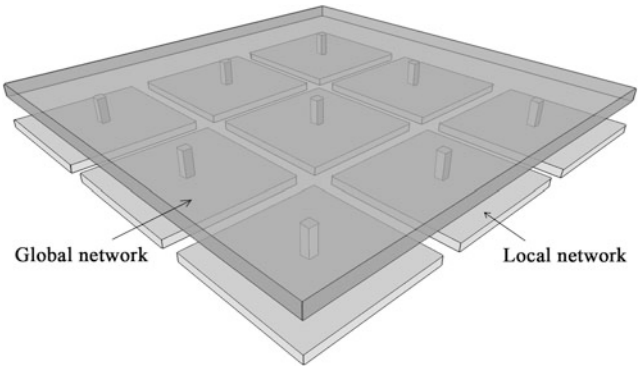


Fig. 38.2 Integrating GIPAC network within an SoC. The GIPAC network is represented by the top layer, while the local separate power and clock networks are located on the bottom layer

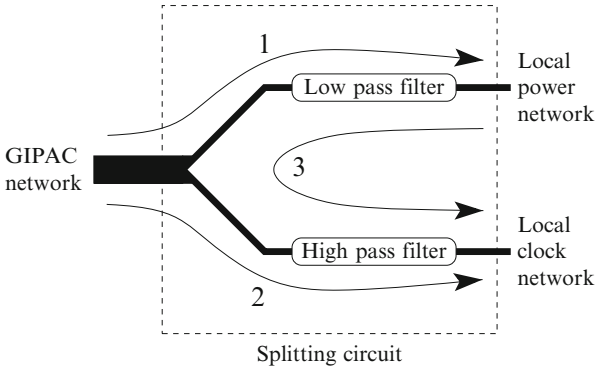


Fig. 38.3 Noise is a major issue for the GIPAC network. The three noise paths are shown. The first path represents noise coupling from the GIPAC network into the local power network, the second path indicates noise coupling from the GIPAC network into the local clock network, and the third path is noise injected from the local power network into the clock network

noise from the power network into the clock network. The different current demands from the power network also affect the clock network. A solution is therefore required to eliminate this noise mechanism.

38.2 GIPAC Splitting Circuit

The GIPAC splitting circuit, called here a splitter, is the primary component of the integrated power and clock network system. The function of the splitter is to separate the clock and power signals using low and high pass *RC* filters, while minimizing the noise between the global and local networks. Generation of the power and clock signals is described in Sect. 38.2.1. The choice of *RC* filter is discussed in Sect. 38.2.2.

38.2.1 Mathematical Perspective

A GIPAC splitter circuit is shown in Fig. 38.4. Each *RC* pair behaves as a low pass filter, where the resistor is a polysilicon resistor and the capacitor is an MOS transistor. The input signal supplied to the GIPAC network as a function of time t is

$$input = A + \alpha \cdot \sin(2\pi f_{clk} t)\,, \tag{38.1}$$

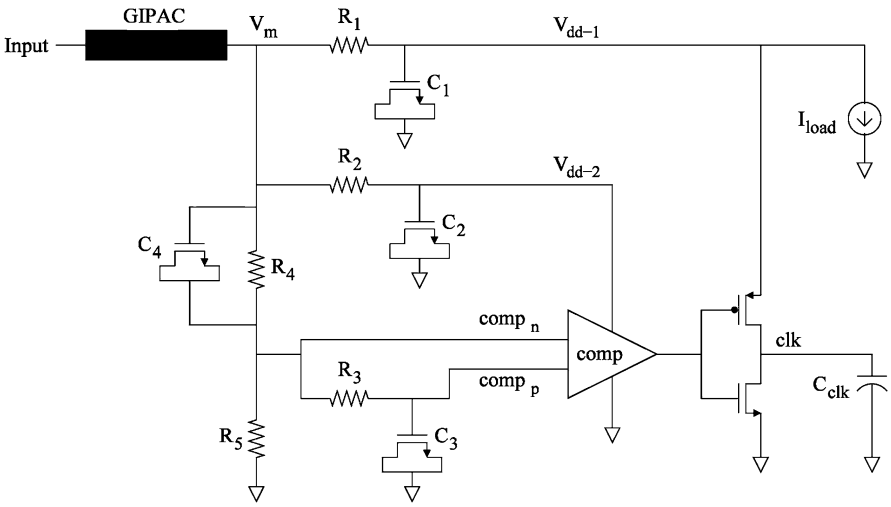


Fig. 38.4 GIPAC splitter circuit. Each *RC* pair behaves as a low pass filter. *R* and *C* are determined by the noise requirements, where the resistors are polysilicon resistors and the capacitors are MOS transistors. C_{clk} and I_{load} are, respectively, the capacitive load of the clock network and total current of the entire circuit

where A is the supply voltage, α is the amplitude of the signal used to generate the clock signal, and f_{clk} is the clock signal frequency. The two subsections below describe, respectively, the generation of the power and clock signals.

(1) *Generating the power signal*. The input signal propagates through the GIPAC network and arrives at the splitter, as illustrated in Fig. 38.4. The R_1C_1 filter is a low pass filter that removes the sinusoidal waveform from the input signal, maintaining only the DC portion of the signal. The output signal from this filter is V_{dd-1}. This filter can also be a higher order filter to improve the quality of the V_{dd-1} signal. Since a significant current typically propagates through this filter, R_1 is small, requiring a higher C_1. A small R_1 is necessary to maintain a high voltage at V_{dd-1}, since a voltage divider is created between R_1 and the local power network.

The R_2C_2 filter is similar to the R_1C_1 low pass filter; however, due to the lower current, R_2 is significantly higher than R_1, permitting C_2 to be smaller. The cutoff frequency for the second filter is lower than the first filter, reducing the noise at the output of the second filter. The output signal of the second low pass filter is called V_{dd-2}. The R_1C_1 and R_2C_2 filters can also be a single low pass filter, reducing overall area. Switching noise on the power lines would however be injected directly into the clock signal, significantly increasing the clock jitter. Separate *RC* filters are therefore used to reduce the noise coupled from the power network into the clock network (the third noise path depicted in Fig. 38.3).

(2) *Generating the clock signal*. The clock signal is produced in several stages. The DC component of the signal V_m is initially divided by two (in Fig. 38.4, labeled as the $comp_n$ signal). The $comp_p$ signal is generated by filtering the $comp_n$ signal with the R_3C_3 low pass filter. This configuration generates two signals with the same DC level. Comparing (or amplifying the difference between) these two signals produces a clock signal with a 50 % duty cycle. The C_4 capacitor passes an AC signal to the input of the comparator, creating a voltage divider at node $comp_n$. By increasing C_4, the AC signal is less attenuated; albeit, requiring more area. The buffer at the output of the comparator adjusts the voltage to V_{dd-1}. This buffer can be designed as cascaded buffers depending upon the load. The comparator utilizes a self-biased structure [590].

38.2.2 RC Filter Values

The value of R and C for the low pass filters is based on the DC and AC noise requirements,

$$noise_{dc} = \frac{R \cdot I_{max}}{V_{dd}} \cdot 100, \tag{38.2}$$

where V_{dd}, I_{max}, and $noise_{dc}$ are, respectively, the required power supply voltage, maximum current, and per cent of the allowed DC noise on the power network.

$$noise_{ac} = \frac{2\alpha \left| \frac{1}{1+j2\pi f_{clk} RC} \right|}{V_{dd}} \cdot 100. \tag{38.3}$$

By increasing C, enhanced noise reduction can be achieved; however, the penalty is the area required for the capacitor, producing a tradeoff between noise and area. The output buffer after the comparator is a cascaded buffer structure to drive a large capacitive load.

38.3 Simulation Results

The GIPAC network and splitter are designed using a 90 nm CMOS technology with a power supply voltage of 1.2 V and a clock signal frequency of 1 GHz. The simulation is evaluated with the current switching with a normal random distribution between 0 and 100 mA. A transient simulation of the input, V_{dd-1} output (to power the entire circuitry), and V_{dd-2} output (to power the clock comparator) are illustrated in Fig. 38.5. Power signal generation is accomplished by propagating the GIPAC output signal V_m through the first low pass filter. Depending upon the current

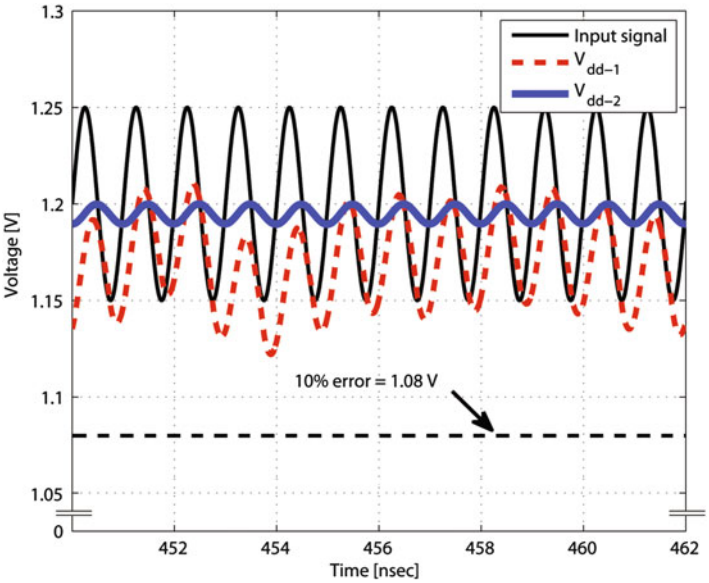


Fig. 38.5 Transient simulation of the GIPAC input, output V_{dd-1}, and output V_{dd-2}. The ripple on the power lines is considered as noise

requirements, the DC level of the power network is shifted due to the resistive voltage drop across the filter. The second low pass filter generating V_{dd-2} only passes a small current to the comparator, attenuating V_m. The V_{dd-1} signal fluctuates between 1.116 and 1.226 V (110 mV), creating less than 10 % error. The V_{dd-2} signal fluctuates between 1.189 and 1.201 V (12 mV), within 1 % error.

To generate the clock signal, the GIPAC output signal V_m is divided by two and filtered by the third low pass filter. The two input signals to the comparators are illustrated in Fig. 38.6a. The $comp_n$ signal swings between 559 and 637 mV, while the $comp_p$ signal only swings between 596 and 598 mV due to the third low pass filter. The DC level of both signals is similar, producing a 50 % duty cycle clock signal. After the comparator, the signal is amplified by the cascaded buffers, generating the clock signal shown in Fig. 38.6b.

An eye diagram of the clock signal is depicted in Fig. 38.7. The impedance of the global ground and power networks is assumed to be equal; therefore, the additional noise on the high power rail shown in the eye diagram is a result of integrating the power and clock signals. As depicted in the eye diagram, the voltage fluctuates on the high power rail between 1.11 and 1.21 V, resulting in about 8 % error and a clock jitter of 33 ps.

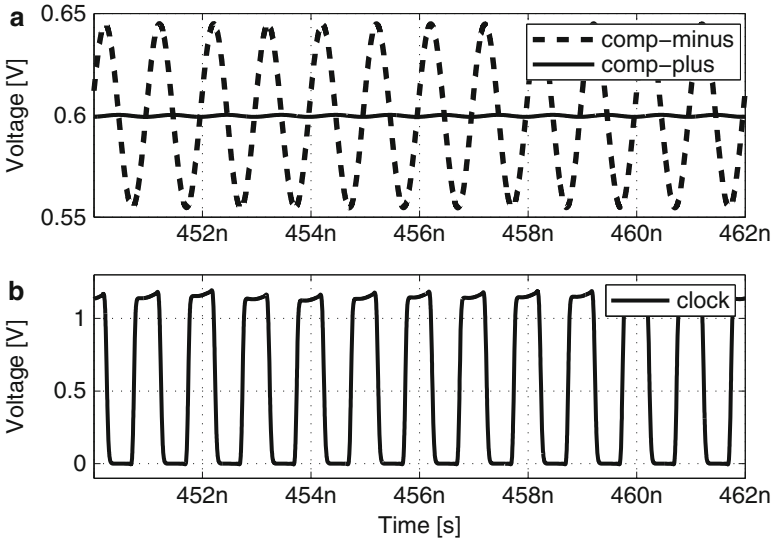


Fig. 38.6 Transient simulation of the GIPAC splitter circuit. (**a**) The two signals at the input of the comparator exhibit the same DC level. (**b**) The generated clock signal drives a 1 pF capacitive load

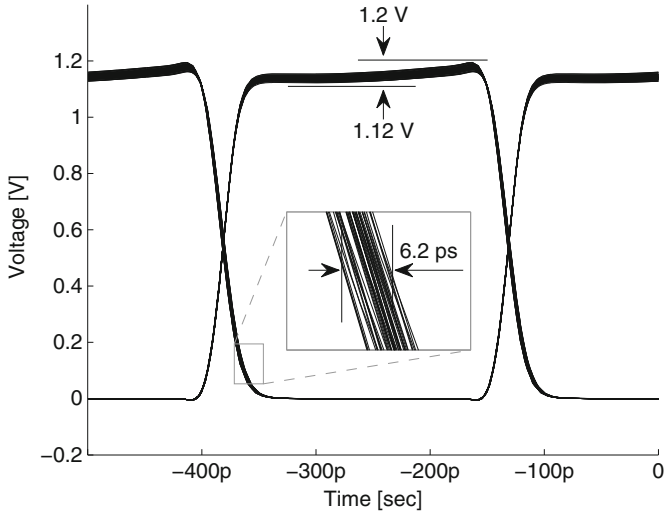


Fig. 38.7 Eye diagram of the simulated clock signal. The global ground and power networks are assumed to be equal. The difference between the noise on the high and low power rails is due to the GIPAC splitter

38.4 Summary

A general approach for combining the global power and clock networks into a single integrated network is discussed in this chapter.

- Replacing two global networks with one integrated network provides increased integration and functionality
- The simulation results, based on a 90 nm CMOS technology, successfully demonstrate splitting the GIPAC output signal into separate clock and power signals
- Noise issues and different tradeoffs are evaluated for this integrated power and clock network

Chapter 39
Conclusions

The focus of Part VII is on multi-layer power distribution networks. Due to differences in the physical dimensions (thickness, width, spacing) among the metal layers within multi-layer systems, the ratio of the grid resistive and inductive impedance varies for each metal layer. The resistive impedance as compared to the inductive impedance of the upper metal layers is typically lower, while at the lower metal layers, the inductive impedance is lower. The high frequency current, therefore, propagates in the lower metal layers, while the low frequency current propagates in the higher metal layers. With advancements in technology and higher frequencies, a greater amount of the current propagates in the lower metal layers, requiring more careful design of the lower metal layers.

To fully utilize a multi-layer system, the current density needs to be maintained relatively equal among the layers, providing higher reliability for every metal layer. This objective can be achieved by widening the power lines of the lower metal layers, producing a power/ground network similar to a pyramid shaped structure. Other design issues, such as routability and heat removal, need to be simultaneously considered when designing power delivery systems.

With recent integration trends in semiconductor technology, the complexity of heterogeneous integrated circuits has significantly increased. Multiple structures, in the past designed and analyzed independently, now need to be merged to more accurately analyze performance. This integrated approach supports the development of design methodologies and architectures for optimizing noise, power, area, and other resources. As an example, a globally integrated power and clock distribution network is described which utilizes a single network to distribute both global signals; thereby reducing metal requirements. The power delivery network, clock distribution network, and substrate need to be modeled as one integrated system to more accurately characterize the noise signature and circuit performance.

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0_39

Part VIII
Multi-voltage Power Delivery Systems

Multi-voltage power supply systems are discussed in Part VIII. These systems are commonly used in heterogeneous mixed-signal integrated circuits, such as systems-on-chip. Design strategies are therefore required for these multi-voltage networks. The interactions among the decoupling capacitances and multi-voltage systems are also reviewed. The following two chapters discuss the design of multi-voltage power systems.

In Chap. 40, systems with multiple power supply voltages are described. Several multi-voltage structures are reviewed. Primary challenges in integrated circuits with multiple power supplies are discussed. The power savings is shown to depend upon the number and magnitude of the available power supply voltages. Rules of thumb are presented to determine the appropriate number and magnitude of the multiple power supplies to lower the power dissipated by the system.

On-chip power distribution grids with multiple power supply voltages are discussed in Chap. 41. A power distribution grid with multiple power supplies and multiple grounds is presented. This power distribution grid structure results in reduced voltage fluctuations as seen at the terminals of the current load, as compared to traditional power distribution grids with multiple supply voltages and a single ground. It is noted that a multi-power and multi-ground power distribution grid can be an alternative to a single supply voltage and single ground power distribution system.

Decoupling capacitors for power distribution systems with multiple power supply voltages is the topic of Chap. 42. With the introduction of a second power supply, the noise at one power supply can propagate to another power supply, producing power and signal integrity issues in the overall system. Interactions between the two power distribution networks should therefore be considered. The dependence of the impedance and magnitude of the voltage transfer function on the parameters of the power distribution system is evaluated. Design techniques to cancel and shift the antiresonant spikes out of the range of the operational frequencies are also presented.

Chapter 40
Multiple On-Chip Power Supply Systems

With recent developments in nanometer CMOS technologies, excessive power dissipation has become a limiting factor in integrating a greater number of transistors onto a single monolithic substrate. With the introduction of systems-on-chip, systems-in-package (SiP), and 3-D integrated technologies, the problem of heat removal has further worsened [591–593]. Unless power consumption is dramatically reduced, packaging and performance of ultra large scale integration (ULSI) circuits will become fundamentally limited by heat dissipation.

Another driving factor behind the push for low power circuits is the growing market for portable electronic devices, such as PDAs, wireless communications, and imaging systems that demand high speed computation and complex functionality while dissipating as little power as possible [594]. Design techniques and methodologies for reducing the power consumed by an IC while providing high speed and high complexity systems are therefore required. These design technologies will support the continued scaling of the minimum feature size, permitting the integration of a greater number of transistors onto a single monolithic substrate.

The most effective way to reduce power consumption is to lower the supply voltage. Dynamic power currently dominates the total power dissipation, quadratically decreasing with supply voltage [595]. Reducing the supply voltage, however, increases the circuit delay. In [596], demonstrated that the increased delay can be compensated by shortening the critical paths using behavioral transformations such as parallelization and pipelining. The resulting circuit consumes less average power while satisfying global throughput constraints; albeit, at the cost of increased circuit area [597].

Power consumption can also be reduced by scaling the threshold voltage while simultaneously reducing the power supply [598]. This approach, however, results in significantly increased standby leakage current. To limit the leakage current during sleep mode, several techniques have been described, such as multi-threshold voltage CMOS [289, 521], variable threshold voltage schemes [599, 600], and circuits with

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0_40

an additional transistor behaving as a sleep switch [601]. These techniques, however, require additional process steps and/or additional circuitry to control the substrate bias or switch off portions of the circuit [600].

The total power dissipation can also be reduced by utilizing multiple power supply voltages [289, 602, 603]. In this scheme, a reduced voltage V_{dd}^{L} is applied to the non-critical paths, while a higher voltage V_{dd}^{H} is provided to the critical paths so as to achieve the specified delay constraints [289]. Multi-voltage schemes result in reduced total power without degrading the overall circuit performance. Multiple on-chip power supply systems are the subject of this chapter. Various circuit techniques exploiting multiple power supply voltages are presented in Sect. 40.1. Challenges of ICs with multiple supply voltages are discussed in Sect. 40.2. Choosing the optimum number and magnitude of the multi-voltage power supplies is discussed in Sect. 40.3. Some conclusions are offered in Sect. 40.4.

40.1 ICs with Multiple Power Supply Voltages

The strategy of exploiting multiple power supply voltages consists of two steps. Those logic gates with excessive slack (the difference between the required time and the arrival time of a signal) is first determined. A reduced supply voltage V_{dd}^{L} is provided to those gates to reduce power. Note that in most practical applications, the number of critical paths is only a small portion of the total number of paths in a circuit. Excess slack therefore exists in the majority of paths within a circuit. Determining those gates with excessive time slack is therefore an important and complex task [289]. A variety of computer-aided design (CAD) algorithms and tools have been developed to evaluate the delay characteristics of high complexity ICs such as microprocessors [604, 605]. Multi-voltage low power techniques are reviewed in this section. A low power technique with multiple power supply voltages is presented in Sect. 40.1.1. Clustered voltage scaling (CVS) is presented in Sect. 40.1.2. Extended clustered voltage scaling (ECVS) is discussed in Sect. 40.1.3.

40.1.1 Multiple Power Supply Voltage Techniques

A critical delay path between flip flops FF_1 and FF_2 in a single supply voltage, synchronous circuit is shown in Fig. 40.1. Since the excessive slack remains in those paths located off the critical path, timing constraints are satisfied if the gates in the non-critical paths use a reduced supply voltage V_{dd}^{L}. A dual supply voltage circuit in which the original power supply voltage V_{dd}^{H} of each of the gates along the non-critical delay paths is replaced by a lower supply voltage V_{dd}^{L} is illustrated in Fig. 40.2. If a low voltage supply is available, the gates with V_{dd}^{L} can be selected to reduce the overall power using conventional algorithms such as gate resizing [606].

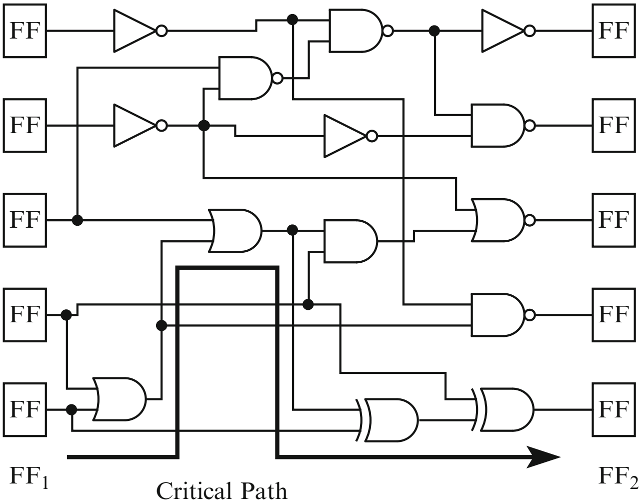


Fig. 40.1 An example single supply voltage circuit

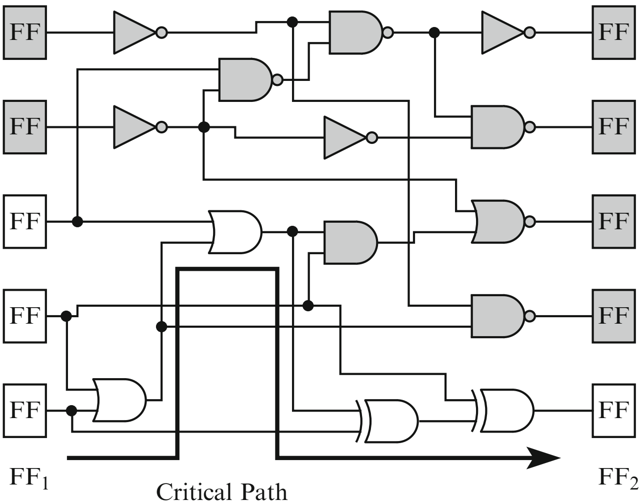


Fig. 40.2 An example dual supply voltage circuit. The gates operating at a lower power supply voltage V_{dd}^{L} (located off the critical delay path) are *shaded*

A circuit with multiple power supply voltages, however, can result in DC current flowing in a high voltage gate due to the direct connection between a low voltage gate and a high voltage gate. If a gate with a reduced supply voltage is directly connected to a gate with the original supply voltage, the "high" level voltage at node *A* is not sufficiently high to turn off the PMOS device in a CMOS circuit, as shown in Fig. 40.3. The PMOS device in the high voltage gate is therefore weakly "ON,"

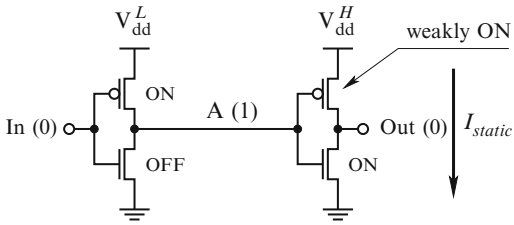


Fig. 40.3 Static current as a result of a direct connection between the V_{dd}^{L} gate and the V_{dd}^{H} gate

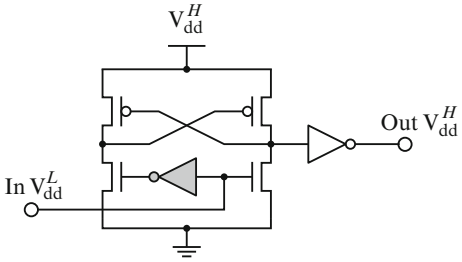


Fig. 40.4 Level converter circuit. The inverter operating at the reduced power supply voltage V_{dd}^{L} is shown in *gray*

conducting static current from the power supply to ground. These static currents significantly increase the overall power consumed by an IC, wasting the savings in power achieved by utilizing a multi-voltage power distribution system.

Level converters are typically inserted at node A to remove the static current path [607]. A simple level converter circuit is illustrated in Fig. 40.4. The level converter restores the full voltage swing from V_{dd}^{L} to V_{dd}^{H}. Note that a great number of level converters is typically required, increasing the area and power overhead. The problem of utilizing a dual power supply voltage scheme is formulated as follows.

Problem formulation: For a given circuit, determine the gates and registers to which a reduced power supply voltage V_{dd}^{L} should be applied such that the overall power and number of level converters are minimized while satisfying system-level timing constraints [608].

40.1.2 Clustered Voltage Scaling (CVS)

The number of level converters can be reduced by minimizing the connections between the V_{dd}^{L} gates and the V_{dd}^{H} gates. The CVS technique, described in [609], results in a circuit structure with a greatly reduced number of level converters, as shown in Fig. 40.5.

To avoid inserting level converters, the CVS technique exploits the specific connectivity patterns among the gates, such as a connection between V_{dd}^{H} gates, between V_{dd}^{L} gates, and between a V_{dd}^{H} gate and a V_{dd}^{L} gate. These connections do not require level converters to remove any static current paths. Level converters are only required at the interface between the output of a V_{dd}^{L} gate and the input of a V_{dd}^{H} gate. The number of required level converters in the CVS structure shown in

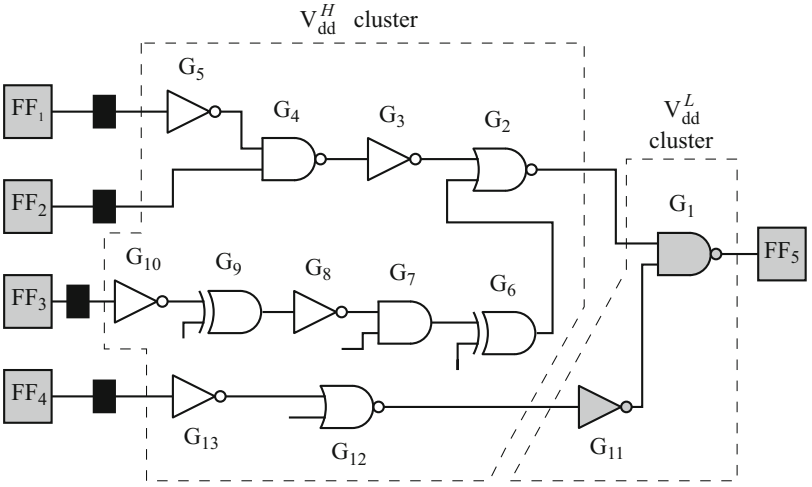


Fig. 40.5 A dual power supply voltage circuit with the clustered voltage scaling (CVS) technique [609]. The gates operating at a lower supply voltage are *shaded*. The level converters are shown as *black rectangles*

Fig. 40.5 is almost the same as the number of V_{dd}^{L} flip flops. The CVS technique therefore results in fewer level converters, reducing the overall power consumed by an integrated circuit.

40.1.3 Extended Clustered Voltage Scaling (ECVS)

The number of gates with a lower power supply voltage can be increased by optimally choosing the insertion points of the level converters, further reducing overall power. As an example, in the CVS structure shown in Fig. 40.5, the path delay from flip flop FF_3 to gate G_2 is longer than the delay from FF_1 to G_2. Moreover, applying a lower power supply to gate G_2 can produce a timing violation. A high power supply should therefore be provided to G_2. From CVS connectivity patterns described in Sect. 40.1.2, note that G_3 also has to be supplied with V_{dd}^{H}. Alternatively, in a CVS structure, G_3 cannot be supplied with V_{dd}^{L} although excessive slack remains in the path from FF_1 to G_2. Similarly, G_4 and G_5 should be connected to V_{dd}^{H} to satisfy existing timing constraints. If the insertion point of the level converter adjacent to FF_1 is moved to the interface between G_3 and G_2, gates G_3, G_4, and G_5 can be connected to V_{dd}^{L}, as illustrated in Fig. 40.6. Note that the structure shown in Fig. 40.6 is obtained from the CVS network by relaxing existing limitations on the insertion positions of the level converters. Such a technique is often referred to as the extended clustered voltage scaling technique [608, 610].

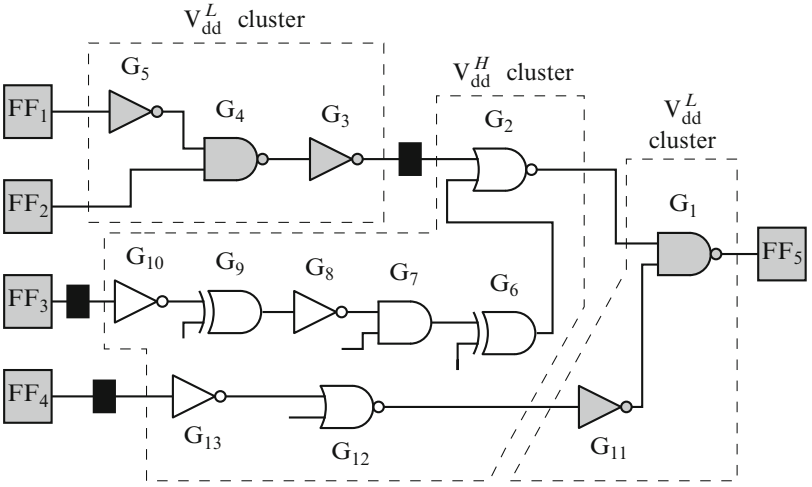


Fig. 40.6 A dual power supply voltage circuit with the extended clustered voltage scaling (ECVS) technique [608]. The gates operating at a lower supply voltage are shaded. The level converters are shown as black rectangles

40.2 Challenges in ICs with Multiple Power Supply Voltages

The application of power reduction techniques with multiple supply voltages in modern high performance ICs is a challenging task. Circuit scheduling algorithms require complex computations, limiting the application of CVS and ECVS techniques to specific paths within an IC. Primary challenges of multi-voltage power reduction schemes are discussed in this section. The issues of area overhead and related tradeoffs are described in Sect. 40.2.1. Power penalties are presented in Sect. 40.2.2. The additional design complexity associated with level converters and integrated DC–DC voltage converters is discussed in Sect. 40.2.3. Several placement and routing strategies are described in Sect. 40.2.4.

40.2.1 Die Area

As described in Sect. 40.1, level converter circuits are inserted at the interface between specific gates in power reduction schemes with multiple power supply voltages to reduce static current. Multi-voltage circuits require additional power connections, significantly increasing routing complexity and die area. Additional area results in greater parasitic capacitance of the signal lines, increasing the dynamic power consumed by an IC. As a result of the increased area, the time slack in the critical paths is often significantly smaller, reducing the power savings of a

multi-voltage scheme. A tradeoff therefore exists between the power savings and area overhead in ICs with multiple power supply voltages. The critical paths should therefore be carefully determined in order to reduce the overall circuit power.

40.2.2 Power Dissipation

Multi-voltage low power techniques require the insertion of level converters to reduce static current. The number of level converters depends upon the connectivity patterns at the interface between each critical and non-critical path. Improper scheduling of the critical paths can lead to an excessive number of level converters, increasing the power. The ECVS technique with relaxed constraints for level converters should therefore be used, resulting in a smaller number of level converters.

Note that the magnitude of the overall reduction in power is determined by the number and voltage of the available power supply voltages, as discussed in Sect. 40.3. It is therefore important to determine the optimum number and magnitude of the power supply voltages to maximize any savings in power. Also note that lower power supply voltages are often generated on-chip from a high voltage power supply using DC–DC voltage converters [611, 612]. The power and area penalties of the on-chip DC–DC voltage converters should therefore be considered to accurately estimate any savings in power.

Several primary factors, such as physical area, the number and magnitude of the power supply voltages, and the number of level converters contribute to the overall power overhead of any multi-voltage low power technique. Complex multi-variable optimization is thus required to determine the proper system parameters in order to achieve the greatest reduction in overall power [613].

40.2.3 Design Complexity

Note that while significantly reducing power, a multiple power supply voltage scheme results in significantly increased design complexity. The complexity overhead of a multi-voltage low power technique is due to two aspects. The level converters not only dissipate power, but also dramatically increase the complexity of the overall design process. A level converter typically consists of both low voltage and high voltage gates, increasing the area and routing resources. Multiple level converters also increase the delay of the critical paths. High speed, low power level converters are therefore required to achieve a significant reduction in overall power while satisfying existing timing constraints [607, 614]. Standard logic gates with embedded level conversion as reported in [614] support the design of circuits without the addition of level converters, substantially reducing power, area, and complexity.

Monolithic DC–DC voltage converters are often integrated on-chip to enhance overall energy efficiency, improve the quality of the voltage regulation, decrease the number of I/O pads dedicated to power delivery, and reduce fabrication costs [314]. To lower the energy dissipated by the parasitic impedance of the circuit board interconnect, the passive components of a low frequency filter (e.g., the filter inductor and filter capacitor) are also placed on-chip, significantly increasing both the required area and design complexity. A great amount of on-chip decoupling capacitance is also often required to improve the quality of the on-chip power supply voltages [186]. The area and power penalty as well as the increased design complexity of the additional on-chip voltage converters should therefore be considered when determining the optimal number and magnitude of the multiple power supply voltages.

40.2.4 Placement and Routing

To achieve the full benefit offered by multiple power supply voltage techniques, various design issues at both the high level and physical level should be simultaneously considered. Existing electronic design automation (EDA) placement and routing tools for conventional circuits with single power supply voltages, however, cannot be directly applied to low power techniques with multiple power supply voltages. Specific CAD tools, capable of placing and routing physical circuits with multiple power supplies based on high level gate assignment information, are therefore required. The placement and routing of ICs with multiple power supply voltages is a complex problem. Three widely utilized layout schemes are described in this section.

Area-by-Area Architecture

The simplest architecture for a circuit with dual power supply voltages is an area-by-area architecture [608], as shown in Fig. 40.7. In this architecture, the V_{dd}^{L} cells are placed in one area, while the V_{dd}^{H} cells are placed in a different area. The area-by-area technique iteratively generates a layout with existing placement and routing tools using one of the available power supply voltages. This architecture, however, results in a degradation in performance due to the substantially increased interconnect length between the V_{dd}^{L} and V_{dd}^{H} cells.

Row-by-Row Architecture

The layout architecture described in [615] is illustrated in Fig. 40.8. In this architecture, the V_{dd}^{L} cells and V_{dd}^{H} cells are placed in different rows. Each row only consists of V_{dd}^{L} cells *or* V_{dd}^{H} cells. This layout technique is therefore a

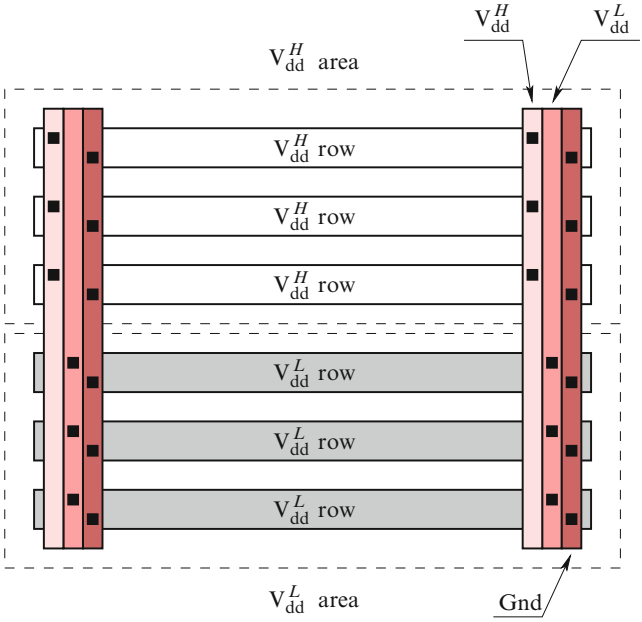


Fig. 40.7 Layout of an area-by-area architecture with a dual power supply voltage. In this architecture, the V_{dd}^{L} cells are placed in one area, while the V_{dd}^{H} cells are separately placed in a different area

row-by-row architecture. Note that in this architecture, a V_{dd}^{L} row is placed next to a V_{dd}^{H} row, reducing the interconnect length between the V_{dd}^{L} cells and the V_{dd}^{H} cells. The performance of a row-by-row layout architecture is therefore higher as compared to the performance of an area-by-area architecture. The row-by-row technique also results in smaller area, further improving system performance. Another advantage of this technique is that an original V_{dd}^{H} cell library can be used for the V_{dd}^{L} cells. Since the layout of the V_{dd}^{L} cells are the same as those of the V_{dd}^{H} cells, the original layout of the V_{dd}^{H} cells can be treated as V_{dd}^{L} cells. A lower power supply voltage can be provided to the V_{dd}^{L} cells.

In-Row Architecture

An improved row-by-row layout architecture is presented in [616]. This architecture is based on a modified cell library [616]. Unlike conventional standard cells, the new standard cell has two power rails and one ground rail. One of the power rails is connected to V_{dd}^{L} and the other power rail is connected to V_{dd}^{H}. The modified library supports the allocation of both V_{dd}^{L} cells and V_{dd}^{H} cells within the same row, as shown in Fig. 40.9. This layout scheme is therefore referred to as an in-row architecture. Note that the width of the power and ground lines in each cell is reduced, slightly

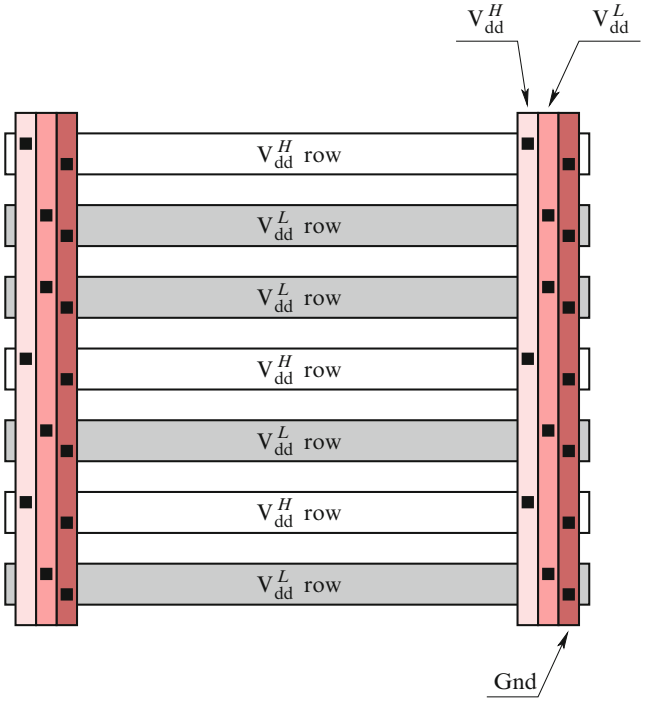


Fig. 40.8 Layout of a row-by-row architecture with a dual power supply voltage. In this architecture, the V_{dd}^L cells and V_{dd}^H cells are placed in different rows. Each row consists of only V_{dd}^L cells *or* V_{dd}^H cells

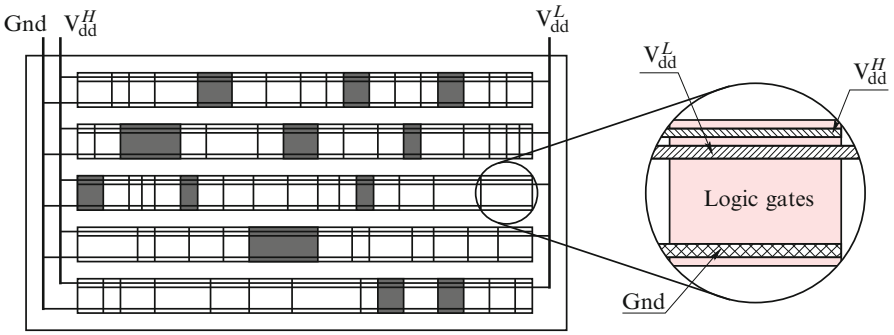


Fig. 40.9 In-row dual power supply voltage scheme. This architecture is based on a modified cell library with two power rails and one ground rail in each cell. The V_{dd}^H cells are shown in *gray* and the V_{dd}^L cells are *white*

increasing the overall area (a 2.7 % area overhead as compared to the original cell) [616]. Since the number of V_{dd}^L cells is typically greater than the number of V_{dd}^H cells, the lower power supply provides higher current. The low voltage power

rail is therefore wider than the high voltage power rail to maintain a similar voltage drop within each power rail. Note that the in-row architecture results in a significant reduction in the interconnect length between the V_{dd}^{L} and V_{dd}^{H} cells, as compared to a row-by-row scheme [616]. An in-row layout scheme should therefore be utilized in high performance, high complexity ICs to reduce overall power with minimal area and complexity penalties.

40.3 Optimum Number and Magnitude of Available Power Supply Voltages

In low power techniques with multiple power supply voltages, the power reduction is primarily determined by the number and magnitude of the available power supply voltages. The trend in power reduction with a multi-voltage scheme as a function of the number of available supply voltages is illustrated in Fig. 40.10. Observe from Fig. 40.10 that if fewer power supplies than the optimum number are available ($n < n_{opt}$), the savings in power can be fairly small. The maximum power savings is achieved with the number of supply voltages close to the optimum number (represented by region $n = n_{opt}$ in Fig. 40.10). If more than the optimum number of power supplies are used, the savings in power becomes smaller, as depicted in Fig. 40.10 for $n > n_{opt}$. This decline in power reduction when the number of supply voltages is greater than the optimum number is due to the increased overhead of the additional power supplies (as a result of the increased area, number of level converters, and design complexity). Any savings in power is also constrained by the magnitude of the available power supplies. A tradeoff therefore exists between the number and magnitude of the available power supplies and the achievable power savings. A methodology is therefore required to estimate the optimum number and

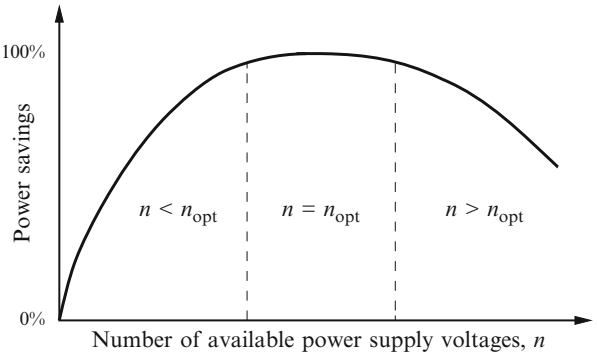


Fig. 40.10 Trend in power reduction with multi-voltage scheme as a function of the number of available supply voltages

magnitude of the available power supply voltages in order to produce the greatest reduction in power. Design techniques for determining the optimum number and magnitude of the available power supplies are the subject of this section.

In systems with multiple power supply voltages (where $V_1 > V_2 > \cdots > V_n$), the power dissipation is [617]

$$P_n = f\left\{\left(C_1 - \sum_{i=2}^{n} C_i\right) V_1^2 + \sum_{i=2}^{n} C_i V_i^2\right\}, \tag{40.1}$$

where C_i is the total capacitance of the logic gates and interconnects operating at a reduced supply voltage V_i and f is the operating frequency. The ratio of the power dissipated by a system with multiple power supply voltages as compared to the power dissipation in a single power supply system is

$$K_{V_{\text{dd}}} \equiv \frac{P_n}{P_1} = 1 - \sum_{i=2}^{n}\left[\left(\frac{C_i}{C_1}\right)\left\{1 - \left(\frac{V_i}{V_1}\right)^2\right\}\right]. \tag{40.2}$$

Since delay is proportional to the total capacitance, $\dfrac{C_i}{C_1}$ is

$$\frac{C_i}{C_1} = \frac{\int_0^1 p(t)\, t_i\, dt}{\int_0^1 p(t)\, t\, dt}, \tag{40.3}$$

where $p(t)$ is the normalized path delay distribution function and t_i is the total delay of the circuits operating at V_i. For a path with a total delay $t_{i,0} < t < t_{i-1,0}$, where $t_{i,0}$ denotes the path delay at V_1 (equal to the cycle time when all of the circuits operate at V_i), the power dissipation is minimum when (V_i, V_{i-1}) are applied. In this case, t_i is

$$t_i = \begin{cases} \dfrac{t_{i,0}}{t_{i,0} - t_{i+1,0}}(t - t_{i+1,0}) & : \quad t_{i+1,0} \le t \le t_{i,0} \\ \dfrac{t_{i,0}}{t_{i-1,0} - t_{i,0}}(t_{i-1,0} - t) & : \quad t_{i,0} \le t \le t_{i-1,0}, \end{cases} \tag{40.4}$$

where $t_{i,0}$ is

$$t_{i,0} = \left(\frac{V_1}{V_i}\right)\left(\frac{V_i - V_{\text{th}}}{V_1 - V_{\text{th}}}\right)^{\alpha}, \tag{40.5}$$

V_{th} is the threshold voltage, and α is the velocity saturation index [618]. Note that $t_{n+1,0} = 0$. $K_{V_{\text{dd}}}$ can be determined from (40.1), (40.2), (40.3), (40.4), and (40.5) for a specific $p(t)$, V_1, V_i, and V_{th}.

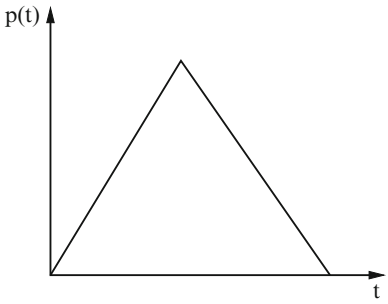


Fig. 40.11 A lambda-shaped normalized path delay distribution function

For a lambda-shaped normalized path delay distribution function $p(t)$ (see Fig. 40.11) as determined from post-layout static timing analysis, approximate rules of thumb for determining the optimum magnitude of the power supply voltages have been determined by Hamada et al. [617],

$$\text{for } \{V_1, V_2\} \qquad \frac{V_2}{V_1} = 0.5 + 0.5\frac{V_{\text{th}}}{V_1}, \tag{40.6}$$

$$\text{for } \{V_1, V_2, V_3\} \qquad \frac{V_2}{V_1} = \frac{V_3}{V_2} = 0.6 + 0.4\frac{V_{\text{th}}}{V_1}, \tag{40.7}$$

$$\text{for } \{V_1, V_2, V_3, V_4\} \qquad \frac{V_2}{V_1} = \frac{V_3}{V_2} = \frac{V_4}{V_3} = 0.7 + 0.3\frac{V_{\text{th}}}{V_1}. \tag{40.8}$$

Criteria (40.6), (40.7), and (40.8) can be used to determine the magnitude of each power supply voltage based on the total number of available power supply voltages. Note that these rules of thumb result in the optimum power supply voltages where the maximum difference in power reduction is less than 1 % as compared to the absolute minimum (as determined from an analytic solution of the system of equations).

Note again that if a greater number of power supplies is used, the total power can be further reduced, reaching a constant power level at some number of power supplies (see Fig. 40.10). As determined in [617], up to three power supply voltages should be utilized to reduce the power consumed by an IC. The reduction in power diminishes as the power supply voltage is scaled and $\frac{V_{\text{th}}}{V_{\text{dd}}}$ increases.

A rule of thumb for two power supply voltages has been evaluated by simulations in [608]. For $V_{\text{dd}}^H = 3.3\,\text{V}$, a V_{dd}^L of 1.9 V is estimated, exhibiting good agreement with (40.6). The dependence of the total power of a dual power supply media processor as a function of the lower power supply V_{dd}^L is depicted in Fig. 40.12. Observe from Fig. 40.12 that the minimum overall power is achieved at $V_{\text{dd}}^L = 1.9\,\text{V}$.

The minimum overall power of a dual power supply system can be explained as follows. In a dual power supply system, the power reduction is determined by two factors: the reduction in power of a single logic gate due to scaling the power supply voltage from V_{dd}^H to V_{dd}^L, and the number of original V_{dd}^H gates replaced with V_{dd}^L gates. At lower V_{dd}^L, the power dissipated by a V_{dd}^L gate decreases, while the number

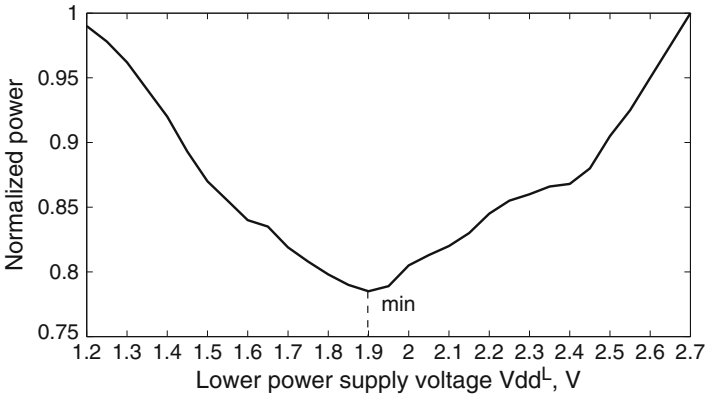


Fig. 40.12 Dependence of the total power of a dual power supply system on a lower power supply voltage V_{dd}^{L} [608]. The original high power supply voltage $V_{\mathrm{dd}}^{H} = 3.3\,\mathrm{V}$

of original V_{dd}^{H} gates replaced with V_{dd}^{L} gates is reduced. This behavior is due to the degradation in performance of the V_{dd}^{L} gates at a lower V_{dd}^{L}. As a result, fewer gates can be replaced with lower voltage gates without violating existing timing constraints. Conversely, at a higher V_{dd}^{L}, the number of gates replaced with V_{dd}^{L} gates increases, while the reduced power in a single V_{dd}^{L} gate decreases. The overall power therefore has a minimum at a specific V_{dd}^{L} voltage, as shown in Fig. 40.12.

Low power techniques with multiple power supply voltages and a single fixed threshold voltage have been discussed in this chapter. Enhanced results are achieved by simultaneously scaling the multiple threshold voltages and the power supply voltages [289, 619, 620]. This approach results in reduced total power with low leakage currents. The total power can also be lowered by simultaneously assigning threshold voltages during gate sizing. Nguyen et al. [621] demonstrated power reductions approaching 32 % on average (57 % maximum) for the ISCAS85 benchmark circuits. CVS with variable supply voltage schemes has been presented in [622]. In this scheme, the power supply voltage is gradually scaled based on an accurate model of the critical path delay. Up to a 70 % power savings has been achieved as compared to the same circuit without these low power techniques. In [623], a column-based dynamic power supply has been integrated into a high frequency SRAM circuit. The power supply voltage is adaptively changed based on the read/write mode of the SRAM, reducing the total power.

As described in this chapter, power dissipation has become a major factor, limiting the performance of high complexity ICs. Multiple low power techniques should therefore be utilized to achieve significant power savings in modern nanoscale ICs.

40.4 Summary

The discussion of multiple on-chip power supply systems and different low power design techniques can be summarized as follows.

- The total power consumed by an IC can be reduced by utilizing multiple power supply voltages
- In multi-voltage low power techniques, a lower power supply voltage is applied to those logic gates with excessive slack to reduce power consumption
- In a multi-voltage scheme, the gates and flip flops with a lower power supply voltage should be determined such that the overall power and number of level converters are minimized while satisfying existing timing constraints
- CVS and ECVS techniques exploit specific connectivity patterns, reducing the number of level converters
- Various penalties, such as area, power, and design complexity, should be considered during the system design process so as to maximize the savings in power
- The in-row layout scheme reduces overall power with minimum area and design complexity
- A maximum of two or three supply voltages should be employed in low power applications
- Rules of thumb have been described for determining the optimum magnitude of the multiple power supply voltages
- A greater savings in power can be achieved by simultaneously scaling the multiple threshold voltages and power supply voltages

Chapter 41
On-Chip Power Grids with Multiple Supply Voltages

With the on-going miniaturization of integrated circuit feature size, the design of power and ground distribution networks has become a challenging task. With technology scaling, the requirements placed on the on-chip power distribution system have significantly increased. These challenges arise from shorter rise/fall times, lower noise margins, higher currents, and increased current densities. Furthermore, the power supply voltage has decreased to lower dynamic power dissipation. A greater number of transistors increases the total current drawn from the power supply. Simultaneously, the higher switching speed of a greater number of smaller transistors produces faster and larger current transients in the power distribution network [286]. The higher currents produce large IR voltage drops. Fast current transients lead to large $L\,dI/dt$ inductive voltage drops (ΔI noise) within the power distribution networks.

The lower voltage of the power supply level can be described as

$$V_{load} = V_{\mathrm{dd}} - IR - L\frac{dI}{dt}, \tag{41.1}$$

where V_{load} is the voltage level seen by a current load, V_{dd} is the power supply voltage, I is the current drawn from the power supply, R and L are the resistance and inductance of the power distribution network, respectively, and dt is the rise time of the current drawn by the load. The power distribution networks must be designed to minimize voltage fluctuations, maintaining the power supply voltage as seen from the load within specified design margins (typically ± 5 % of the power supply level). If the power supply voltage drops too low, the performance (delay) and functionality of the circuit will be severely compromised. Excessive overshoots of the supply voltage can also affect circuit reliability and should therefore be reduced.

With a new era of nanometer scale CMOS circuits, power dissipation has become perhaps the critical design criterion. As described in Chap. 40, to manage the problem of high power dissipation, multiple on-chip power supply voltages have become commonplace [609]. This strategy has the advantage of permitting those

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0_41

modules along the critical paths to operate with the highest available voltage level (in order to satisfy target timing constraints) while permitting modules along the noncritical paths to use a lower voltage (thereby reducing energy consumption). In this manner, the energy consumption is decreased without affecting the circuit speed. This scheme is used to enhance speed in a smaller area as compared to the use of parallel architectures. Using multiple supply voltages for reducing power requirements has been evaluated in the area of high level synthesis for low power [605, 624]. While it is possible to provide multiple supply voltages, in practical applications, such a scenario is expensive. Practically, a small number of voltage supplies (two or three) can be effective [289].

Power distribution networks in high performance ICs are commonly structured as a multi-layer grid [30]. In such a grid, straight power/ground lines in each metalization layer can span an entire die and are orthogonal to the lines in adjacent layers. Power and ground lines typically alternate in each layer. Vias connect a power (ground) line to another power (ground) line at the overlap sites. A typical on-chip power grid is illustrated in Fig. 41.1, where three layers of interconnect are depicted with the power lines shown in *dark gray* and the ground lines shown in *light gray*.

An on-chip power distribution grid in modern high performance ICs is a complex multi-level system. The design of on-chip power distribution grids with multiple supply voltages is the primary focus of this chapter. This chapter is organized as follows. Existing work on power distribution grids and related power distribution systems with multiple supply voltages is reviewed in Sect. 41.1. The structure of a power distribution grid and the simulation setup are reviewed in Sect. 41.2. The structure of a power distribution grid with dual supply voltages and dual grounds (DSDG) is discussed in Sect. 41.3. Interdigitated power distribution grids with DSDG are described in Sect. 41.4. Paired power distribution grids with DSDG are analyzed in Sect. 41.5. Simulation results are presented in Sect. 41.6. Circuit design implications are discussed in Sect. 41.7. Some specific conclusions are summarized in Sect. 41.8.

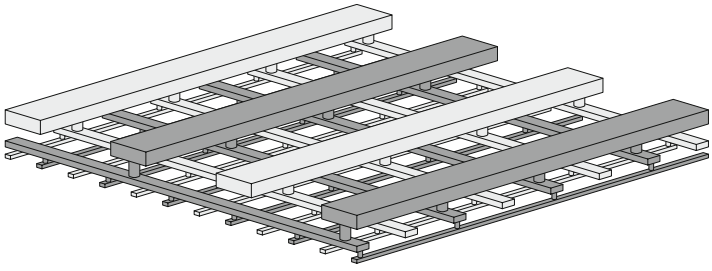


Fig. 41.1 A multi-layer on-chip power distribution grid [625]. The ground lines are *light gray*, the power lines are *dark gray*. The signal lines are not shown

41.1 Background

On-chip power distribution grids have traditionally been analyzed as purely resistive networks [116]. In this early work, a simple model is presented to estimate the maximum on-chip *IR* drop as a function of the number of metal layers and the metal layer thickness. The optimal thickness of each layer produces the minimum *IR* drops. Design techniques are provided to maximize the available signal wiring area while maintaining a constant *IR* drop. These guidelines, however, have limited application to modern, high complexity power distribution networks. The inductive behavior of the on-chip power distribution networks has historically been neglected because the network inductance has been to date dominated by the off-chip parasitic inductance of the package. With the introduction of advanced packaging techniques and the increased switching speed of integrated circuits, this situation has changed. As noted in [150], by replacing wider power and ground lines with narrower interdigitated power and ground lines, the partial self-inductance of the power supply network can be reduced. The authors in [151] propose replacing the wide power and ground lines with an array of interdigitated narrow power and ground lines to decrease the characteristic impedance of the power grid. The dependence of the characteristic impedance on the separation between the metal lines and the metal ground plane is considered. The application of the power delivery scheme, however, is limited to interdigitated structures.

Several design methodologies using multiple power supply voltages have been described in the literature. A row-by-row optimized power supply scheme, providing a different supply voltage to each cell row, is described in [615]. The original circuit is partitioned into two subcircuits by conventional layout methods. Another technique, presented in [616], decreases the total length of the on-chip power and ground lines by applying a multiple supply voltage scheme. A layout architecture exploiting multiple supply voltages in cell-based arrays is described in [608]. Three different layout architectures are analyzed. The authors show that the power consumed by an IC can be reduced, albeit with an increase in area. In previously reported publications, only power distribution systems with two power supply voltages and one common ground have been described. On-chip power distribution grids with multiple power supply voltages and multiple grounds are discussed in this chapter.

41.2 Simulation Setup

The inductance extraction program FastHenry [70] is used to analyze the inductive properties of on-chip power grids. FastHenry efficiently calculates the frequency dependent self- and mutual impedances, $R(\omega) + \omega L(\omega)$, in complex three-dimensional interconnect structures. A quasi-magnetostatic approximation is utilized, meaning the distributed capacitance of the line and any related

displacement currents associated with the capacitances are ignored. The accelerated solution algorithm employed in FastHenry provides approximately a 1 % worst case accuracy as compared to directly solving the system of linear equations characterizing the system.

Copper is assumed as the interconnect material with a conductivity of $(1.72\,\mu\Omega \cdot \text{cm})^{-1}$. A line thickness of 1 μm is assumed for each of the lines in the grids. In the analysis, the lines are split into multiple filaments to account for the skin affect. The number of filaments are estimated to be sufficiently large so as to achieve a 1 % accuracy. Simulations are performed assuming a 1 GHz signal frequency (modeling the low frequency case) and a 100 GHz signal frequency (modeling the high frequency case). The interconnect structures are composed of interdigitated and paired power and ground lines. Three different types of interdigitated power distribution grids are shown in Fig. 41.2. The total number of lines in each power grid is 24. Each of the lines is incorporated into a specific power distribution network and distributed equally between the power and ground networks. The maximum simulation time is under 5 min on a Sun Blade 100 workstation.

41.3 Power Distribution Grid with Dual Supply and Dual Ground

Multiple power supply voltages have been widely used in modern high performance ICs, such as microprocessors, to decrease power dissipation. Only power distribution schemes with dual supply voltages and a single ground (DSSG) have been reported in the literature [28, 30, 276, 608, 615, 616]. In such networks, both power supplies share the one common ground. The ground bounce produced by one of the power supplies therefore adds to the power noise in the other power supply. As a result, voltage fluctuations are significantly increased. To address this problem, an on-chip power distribution scheme with DSDG is presented. In this way, the power distribution system consists of two independent power delivery networks.

A power distribution grid with DSDG consists of two separate subnetworks with independent power and ground supply voltages and current loads. No electrical connection exists between the two power delivery subnetworks. In such a structure, the two power distribution systems are only coupled through the mutual inductance of the ground and power paths, as shown in Fig. 41.3.

The loop inductance of the current loop formed by the two parallel paths is

$$L_{loop} = L_{pp} + L_{gg} - 2M, \tag{41.2}$$

where L_{pp} and L_{gg} are the partial self-inductance of the power and ground paths, respectively, and M is the mutual inductance between these paths. The current in the power and ground lines is assumed to always flow in opposite directions (a reasonable and necessary assumption in large power grids). The inductance of the

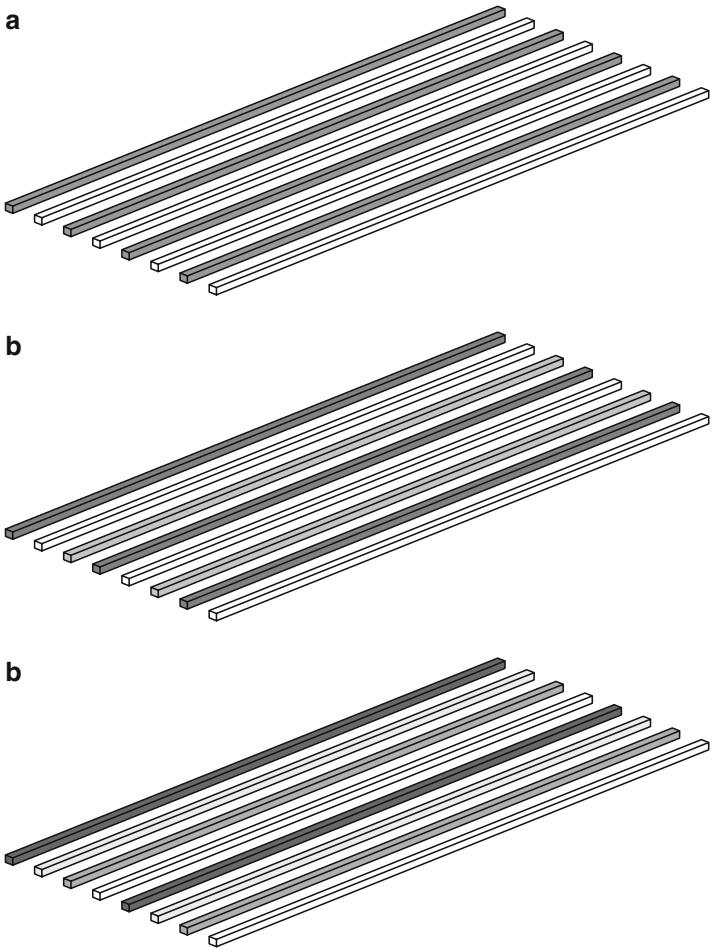


Fig. 41.2 Interdigitated power distribution grids under investigation. In all of the power distribution structures, the power lines are interdigitated with the ground lines; (**a**) reference power distribution grid with a single supply voltage and a single ground (SSSG) (the power lines are *gray colored* and the ground lines are *white colored*), (**b**) power distribution grid with DSSG (the power lines are *light* and *dark gray colored* and the ground lines are *white colored*), (**c**) the power distribution grid with DSDG (the power lines are shown in *black* and *dark gray colors* and the ground lines are shown in *white* and *light gray colors*)

current loop formed by the power and ground lines is therefore reduced by $2M$. The loop inductance of the power distribution grid can be further reduced by increasing the mutual inductive coupling between the power and ground lines. As described by Rosa in 1908 [46], the mutual inductance between two parallel straight lines of equal length is

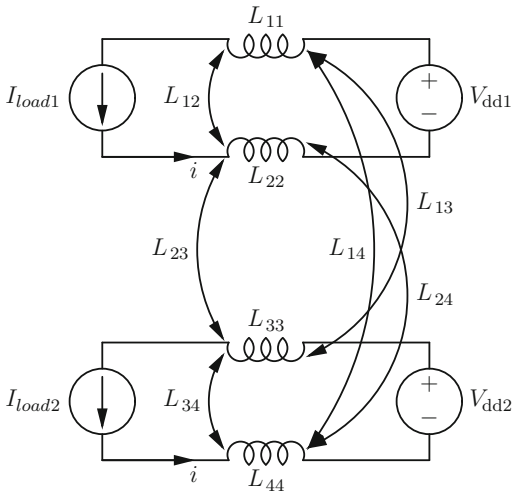


Fig. 41.3 Circuit diagram of the mutual inductive coupling of the DSDG power distribution grid. L_{11} and L_{33} denote, respectively, the partial self-inductances of the power lines and L_{22} and L_{44} denote the partial self-inductances of the ground lines

$$M_{loop} = 0.2l\left(\ln\frac{2l}{d} - 1 + \frac{d}{l} - \ln\gamma + \ln k\right)\mu\text{H}, \tag{41.3}$$

where l is the line length, and d is the distance between the line centers. This expression is valid for the case where $l \gg d$. The mutual inductance of two straight lines is a weak function of the distance between the lines [30].

Analogous to inductive coupling between two parallel loop segments as described in [71], the mutual loop inductance of the two power distribution grids with DSDG is

$$M_{loop} = L_{13} - L_{14} + L_{24} - L_{23}. \tag{41.4}$$

Note that the two negative signs before the mutual inductance components in (41.4) correspond to the current in the power and ground paths flowing in opposite directions. Also note that since the mutual inductance M in (41.2) is negative, M_{loop} should be negative to lower the loop inductance. If M_{loop} is positive, the mutual inductive coupling between the power/ground paths is reduced and the effective loop inductance is therefore increased. If the distance between the lines making a loop is much smaller than the separation between the two loops, $L_{13} \approx L_{14}$ and $L_{23} \approx L_{24}$. This situation is the case for paired power distribution grids. In such grids, the power and ground lines are located in pairs in close proximity. For the interdigitated grid structure shown in Fig. 41.2c, the distance between the lines d_{12} is the same as an offset between the two loops d_{23}, as illustrated in Fig. 41.4. In this case, assuming $d_{12} = d_{23} = d$, from (41.3), M_{loop} between the two grids is approximately

$$M_{loop} = 0.2l\ln\frac{3}{4}\ \mu\text{H}. \tag{41.5}$$

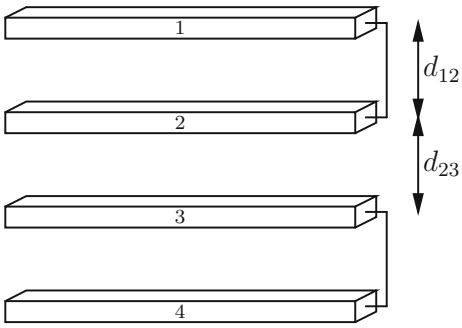


Fig. 41.4 Physical structure of an interdigitated power distribution grid with DSDG. The power delivery scheme consists of two independent power delivery networks

Thus, M_{loop} between the two grids is negative (with an absolute value greater than zero) in DSDG grids. The loop inductance of the particular power distribution grid, therefore, can be further lowered by $2M$. Conversely, in grids with DSSG, currents in both power paths flow in the same direction. In this case, the resulting partial inductance of the current path formed by the two power paths is

$$L_{||} = \frac{L_{pp}^1 L_{pp}^2 - M^2}{L_{pp}^1 + L_{pp}^2 - 2M}, \tag{41.6}$$

where L_{pp}^1 and L_{pp}^2 are the partial self-inductance of the two power paths, respectively, and M is the mutual inductance between these paths. The mutual inductance between the two loops is therefore increased. Thus, the loop inductance seen from a particular current load increases, producing larger power/ground $L\, dI/dt$ voltage fluctuations.

41.4 Interdigitated Grids with DSDG

As shown in Sect. 41.3, by utilizing the power distribution scheme with DSDG, the loop inductance of the particular power delivery network is reduced. In power distribution grids with DSDG, the mutual inductance M between the power and ground paths in (41.2) includes two terms. One term accounts for the increase (or decrease) in the mutual coupling between the power and ground paths in a particular power delivery network due to the presence of the second power delivery network. The other term is the mutual inductance in the loop formed by the power and ground paths of the particular power delivery network. Thus, the mutual inductance in power distribution grids with DSDG is

$$M = M' + M_{loop}, \tag{41.7}$$

where M' is the mutual inductance in the loop formed by the power and ground lines of the particular power delivery network and M_{loop} is the mutual inductance between the two power delivery networks. M' is always negative. M_{loop} can be either negative or positive.

The loop inductance of a conventional interdigitated power distribution grid with DSSG has recently been compared to the loop inductance of an example interdigitated power distribution grid with DSDG [626]. In general, multiple interdigitated power distribution grids with DSDG can be utilized, satisfying different design constraints in high performance ICs. Exploiting the symmetry between the power supply and ground networks, all of the possible interdigitated power distribution grids with DSDG can be characterized by two primary power delivery schemes. Two types of interdigitated power distribution grids with DSDG are described in this section. The loop inductance in the first type of power distribution grids is presented in Sect. 41.4.1. The loop inductance in the second type of power distribution grids is discussed in Sect. 41.4.2.

41.4.1 Type I Interdigitated Grids with DSDG

In the first type of interdigitated power distribution grid, the power and ground lines in each power delivery network and in different voltage domains (power and ground supply voltages) are alternated and equidistantly spaced, as shown in Fig. 41.5. In such power distribution grids, the distance between the lines inside the loop d_I^i is equal to the separation between the two loops s_I^i. Such power distribution grids are described here as *fully interdigitated* power distribution grids with DSDG.

Consistent with (41.4), the mutual inductive coupling of two current loops in fully interdigitated grids with DSDG is

$$M_{loop}^{\text{intI}} = L_{\text{Vdd1}-\text{Vdd2}} - L_{\text{Vdd1}-\text{Gnd2}} + L_{\text{Gnd1}-\text{Gnd2}} - L_{\text{Vdd2}-\text{Gnd1}}, \tag{41.8}$$

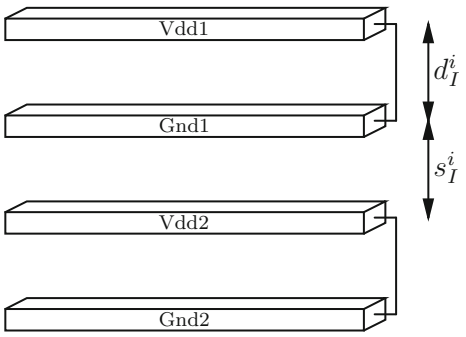


Fig. 41.5 Physical structure of a fully interdigitated power distribution grid with DSDG. The distance between the lines making the loops d_I^i is equal to the separation between the two loops s_I^i

where L_{ij} is the mutual inductance between the power and ground paths in the two power distribution networks. In general, a power distribution grid with DSDG should be designed such that M_{loop} is negative with the absolute maximum possible value. Alternatively,

$$|L_{\text{Vdd1}-\text{Gnd2}}| + |L_{\text{Vdd2}-\text{Gnd1}}| > |L_{\text{Vdd1}-\text{Vdd2}}| + |L_{\text{Gnd1}-\text{Gnd2}}|. \tag{41.9}$$

For fully interdigitated power distribution grids with DSDG, the distance between the power and ground lines inside each loop d_I^i is the same as an offset between the two loops s_I^i. In this case, substituting the mutual inductances between the power and ground paths in the two voltage domains into (41.8), M_{loop}^{intI} between the two grids is determined by (41.5). Observe that M_{loop}^{intI} is negative. A derivation of the mutual coupling between the two current loops in fully interdigitated power distribution grids with DSDG is provided in Appendix D.

41.4.2 Type II Interdigitated Grids with DSDG

In the second type of interdigitated power distribution grid, a power/ground line from one voltage domain is placed next to a power/ground line from the other voltage domain. Groups of power/ground lines are alternated and equidistantly spaced, as shown in Fig. 41.6. In such power distribution grids, the distance between the lines inside the loop d_{II}^i is two times greater than the separation between the lines. Since one loop is located inside the other loop, the separation between the two loops s_{II}^i is negative. Such power distribution grids are described here as *pseudo-interdigitated* power distribution grids with DSDG.

The mutual inductive coupling of two current loops in pseudo-interdigitated grids with DSDG is determined by (41.8). For pseudo-interdigitated power distribution grids with DSDG, the distance between the power and ground lines inside each loop d_{II}^i is two time greater than the offset between the two loops s_{II}^i. In this case, substituting the mutual inductances between the power and ground paths in the different voltage domains into (41.8), the mutual inductive coupling between the two networks M_{loop}^{intII} is

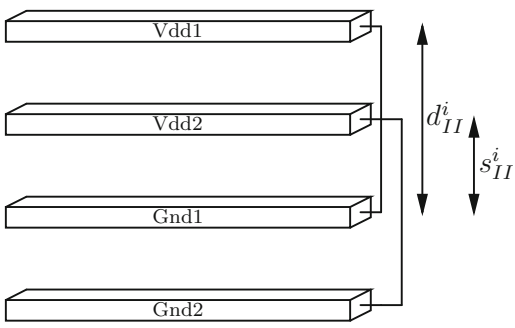


Fig. 41.6 Physical structure of a pseudo-interdigitated power distribution grid with DSDG. The distance between the lines making the loops d_{II}^i is two times greater than the separation between the lines

$$M_{loop}^{\text{intII}} = 0.2l\left(\ln 3 - \frac{2d}{l}\right), \tag{41.10}$$

where d is the distance between the two adjacent lines. Observe that M_{loop}^{intII} is positive for $l \gg d$. The derivation of the mutual coupling between the two current loops in pseudo-interdigitated power distribution grids with DSDG is presented in Appendix E.

In modern high performance ICs, the inductive component of the power distribution noise has become comparable to the resistive noise [129]. In future nanoscale ICs, the inductive $L\ dI/dt$ voltage drop will dominate the resistive IR voltage drop, becoming the major component in the overall power noise. The partial self-inductance of the metal lines comprising the power distribution grid is constant for fixed parameters of a power delivery system (i.e., the line width, line thickness, and line length). In order to reduce the power distribution noise, the total mutual inductance of a particular power distribution grid should therefore be negative with an absolute maximum value.

Comparing (41.5) to (41.10), note that for a line separation d much smaller than line length l, the mutual inductive coupling between different voltage domains in fully interdigitated grids M_{loop}^{intI} is negative with a nonzero absolute value, whereas the mutual inductive coupling between two current loops in pseudo-interdigitated grids M_{loop}^{intII} is positive. Moreover, since the distance between the lines comprising the loop in fully interdigitated power distribution grids is two times smaller than the line separation inside each current loop in pseudo-interdigitated power distribution grids, the mutual inductance inside the loop M'_{intI} is larger than M'_{intII}. Thus, the total mutual inductance as described by (41.7) in fully interdigitated grids is further increased by M_{loop}^{intI}. Conversely, the total mutual inductance in pseudo-interdigitated grids is reduced by M_{loop}^{intII}, as shown in Fig. 41.7. The total mutual inductance in fully interdigitated power distribution grids with DSDG is therefore greater than the total mutual inductance in pseudo-interdigitated grids with DSDG.

41.5 Paired Grids with DSDG

Another type of power distribution grid with alternating power and grounds lines is paired power distribution grids [30, 73]. Similar to interdigitated grids, the power and ground lines in paired grids are alternated, but rather than placed equidistantly, the lines are placed in equidistantly spaced pairs of adjacent power and ground lines. Analogous to the concepts presented in Sect. 41.3, the loop inductance of a particular power distribution network in paired power distribution grids with DSDG is affected by the presence of the other power distribution network.

In general, multiple paired power distribution grids with DSDG can be designed to satisfy different design constraints in high performance ICs. Exploiting the symmetry between the power and ground networks, each of the possible paired

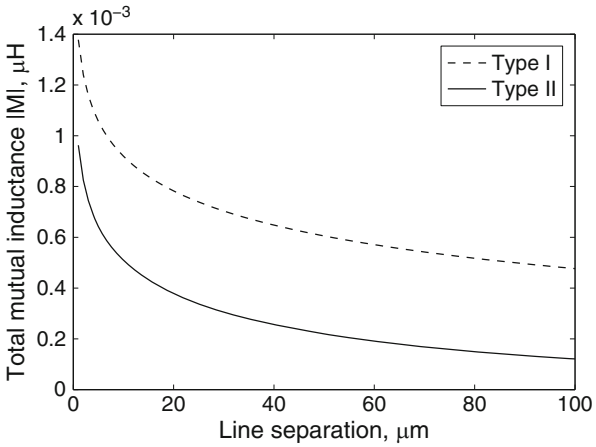


Fig. 41.7 Total mutual inductance of interdigitated power distribution grids with DSDG as a function of line separation. The length of the lines is 1000 μm

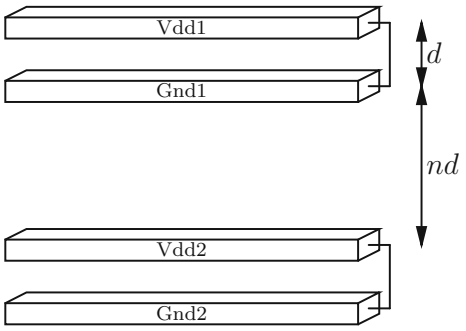


Fig. 41.8 Physical structure of a fully paired power distribution grid with DSDG. In such a grid, each pair is composed of power and ground lines for a particular voltage domain. The separation between the pairs is *n* times larger than the distance between the lines making up the loop *d*

power distribution grids with DSDG can be characterized by the two main power delivery schemes. Two types of paired power distribution grids with DSDG are presented in this section. The loop inductance in the first type of power distribution grid is described in Sect. 41.5.1. The loop inductance in the second type of power distribution grid is discussed in Sect. 41.5.2.

41.5.1 Type I Paired Grids with DSDG

In the first type of paired power distribution grid with DSDG, the power and ground lines of a particular power delivery network are placed in equidistantly spaced pairs. The group of adjacent power and ground lines from one voltage domain is alternated with the group of power and ground lines from the other voltage domain, as shown in Fig. 41.8. In such power distribution grids, the power and ground lines from a

specific power delivery network are placed in pairs. The separation between the pairs is n times (where $n \geq 1$) larger than the separation between the lines inside each pair. Such power distribution grids are described here as *fully paired* power distribution grids with DSDG. Note that in the case of $n = 1$, fully paired grids degenerate to fully interdigitated grids.

Similar to the mutual inductance between the two loops in interdigitated power distribution grids as discussed in Sect. 41.4, the mutual inductive coupling of the two current loops in fully paired grids with DSDG is determined by (41.8). In fully paired power distribution grids with DSDG, the distance between the pairs is n times greater than the separation d between the power and ground lines making up the pair. Thus, substituting the mutual inductance between the power and ground lines for the different voltage domains into (41.8), the mutual inductive coupling between the two networks M_{loop}^{prdI} is

$$M_{loop}^{\text{prdI}} = 0.2l \ln \left[\frac{(n+2)n}{(n+1)^2} \right]. \tag{41.11}$$

A derivation of the mutual coupling between the two current loops in fully paired power distribution grids with DSDG is presented in Appendix F. Note that M_{loop}^{prdI} is negative for $n \geq 1$ with an absolute value slightly greater than zero. Also note that the mutual inductance inside each current loop M'_{prdI} does not depend on n and is determined by (41.3).

41.5.2 Type II Paired Grids with DSDG

In the second type of paired power distribution grid with DSDG, a power/ground line from one voltage domain is placed in a pair with a power/ground line from the other voltage domain. The group of adjacent power lines alternates with the group of ground lines from different voltage domains, as shown in Fig. 41.9. In such power distribution grids, the power and ground lines from different power delivery networks are placed in pairs. The separation between the pairs is n times (where $n \geq 1$) larger than the separation between the lines within each pair. Such power distribution grids are described here as *pseudo-paired* power distribution grids with DSDG. Note that in the case of $n = 1$, pseudo-paired grids are identical to pseudo-interdigitated grids.

As discussed in Sect. 41.5.1, the mutual inductive coupling between the two power delivery networks in pseudo-paired grids with DSDG is determined by (41.8). In pseudo-paired power distribution grids with DSDG, the distance between the pairs is n times greater than the separation d between the power/ground lines making up the pair. The effective distance between the power and ground lines in a particular power delivery network is therefore $(n + 1)d$. Substituting the mutual inductance

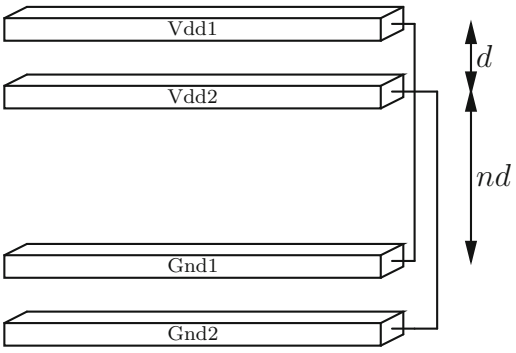


Fig. 41.9 Physical structure of a pseudo-paired power distribution grid with DSDG. In such a grid, each pair is composed of power or ground lines from the two voltage domains. The separation between the pairs is n times larger than the distance between the lines making up the loop d. The effective distance between the power and ground lines in a particular power delivery network is $(n+1)d$

between the power and ground lines in the two different voltage domains into (41.8), the mutual inductive coupling between the two networks M_{loop}^{prdII} is

$$M_{loop}^{\text{prdII}} = 0.2l\left[\ln\left(n^2+2n\right) - \frac{2nd}{l}\right]. \tag{41.12}$$

A derivation of the mutual coupling between the two current loops in pseudo-paired power distribution grids with DSDG is provided in Appendix G. Note that M_{loop}^{prdII} is positive for $n \geq 1$. In contrast to fully paired grids, in pseudo-paired power distribution grids, the mutual inductance inside each current loop M'_{prdII} is a function of n,

$$M'_{\text{prdII}} = 0.2l\left[\ln\frac{2l}{(n+1)d} - 1 + \frac{(n+1)d}{l} - \ln\gamma + \ln k\right]. \tag{41.13}$$

Note that M'_{prdII} decreases with n, approaching zero for large n.

Comparing Figs. 41.8 to 41.9, note that the line separation inside each pair in the pseudo-paired power distribution grid is n times greater than the line separation between the power and ground lines making up a pair in fully paired power distribution grids. The mutual inductance within the power delivery network in fully paired power distribution grids M'_{prdI} is therefore greater than the mutual inductance within the power delivery network in pseudo-paired power distribution grids M'_{prdII}. Moreover, the distance between the lines in the particular voltage domain in fully paired power distribution grids does not depend on the separation between the pairs (no dependence on n). Thus, M'_{prdI} is a constant. The distance between the power/ground lines from the different voltage domains in pseudo-paired power distribution grids is smaller, however, than the distance between the power/ground

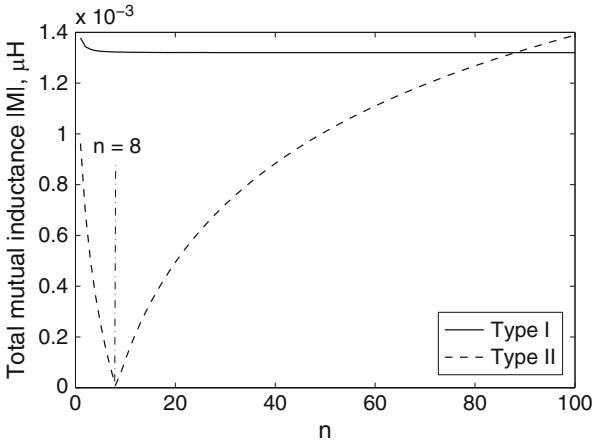


Fig. 41.10 Total mutual inductance of paired power distribution grids with DSDG as a function of the ratio of the distance between the pairs to the line separation inside each pair (n). The length of the lines is 1000 μm and the line separation inside each pair d is 1 μm. Note that the total mutual inductance in pseudo-paired power distribution grids becomes zero at $n = 8$

lines from the different power delivery networks in fully paired power distribution grids. The magnitude of the mutual inductive coupling between the two current loops in pseudo-paired grids M_{loop}^{prdII} is therefore larger than the magnitude of the mutual inductive coupling between the two power delivery networks in fully paired grids M_{loop}^{prdI}. Note that the magnitude of M_{loop}^{prdII} increases with n and becomes much greater than zero for large n. Also note that M_{loop}^{prdI} is negative while M_{loop}^{prdII} is positive for all $n \geq 1$.

The total mutual inductance M as determined by (41.7) for two types of paired power distribution grids with DSDG is plotted in Fig. 41.10. Note that the total mutual inductance in fully paired grids is primarily determined by the mutual inductance inside each power delivery network M'_{prdI}. The absolute value of the total mutual inductance in fully paired grids is further increased by M_{loop}^{prdI}. As the separation between the pairs n increases, the mutual inductive coupling between the two current loops M_{loop}^{prdI} decreases, approaching zero at large n. Thus, the magnitude of the total mutual inductance in fully paired power distribution grids slightly drops with n. In pseudo-paired grids, however, the total mutual inductance is a non-monotonic function of n and can be divided into two regions. The total mutual inductance is determined by the mutual inductance inside each current loop M'_{prdII} for small n and by the mutual inductive coupling between the two voltage domains M_{loop}^{prdII} for large n. Since M'_{prdII} is negative and M_{loop}^{prdII} is positive for all n, the total mutual inductance in pseudo-paired grids is negative with a decreasing absolute value for small n. As n increases, M_{loop}^{prdII} begins to dominate and, at some n ($n = 8$ in Fig. 41.10), the total mutual inductance becomes positive with increasing

absolute value. For large n, pseudo-paired grids with DSDG become identical to power distribution grids with DSSG. Similar to grids with DSSG, power and ground paths in both voltage domains are strongly coupled, increasing the loop inductance as seen from a specific power delivery network. The resulting voltage fluctuations are therefore larger.

41.6 Simulation Results

To characterize the voltage fluctuations as seen at the load, both power distribution grids are modeled as ten series *RL* segments. It is assumed that both power delivery subnetworks are similar and source similar current loads. Two equal current loads are applied to the power grid with a single supply voltage and single ground. A triangular current source with 50 mA amplitude, 100 ps rise time, and 150 ps fall time is applied to each grid within the power distribution network. No skew between the two current loads is assumed, modeling the worst case scenario with the maximum power noise. For each grid structure, the width of the lines varies from 1 to 10 μm, maintaining the line pair pitch P at a constant value of 40 μm (80 μm in the case of paired grids). In paired power distribution grids, the line separation inside each pair is 1 μm. The decrease in the maximum voltage drop (or the voltage sag) from V_{dd} is estimated from SPICE for different line widths.

The resistance and inductance for the power distribution grids with SSSG operating at 1 and 100 GHz are listed in Table 41.1. The resistance and inductance for the power distribution grids with DSSG operating at 1 and 100 GHz are listed in Table 41.2. Note that in the case of DSSG, only interdigitated grids can be used. The power grids with DSSG lack symmetry in both voltage domains which is necessary for paired grids. Also note that two types of interdigitated power distribution grids with DSSG can be used. Both types of interdigitated grids with DSSG are identical except for those power/ground lines located at the periphery of the power grid. Thus, the difference in loop inductance in both interdigitated grids with DSSG is negligible for a large number of power/ground lines comprising the grid. Only one interdigitated power distribution grid with DSSG is therefore analyzed. The impedance characteristics of the interdigitated and paired power distribution grids with DSDG are listed in Tables 41.3, 41.4, and 41.5. The results listed in Tables 41.1, 41.2, 41.3, 41.4, and 41.5 are discussed in Sects. 41.6.1, 41.6.2, 41.6.3, and 41.6.4.

The performance of interdigitated power distribution grids is quantitatively compared to the power noise of a conventional power distribution scheme with DSSG in Sect. 41.6.1. The maximum voltage drop from V_{dd} for paired power distribution grids is evaluated in Sect. 41.6.2. Both types of power distribution grids are compared to the reference power distribution grid with SSSG. Power distribution schemes with decoupling capacitors are compared in Sect. 41.6.3. The dependence of the power noise on the switching frequency of the current loads is discussed in Sect. 41.6.4.

Table 41.1 Impedance characteristics of power distribution grids with SSSG

	1 GHz				100 GHz			
Line cross section (μm × μm)	R_{pp}, R_{gg} (Ω)	L_{pp}, L_{gg} (nH)	L_{pg} (nH)	k	R_{pp}, R_{gg} (Ω)	L_{pp}, L_{gg} (nH)	L_{pg} (nH)	k
Interdigitated								
1 × 1	1.478	0.357	0.289	0.810	2.514	0.351	0.284	0.809
2 × 1	0.763	0.348	0.286	0.822	1.652	0.343	0.284	0.828
3 × 1	0.519	0.341	0.285	0.835	1.217	0.337	0.283	0.840
4 × 1	0.395	0.337	0.285	0.846	0.944	0.333	0.283	0.850
5 × 1	0.320	0.333	0.284	0.853	0.764	0.330	0.283	0.858
6 × 1	0.269	0.330	0.284	0.859	0.643	0.327	0.283	0.865
7 × 1	0.233	0.328	0.283	0.863	0.555	0.325	0.283	0.871
8 × 1	0.206	0.326	0.283	0.868	0.489	0.323	0.283	0.876
9 × 1	0.184	0.324	0.283	0.873	0.438	0.321	0.283	0.882
10 × 1	0.167	0.322	0.283	0.879	0.397	0.319	0.282	0.884
Paired								
1 × 1	1.467	0.357	0.332	0.930	2.652	0.352	0.329	0.935
2 × 1	0.747	0.349	0.324	0.928	1.728	0.344	0.323	0.939
3 × 1	0.504	0.343	0.319	0.930	1.274	0.338	0.319	0.944
4 × 1	0.382	0.339	0.315	0.929	0.987	0.333	0.315	0.846
5 × 1	0.309	0.335	0.312	0.931	0.798	0.330	0.312	0.845
6 × 1	0.260	0.332	0.309	0.931	0.671	0.327	0.310	0.948
7 × 1	0.225	0.330	0.307	0.930	0.580	0.325	0.308	0.948
8 × 1	0.199	0.328	0.305	0.930	0.510	0.322	0.306	0.950
9 × 1	0.179	0.326	0.303	0.929	0.456	0.321	0.304	0.949
10 × 1	0.163	0.324	0.301	0.929	0.413	0.319	0.303	0.950

Line pair pitch—40 μm, grid length—1000 μm, and $k = \frac{L_{pg}}{\sqrt{L_{pp}L_{gg}}}$—coupling coefficient

Table 41.2 Impedance characteristics of interdigitated power distribution grids with DSSG

Line cross section (μm × μm)	R_{pp}, R_{gg} (Ω)	L^{a}_{pp}, L^{a}_{gg} (nH)	L^{a}_{pg} (nH)	k^{a}	L^{b}_{pp}, L^{b}_{gg} (nH)	L^{b}_{pg} (nH)	k^{b}
1 GHz							
1 × 1	2.180	0.397	0.289	0.728	0.396	0.285	0.720
2 × 1	1.109	0.385	0.287	0.745	0.383	0.283	0.738
3 × 1	0.748	0.377	0.286	0.759	0.375	0.282	0.752
4 × 1	0.566	0.370	0.286	0.773	0.368	0.281	0.764
5 × 1	0.456	0.365	0.285	0.781	0.363	0.281	0.774
6 × 1	0.383	0.361	0.285	0.789	0.359	0.280	0.780
7 × 1	0.330	0.358	0.285	0.796	0.355	0.280	0.789
8 × 1	0.290	0.355	0.285	0.804	0.352	0.280	0.795
9 × 1	0.260	0.352	0.285	0.810	0.349	0.280	0.802
10 × 1	0.235	0.349	0.285	0.817	0.346	0.279	0.806
100 GHz							
1 × 1	3.603	0.391	0.285	0.729	0.389	0.281	0.722
2 × 1	2.357	0.379	0.285	0.752	0.377	0.280	0.743
3 × 1	1.730	0.372	0.285	0.766	0.369	0.280	0.759
4 × 1	1.338	0.366	0.285	0.779	0.363	0.280	0.771
5 × 1	1.081	0.361	0.285	0.789	0.358	0.280	0.782
6 × 1	0.908	0.357	0.284	0.796	0.354	0.279	0.788
7 × 1	0.784	0.354	0.284	0.802	0.350	0.279	0.796
8 × 1	0.691	0.351	0.284	0.809	0.347	0.279	0.803
9 × 1	0.618	0.348	0.284	0.816	0.345	0.279	0.809
10 × 1	0.560	0.346	0.284	0.821	0.342	0.279	0.816

Line pair pitch—40 μm, grid length—1000 μm

[a] denotes coupling between V_{dd1} (V_{dd2}) and Gnd

[b] denotes coupling between V_{dd1} and V_{dd2}

Table 41.3 Impedance characteristics of interdigitated power distribution grids with DSDG

Grid type	Cross section (μm×μm)	R_{pp}, R_{gg} (Ω)	L^{a}_{pp}, L^{a}_{gg} (nH)	L^{a}_{pg} (nH)	k^{a}	L^{b}_{pp}, L^{b}_{gg} (nH)	L^{b}_{pg} (nH)	k^{b}	L^{c}_{pp}, L^{c}_{gg} (nH)	L^{c}_{pg} (nH)	k^{c}
Type I	1 GHz										
	1 × 1	2.887	0.439	0.293	0.667	0.439	0.279	0.636	0.438	0.284	0.648
	2 × 1	1.458	0.424	0.292	0.689	0.423	0.277	0.654	0.422	0.282	0.668
	3 × 1	0.979	0.414	0.291	0.703	0.413	0.276	0.668	0.410	0.281	0.685
	4 × 1	0.738	0.406	0.291	0.717	0.405	0.276	0.681	0.402	0.280	0.697
	5 × 1	0.594	0.400	0.290	0.725	0.398	0.275	0.691	0.395	0.280	0.709
	6 × 1	0.497	0.394	0.290	0.736	0.393	0.275	0.700	0.389	0.279	0.717
	7 × 1	0.428	0.390	0.290	0.744	0.388	0.275	0.709	0.384	0.279	0.727
	8 × 1	0.376	0.385	0.290	0.753	0.384	0.275	0.716	0.380	0.279	0.734
	9 × 1	0.336	0.382	0.290	0.759	0.380	0.275	0.724	0.376	0.279	0.742
	10 × 1	0.304	0.379	0.290	0.766	0.376	0.274	0.728	0.372	0.278	0.747
	100 GHz										
	1 × 1	4.703	0.434	0.290	0.668	0.432	0.275	0.637	0.429	0.279	0.650
	2 × 1	3.070	0.419	0.290	0.692	0.417	0.275	0.659	0.413	0.279	0.676
	3 × 1	2.251	0.408	0.290	0.711	0.406	0.275	0.677	0.403	0.279	0.692
	4 × 1	1.739	0.401	0.290	0.723	0.399	0.275	0.689	0.395	0.279	0.706
	5 × 1	1.406	0.394	0.290	0.736	0.392	0.274	0.699	0.388	0.278	0.716
	6 × 1	1.179	0.389	0.290	0.746	0.387	0.274	0.708	0.383	0.278	0.726
	7 × 1	1.017	0.385	0.289	0.751	0.383	0.274	0.715	0.378	0.278	0.735
	8 × 1	0.896	0.381	0.289	0.759	0.379	0.274	0.723	0.374	0.278	0.743
	9 × 1	0.802	0.377	0.289	0.767	0.375	0.274	0.731	0.370	0.278	0.751
	10 × 1	0.727	0.374	0.289	0.773	0.372	0.274	0.737	0.367	0.278	0.757

Type II	1 GHz									
	1 × 1	2.893	0.439	0.279	0.636	0.439	0.293	0.667	0.438	0.284	0.648
	2 × 1	1.466	0.423	0.277	0.655	0.424	0.292	0.689	0.422	0.282	0.668
	3 × 1	0.987	0.413	0.276	0.668	0.414	0.291	0.703	0.410	0.281	0.685
	4 × 1	0.747	0.405	0.276	0.681	0.406	0.291	0.717	0.402	0.280	0.697
	5 × 1	0.601	0.398	0.275	0.691	0.400	0.290	0.725	0.395	0.280	0.709
	6 × 1	0.504	0.393	0.275	0.700	0.394	0.290	0.736	0.389	0.279	0.717
	7 × 1	0.435	0.388	0.275	0.709	0.390	0.290	0.744	0.384	0.279	0.727
	8 × 1	0.383	0.384	0.275	0.716	0.386	0.290	0.751	0.380	0.279	0.734
	9 × 1	0.342	0.380	0.275	0.724	0.382	0.290	0.759	0.376	0.279	0.742
	10 × 1	0.310	0.377	0.274	0.727	0.379	0.290	0.765	0.372	0.278	0.747
	100 GHz									
	1 × 1	4.756	0.432	0.275	0.637	0.434	0.290	0.668	0.429	0.279	0.650
	2 × 1	3.109	0.417	0.275	0.659	0.419	0.290	0.692	0.413	0.279	0.676
	3 × 1	2.281	0.406	0.275	0.677	0.408	0.290	0.711	0.403	0.279	0.692
	4 × 1	1.764	0.399	0.275	0.689	0.401	0.290	0.723	0.395	0.279	0.706
	5 × 1	1.425	0.392	0.274	0.699	0.394	0.290	0.736	0.388	0.278	0.716
	6 × 1	1.196	0.387	0.274	0.708	0.389	0.290	0.746	0.383	0.278	0.726
	7 × 1	1.031	0.383	0.274	0.715	0.385	0.290	0.753	0.378	0.278	0.735
	8 × 1	0.907	0.379	0.274	0.723	0.381	0.289	0.759	0.374	0.278	0.743
	9 × 1	0.812	0.375	0.274	0.731	0.377	0.289	0.767	0.370	0.278	0.751
	10 × 1	0.735	0.372	0.274	0.737	0.374	0.289	0.773	0.367	0.278	0.757

Line pair pitch—40 μm, grid length—1000 μm

[a] denotes coupling between V_{dd1} (V_{dd2}) and Gnd_1 (Gnd_2)

[b] denotes coupling between V_{dd1} (Gnd_1) and V_{dd2} (Gnd_2)

[c] denotes coupling between Gnd_1 and V_{dd2}

Table 41.4 Impedance characteristics of Type I paired power distribution grids with DSDG

Line cross section (μm × μm)	R_{pp}, R_{gg} (Ω)	L^{a}_{pp}, L^{a}_{gg} (nH)	L^{a}_{pg} (nH)	k^{a}	L^{b}_{pp}, L^{b}_{gg} (nH)	L^{b}_{pg} (nH)	k^{b}	L^{c}_{pp}, L^{c}_{gg} (nH)	L^{c}_{pg} (nH)	k^{c}	L^{d}_{pp}, L^{d}_{gg} (nH)	L^{d}_{pg} (nH)	k^{d}
1 GHz													
1 × 1	2.883	0.439	0.389	0.886	0.439	0.279	0.636	0.439	0.279	0.636	0.439	0.278	0.633
2 × 1	1.450	0.425	0.376	0.885	0.423	0.277	0.655	0.424	0.278	0.656	0.423	0.277	0.655
3 × 1	0.972	0.415	0.366	0.882	0.413	0.276	0.668	0.413	0.277	0.671	0.413	0.276	0.668
4 × 1	0.733	0.407	0.359	0.882	0.405	0.276	0.681	0.405	0.276	0.681	0.404	0.275	0.681
5 × 1	0.590	0.400	0.353	0.883	0.398	0.275	0.691	0.398	0.276	0.693	0.398	0.275	0.691
6 × 1	0.495	0.395	0.348	0.881	0.392	0.275	0.702	0.393	0.276	0.702	0.392	0.274	0.699
7 × 1	0.428	0.390	0.344	0.882	0.388	0.275	0.709	0.388	0.276	0.711	0.387	0.274	0.708
8 × 1	0.378	0.386	0.340	0.881	0.383	0.275	0.718	0.384	0.276	0.719	0.383	0.274	0.715
9 × 1	0.339	0.382	0.336	0.880	0.379	0.274	0.723	0.380	0.276	0.726	0.379	0.274	0.723
10 × 1	0.308	0.379	0.333	0.879	0.376	0.274	0.729	0.377	0.276	0.732	0.375	0.273	0.728
100 GHz													
1 × 1	5.121	0.434	0.388	0.894	0.431	0.275	0.638	0.431	0.275	0.638	0.431	0.275	0.638
2 × 1	3.324	0.417	0.376	0.902	0.414	0.275	0.664	0.414	0.275	0.664	0.413	0.275	0.666
3 × 1	2.441	0.405	0.367	0.906	0.402	0.275	0.684	0.402	0.275	0.684	0.402	0.274	0.682
4 × 1	1.887	0.397	0.361	0.909	0.394	0.274	0.695	0.394	0.275	0.698	0.393	0.274	0.697
5 × 1	1.525	0.390	0.355	0.910	0.387	0.274	0.708	0.387	0.275	0.711	0.387	0.274	0.708
6 × 1	1.279	0.385	0.350	0.909	0.381	0.274	0.719	0.382	0.275	0.720	0.381	0.274	0.719
7 × 1	1.102	0.380	0.246	0.911	0.377	0.274	0.727	0.377	0.275	0.729	0.376	0.274	0.729
8 × 1	0.970	0.376	0.343	0.912	0.372	0.274	0.737	0.373	0.275	0.737	0.372	0.273	0.734
9 × 1	0.867	0.372	0.339	0.911	0.369	0.274	0.743	0.369	0.275	0.745	0.368	0.273	0.742
10 × 1	0.785	0.369	0.336	0.911	0.365	0.274	0.751	0.366	0.275	0.751	0.365	0.273	0.748

Pairs pitch—80 μm, grid length—1000 μm

[a] denotes coupling between V_{dd1} and Gnd_1

[b] denotes coupling between V_{dd1} and V_{dd2}

[c] denotes coupling between V_{dd1} and Gnd_2

[d] denotes coupling between Gnd_1 and V_{dd2}

Table 41.5 Impedance characteristics of Type II paired power distribution grids with DSDG

Line cross section (μm × μm)	R_{pp}, R_{gg} (Ω)	L^{a}_{pp}, L^{a}_{gg} (nH)	L^{a}_{pg} (nH)	k^{a}	L^{b}_{pp}, L^{b}_{gg} (nH)	L^{b}_{pg} (nH)	k^{b}	L^{c}_{pp}, L^{c}_{gg} (nH)	L^{c}_{pg} (nH)	k^{c}	L^{d}_{pp}, L^{d}_{gg} (nH)	L^{d}_{pg} (nH)	k^{d}
1 GHz													
1 × 1	2.883	0.439	0.389	0.886	0.439	0.279	0.636	0.439	0.279	0.636	0.439	0.278	0.633
2 × 1	1.450	0.425	0.376	0.885	0.423	0.277	0.655	0.424	0.278	0.656	0.423	0.277	0.655
3 × 1	0.972	0.415	0.366	0.882	0.413	0.276	0.668	0.413	0.277	0.671	0.413	0.276	0.668
4 × 1	0.733	0.407	0.359	0.882	0.405	0.276	0.681	0.405	0.276	0.681	0.404	0.275	0.681
5 × 1	0.590	0.400	0.353	0.883	0.398	0.275	0.691	0.398	0.276	0.693	0.398	0.275	0.691
6 × 1	0.495	0.395	0.348	0.881	0.392	0.275	0.702	0.393	0.276	0.702	0.392	0.274	0.699
7 × 1	0.428	0.390	0.344	0.882	0.388	0.275	0.710	0.388	0.276	0.711	0.387	0.274	0.708
8 × 1	0.378	0.386	0.340	0.881	0.383	0.275	0.718	0.384	0.276	0.719	0.383	0.274	0.715
9 × 1	0.339	0.382	0.336	0.880	0.379	0.275	0.726	0.380	0.276	0.726	0.379	0.274	0.723
10 × 1	0.308	0.379	0.333	0.879	0.376	0.274	0.729	0.377	0.276	0.732	0.375	0.273	0.728
100 GHz													
1 × 1	5.122	0.434	0.388	0.894	0.431	0.275	0.638	0.431	0.275	0.638	0.431	0.275	0.638
2 × 1	3.323	0.417	0.376	0.902	0.414	0.275	0.664	0.414	0.275	0.664	0.413	0.275	0.666
3 × 1	2.442	0.405	0.367	0.906	0.402	0.275	0.684	0.402	0.275	0.684	0.402	0.274	0.682
4 × 1	1.887	0.397	0.361	0.909	0.394	0.274	0.695	0.394	0.275	0.698	0.393	0.274	0.697
5 × 1	1.522	0.390	0.355	0.910	0.387	0.274	0.708	0.387	0.275	0.711	0.387	0.274	0.708
6 × 1	1.279	0.385	0.350	0.909	0.381	0.274	0.719	0.382	0.275	0.720	0.381	0.274	0.719
7 × 1	1.103	0.380	0.346	0.911	0.377	0.274	0.728	0.377	0.275	0.729	0.376	0.274	0.729
8 × 1	0.971	0.376	0.343	0.912	0.372	0.274	0.737	0.373	0.275	0.737	0.372	0.273	0.734
9 × 1	0.868	0.372	0.339	0.911	0.369	0.274	0.743	0.369	0.275	0.745	0.368	0.273	0.742
10 × 1	0.786	0.369	0.336	0.911	0.365	0.274	0.751	0.366	0.275	0.751	0.365	0.273	0.748

Pairs pitch—80 μm, grid length—1000 μm

[a] denotes coupling between V_{dd1}—V_{dd2}

[b] denotes coupling between V_{dd1}—Gnd_1

[c] denotes coupling between V_{dd1} and Gnd_2

[d] denotes coupling between Gnd_1 and V_{dd2}

41.6.1 Interdigitated Power Distribution Grids Without Decoupling Capacitors

The maximum voltage drop for four interdigitated power distribution grids without decoupling capacitors is depicted in Fig. 41.11. For each of the power distribution grids, the maximum voltage drop decreases sublinearly as the width of the lines is increased. This noise voltage drop is caused by the decreased loop impedance. The resistance of the metal lines decreases linearly with an increase in the line width. The loop inductance increases slowly with increasing line width. As a result, the total impedance of each of the power distribution schemes decreases sublinearly, approaching a constant impedance as the lines become very wide.

As described in Sect. 41.3, the power distribution scheme with DSDG outperforms power distribution grids with DSSG. Fully interdigitated grids with DSDG produce, on average, a 15.3 % lower voltage drop as compared to the scheme with DSSG. Pseudo-interdigitated grids with DSDG produce, on average, a close to negligible 0.3 % lower voltage drop as compared to the scheme with DSSG. The maximum reduction in noise is 16.5 % for an 8 μm wide line, and 7.1 % for a 1 μm wide line, for, respectively, fully- and pseudo-interdigitated grids with DSDG. Note that pseudo-interdigitated power grids with DSDG outperform conventional power delivery schemes with DSSG for narrow lines. For wide lines, however, the power delivery scheme with DSSG results in a lower voltage drop. From the results depicted in Fig. 41.11, observe that the power delivery schemes with both DSDG and SSSG outperform the power grid with DSSG. The fully interdigitated power distribution grid with DSDG outperforms the reference power grid with

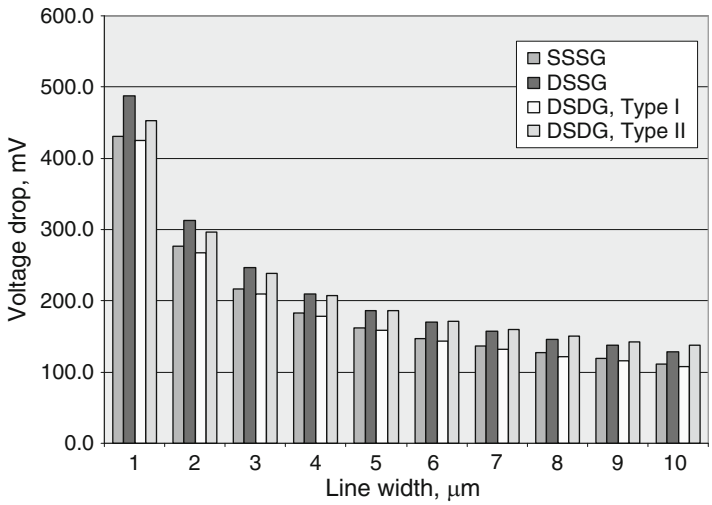


Fig. 41.11 Maximum voltage drop for the four interdigitated power distribution grids under investigation. No decoupling capacitors are added

SSSG by 2.7 %. This behavior can be explained as follows. Since the number of lines dedicated to each power delivery network in the grid with DSDG is two times smaller than the total number of lines in the reference grid, the resistance of each subnetwork is two times greater than the resistance of the reference power grid. The loop inductance of an interdigitated power distribution grid depends inversely linearly on the number of lines in the grid [73]. The loop inductance of each subnetwork is two times greater than the overall loop inductance of the grid with SSSG. Given two similar current loads applied to the reference power distribution scheme, the maximum voltage drop for both systems should be the same. However, from (41.4), the mutual inductive coupling in the power grid with DSDG increases due to the presence of the second subnetwork. As a result, the overall loop inductance of each network comprising the power grid with DSDG is lower, resulting in a lower power noise as seen from the current load of each subnetwork. Note from Fig. 41.7 that in pseudo-interdigitated power distribution grids with DSDG, the mutual inductance between two current loops M_{loop}^{intII} is positive, reducing the overall mutual inductance. The resulting loop inductance as seen from the load of the particular network is therefore increased, producing a larger inductive voltage drop. In many applications such as high performance microprocessors, mixed-signal circuits, and systems-on-chip, a power distribution network with DSDG is often utilized. In other applications, however, a fully interdigitated power distribution system with multiple voltages and multiple grounds can be a better alternative than distributing power with SSSG.

41.6.2 Paired Power Distribution Grids Without Decoupling Capacitors

The maximum voltage drop for three paired power distribution grids without decoupling capacitors is depicted in Fig. 41.12. Similar to interdigitated grids, the maximum voltage drop decreases sublinearly with increasing line width. Observe that fully paired power distribution grids with DSDG outperform conventional paired power distribution grids with SSSG by, on average, 2.3 %. Note the information shown in Fig. 41.12, the ratio of the separation between the pairs to the distance between the lines in each pair (n) is eighty. Also note from Fig. 41.10 that the total mutual inductance in fully paired grids increases as n is decreased (the pairs are placed physically closer). Thus, better performance is achieved in fully paired grids with DSDG for densely placed pairs. In contrast to fully paired grids, in pseudo-paired grids with DSDG, the total mutual inductance is reduced by inductive coupling between the two current loops M_{loop}^{prdII}. For $n > 8$ (see Fig. 41.10), the mutual inductive coupling between the two current loops in pseudo-paired grid becomes comparable to the mutual inductive coupling between the two current loops in the conventional power grid with DSSG (the $-2M$ term in (41.2) becomes positive). As n further increases, the power and ground paths within the two voltage domains become strongly coupled, increasing the loop inductance.

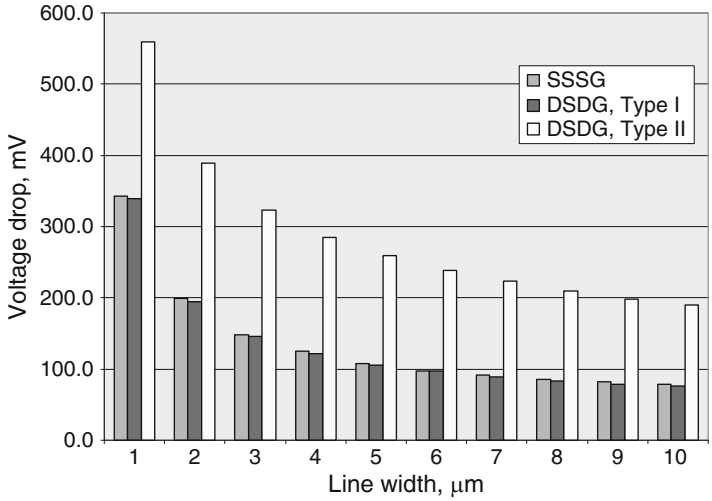


Fig. 41.12 Maximum voltage drop for the three paired power distribution grids under investigation. No decoupling capacitors are added

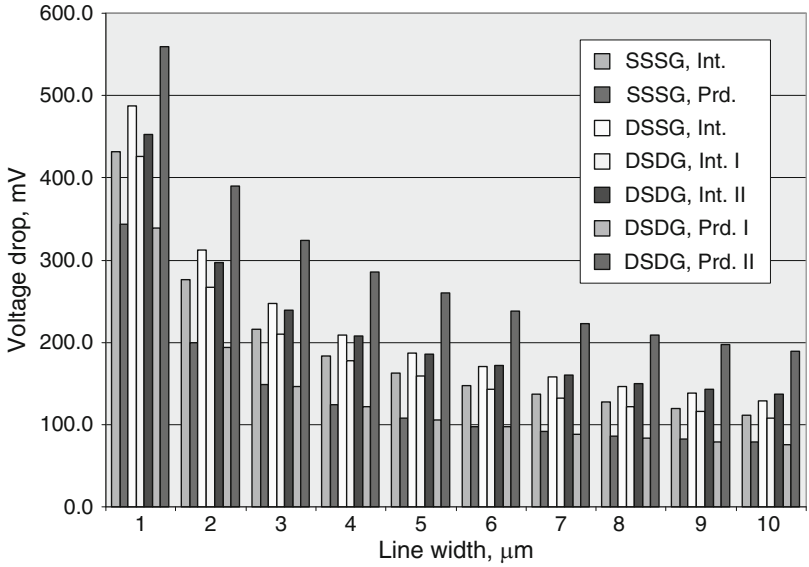


Fig. 41.13 Maximum voltage drop for interdigitated and paired power distribution grids under investigation. No decoupling capacitors are added

To quantitatively compare interdigitated grids to paired grids, the maximum voltage drop for seven different types of power distribution grids without decoupling capacitors is plotted in Fig. 41.13. Note in Fig. 41.13 that the conventional power delivery scheme with DSSG results in larger voltage fluctuations as compared to

fully interdigitated grids with DSDG. The performance of pseudo-interdigitated grids with DSDG is comparable to the performance of the conventional delivery scheme with DSSG. In pseudo-interdigitated grids, the positive mutual inductance between two current loops lowers the overall negative mutual inductance. The loop inductance in the specific power delivery network is therefore increased, resulting in greater power noise. Analogous to the conventional scheme, in pseudo-paired grids, the power and ground paths in different voltage domains are strongly coupled, producing the largest voltage drop. Both fully interdigitated and fully paired power distribution grids with DSDG produce the lowest voltage fluctuations, slightly outperforming the reference power delivery network with SSSG. In these grids, the resulting loop inductance is reduced due to strong coupling between the power/ground pairs from different voltage domains (with currents flowing in opposite directions). Alternatively, the total mutual inductance is negative with large magnitude, reducing the loop inductance. Both fully interdigitated and fully paired power distribution grids with DSDG should be used in those systems with multiple power supply voltages. Fully interdigitated and fully paired power distribution grids with DSDG can also be a better alternative than a power distribution grid with SSSG.

41.6.3 Power Distribution Grids with Decoupling Capacitors

To lower the voltage fluctuations of on-chip power delivery systems, decoupling capacitors are placed on ICs to provide charge when the voltage drops [30]. The maximum voltage drop of seven power distribution schemes with decoupling capacitors operating at 1 GHz is shown in Fig. 41.14. All of the decoupling capacitors are assumed to be ideal, i.e., no parasitic resistances and inductances are associated with the capacitor. Also, all of the decoupling capacitors are assumed to be effective (located inside the effective radius of an on-chip decoupling capacitor, as described in Chap. 12 [277]). The total budgeted capacitance is divided equally between the two supply voltages. The decoupling capacitor added to the power distribution grid with SSSG is two times larger than the decoupling capacitor in each subnetwork of the power delivery scheme with dual voltages. As shown in Fig. 41.14, the maximum voltage drop decreases as the lines become wider. The maximum voltage drop of the fully interdigitated power distribution scheme with DSDG is reduced by, on average, 9.2 % (13.6 % maximum) for a 30 pF decoupling capacitance as compared to a conventional power distribution scheme with DSSG. For a 20 pF decoupling capacitance, however, a fully interdigitated power distribution grid with DSDG produces about 55 % larger power noise as compared to a conventional power distribution scheme with DSSG. This performance degradation is caused by on-chip resonances, as explained below.

Comparing the data shown in Fig. 41.13 to that shown in Fig. 41.14, note that the voltage drop of the power distribution grids with decoupling capacitors as compared to the case with no decoupling capacitances is greatly reduced for narrow lines and is higher for wider lines. This behavior can be explained as follows. For narrow

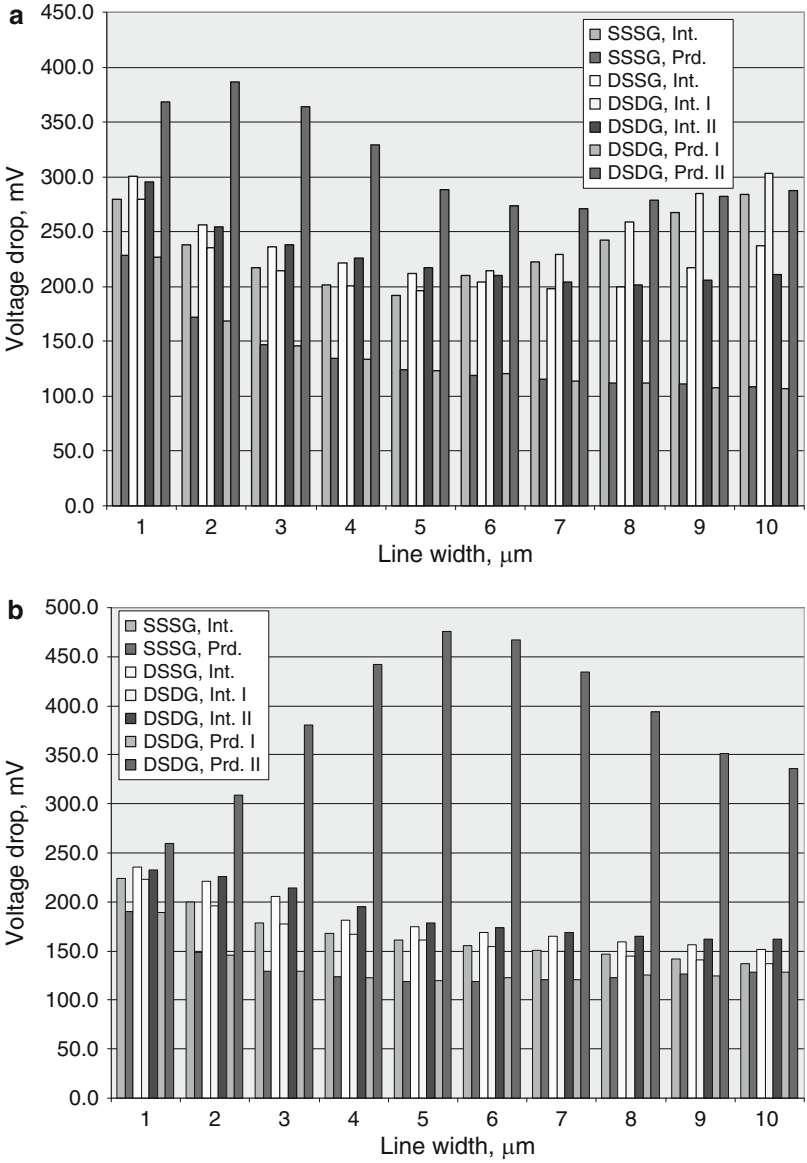


Fig. 41.14 Maximum voltage drop for seven types of power distribution grids with a decoupling capacitance of (**a**) 20 pF and (**b**) 30 pF added to each power supply. The switching frequency of the current loads is 1 GHz

lines, the grid resistance is high and the loop inductance is low. The grid impedance, therefore, is primarily determined by the resistance of the lines. Initially, the system with an added decoupling capacitor is overdamped. As the lines become wider,

the grid resistance decreases faster than the increase in the loop inductance and the system becomes less damped. As the loop inductance increases, the resonant frequency of an *RLC* circuit, formed by the on-chip decoupling capacitor and the parasitic *RL* impedance of the grid, decreases. This resonant frequency moves closer to the switching frequency of the current load. As a result, the voltage response of the overall system oscillates. Since the decoupling capacitance added to the power grid with SSSG is two times larger than the decoupling capacitance added to each power supply voltage in the dual voltage schemes, the system with a single supply voltage is more highly damped and the self-resonant frequency is significantly lower. Furthermore, the resonant frequency is located far from the switching frequency of the circuit.

For narrow lines propagating a signal with 1 GHz harmonics, the resulting power noise in fully interdigitated power grids with DSDG with 20 pF added on-chip decoupling capacitance is smaller than the power noise of the power distribution scheme with SSSG, as shown in Fig. 41.14a. With increasing line width, the inductance of the power grids increases more slowly than the decrease in the grid resistance. An *RLC* system formed by the *RL* impedance of the power grid and the decoupling capacitance, therefore, is less damped. Both of the power distribution grids with DSDG and the conventional power distribution grid with SSSG result in larger voltage fluctuations as the line width increases. The self-resonant frequency of the fully interdigitated grid with DSDG is almost coincident with the switching frequency of the current load. The self-resonant frequency of the power grid with SSSG however is different from the switching frequency of the current source. Thus, for wide lines, a conventional power delivery scheme with SSSG outperforms the fully interdigitated power distribution grid with DSDG. Note that the loop inductance in pseudo-interdigitated power distribution grids with DSDG is greater than the loop inductance in fully interdigitated grids. As a result, the self-resonant frequency of a pseudo-interdigitated grid with DSDG is smaller than the switching frequency of the current load, resulting in smaller power noise as compared to power grids with SSSG and fully interdigitated grids with DSDG. Also note that the loop inductance in paired power distribution grids is further reduced as compared to interdigitated grids. In this case, the self-resonant frequency of all of the paired power distribution grids is greater than the circuit switching frequency. Thus, the power noise in paired power distribution grids gradually decreases as the line width increases (and is slightly higher in wide lines in the case of pseudo-paired grids).

Increasing the on-chip decoupling capacitance from 20 to 30 pF further reduces the voltage drop. For a 30 pF decoupling capacitance in a pseudo-paired power delivery scheme with DSSG, the self-resonant frequency is close to the switching frequency of the current load. Simultaneously, the grid resistance decreases much faster with increasing line width than the increase in the loop inductance. The system becomes underdamped with the self-resonant frequency equal to the circuit switching frequency. As a result, the system produces high amplitude voltage fluctuations. The maximum voltage drop in the case of a pseudo-paired power grid with DSDG therefore increases as the lines become wider. This phenomenon is illustrated in Fig. 41.14b for a line width of 5 μm.

With decoupling capacitors, the self-resonant frequency of an on-chip power distribution system is lowered. If the resonant frequency of an *RLC* system with intentionally added decoupling capacitors is sufficiently close to the circuit switching frequency, the system will produce high amplitude voltage fluctuations. Voltage sagging will degrade system performance and may cause significant failure. An excessively high power supply voltage can degrade the reliability of a system. The decoupling capacitors for power distribution systems with multiple supply voltages therefore have to be carefully designed. Improper choice (magnitude and location) of the on-chip decoupling capacitors can therefore worsen the power noise, further degrading system performance [28, 276].

41.6.4 Dependence of Power Noise on the Switching Frequency of the Current Loads

To model the dependence of the power noise on the switching frequency, the power grids are stimulated with triangular current sources with a 50 mA amplitude, 20 ps rise times, and 30 ps fall times. The switching frequency of each current source varies from 1 to 10 GHz to capture the resonances in each power grid. For each grid structure, the width of the line is varied from 1 to 10 μm. The maximum voltage drop is determined from SPICE for different line widths at each frequency.

The maximum voltage drop for the power distribution grid with SSSG is illustrated in Fig. 41.15. The maximum voltage drop decreases slightly for wider lines. Note that with decoupling capacitors, the voltage drop is lower except for two regions. The significant increase in power noise at specific frequencies and line widths is due to the following two effects. As lines become wider, the resistance of the power grid is lower, whereas the inductance is slightly increased, decreasing the damping of the entire system. When the switching frequency of a current load approaches the self-resonant frequency of the power grid, the voltage drop due to the *RLC* system increases (due to resonances). As the width of the lines increases, the system becomes more underdamped, resulting in a sharper resonant peak. The amplitude of the resonant peak increases rapidly as the system becomes less damped. The maximum voltage drop occurs between 6 and 7 GHz for a power grid with a 20 pF decoupling capacitance, as shown in Fig. 41.15a.

The maximum voltage drop also increases at high frequencies in narrow lines. Decoupling capacitors are effective only if the capacitor is fully charged within one clock cycle. The effectiveness of the decoupling capacitor is related to the *RC* time constant, where *R* is the resistance of the interconnect connecting the capacitor to the power supply. For narrow resistive lines, the time constant is prohibitively large at high frequencies, i.e., the decoupling capacitor cannot be fully charged within one clock period. The effective magnitude of the decoupling capacitor is therefore reduced. The capacitor has the same effect on the power noise as a smaller capacitor [277].

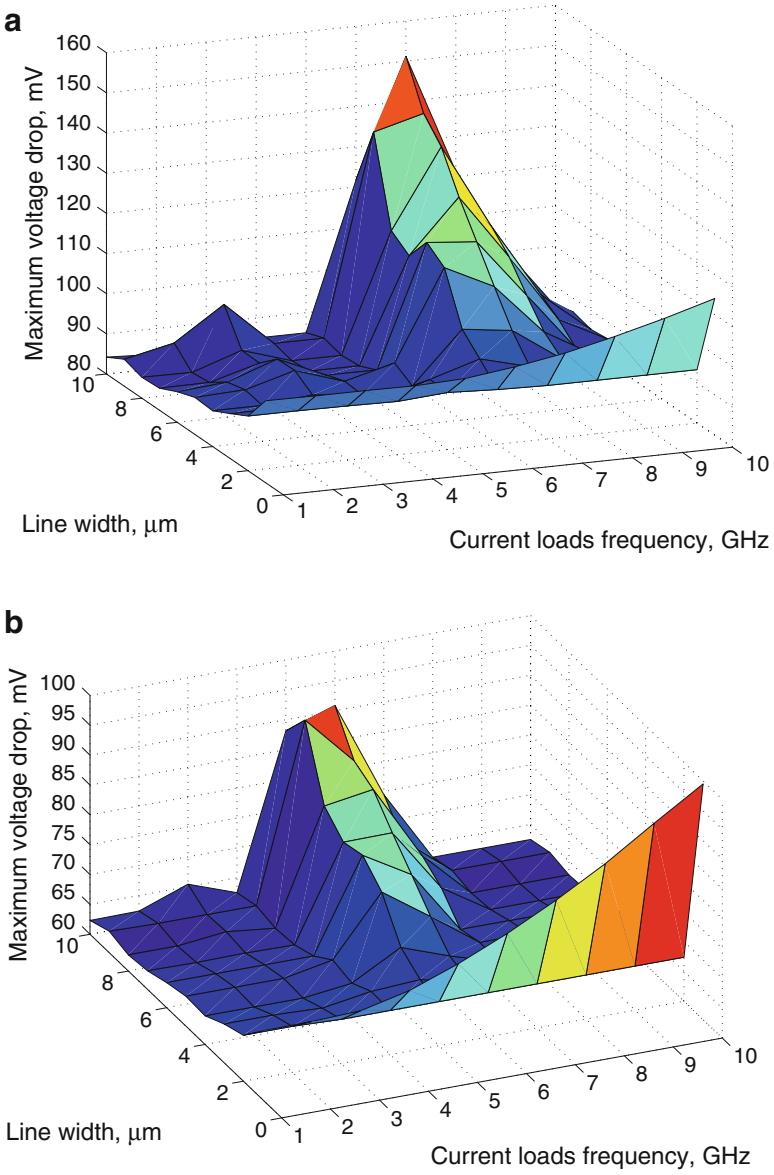


Fig. 41.15 Maximum voltage drop for the power distribution grid with SSSG as a function of frequency and line width for different values of decoupling capacitance; (**a**) decoupling capacitance budget of 20 pF, (**b**) decoupling capacitance budget of 30 pF

By increasing the magnitude of the decoupling capacitor, the overall power noise can be further reduced, as shown in Fig. 41.15b. Moreover, the system becomes more damped, producing a resonant peak with a smaller amplitude. The self-resonant frequency of the power delivery system is also lowered. Comparing Fig. 41.15a to b, note that the resonant peak shifts in frequency from approximately 6–7 GHz for a 20 pF decoupling capacitance to 5–6 GHz for a 30 pF decoupling capacitance. Concurrently, increasing the decoupling capacitor increases the *RC* time constant, making the capacitor less effective at high frequencies in narrow resistive lines. Note the significant increase in the maximum voltage drop for a 1 μm wide line for a 30 pF decoupling capacitance as compared to the case of a 20 pF decoupling capacitance. Power distribution grids with DSSG and DSDG behave similarly. For the same decoupling capacitance and for the non-resonant case, both the fully- and pseudo-interdigitated power distribution schemes with DSDG result in a lower voltage drop than a power distribution scheme with DSSG. The magnitude of the decoupling capacitance needs to be carefully chosen to guarantee that the two prohibited regions are outside the operating frequency of the system for a particular line width. Also, for narrow lines, the magnitude of the decoupling capacitor is limited by the *RC* time constant. The amplitude of the resonant peak can be lowered by increasing the parasitic resistance of the decoupling capacitors.

41.7 Design Implications

Historically, due to low switching frequencies and the high resistance of on-chip interconnects, resistive voltage drops have dominated the overall power noise. In modern high performance ICs, the inductive component of the power distribution noise has become comparable to the resistive noise [30]. It is expected that in future nanoscale ICs, the inductive $L\ dI/dt$ voltage drop will dominate the resistive *IR* voltage drop, becoming the primary component of the overall power noise [129]. As shown previously, the performance of the power delivery schemes with DSDG depends upon the switching frequency of the current load, improving with frequency (due to increased mutual coupling between the power and ground lines). It is expected that the performance of the power distribution grids with DSDG will increase with technology scaling.

As discussed in Sect. 41.6, fully interdigitated power distribution grids with DSDG outperform pseudo-interdigitated grids with DSDG. Moreover, in pseudo-interdigitated grids, the power/ground lines from different voltage domains are placed next to each over, increasing the coupling between the different power supply voltages. Pseudo-interdigitated power distribution grids with DSDG should therefore not be used in those ICs where high isolation is required between the power supply voltages (e.g., mixed-signal ICs, systems-on-chip). Rather, fully interdigitated power distribution grids with DSDG should be utilized.

Similar to interdigitated grids, fully paired power distribution grids with DSDG produce smaller power noise as compared to pseudo-paired power distribution grids

with DSDG. In pseudo-paired grids, the separation between the power/ground lines from different voltage domains is much smaller than the distance between the power and ground lines inside each power delivery network (current loop). Different power supply voltages are therefore strongly coupled in pseudo-paired grids. Note that pseudo-paired grids have the greatest coupling between different power supplies among all of the power distribution schemes described in this chapter. Such grids, therefore, are not a good choice for distributing power in mixed-signal ICs. Later in the design flow, when it is prohibitively expensive to redesign the power distribution system, the spacing between the pairs in pseudo-paired grids with DSDG should be decreased. If the pairs are placed close to each over (n is small), as illustrated in Fig. 41.10, the loop inductance of a particular current loop is lowered, approaching the loop inductance in pseudo-interdigitated grids.

The self-resonant frequency of a system is determined by the power distribution network. For example, in power distribution grids with DSDG, the decoupling capacitance added to each power delivery network is two times smaller than the decoupling capacitance in the power delivery scheme with SSSG. The loop inductance of power distribution grids with DSDG is comparable however to the loop inductance of power distribution grids with SSSG. Assuming the same decoupling capacitance, the self-resonant frequency of power distribution grids with DSDG is higher than the self-resonant frequency of the reference power delivery scheme with SSSG, increasing the maximum operating frequency of the overall system. Note that for comparable resonant frequencies, the resistance of the power distribution grid with DSDG is two times greater than the resistance of a conventional power grid with SSSG. Thus, power distribution grids with DSDG are more highly damped, resulting in reduced voltage fluctuations at the resonant frequency. Also note that on-chip decoupling capacitors lower the resonant frequency of the system. On-chip power distribution grids with decoupling capacitors should therefore be carefully designed to avoid (and control) any on-chip resonances.

Power distribution grids operating at 1 GHz (the low frequency case) have been analyzed in this chapter. Comparing the results listed in Tables 41.1, 41.2, 41.3, 41.4, and 41.5, the mutual inductive coupling at 100 GHz (the high frequency case) increases, reducing the loop inductance. Thus, for future generations of ICs operating at high frequencies [286], the performance of power distribution grids with DSDG is expected to improve by reducing the power distribution noise.

41.8 Summary

Power distribution grids with multiple power supply voltages are analyzed in this chapter. The primary results can be summarized as follows.

- Two types of interdigitated and paired on-chip power distribution grids with DSDG are presented

- Closed-form expressions to estimate the loop inductance in four types of power distribution grids with DSDG have been developed
- With no decoupling capacitors placed between the power supply and ground, fully- and pseudo-interdigitated power distribution grids outperform a conventional interdigitated power distribution grid with DSSG by 15.3 % and 0.3 %, respectively, in terms of lower power noise
- In the case of power grids with decoupling capacitors, the voltage drop is reduced by about 9.2 % for fully interdigitated grids with a 30 pF additional decoupling capacitance and is higher by 55.4 % in the case of an added 20 pF decoupling capacitance
- If no decoupling capacitors are added, the voltage drop of a fully interdigitated power distribution grid with DSDG is reduced by 2.7 %, on average, as compared to the voltage drop of an interdigitated power distribution grid with SSSG
- In the case of a fully paired grid, the power noise is reduced by about 2.3 % as compared to the reference paired power distribution grid with SSSG
- With on-chip decoupling capacitors added to the power delivery networks, both fully interdigitated and fully paired power distribution grids with DSDG slightly outperform the reference power distribution scheme with SSSG
- On-chip decoupling capacitors are shown to lower the self-resonant frequency of the on-chip power distribution grid, producing resonances. An improper choice of the on-chip decoupling capacitors can therefore degrade the overall performance of a system
- Fully interdigitated and fully paired power distribution grids with DSDG should be utilized in those ICs where high isolation is required between the power supply voltages so as to effectively decouple the power supplies

Chapter 42
Decoupling Capacitors for Multi-Voltage Power Distribution Systems

Power dissipation has become a critical design issue in high performance microprocessors as well as battery powered and wireless electronics, multimedia and digital signal processors, and high speed networking. The most effective way to reduce power consumption is to lower the supply voltage. Reducing the supply voltage, however, increases the circuit delay [596, 598, 627]. The increased delay can be compensated by changing the critical paths with behavioral transformations such as parallelization or pipelining [628]. The resulting circuit consumes less power while satisfying global throughput constraints at the cost of increased circuit area.

Recently, the use of multiple on-chip supply voltages has become common practice [609]. This strategy has the advantage of permitting modules along the critical paths to operate with the highest available voltage level (in order to satisfy target timing constraints) while permitting modules along the non-critical paths to use a lower voltage (thereby reducing the energy consumption). A multi-voltage scheme lowers the speed of those circuits operating at a lower power supply voltage without affecting the overall frequency, thereby reducing power without decreasing the system frequency. In this manner, the energy consumption is decreased without affecting circuit speed. This scheme results in a smaller area as compared to parallel architectures. The problem of using multiple supply voltages for reducing the power requirements has been evaluated in the area of high level synthesis for low power [605, 624]. While it is possible to provide many supply voltages, in practice such a scenario is expensive. Practically, the availability of a small number of voltage supplies (two or three) is reasonable, as discussed in Chap. 40.

The design of the power distribution system has become an increasingly difficult challenge in modern CMOS circuits [30]. As CMOS technologies are scaled, the power supply voltage is lowered. As clock rates rise and more functions are integrated on-chip, the power consumed has greatly increased. Assuming that only

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0_42

a small per cent of the power supply voltage (about 10 %) is permitted as ripple voltage (noise), a target impedance for an example power distribution system is [136]

$$Z_{\text{target}} = \frac{V_{\text{dd}} \times \zeta}{I} = \frac{1.8\ \text{V} \times 10\,\%}{100\ \text{A}} \approx 0.002\ \Omega, \tag{42.1}$$

where V_{dd} is the power supply voltage, ζ is the allowed ripple voltage, and I is the current. With general scaling theory [120], the current I is increasing and the power supply voltage is decreasing. The impedance of a power distribution system should therefore be decreased to satisfy power noise constraints. The target impedance of a power distribution system is falling at an alarming rate, a factor of 5 per computer generation [629]. The target impedance must be satisfied not only at DC, but also at all frequencies where current transients exist [213]. Several major components of a power delivery system are used to satisfy a target impedance over a broad frequency range. A voltage regulator module is effective up to about 1 kHz. Bulk capacitors supply current and maintain a low power distribution system impedance from 1 kHz to 1 MHz. High frequency ceramic capacitors maintain the power distribution system impedance from 1 MHz to several hundred MHz. On-chip decoupling capacitors can be effective above 100 MHz.

By introducing a second power supply, the power supplies are coupled through a decoupling capacitor effectively placed between the two power supply networks. Assuming a power delivery system with dual power supplies and only a small per cent of the power supply voltage is permitted as ripple voltage (noise), the following inequality for the magnitude of a voltage transfer function K_V should be satisfied,

$$|K_V| \leq \frac{\chi\, V_{\text{dd1}}}{V_{\text{dd2}}}, \tag{42.2}$$

where V_{dd1} is a lower voltage power supply, χ is the allowed ripple voltage on a lower voltage power supply, and V_{dd2} is a higher voltage power supply. Since the higher voltage power supply is applied to the high speed paths, as for example a clock distribution network, V_{dd2} can be noisy. To guarantee that noise from the higher voltage supply does not affect the quiet power supply, (42.2) should be satisfied. For typical values of the power supply voltages and allowed ripple voltage for a CMOS 0.18 μm technology, $|K_V|$ is chosen to be less than or equal to 0.1 to effectively decouple a noisy power supply from a quiet power supply. The design of a power distribution system with multiple supply voltages is the primary focus of this chapter. The influence of a second supply voltage on a system of decoupling capacitors is evaluated. Noise coupling among multiple power distribution systems is also discussed in this chapter. A criterion for producing an overshoot-free voltage response is determined. It is shown that to satisfy a target specification in order to decouple multiple power supplies, it is necessary to maintain the magnitude of the voltage transfer function below 0.1. In certain cases, it is difficult to satisfy this criterion over the entire range of operating frequencies.

In such a scenario, the frequency range of an overshoot-free voltage response can be traded off with the magnitude of the response [28]. Case studies are also presented in the chapter to quantitatively illustrate this methodology for designing a system of decoupling capacitors.

This chapter is organized as follows. The impedance of a power distribution system with multiple supply voltages is described in Sect. 42.1. A case study of the dependence of the impedance on the power distribution system parameters is presented in Sect. 42.2. The voltage transfer function of a power distribution system with multiple supply voltages is discussed in Sect. 42.3. Case studies examining the dependence of the magnitude of the voltage transfer function on the parameters of the power distribution system are illustrated in Sect. 42.4. Some specific conclusions are summarized in Sect. 42.5.

42.1 Impedance of a Power Distribution System

The impedance of a power distribution network is an important issue in modern high performance ICs such as microprocessors. The impedance should be maintained below a target level to guarantee the power and signal integrity of a system, as described in Chap. 7. The impedance of a power distribution system with multiple power supplies is described in Sect. 42.1.1. The antiresonance of capacitors connected in parallel is addressed in Sect. 42.1.2. The dependence of the impedance on the power delivery system is evaluated in Sect. 42.1.3.

42.1.1 Impedance of a Power Distribution System

A model of the impedance of a power distribution system with two supply voltages is shown in Fig. 42.1. The impedance seen from the load of the power supply V_{dd1} is illustrated. The model of the impedance is applicable for the load of the power supply V_{dd2} if Z_1 is substituted for Z_2. The impedance of the power distribution system shown in Fig. 42.1 can be modeled as

$$Z = \frac{Z_1 Z_{12} + Z_1 Z_2}{Z_1 + Z_{12} + Z_2}. \tag{42.3}$$

Decoupling capacitors have traditionally been modeled as a series *RLC* network [114]. A schematic representation of a power distribution network with two supply voltages and the decoupling capacitors represented by *RLC* series networks is shown in Fig. 42.2.

In this case, the impedance of the power distribution network is

$$Z = \frac{a_4 s^4 + a_3 s^3 + a_2 s^2 + a_1 s + a_0}{b_3 s^3 + b_2 s^2 + b_1 s}, \tag{42.4}$$

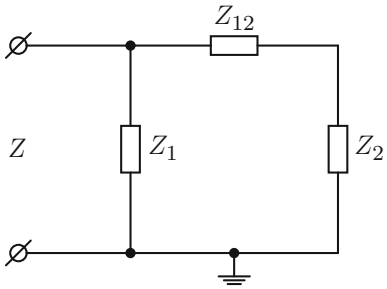


Fig. 42.1 Impedance of power distribution system with two supply voltages seen from the load of the power supply V_{dd1}

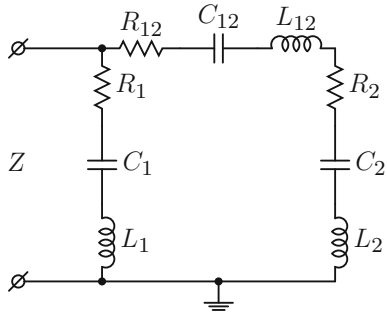


Fig. 42.2 Impedance of power distribution system with two supply voltages and the decoupling capacitors represented as series *RLC* networks

where

$$a_4 = L_1(L_{12} + L_2), \tag{42.5}$$

$$a_3 = R_1L_{12} + R_{12}L_1 + R_1L_2 + R_2L_1, \tag{42.6}$$

$$a_2 = R_1R_{12} + R_1R_2 + \frac{L_1}{C_{12}} + \frac{L_{12}}{C_1} + \frac{L_1}{C_2} + \frac{L_2}{C_1}, \tag{42.7}$$

$$a_1 = \frac{R_1}{C_2} + \frac{R_2}{C_1} + \frac{R_1}{C_{12}} + \frac{R_{12}}{C_1}, \tag{42.8}$$

$$a_0 = \frac{C_{12} + C_2}{C_1C_{12}C_2}, \tag{42.9}$$

$$b_3 = L_1 + L_{12} + L_2, \tag{42.10}$$

$$b_2 = R_1 + R_{12} + R_2, \tag{42.11}$$

$$b_1 = \frac{1}{C_1} + \frac{1}{C_{12}} + \frac{1}{C_2}, \tag{42.12}$$

and $s = j\omega$ is a complex frequency.

The frequency dependence of the closed-form expression for the impedance of a power distribution system with dual power supply voltages is illustrated in Fig. 42.3. The minimum power distribution system impedance is limited by the ESR of the decoupling capacitors. For on-chip applications, the ESR includes the parasitic

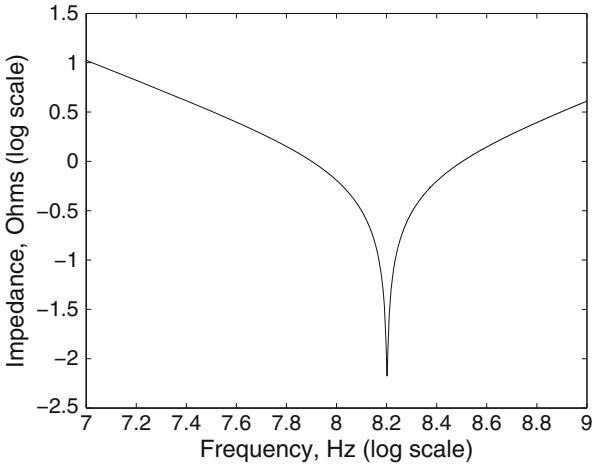


Fig. 42.3 Frequency dependence of the impedance of a power distribution system with dual supply voltages, $R_1 = R_{12} = R_2 = 10\,\text{m}\Omega$, $C_1 = C_{12} = C_2 = 1\,\text{nF}$, and $L_1 = L_{12} = L_2 = 1\,\text{nH}$. Since all of the parameters of a power distribution system are identical, the system behaves as a single capacitor with one minimum at the resonant frequency. The minimum power distribution system impedance is limited by the ESR of the decoupling capacitors

resistance of the decoupling capacitor and the resistance of the power distribution network connecting a decoupling capacitor to a load. The resistance of the on-chip power distribution network is greater than the parasitic resistance of the on-chip decoupling capacitors. For on-chip applications, therefore, the ESR is represented by the resistance of the power delivery system. Conversely, for printed circuit board applications, the resistance of the decoupling capacitors dominates the resistance of the power delivery system. In this case, therefore, the ESR is primarily the resistance of the decoupling capacitors. In order to achieve a target impedance as described by (42.1), multiple decoupling capacitors are placed at different levels of the power grid hierarchy [136].

As described in [138], the ESR of the decoupling capacitors does not change the location of the poles and zeros of the power distribution system impedance, only the damping factor of the *RLC* system formed by the decoupling capacitor is affected. Representing a decoupling capacitor with a series *LC* network, the impedance of the power distribution system with dual power supply voltages is

$$Z = \frac{a_4 s^4 + a_2 s^2 + a_0}{b_3 s^3 + b_1 s}, \tag{42.13}$$

where

$$a_4 = L_1(L_{12} + L_2), \tag{42.14}$$

$$a_2 = \frac{L_1}{C_{12}} + \frac{L_{12}}{C_1} + \frac{L_1}{C_2} + \frac{L_2}{C_1}, \tag{42.15}$$

$$a_0 = \frac{C_{12} + C_2}{C_1 C_{12} C_2}, \tag{42.16}$$

$$b_3 = L_1 + L_{12} + L_2, \tag{42.17}$$

$$b_1 = \frac{1}{C_1} + \frac{1}{C_{12}} + \frac{1}{C_2}. \tag{42.18}$$

42.1.2 Antiresonance of Parallel Capacitors

To maintain the impedance of a power distribution system below a specified level, multiple decoupling capacitors are placed in parallel at different levels of the power grid hierarchy. The ESR affects the quality factor of the *RLC* system by acting as a damping element. The influence of the ESR on the impedance is therefore ignored. If all of the parameters of the circuit shown in Fig. 42.2 are equal, the impedance of the power distribution system can be described as a series *RLC* circuit. Expression (42.13) has four zeros and three poles. Two zeros are located at the same frequency as the pole when all of the parameters of the circuit are equal. The pole is therefore canceled for this special case and the circuit behaves as a series *RLC* circuit with one resonant frequency.

If the parameters of the power distribution system are not equal, the zeros of (42.13) are not paired. In this case, the pole is not canceled by a zero. For instance, in the case of two capacitors connected in parallel as shown in Fig. 42.4, in the frequency range from f_1 to f_2, the impedance of the capacitor C_1 has become inductive whereas the impedance of the capacitor C_2 remains capacitive. In this case, an *LC* tank will produce a peak at a resonant frequency located between f_1 and f_2. Such a phenomenon is called *antiresonance* [136] and is described in greater detail in Chap. 11.

The location of the antiresonant spike depends on the ratio of the ESL of the decoupling capacitors. Depending upon the parasitic inductance, the peak impedance caused by the decoupling capacitor is shifted to a different frequency, as shown in Fig. 42.4. For instance, if the parasitic inductance of C_1 is greater than the

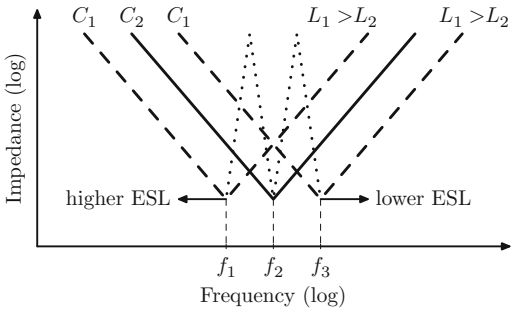


Fig. 42.4 Antiresonance of the two capacitors connected in parallel, $C_2 = C_1$. Two antiresonant spikes appear between frequencies f_1 and f_2 and f_2 and f_3 (*dotted lines*)

parasitic inductance of C_2, the antiresonance will appear at a frequency ranging from f_1 to f_2, i.e., before the self-resonant frequency f_2 of the capacitor C_2. If the parasitic inductance of C_1 is lower than the parasitic inductance of C_2, the antiresonance will appear at a frequency ranging from f_2 to f_3, i.e., after the self-resonant frequency of the capacitor C_2. The ESL of the decoupling capacitors, therefore, determines the frequency (location) of the antiresonant spike of the system [29].

42.1.3 Dependence of Impedance on Power Distribution System Parameters

In practical applications, a capacitor C_{12} placed between V_{dd1} and V_{dd2} exists either as a parasitic capacitance or as a decoupling capacitor. Intuitively, from Fig. 42.2, by decreasing the impedance Z_{12} (increasing C_{12}), the greater part of Z_2 is connected in parallel with Z_1, reducing the impedance of the power distribution system as seen from the load of the power supply V_{dd1}. The value of a parasitic capacitance is typically much smaller than a decoupling capacitor such as C_1 and C_2. The decoupling capacitor C_{12} can be chosen to be equal to or greater than C_1 and C_2. Depending upon the placement of the decoupling capacitors, ESL can vary from 50 nH at the power supply to almost negligible values on-chip. The ESL includes both the parasitic inductance of the decoupling capacitors and the inductance of the power delivery system. For on-chip applications, the inductance of the decoupling capacitors is much smaller than the inductance of the power distribution network and can be ignored. At the board level, however, the parasitic inductance of the decoupling capacitors dominates the overall inductance of a power delivery system. For these reasons, the model depicted in Fig. 42.2 is applicable to any hierarchical level of a power distribution system from the circuit board to on-chip.

Assuming $C_1 = C_2$, if $C_{12} > C_1$, an antiresonance spike occurs at a lower frequency than the resonance frequency of an *RLC* series circuit. If $C_{12} < C_1$, the antiresonance spike occurs at a higher frequency than the resonance frequency of an *RLC* series circuit. This phenomenon is illustrated in Fig. 42.5.

Antiresonance is highly undesirable because at a particular frequency, the impedance of a power distribution network can become unacceptably high. To cancel the antiresonance at a given frequency, a smaller decoupling capacitor is placed in parallel, shifting the antiresonance spike to a higher frequency. This procedure is repeated until the antiresonance spike appears at a frequency out of range of the operating frequencies of the system, as shown in Fig. 42.6.

Another technique for shifting the antiresonance spike to a higher frequency is to decrease the ESL of the decoupling capacitor. The dependence of the impedance of a power distribution system on the ESL is discussed below.

To determine the location of the antiresonant spikes, the roots of the denominator of (42.13) are evaluated. One pole is located at $\omega = 0$. Two other poles are located at frequencies,

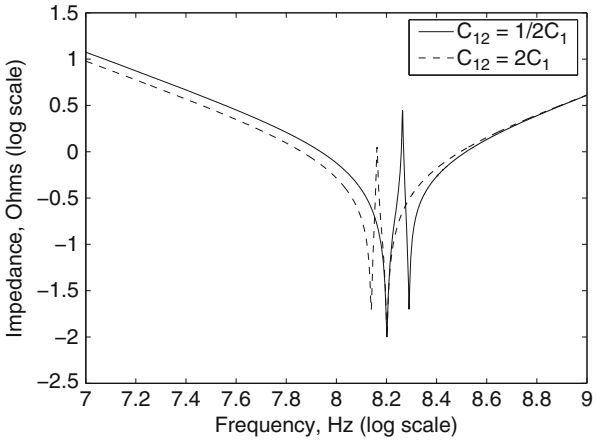


Fig. 42.5 Antiresonance of a power distribution system with dual power supply voltages, $R_1 = R_{12} = R_2 = 10\,\text{m}\Omega$, $C_1 = C_2 = 1\,\text{nF}$, and $L_1 = L_{12} = L_2 = 1\,\text{nH}$. Depending upon the ratio of C_{12} to C_1, the antiresonance appears before or after the resonant frequency of the system (the impedance minimum)

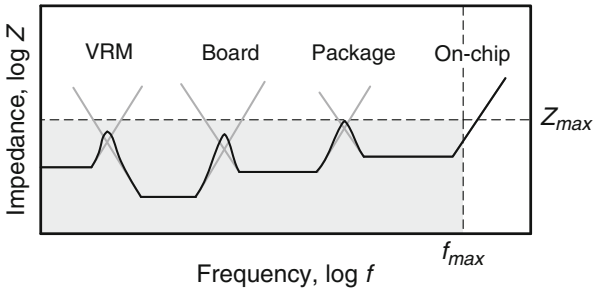


Fig. 42.6 Impedance of the power distribution system as a function of frequency. Decoupling capacitors are placed at different hierarchical levels to shift an antiresonant spike above the maximum operating frequency of the system

$$\omega = \pm\sqrt{\frac{C_2 + C_1C_2/C_{12} + C_1}{C_1C_2(L_1 + L_{12} + L_2)}}. \tag{42.19}$$

To shift the poles to a higher frequency, the ESL of the decoupling capacitors must be decreased. If the ESL of the decoupling capacitors is close to zero, the impedance of a power delivery network will not produce overshoots over a wide range of operating frequencies. Expression (42.19) shows that by minimizing the decoupling capacitor C_{12} between the two supply voltages, the operating frequency of the overshoot-free impedance of a power delivery network can be increased.

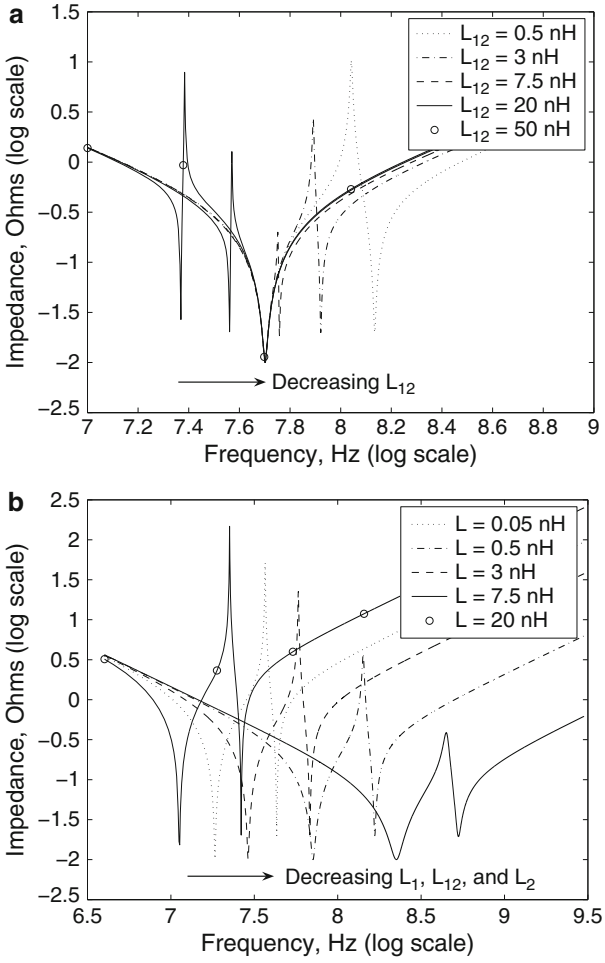


Fig. 42.7 Dependence of a dual V_{dd} power distribution system impedance on frequency for different ESL of the decoupling capacitors. The ESL of capacitors C_1, C_{12}, and C_2 is represented by, respectively, L_1, L_{12}, and L_2. (**a**) $R_1 = R_{12} = R_2 = 10\,\text{m}\Omega$, $C_1 = C_2 = 10\,\text{nF}$, $C_{12} = 1\,\text{nF}$, and $L_1 = L_2 = 1\,\text{nH}$. (**b**) $R_1 = R_{12} = R_2 = 10\,\text{m}\Omega$, $C_1 = C_2 = 10\,\text{nF}$, $C_{12} = 1\,\text{nF}$, and $L_1 = L_{12} = L_2 = L$

The dependence of the power distribution system impedance on the ESL of C_{12} is shown in Fig. 42.7a. Note the strong dependence of the antiresonant frequency on the ESL of the decoupling capacitor located between V_{dd1} and V_{dd2}. As discussed above, the location of the antiresonant spike is determined by the ESL ratio of the decoupling capacitors. The magnitude of the antiresonance spike is determined by the total ESL of C_1, C_{12}, and C_2, as shown in Fig. 42.7b.

By lowering the system inductance, the quality factor is decreased. The peaks become wider in frequency and lower in magnitude. The amplitude of the antiresonant spikes can be decreased by lowering the ESL of all of the decoupling capacitors within the power distribution system. As shown in Fig. 42.7b, decreasing the parasitic inductance of all of the decoupling capacitors of the system reduces the peak magnitude. When the parasitic inductance of C_{12} is similar in magnitude to the other decoupling capacitors, from (42.4), the poles and zeros do not cancel, affecting the behavior of the circuit. The zero at the resonant frequency of a system (the minimum value of the impedance) decreases the antiresonant spike. The closer the location of an antiresonant spike is to the resonant frequency of a system, the greater the influence of a zero on the antiresonance behavior. From a circuits perspective, the more similar the ESL of each capacitor, the smaller the amplitude of the antiresonant spike. Decreasing the inductance of the decoupling capacitors has the same effect as increasing the resistance. Increasing the parasitic resistance of a decoupling capacitor is limited by the target impedance of the power distribution system. Decreasing the inductance of a power distribution system is highly desirable and, if properly designed, the inductance of a power distribution system can be significantly reduced [73].

42.2 Case Study of the Impedance of a Power Distribution System

The dependence of the impedance on the power distribution system parameters is described in this section to quantitatively illustrate the concepts presented in Sect. 42.1. An on-chip power distribution system is assumed in this example. The total budgeted on-chip decoupling capacitance is distributed among the low voltage power supply ($C_1 = 10\,\text{nF}$), high voltage power supply ($C_2 = 10\,\text{nF}$), and the capacitance placed between the two power supplies ($C_{12} = 1\,\text{nF}$). The ESR and ESL of the power distribution network are chosen to be equal to, respectively, $0.1\,\Omega$ and 1 nH. The target impedance is $0.4\,\Omega$.

For typical values of an example power distribution system, an antiresonant spike is produced at approximately 100 MHz with a magnitude greater than the target impedance, as shown in Fig. 42.8. According to (42.19), to shift the antiresonant spike to a higher frequency, the capacitor C_{12} should be decreased. As C_{12} is decreased to 0.3 nF, the antiresonant spike appears at a higher frequency, approximately 158 MHz, and is of higher magnitude. To further decrease the impedance of a power distribution system with multiple power supply voltages, the total ESL of the decoupling capacitors should be decreased. As the total ESL of the system is decreased to 0.1 nH, the impedance of the power distribution system is below the target impedance over a wide frequency range, from approximately 40 MHz to 1 GHz. Three different tradeoff scenarios similar to the case study illustrated in Fig. 42.8 are summarized in Table 42.1. The design parameters for each scenario

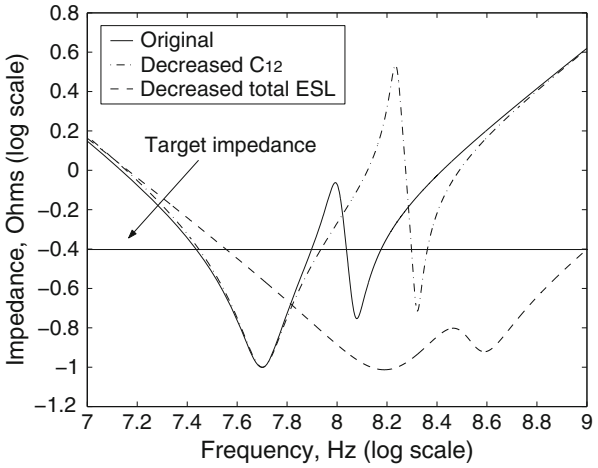


Fig. 42.8 The impedance of a power distribution system with dual power supply voltages as a function of frequency, $R_1 = R_{12} = R_2 = 100\,\text{m}\Omega$, $C_1 = C_2 = 10\,\text{nF}$, $C_{12} = 1\,\text{nF}$, and $L_1 = L_{12} = L_2 = 1\,\text{nH}$. The impedance of the example power distribution network produces an antiresonant spike with a magnitude greater than the target impedance (*the solid line*). The antiresonant spike is shifted to a higher frequency with a larger magnitude by decreasing C_{12} to 0.3 nF (*the dashed-dotted line*). By decreasing the total ESL of the system, the impedance can be maintained below the target impedance over a wide frequency range, from approximately 40 MHz to 1 GHZ (*the dashed line*)

represent typical values of board, package, and on-chip power distribution systems with decoupling capacitors, as shown in Fig. 42.9. The minimum and maximum frequencies denote the frequency range in which the impedance of a power delivery network seen from the load of V_{dd1} does not exceed the target level of 400 mΩ. Note that by decreasing the decoupling capacitor placed between V_{dd1} and V_{dd2}, the range of operating frequencies, where the target impedance is met, is slightly increased. Alternatively, if the total ESL of the system is lowered by an order of magnitude, the frequency range Δf is increased by significantly more than an order of magnitude (for tradeoff scenario III, Δf increases from 560 MHz to 7.01 GHz).

The design of a power distribution system with multiple power supply voltages is a complex task and requires many iterative steps. In general, to maintain the impedance of a power delivery system below a target level, the proper combination of design parameters needs to be determined. In on-chip applications, the ESL and C_{12} can be chosen to satisfy specific values. At the board level, the ESR and C_{12} can be adjusted to satisfy target impedance specifications. At the package level, the ESL, C_{12}, and ESR are the primary design parameters of the system. Usually, the total decoupling capacitance is constrained by the technology and application. In certain cases, it is possible to increase the decoupling capacitance. From (42.13), note that by increasing the decoupling capacitance, the overall impedance of a power distribution system with multiple power supply voltages can be significantly decreased.

Table 42.1 Case study of the impedance of a power distribution system

Tradeoff scenario	Power distribution system	Minimum frequency	Maximum frequency	Frequency range Δf
I	Original	4 kHz	35.48 kHz	31.48 kHz
	Decreased C_{12}	4 kHz	50.1 kHz	46.1 kHz
	Decreased L_1, L_{12}, L_2	4 kHz	1.26 MHz	1.256 MHz
II	Original	100 kHz	1 MHz	900 kHz
	Decreased C_{12}	100 kHz	2.82 MHz	2.72 MHz
	Decreased L_1, L_{12}, L_2	100 kHz	79 MHz	78.9 MHz
III	Original	560 MHz	1 GHz	440 MHz
	Decreased C_{12}	560 MHz	1.12 GHz	560 MHz
	Decreased L_1, L_{12}, L_2	890 MHz	7.9 GHz	7.01 GHz
Scenario I board	Original system: $R_1 = R_{12} = R_2 = 1\,\text{m}\Omega$, $L_1 = L_{12} = L_2 = 50\,\text{nH}$, $C_{12} = 100\,\mu\text{F}$, $C_1 = C_2 = 1\,\text{mF}$			
	Decreased C_{12}: $R_1 = R_{12} = R_2 = 1\,\text{m}\Omega$, $L_1 = L_{12} = L_2 = 50\,\text{nH}$, $C_{12} = 20\,\mu\text{F}$, $C_1 = C_2 = 1\,\text{mF}$			
	Decreased L_1, L_{12}, L_2: $R_1 = R_{12} = R_2 = 1\,\text{m}\Omega$, $L_1 = L_{12} = L_2 = 5\,\text{nH}$, $C_{12} = 100\,\mu\text{F}$, $C_1 = C_2 = 1\,\text{mF}$			
Scenario II package	Original system: $R_1 = R_{12} = R_2 = 1\,\text{m}\Omega$, $L_1 = L_{12} = L_2 = 1\,\text{nH}$, $C_{12} = 3\,\mu\text{F}$, $C_1 = C_2 = 50\,\mu\text{F}$			
	Decreased C_{12}: $R_1 = R_{12} = R_2 = 1\,\text{m}\Omega$, $L_1 = L_{12} = L_2 = 1\,\text{nH}$, $C_{12} = 1\,\mu\text{F}$, $C_1 = C_2 = 50\,\mu\text{F}$			
	Decreased L_1, L_{12}, L_2: $R_1 = R_{12} = R_2 = 1\,\text{m}\Omega$, $L_1 = L_{12} = L_2 = 100\,\text{pH}$, $C_{12} = 3\,\mu\text{F}$, $C_1 = C_2 = 50\,\mu\text{F}$			
Scenario III on-chip	Original system: $R_1 = R_{12} = R_2 = 10\,\text{m}\Omega$, $L_1 = L_{12} = L_2 = 10\,\text{pH}$, $C_{12} = 1\,\text{nF}$, $C_1 = C_2 = 4\,\text{nF}$			
	Decreased C_{12}: $R_1 = R_{12} = R_2 = 10\,\text{m}\Omega$, $L_1 = L_{12} = L_2 = 10\,\text{pH}$, $C_{12} = 0.3\,\text{nF}$, $C_1 = C_2 = 4\,\text{nF}$			
	Decreased L_1, L_{12}, L_2: $R_1 = R_{12} = R_2 = 10\,\text{m}\Omega$, $L_1 = L_{12} = L_2 = 1\,\text{pH}$, $C_{12} = 1\,\text{nF}$, $C_1 = C_2 = 4\,\text{nF}$			

42.3 Voltage Transfer Function of Power Distribution System

Classical methodologies for designing power distribution systems with a single power supply voltage typically only consider the target output impedance of the network. By introducing a second power supply voltage, a decoupling capacitor is effectively placed between the two power supply voltages [28, 276]. The problem of noise propagating from one power supply to the other power supply is aggravated if multiple power supply voltages are employed in a power distribution system. Since multiple power supplies are naturally coupled, the voltage transfer function of a multi-voltage power distribution network should be considered [279, 630]. The voltage transfer function of a power distribution system with dual power supplies is

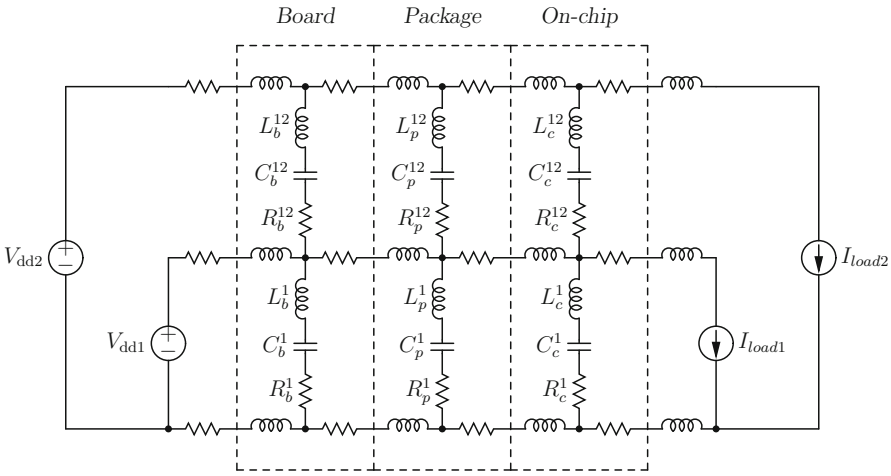


Fig. 42.9 Hierarchical model of a power distribution system with dual supply voltages and a single ground. The decoupling capacitors are represented by the series connected resistance, capacitance, and inductance. For simplicity, the decoupling capacitors placed between V_{dd2} and ground are not illustrated. Subscripts b, p, and c denote, respectively, the board, package, and on-chip power delivery systems. Superscript 1 denotes the decoupling capacitors placed between V_{dd1} and ground and superscript 12 denotes the decoupling capacitors placed between V_{dd1} and V_{dd2}

described in Sect. 42.3.1. The dependence of the magnitude of the voltage transfer function on certain parameters of the power distribution system is described in Sect. 42.3.2.

42.3.1 Voltage Transfer Function of a Power Distribution System

A power distribution system with two power supply voltages and the decoupling capacitors represented by an *RLC* series network is shown in Fig. 42.10. All of the following formulae describing this system are symmetric in terms of the power supply voltages. The ESR and ESL of the three decoupling capacitors are represented by, respectively, R_1, R_{12}, R_2 and L_1, L_{12}, L_2.

The voltage transfer function K_V of a power distribution system with two power supply voltages and decoupling capacitors, represented by an *RLC* network, is

$$K_V = \frac{a_2 s^2 + a_1 s + a_0}{b_2 s^2 + b_1 s + b_0}, \tag{42.20}$$

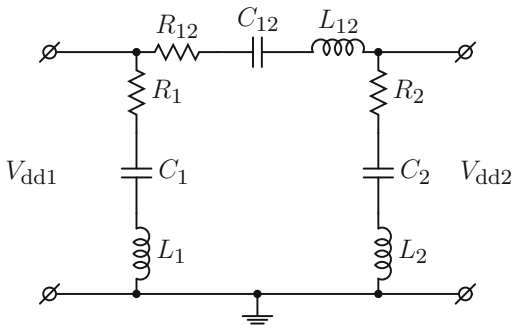


Fig. 42.10 Voltage transfer function of a power distribution network with two supply voltages and the decoupling capacitors represented as series *RLC* networks

where

$$a_2 = L_2C_2, \tag{42.21}$$

$$a_1 = R_2C_2, \tag{42.22}$$

$$a_0 = C_{12}, \tag{42.23}$$

$$b_2 = C_{12}C_2(L_{12} + L_2), \tag{42.24}$$

$$b_1 = C_{12}C_2(R_{12} + R_2), \tag{42.25}$$

$$b_0 = C_{12} + C_2. \tag{42.26}$$

Rearranging, (42.20) can be written as

$$K_V = \frac{1}{\dfrac{a_2s^2 + a_1s + a_0}{b_2s^2 + b_1s + b_0} + 1}, \tag{42.27}$$

where

$$a_2 = L_{12}C_{12}C_2, \tag{42.28}$$

$$a_1 = R_{12}C_{12}C_2, \tag{42.29}$$

$$a_0 = C_2, \tag{42.30}$$

$$b_2 = L_2C_{12}C_2, \tag{42.31}$$

$$b_1 = R_2C_{12}C_2, \tag{42.32}$$

$$b_0 = C_{12}. \tag{42.33}$$

Equations (42.20) and (42.27) are valid only for non-zero frequency, i.e., for $s > 0$. Note from (42.20) that if all of the parameters of a power distribution system are identical, the transfer function equals 0.5 and is independent of frequency. The dependence of the voltage transfer function on the parameters of the power distribution system is discussed below.

42.3.2 Dependence of Voltage Transfer Function on Power Distribution System Parameters

In power distribution systems with two supply voltages, the higher power supply is usually provided for the high speed circuits while the lower power supply is used in the non-critical paths [289, 314]. The two power supplies are often strongly coupled, implying that voltage fluctuations on one power supply propagate to the other power supply. The magnitude of the voltage transfer function should be sufficiently small in order to decouple the noisy power supply from the quiet power supply. The objective is therefore to achieve a transfer function K_V such that the two power supplies are effectively decoupled.

The dependence of the magnitude of the voltage transfer function on frequency for different values of the ESR of the power distribution network with decoupling capacitors is shown in Fig. 42.11. Reducing the ESR of a decoupling capacitor decreases the magnitude and range of the operating frequency of the transfer function. Note that to maintain $|K_V|$ below or equal to 0.5, the following inequality has to be satisfied,

$$R_2 \leq R_{12}. \tag{42.34}$$

This behavior can be explained as follows. From (42.27), to maintain $|K_V|$ below or equal to 0.5,

$$\frac{L_{12}C_{12}C_2s^2 + R_{12}C_{12}C_2s + C_2}{L_2C_{12}C_2s^2 + R_2C_{12}C_2s + C_{12}} + 1 \geq 2. \tag{42.35}$$

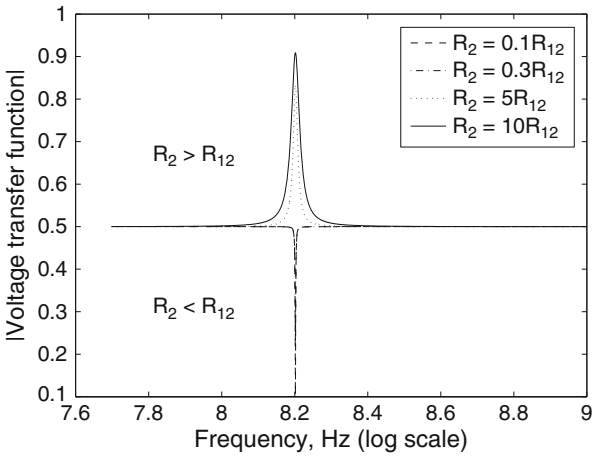


Fig. 42.11 Dependence of the magnitude of the voltage transfer function on frequency of a dual V_{dd} power distribution system for different values of ESR of the decoupling capacitors, $R_{12} = 10\,\text{m}\Omega$, $C_{12} = C_2 = 1\,\text{nF}$, and $L_{12} = L_2 = 1\,\text{nH}$

For equal decoupling capacitors and parasitic inductances, (42.35) leads directly to (42.34). Generally, to maintain $|K_V|$ below or equal to 0.5,

$$L_2C_2C_3s^2 + R_2C_2C_3s + C_3 \geq L_3C_2C_3s^2 + R_3C_2C_3s + C_2. \tag{42.36}$$

From (42.36), in order to maintain the magnitude of the voltage transfer function below or equal to 0.5, the ESR and ESL of the decoupling capacitors should be chosen to satisfy (42.36).

To investigate the dependence of the magnitude of the voltage transfer function on the decoupling capacitors and associated parasitic inductances, the roots of the characteristic equation, the denominator of (42.20), should be analyzed. To produce an overshoot-free response, the roots of the characteristic equation must be real, yielding

$$R_{12} + R_2 \geq 2\sqrt{\frac{(L_{12} + L_2)(C_{12} + C_2)}{C_{12}C_2}}. \tag{42.37}$$

In the case where $R_{12} = R_2 = R$, $L_{12} = L_2 = L$, and $C_{12} = C_2 = C$, (42.37) reduces to the well-known formula [467],

$$R \geq 2\sqrt{\frac{L}{C}}. \tag{42.38}$$

The dependence of the magnitude of the voltage transfer function on the ESL of a power distribution system is shown in Fig. 42.12. For the power distribution system

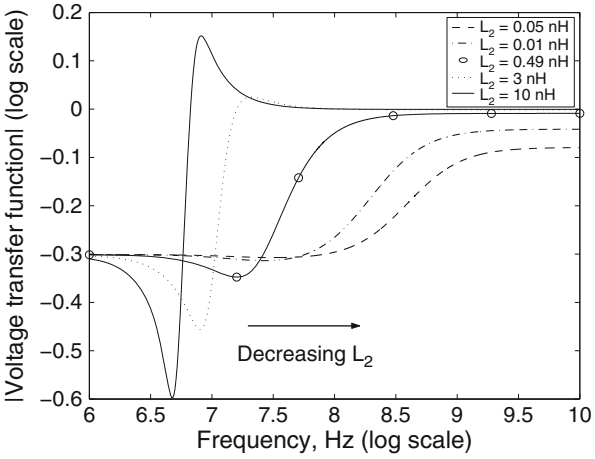


Fig. 42.12 Frequency dependence of the voltage transfer function of a dual V_{dd} power distribution system for different values of ESL of the decoupling capacitors, $R_{12} = R_2 = 100\,\text{m}\Omega$, $C_{12} = C_2 = 100\,\text{nF}$, and $L_{12} = 10\,\text{pH}$

parameters listed in Fig. 42.12, the critical value of L_2 to ensure an overshoot-free response is 0.49 nH. Therefore, in order to produce an overshoot-free response, the ESL of C_2 should be smaller than or equal to 0.49 nH.

Intuitively, if the ESL of a system is large, the system is underdamped and produces an undershoot and an overshoot. By decreasing L_2, the resulting inductance of the system in (42.37) is lowered and the system becomes more damped. As a result, the undershoots and overshoots of the voltage response are significantly smaller. If L_2 is decreased to the critical value, the system becomes overdamped, producing an overshoot-free voltage response.

As shown in Fig. 42.12, the magnitude of the voltage transfer function is strongly dependent on the ESL, decreasing with smaller ESL. It is highly desirable to maintain the ESL as low as possible to achieve a small overshoot-free response characterizing a dual V_{dd} power distribution system over a wide range of operating frequencies. Criterion (42.37) is strict and produces an overshoot-free voltage response. In most applications, if small overshoots (about 1 %) are permitted, (42.37) is less strict, permitting the parameters of a power distribution network to vary over a wider range.

For the parameters listed in Fig. 42.12, the minimum overshoot-free voltage response equals 0.5. It is often necessary to maintain an extremely low magnitude voltage transfer function over a specific frequency range. This behavior can be achieved by varying one of the three design parameters (ESR, ESL or C) characterizing a decoupling capacitor while maintaining the other parameters at predefined values. In this case, for different decoupling capacitors, the magnitude of the voltage transfer function is maintained as low as 0.1 over the frequency range from DC to the self-resonant frequency of the decoupling capacitor induced by the *RLC* series circuit (hereafter called the *break frequency*).

The inductance of the decoupling capacitor has an opposite effect on the magnitude of the voltage transfer function. By increasing the ESL of a dual V_{dd} power distribution system, the magnitude of the voltage transfer function can be maintained below 0.1 from the self-resonant frequency (or break frequency) of the decoupling capacitor to the maximum operating frequency. From (42.27), for frequencies smaller than the break frequency, the magnitude of the voltage transfer function is approximately $\frac{C_{12}}{C_2}$. For frequencies greater than the break frequency, the magnitude of the voltage transfer function is approximately $\frac{L_2}{L_{12}}$. To maintain $|K_V|$ below 0.1, it is difficult to satisfy (42.37), and the range of operating frequency is divided by the break frequency into two ranges. This phenomenon is illustrated in Fig. 42.13a, b.

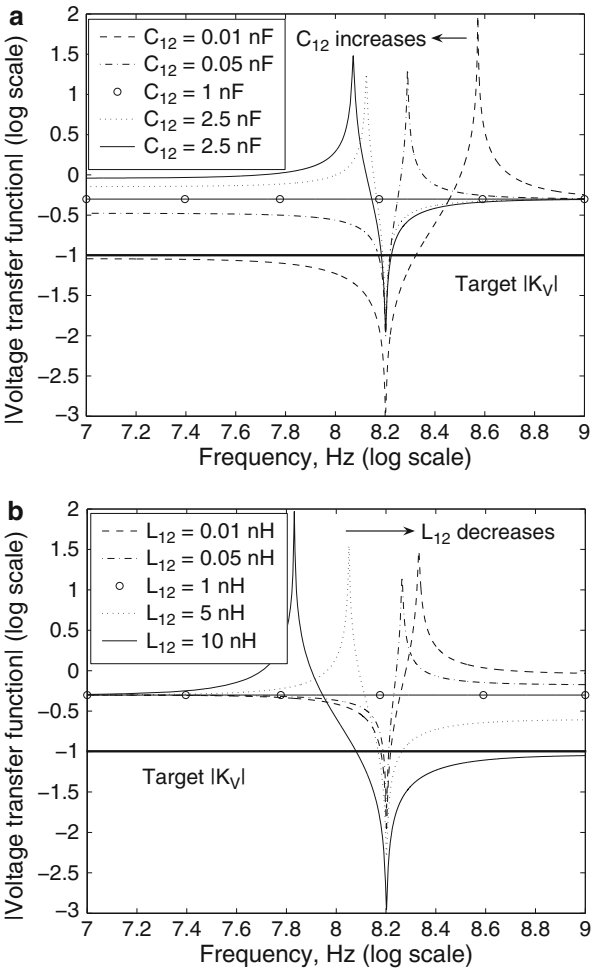


Fig. 42.13 Frequency dependence of the voltage transfer function of a dual V_{dd} power distribution system. The ESR and ESL of the decoupling capacitors for each power supply are represented by, respectively, R_{12}, R_2 and L_{12}, L_2. (**a**) $R_{12} = R_2 = 10\,\text{m}\Omega$, $C_2 = 1\,\text{nF}$, and $L_{12} = L_2 = 1\,\text{nH}$. (**b**) $R_{12} = R_2 = 10\,\text{m}\Omega$, $C_{12} = C_2 = 1\,\text{nF}$, and $L_2 = 1\,\text{nH}$

42.4 Case Study of the Voltage Response of a Power Distribution System

The dependence of the voltage transfer function on the parameters of a power distribution system is described in this section to quantitatively illustrate the concepts presented in Sect. 42.3. An on-chip power distribution system is assumed in this example. In modern high performance ICs, the total on-chip decoupling capacitance can exceed 300 nF, occupying about 20 % of the total area of an IC [147]. In this example, the on-chip decoupling capacitance is assumed to be 160 nF. The total budgeted on-chip decoupling capacitance is arbitrarily distributed among the low voltage power supply ($C_1 = 100\,\text{nF}$), high voltage power supply ($C_2 = 40\,\text{nF}$), and

the capacitance placed between the two power supplies ($C_{12} = 20\,\text{nF}$). The ESR and ESL of the decoupling capacitor are chosen to be, respectively, $0.1\,\Omega$ and 1 nH.

In designing a power distribution system with dual power supply voltages, it is crucial to produce an overshoot-free voltage response over the range of operating frequencies. Depending on the system parameters, it can be necessary to further decouple the power supplies, requiring the magnitude of the voltage transfer function to be decreased. In this case, it is difficult to satisfy (42.37) and the range of operating frequencies is therefore divided into two. There are two possible scenarios: (1) the two power supplies should be decoupled as much as possible from DC to the break frequency, and (2) the two power supplies should be decoupled as much as possible from the break frequency to infinity.

Note that infinite frequency is constrained by the maximum operating frequency of a specific system. Also note that the ESR, ESL, and magnitude of the decoupling capacitors can be considered as design parameters. The ESR is limited by the target impedance of the power distribution network. The ESL, however, can vary significantly. The total budgeted decoupling capacitance is distributed among C_1, C_{12}, and C_2. Note that C_{12} can range from zero (no decoupling capacitance between the two power supplies) to $C_{12} = C_{total} - C_1 - C_2$ (the maximum available decoupling capacitance between the two power supplies), where C_{total} is the total budgeted decoupling capacitance.

42.4.1 Overshoot-Free Magnitude of a Voltage Transfer Function

For typical values of an example power distribution system, (42.37) is not satisfied and the response of the voltage transfer function produces an overshoot as shown in Fig. 42.14. To produce an overshoot-free voltage response, the capacitor placed between the two power supplies should be significantly increased, permitting the ESR and ESL to be varied. Increasing the ESR of the decoupling capacitors to $0.5\,\Omega$ produces an overshoot-free response. By decreasing the ESL of C_2, the overshoot-free voltage response can be further decreased, also shown in Fig. 42.14. As described in Sect. 42.3.2, at low frequency the magnitude of the voltage transfer function is approximately $\frac{C_{12}}{C_2}$. Note that all curves start from the same point. By increasing the ESR, the system becomes overdamped and produces an overshoot-free voltage response. Since the ESR does not change the $\frac{L_2}{L_{12}}$ ratio, the voltage response of the overdamped system is the same as the voltage response of the initial underdamped system. Note that the dashed line and solid line converge to the same point at high frequencies, where the magnitude of the voltage transfer function is approximately $\frac{L_2}{L_{12}}$. By decreasing L_2, the total ESL of the system is lowered and the system becomes overdamped, producing an overshoot-free voltage response. Also, since the $\frac{L_2}{L_{12}}$ ratio is lowered, the magnitude of the voltage response is significantly reduced at high frequencies.

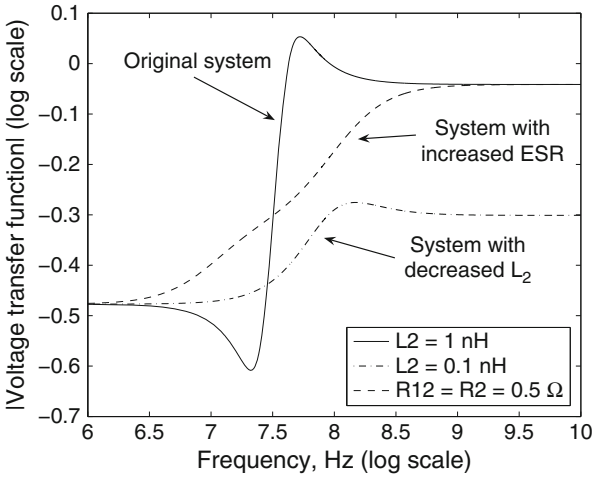


Fig. 42.14 Dependence of the magnitude of the voltage transfer function of a dual V_{dd} power distribution system on frequency for different values of the ESR and ESL of the decoupling capacitors, $R_{12} = R_2 = 0.1\,\Omega$, $C_{12} = 20\,\text{nF}$, $C_2 = 40\,\text{nF}$, and $L_{12} = L_2 = 1\,\text{nH}$. The initial system with $L_2 = 1\,\text{nH}$ produces an overshoot (*solid line*). To produce an overshoot-free voltage response, either the ESR of the system should be increased (*dashed line*) or the ESL should be decreased (*dash-dotted line*)

In general, a design methodology for producing an overshoot-free response of a power distribution system with dual power supply voltages is as follows. Based on the available decoupling capacitance for each power supply, the value of the decoupling capacitor placed between the two power supplies is determined by $C_{12} = C_{total} - C_1 - C_2$. The ESR is chosen to be less than or equal to the target impedance to satisfy the impedance constraint. The critical ESL of the capacitors C_{12} and C_2 is determined from (42.37). If the parasitic inductance of C_{12} and C_2 is less than or equal to the critical ESL, the system will produce an overshoot-free voltage response and no adjustment is required. Otherwise, the total decoupling capacitance budget should be redistributed among C_1, C_{12}, and C_2 until (42.37) is satisfied. In certain cases, the total budgeted decoupling capacitance should be increased to satisfy (42.37).

42.4.2 Tradeoff Between the Magnitude and Frequency Range

If it is necessary to further decouple the power supplies, the frequency range of the overshoot-free voltage response can be traded off with the magnitude of the voltage response, as described in Sect. 42.3.2. There are two ranges of interest. The magnitude of the voltage transfer function can be decreased over the frequency range from DC to the break frequency or from the break frequency to the highest

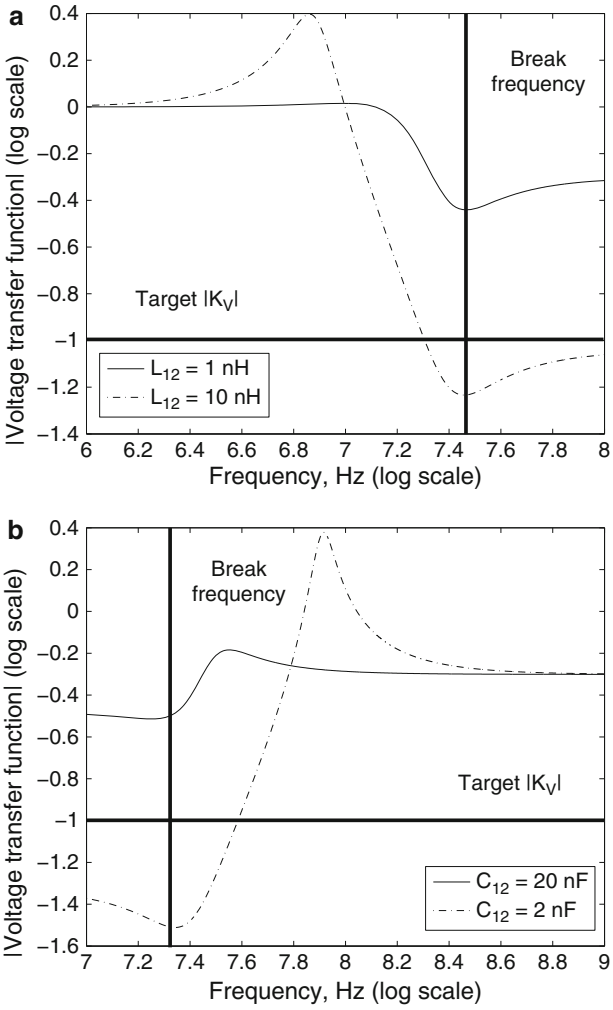


Fig. 42.15 Magnitude of the voltage transfer function of an example dual V_{dd} power distribution system as a function of frequency. The ESR and ESL of the decoupling capacitors are represented by, respectively, R_{12} and R_2 and L_{12} and L_2. (**a**) $R_{12} = R_2 = 0.1\,\Omega$, $C_{12} = 20\,\text{nF}$, $C_2 = 40\,\text{nF}$, and $L_2 = 1\,\text{nH}$. (**b**) $R_{12} = R_2 = 0.1\,\Omega$, $C_2 = 40\,\text{nF}$, and $L_{12} = L_2 = 1\,\text{nH}$

operating frequency [28]. For the example power distribution system, as shown in Fig. 42.15a, the magnitude of the voltage transfer function is overshoot-free from the break frequency to the highest operating frequency. To further decrease the magnitude of the voltage transfer function over a specified frequency range, the ESL of the decoupling capacitor placed between the two power supply voltages should be increased and C_{12} should be the maximum available decoupling capacitance, $C_{12} = C_{total} - C_1 - C_2$.

To decrease the magnitude of the voltage transfer function of a power distribution system with dual power supply voltages for frequencies less than the break frequency, the ESL of all of the decoupling capacitors and the value of C_{12} should be decreased, as shown in Fig. 42.15b. If it is necessary to completely decouple the two power supply voltages, C_{12} should be minimized. This behavior can be explained as follows. The initial system produces an overshoot-free voltage response in the frequency range from DC to the highest operating frequency of the system. In order to satisfy the target $|K_V|$ at high frequencies, L_{12} should be increased in order to decrease the $\frac{L_2}{L_{12}}$ ratio. By increasing L_{12}, the magnitude of the voltage response falls below the target $|K_V|$ in the frequency range from the break frequency to the highest operating frequency of the system. At the same time, the system becomes underdamped and produces an overshoot as shown in Fig. 42.15a. Similarly, by decreasing C_{12}, the $\frac{C_{12}}{C_2}$ ratio is lowered and the magnitude of the voltage response falls below the target $|K_V|$ in the frequency range from DC to the break frequency. The system becomes underdamped and produces an overshoot as shown in Fig. 42.15b.

Three different tradeoff scenarios similar to the case study shown in Fig. 42.14 are summarized in Table 42.2. The design parameters for each scenario represent typical values of board, package, and on-chip decoupling capacitors, as shown in Fig. 42.9. The original system in each scenario produces an overshoot-free voltage response over a wide range of operating frequencies from DC to the highest operating frequency of the system. By increasing the ESL of the decoupling capacitor placed between the two power supplies, the system produces an overshoot and the range of operating frequencies is divided by two. The same phenomenon takes place if the value of the decoupling capacitor placed between the two power supplies is decreased. In the first case, when the ESL is increased by an order of magnitude, the magnitude of the voltage response is lowered by more than an order of magnitude from the break frequency to infinity. When C_{12} is decreased by an order of magnitude, the magnitude of the voltage response is lowered by more than an order of magnitude from DC to the break frequency. Note from the table that the location of the break point depends upon the particular system parameters. The break frequency of the board system occurs at a lower frequency as compared to the break frequency of the package power delivery network. Similarly, the break frequency of the package power distribution system is lower than the break frequency of the on-chip system. As previously mentioned, for typical power supplies values and allowed ripple voltage, $|K_V|$ should be less than 0.1 to decouple a noisy power supply from a quiet power supply. As listed in Table 42.2, this requirement is satisfied for the power distribution system if L_{12} is increased or C_{12} is decreased. The magnitude of the overshoot falls rapidly with decreasing ESL of the decoupling capacitors. Due to the extremely low value of the ESL in an on-chip power network, typically several hundred femtohenrys, the magnitude of the overshoot does not exceed the maximum magnitude of the overshoot-free voltage response.

Unlike the design methodology for producing an overshoot-free response as described in Sect. 42.4.1, a design methodology to trade off the magnitude of the

Table 42.2 Tradeoff between the magnitude and frequency range of the voltage response

Tradeoff scenario	Power distribution system	Minimum $\lvert K_V \rvert$	Maximum $\lvert K_V \rvert$	Minimum frequency	Maximum frequency
I	Original	0.30	0.50	DC	∞
	Increased L_{12}	0.09	0.56	63 kHz	∞
	Decreased C_{12}	0.05	0.60	DC	63 kHz
II	Original	0.20	0.50	DC	∞
	Increased L_{12}	0.09	0.50	3 MHz	∞
	Decreased C_{12}	0.03	0.60	DC	3 MHz
III	Original	0.20	0.50	DC	∞
	Increased L_{12}	0.09	0.50	3 GHz	∞
	Decreased C_{12}	0.05	0.45	DC	3 GHz
Scenario I board	Original circuit: $R_{12} = R_2 = 2\,\text{m}\Omega$, $L_{12} = L_2 = 1\,\text{nH}$, $C_{12} = 2\,\text{mF}$, $C_2 = 4\,\text{mF}$				
	Increased L_{12}: $R_{12} = R_2 = 2\,\text{m}\Omega$, $L_{12} = 10\,\text{nH}$, $L_2 = 1\,\text{nH}$, $C_{12} = 2\,\text{mF}$, $C_2 = 4\,\text{mF}$				
	Decreased C_{12}: $R_{12} = R_2 = 2\,\text{m}\Omega$, $L_{12} = L_2 = 1\,\text{nH}$, $C_{12} = 200\,\mu\text{F}$, $C_2 = 4\,\text{mF}$				
Scenario II package	Original circuit: $R_{12} = R_2 = 10\,\text{m}\Omega$, $L_{12} = L_2 = 100\,\text{pH}$, $C_{12} = 10\,\mu\text{F}$, $C_2 = 40\,\mu\text{F}$				
	Increased L_{12}: $R_{12} = R_2 = 10\,\text{m}\Omega$, $L_{12} = 1\,\text{nH}$, $L_2 = 100\,\text{pH}$, $C_{12} = 10\,\mu\text{F}$, $C_2 = 40\,\mu\text{F}$				
	Decreased C_{12}: $R_{12} = R_2 = 10\,\text{m}\Omega$, $L_{12} = L_2 = 100\,\text{pH}$, $C_{12} = 1\,\mu\text{F}$, $C_2 = 40\,\mu\text{F}$				
Scenario III on-chip	Original circuit: $R_{12} = R_2 = 10\,\text{m}\Omega$, $L_{12} = L_2 = 100\,\text{fH}$, $C_{12} = 20\,\text{nF}$, $C_2 = 40\,\text{nF}$				
	Increased L_{12}: $R_{12} = R_2 = 10\,\text{m}\Omega$, $L_{12} = 1\,\text{pH}$, $L_2 = 100\,\text{fH}$, $C_{12} = 20\,\text{nF}$, $C_2 = 40\,\text{nF}$				
	Decreased C_{12}: $R_{12} = R_2 = 10\,\text{m}\Omega$, $L_{12} = L_2 = 100\,\text{fH}$, $C_{12} = 2\,\text{nF}$, $C_2 = 40\,\text{nF}$				

voltage response of the power distribution system with the frequency range of an overshoot-free response is as follows. Based upon the available decoupling capacitance, the decoupling capacitances for each power supply are determined. Depending upon the target frequency range with respect to the break frequency, the ESL of the capacitor placed between the two power supplies and the decoupling capacitors should both be increased (above the break frequency). Otherwise, the capacitor placed between the two power supplies and the ESL of all of the decoupling capacitors should both be decreased (below the break frequency).

42.5 Summary

A system of decoupling capacitors used in power distribution systems with multiple power supply voltages is described in this chapter. The primary conclusions are summarized as follows.

- Multiple on-chip power supply voltages are often utilized to reduce power dissipation without degrading system speed
- To maintain the impedance of a power distribution system below a specified impedance, multiple decoupling capacitors are placed at different levels of the power grid hierarchy
- The decoupling capacitors should be placed both with progressively decreasing value to shift the antiresonance spike beyond the maximum operating frequency and with increasing ESR to control the damping characteristics
- The magnitude of the antiresonant spikes can also be limited by reducing the ESL of each of the decoupling capacitors
- To maintain the magnitude of the voltage transfer function below 0.5, the ESR and ESL of the decoupling capacitors should be carefully chosen to satisfy the overshoot-free voltage response criterion
- To further decouple the power supplies in frequencies ranging from DC to the break frequency, both the capacitor placed between the two power supply voltages and the ESL of each of the decoupling capacitors should be decreased
- To decouple the power supplies in frequencies ranging from the break frequency to infinity, both the ESL of the capacitor placed between the two power supply voltages and the decoupling capacitors should be increased
- The frequency range of an overshoot-free voltage response can be traded off with the magnitude of the response

Chapter 43
Conclusions

Power consumption can be reduced by applying a multi-voltage scheme while enhancing the overall performance of an integrated circuit. Providing a lower voltage for the non-critical data paths can save power while maintaining the speed. Tradeoffs among area, power, and design complexity are critical in multi-voltage systems. The power savings can be enhanced by simultaneously utilizing devices with different threshold voltages. Design complexity however increases since multiple networks are required to support different power supply voltages.

The placement of the decoupling capacitors is an important factor in multi-voltage systems. Multiple power supplies are naturally coupled through the power grids, affecting the characteristics of a multi-voltage power delivery system. The overall power noise propagating from one power supply to the other power supply increases with the number of on-chip power supplies.

The voltage transfer function and key characteristics of a power distribution system with dual power supplies are described. The placement and magnitude of the decoupling capacitors in multi-voltage systems need to be chosen carefully considering these specific characteristics. A significant decrease in noise is exhibited for multi-voltage networks with decoupling capacitors.

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0_43

Part IX
Final Comments

Chapter 44
Closing Remarks

The continued advance of societal and emerging market segments require functionally diverse semiconductors. In these modern heterogeneous systems, functionally diverse circuits are integrated on-chip, requiring a wide range of high quality DC voltages. In addition, as the need for portable, high performance ICs increases, intelligent management of the energy budget becomes a primary concern. The future of heterogeneous, high performance systems is strongly dependent upon the power delivery system and deeply affected by the quality and efficiency of the on-chip power, stability of distributed parallel connected power supplies, availability of fine grain dynamically controlled voltage levels, and the ability to manage power in real-time. On-chip integration of several power supplies is no longer sufficient to address these power delivery challenges.

To satisfy evolving power delivery requirements, the classical approach for power generation and distribution has changed over the last decade. Power generation circuits are physically closer to the loads to provide enhanced control over the quality of the delivered power. Heterogeneous power delivery systems with different types of off-chip, in-package, and distributed on-chip power converters must be considered. Software and firmware solutions are necessary to address the increased design complexity of these hierarchical nonlinear power delivery systems with thousands of power delivery components and billions of loads. Dynamic voltage and frequency scaling increases energy efficiency over time. Fine grain power management schemes optimize the delivery of power in terms of quality, area, efficiency, and design complexity. An effective power delivery solution should provide a systematic methodology, distributed scalable architectures, algorithms for distributed power management, and specialized circuit structures. Intelligent power will become an integrated part of next generation power delivery systems.

A platform for scalable power delivery and management has been developed and is described in this book. The key concept of this platform is to manage the overall energy budget with fine grain distributed on-chip power networks, providing local feedback paths from the billions of loads to multiple, locally intelligent

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0_44

power routers. This PNoC approach addresses the issues of design complexity and scalability by providing a modular architecture that supports the integration of functional blocks and power features without requiring re-design of the power delivery system. Architectural, algorithmic, and circuit level requirements for intelligent PNoC-based power delivery are necessary, and some possible solutions are described within this book. A power delivery topology and related algorithms to co-design power supplies within the PNoC framework are presented. Circuit and design level solutions that address the challenges of distributed on-chip power delivery and intelligent power management such as on-chip area limitations, dynamic power control, and stability are also demonstrated. Integrating emerging technologies (e.g., mitigation of invasive and non-invasive power attacks) within the PNoC framework is discussed.

Advanced circuit solutions, such as specialized power routers, switches, programmable control logic, and ultra-small voltage regulators, are required to provide efficient high quality dynamically managed power within the PNoC framework. To demonstrate the feasibility of a distributed, dynamically controllable, intelligent power delivery systems, several types of power delivery circuits are described. To provide a circuit level means for dynamically scaling the voltage in adaptive systems, a digitally controlled pulse width modulator is described, including closed-form expressions characterizing the duty cycle and validated under PVT variations. Another key component of distributed power delivery systems is an ultra-small power efficient linear regulator. To demonstrate these design concepts and techniques, a distributed power delivery system with six identical ultra-small fully integrated low dropout regulators, designed and fabricated in a 28 nm CMOS process, is reviewed. This system of distributed parallel LDO regulators exhibits a stable system response over a wide range of temperatures and voltage variations. The system is believed to be the first successful silicon demonstration of stable parallel analog linear regulators.

While intelligent power networks may potentially become the focus of future power delivery and management systems, the power distribution network remains an essential element in efficiently delivering power to high speed integrated circuits. In modern ICs, many hundreds of amperes must be efficiently distributed to supply power to the on-chip circuits. Despite high currents and frequencies, the impedance of a power distribution system should be maintained sufficiently low over a wide range of frequencies to limit voltage variations at the power load—the billions of on-chip transistors. Maintaining a low impedance over a wide range of frequencies is a complex task. Decoupling capacitors effectively reduce the impedance of a power distribution system near the resonant frequency by allowing the high frequency current to bypass the high inductance interconnect structures. The decoupling capacitance and interconnect inductance, however, create resonant modes within a power distribution system, increasing the impedance near the tank resonant frequency. The magnitude of the tank resonance is controlled by maintaining appropriate damping characteristics within the system. The design of a low impedance power distribution system therefore requires a careful balance among the resistive, inductive, and capacitive impedances of the comprising elements.

This balance should be maintained throughout the hierarchical structure of the system—at the board, package, and integrated circuit levels. The low impedance characteristics of the entire power distribution system, from the system-level voltage regulator through the printed circuit board and package onto the integrated circuit to the power terminals of the on-chip circuitry, are maintained using a hierarchy of decoupling capacitors. In a hierarchical decoupling scheme, the power current loop is terminated progressively closer to the load with increasing frequency. The capacitance at each decoupling stage is constrained by the inductive and resistive characteristics of the capacitors and the power distribution network. When designing a power delivery system, these physical characteristics should be modeled and carefully considered at higher levels of abstraction. Power management policies should also consider these physical phenomena.

To maintain the impedance of a power distribution system below a target impedance, multiple decoupling capacitors are placed in parallel at different levels of the power grid hierarchy. Two capacitors with different magnitudes connected in parallel produce antiresonance—an increase in the impedance of the power distribution system over a specific frequency range. If not properly controlled, the antiresonant peak may exceed the target impedance, jeopardizing the signal integrity of the system. The frequency of the antiresonant spike depends upon the effective series inductance of the decoupling capacitors. As the parasitic inductance of the decoupling capacitors is reduced, the antiresonant spike is shifted to a higher frequency. A power distribution system with decoupling capacitors should therefore be carefully designed to control the effective series inductance of the capacitors. Alternatively, multiple decoupling capacitors with progressively decreasing magnitude should be allocated to lower the antiresonance, shifting the antiresonant spikes to a frequency greater than the maximum operating frequency of the system.

Maintaining balanced impedance characteristics at the integrated circuit level is particularly challenging. The power current requirements and impedance characteristics can vary significantly across the die area. Electromigration reliability considerations place additional constraints on the design of the power delivery system. The design of the on-chip interconnect within the power delivery system, placement of the on-chip voltage regulators, allocation of the on-chip decoupling capacitors, and analysis of the chip-package interface characteristics should all be carefully choreographed to achieve the target power noise characteristics.

Controlling the inductive characteristics of the interconnect comprising a power distribution network in a complex on-chip environment is of significant importance in high speed circuits. The inductive behavior of an on-chip power delivery system makes the power supply noise difficult to predict, exacerbating the analysis and verification process. The grid inductance can be effectively reduced with a moderate penalty in either the area or resistance of the grid. Historically, the impedance characteristics of multi-layer grids resulted in the efficient distribution of power in conventional high speed circuits with relatively thick and wide lines in the upper layers and fine lines in the lower layers. The upper layers provide a low impedance current path at low frequencies, while the lower layers serve as a low impedance

path at higher frequencies. At higher frequencies, the impedance and reliability of the power grid can be enhanced by designing a multi-layer system with equal current density within each metal layer. This objective can be achieved by widening the power lines of the lower metal layers, producing a power/ground network similar to a pyramid shaped structure. Characterization of the power network impedance as a multi-layer grid is therefore necessary to efficiently deliver power to high speed integrated circuits. The effects of modular architectures that support the integration of functional blocks, emerging technologies, and system scalability on the power delivery system should be considered.

Despite recent advancements in integrated circuit technologies and packaging solutions, on-chip decoupling capacitors remain an attractive and cost effective solution for supplying current over a wide range of frequencies. A decoupling capacitor acts as a local reservoir of charge, where the charge is released when the power supply voltage across a particular current load drops below some tolerable level. MOS transistors have historically been used as on-chip decoupling capacitors, exploiting the relatively high gate capacitance of these structures. In advanced nanometer technologies, however, the application of on-chip MOS decoupling capacitors has become undesirable due to prohibitively high leakage currents. Occupying up to 40 % of the physical area, on-chip MOS decoupling capacitors can contribute more than half of the total leakage power in modern high speed, high complexity ICs. Different types of on-chip decoupling capacitors, such as MIM and lateral flux capacitors, have emerged as better candidates for on-chip decoupling capacitors.

On-chip decoupling capacitors have traditionally been allocated within the white space available on the die based on an unsystematic ad hoc approach. Conventional approaches for placing on-chip decoupling capacitors result in oversized capacitors often placed at a significant physical distance from the current loads. As a result, the power noise increases, compromising the signal integrity of the entire system. The efficacy of the decoupling capacitors depends upon the impedance of the conductors connecting the capacitors to the current loads and power supplies. To be effective, an on-chip decoupling capacitor should be placed to ensure that both the power supply and the current load are located within the appropriate effective radii of each decoupling capacitor. The size of an on-chip decoupling capacitor, however, is directly proportional to the area occupied by the capacitor and can require significant on-chip area.

While improving the characteristics of the individual voltage regulators, decoupling capacitors, and power grids is important, optimizing system-wide power efficiency is critical in high performance ICs. To maintain high quality on-chip power, the power should be converted and regulated within a hierarchical structure composed of different types of power supplies. To convert power with minimum power losses while avoiding area consuming on-chip passive components, power efficient switching mode power supplies should be placed off-chip or in-package. A system of distributed on-chip decoupling capacitors should therefore be utilized within nanoscale ICs to satisfy technological and performance constraints. The methodologies for placing on-chip decoupling capacitors and co-designing voltage

regulators within a heterogeneous system, as presented in this book, provide a computationally efficient method for allocating on-chip power resources to support expected current and quality of power demands.

The efficient analysis of on-chip power delivery systems is an essential step in the design process. The computational time and memory required to analyze and design these power networks is extremely high due to the large number of interconnects and the significant area occupied by the power distribution network. Impedance models of a power distribution network are therefore required to be both computationally efficient and accurate. The computational complexity can be reduced by utilizing closed-form expressions to model the impedance characteristics of the network. The accuracy and efficiency of these models have historically been a primary tradeoff within the research community. A variety of design and analysis methodologies and tools have been developed based on different models of power delivery systems.

An important challenge in the realization of distributed power delivery systems is maintaining the stability of multi-feedback structures. A distributed system with multiple parallel connected power supplies and dependent feedback paths may exhibit degraded stability due to complex interactions among the voltage regulators, power distribution network, and current loads. To provide a stable, distributed power delivery system, a passivity-based stability criterion is described in this book. Based on this criterion, a distributed power delivery system is stable if and only if the output impedance of the parallel connected power supplies exhibits no right half plane poles and a phase between -90° and $+90^\circ$. This criterion can be used to evaluate the stability of complex distributed power delivery systems and integrated within design automation environments.

The design of power delivery and management systems, particularly in high complexity, high performance integrated circuits, remains a challenging task. The integration of diverse circuit structures within complex systems requires a thorough understanding of the electrical behavior of on-chip power delivery systems. On-chip decoupling capacitors are an efficient solution for reducing power/ground voltage fluctuations in nanoscale ICs. As technologies continue to scale, determining the proper amount of on-chip decoupling capacitance will become increasingly important in reducing leakage currents. The design and analysis of these large scale and complex on-chip power delivery and management systems is expected to remain of high interest to both the academic and industrial communities.

The topics presented in *On-Chip Power Delivery and Management, 4th edition* are intended to provide insight into the electrical behavior and design principles of high performance nanoscale systems. A thorough understanding of the electrical phenomena in complex heterogeneous intelligent multi-layer power delivery systems is therefore essential for applying effective design and analysis methodologies, techniques, and computer-aided tools for developing the next generation of high complexity, nanoscale integrated systems.

Appendices

Appendix A
Estimate of Initial Optimal Width for Interdigitated Power/Ground Network

Since the effective inductance is a transcendental function of width, no closed-form analytic solution can be determined for the wire width that minimizes the effective impedance. The line thickness t is replaced with t_{ind} to simplify the effective inductance model when determining the optimal width. The effective inductance for an interdigitated structure where the distance between the power and ground wires is equal to the thickness of the metal is

$$\left\langle L_{eff}\right\rangle_{s=t_{int}} = \frac{2l(w+s)}{A}\frac{\mu_0 l}{\pi}\left[\frac{3}{2} + ln\left(\frac{2}{\pi}\right)\right]. \tag{A.1}$$

The minimum impedance is determined by solving for the root of the derivative of $\left|Z_{eff}\ (w)\right|_{s=t_{int}}$,

$$\frac{\partial\left|\left[Z_{eff}(w)\right]_{s=t_{int}}\right|}{\partial w} = 0. \tag{A.2}$$

A closed-form solution for the wire width that produces the minimum impedance assuming $s = t_{int}$ is

$$w_{opt} = \left(\frac{1}{\left[\frac{3}{2} + ln\left(\frac{2}{\pi}\right)\right]^2}\right)^{1/3}\sqrt[3]{\frac{s\rho^2}{\mu_o^2 t^2 f^2}}. \tag{A.3}$$

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0

Appendix B
First Optimization Approach for Multi-Layer Interdigitated Power Distribution Network

The input to the EQUAL-CURRENT-DENSITY algorithm, illustrated in Fig. B.1, is the technology parameters for each metal layer in the system, the physical dimensions, and the total current. At line 1, the width of the top metal layer is determined. The process is initiated from the top metal layer since this layer is thickest, permitting a solution for the width of the remaining metal layers. If the current density determined in line 3 is greater than the maximum current density allowed by the technology, additional metal layers should be allocated for the P/G distribution network. The width of the additional metal layers is determined from (37.19) to lower the limiting current density within the P/G network. At higher frequencies, the skin depth is considered when evaluating the current density. n represents the minimum number of metal layers required to effectively distribute power and ground.

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0

EQUAL-CURRENT-DENSITY
1. Optimize the top metal layer width, based on (37.10) and (37.11).
2. Determine R_1, L_1, and Z_1.
3. Determine the current density for a single layer, $n = 1$.
4. while (allowed maximum current density $<$ limiting current density)
5. Increase a number of metal layers, $n = n + 1$.
6. Determine $width_n$, based on (37.19).
7. Determine R_n, L_n, and Z_n.
8. Determine the current density for each layer.
9. end

Fig. B.1 Pseudo-code for the first optimization approach. The widths are chosen to maintain equal current density among each of the layers

Appendix C
Second Optimization Approach for Multi-Layer Interdigitated Power Distribution Network

The MINIMUM-IMPEDANCE pseudo-code, presented in Fig. C.1, is based on minimizing the impedance of each metal layer within a multi-layer P/G system. Note the optimization algorithm begins from the highest metal layer and decreases as required. A specific metal width is determined in line 3. In line 4, the impedance of the current metal layer is determined. The current density is recalculated for each metal layer in line 5. If the maximum current density allowed by the technology is lower than the limiting current density, the algorithm returns to line 2, assigning an additional metal layer for the P/G structure. n represents the minimum number of metal layers required for the P/G network.

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0

MINIMUM-IMPEDANCE
1. $n = 0$.
2. $n = n + 1$.
3. Optimize width of the n-layer based on (37.10) and (37.11).
4. Determine R_n, L_n, and Z_n.
5. Determine current density for every layer.
6. Limiting current density is the highest current density.
7. if (allowed maximum current density $<$ limiting current density)
   goto 2.

Fig. C.1 Pseudo-code for the second optimization approach. The widths are determined to achieve the minimum impedance for each individual metal layer

Appendix D
Mutual Loop Inductance in Fully Interdigitated Power Distribution Grids with DSDG

Assuming $d_I^i = s_I^i = d$, from (41.3), the mutual inductance between the power and ground paths of the different voltage domains for a fully interdigitated power distribution grid with DSDG is

$$L_{V\mathrm{dd1}-V\mathrm{dd2}} = 0.2l\left(\ln\frac{2l}{2d} - 1 + \frac{2d}{l} - \ln\gamma + \ln k\right), \tag{D.1}$$

$$L_{V\mathrm{dd1}-G\mathrm{nd2}} = 0.2l\left(\ln\frac{2l}{3d} - 1 + \frac{3d}{l} - \ln\gamma + \ln k\right), \tag{D.2}$$

$$L_{G\mathrm{nd1}-G\mathrm{nd2}} = 0.2l\left(\ln\frac{2l}{2d} - 1 + \frac{2d}{l} - \ln\gamma + \ln k\right), \tag{D.3}$$

$$L_{V\mathrm{dd2}-G\mathrm{nd1}} = 0.2l\left(\ln\frac{2l}{d} - 1 + \frac{d}{l} - \ln\gamma + \ln k\right). \tag{D.4}$$

Substituting (D.1), (D.2), (D.3), and (D.4) into (41.8), the mutual inductive coupling M_{loop}^{intI} between the two current loops in a fully interdigitated power distribution grid with DSDG is

$$\begin{aligned} M_{loop}^{\mathrm{intI}} = 0.2l\Bigg(&\ln\frac{2l}{2d} - 1 + \frac{2d}{l} - \ln\gamma + \ln k - \ln\frac{2l}{3d} + 1 - \frac{3d}{l} \\ &+\ln\gamma - \ln k + \ln\frac{2l}{2d} - 1 + \frac{2d}{l} - \ln\gamma + \ln k - \ln\frac{2l}{d} + 1 \\ &- \frac{d}{l} + \ln\gamma - \ln k\Bigg). \end{aligned} \tag{D.5}$$

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0

Simplifying (D.5) and considering that $\ln \gamma$ and $\ln k$ are approximately the same for different distances between the lines, M_{loop}^{intI} is

$$\begin{aligned} M_{loop}^{\text{intI}} &= 0.2l \left(\ln \frac{2l}{2d} - \ln \frac{2l}{3d} + \ln \frac{2l}{2d} - \ln \frac{2l}{d} \right) \\ &= 0.2l \ln \frac{2l \times 3d \times 2l \times d}{2d \times 2l \times 2d \times 2l} \\ &= 0.2l \ln \frac{3}{4} < 0. \end{aligned} \tag{D.6}$$

Appendix E
Mutual Loop Inductance in Pseudo-Interdigitated Power Distribution Grids with DSDG

Assuming $d_{II}^{i} = 2d$ and $s_{II}^{i} = d$, from (41.3), the mutual inductance between the power and ground paths of the different voltage domains for a pseudo-interdigitated power distribution grid with DSDG is

$$L_{V\mathrm{dd1}-V\mathrm{dd2}} = 0.2l\left(\ln\frac{2l}{d} - 1 + \frac{d}{l} - \ln\gamma + \ln k\right), \tag{E.1}$$

$$L_{V\mathrm{dd1}-G\mathrm{nd2}} = 0.2l\left(\ln\frac{2l}{3d} - 1 + \frac{3d}{l} - \ln\gamma + \ln k\right), \tag{E.2}$$

$$L_{G\mathrm{nd1}-G\mathrm{nd2}} = 0.2l\left(\ln\frac{2l}{d} - 1 + \frac{d}{l} - \ln\gamma + \ln k\right), \tag{E.3}$$

$$L_{V\mathrm{dd2}-G\mathrm{nd1}} = 0.2l\left(\ln\frac{2l}{d} - 1 + \frac{d}{l} - \ln\gamma + \ln k\right). \tag{E.4}$$

Substituting (E.1), (E.2), (E.3), and (E.4) into (41.8), the mutual inductive coupling $M_{loop}^{\mathrm{intII}}$ between the two current loops in a pseudo-interdigitated power distribution grid with DSDG is

$$\begin{aligned} M_{loop}^{\mathrm{intII}} = 0.2l\Bigg(&\ln\frac{2l}{d} - 1 + \frac{d}{l} - \ln\gamma + \ln k - \ln\frac{2l}{3d} + 1 - \frac{3d}{l} \\ &+\ln\gamma - \ln k + \ln\frac{2l}{d} - 1 + \frac{d}{l} - \ln\gamma + \ln k - \ln\frac{2l}{d} + 1 \\ &- \frac{d}{l} + \ln\gamma - \ln k\Bigg). \end{aligned} \tag{E.5}$$

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0

Simplifying (E.5) and considering that $\ln\gamma$ and $\ln k$ are approximately the same for different distances between the lines, M_{loop}^{intII} is

$$\begin{aligned} M_{loop}^{\text{intII}} &= 0.2l\left(\ln\frac{2l}{d} - \ln\frac{2l}{3d} + \ln\frac{2l}{d} - \ln\frac{2l}{d} - \frac{2d}{l}\right) \\ &= 0.2l\left(\ln\frac{2l\times 3d\times 2l\times d}{d\times 2l\times d\times 2l} - \frac{2d}{l}\right) \\ &= 0.2l\left(\ln 3 - \frac{2d}{l}\right) > 0. \end{aligned} \tag{E.6}$$

Appendix F
Mutual Loop Inductance in Fully Paired Power Distribution Grids with DSDG

Assuming the separation between the pairs is n times larger than the distance between the power and ground lines inside each pair d (see Fig. 41.8), from (41.3), the mutual inductance between the power and ground paths of the different voltage domains for a fully paired power distribution grid with DSDG is

$$L_{V\mathrm{dd1}-V\mathrm{dd2}} = 0.2l\left[\ln\frac{2l}{(n+1)d} - 1 + \frac{(n+1)d}{l} - \ln\gamma + \ln k\right], \tag{F.1}$$

$$L_{V\mathrm{dd1}-G\mathrm{nd2}} = 0.2l\left[\ln\frac{2l}{(n+2)d} - 1 + \frac{(n+2)d}{l} - \ln\gamma + \ln k\right], \tag{F.2}$$

$$L_{G\mathrm{nd1}-G\mathrm{nd2}} = 0.2l\left[\ln\frac{2l}{(n+1)d} - 1 + \frac{(n+1)d}{l} - \ln\gamma + \ln k\right], \tag{F.3}$$

$$L_{V\mathrm{dd2}-G\mathrm{nd1}} = 0.2l\left(\ln\frac{2l}{nd} - 1 + \frac{nd}{l} - \ln\gamma + \ln k\right). \tag{F.4}$$

Substituting (F.1), (F.2), (F.3), and (F.4) into (41.8), the mutual inductive coupling M_{loop}^{prdI} between the two current loops in a fully paired power distribution grid with DSDG is

$$\begin{aligned} M_{loop}^{\mathrm{prdI}} = 0.2l\Bigg[&\ln\frac{2l}{(n+1)d} - 1 + \frac{(n+1)d}{l} - \ln\gamma + \\ &\ln k - \ln\frac{2l}{(n+2)d} + 1 - \frac{(n+2)d}{l} + \ln\gamma - \\ &\ln k + \ln\frac{2l}{(n+1)d} - 1 + \frac{(n+1)d}{l} - \ln\gamma + \\ &\ln k - \ln\frac{2l}{nd} + 1 - \frac{nd}{l} + \ln\gamma - \ln k\Bigg]. \end{aligned} \tag{F.5}$$

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0

Simplifying (F.5) and considering that $\ln\gamma$ and $\ln k$ are approximately the same for different distances between the lines, M_{loop}^{prdI} is

$$\begin{aligned} M_{loop}^{\text{prdI}} &= 0.2l\left[\ln\frac{2l}{(n+1)d}+\frac{(n+1)d}{l}-\ln\frac{2l}{(n+2)d}\right. \\ &\quad \left.-\frac{(n+2)d}{l}+\ln\frac{2l}{(n+1)d}+\frac{(n+1)d}{l}-\ln\frac{2l}{nd}-\frac{nd}{l}\right] \\ &= 0.2l\left[\ln\frac{2l\times(n+2)d\times 2l\times nd}{(n+1)d\times 2l\times(n+1)d\times 2l}\right. \\ &\quad \left.+\frac{(n+1)d-(n+2)d+(n+1)d-nd}{l}\right] \\ &= 0.2l\ln\left[\frac{(n+2)n}{(n+1)^2}\right]<0 \ \text{ for } n\geq 1. \end{aligned} \tag{F.6}$$

Appendix G
Mutual Loop Inductance in Pseudo-Paired Power Distribution Grids with DSDG

Observing that the effective distance between the power and ground lines in a specific power delivery network is $n+1$ times greater than the separation d between the lines making up the pair (see Fig. 41.9), from (41.3), the mutual inductance between the power and ground paths of the different voltage domains for a pseudo-paired power distribution grid with DSDG is

$$L_{V\mathrm{dd1}-V\mathrm{dd2}} = 0.2l\left(\ln\frac{2l}{d} - 1 + \frac{d}{l} - \ln\gamma + \ln k\right), \tag{G.1}$$

$$L_{V\mathrm{dd1}-G\mathrm{nd2}} = 0.2l\left[\ln\frac{2l}{(n+2)d} - 1 + \frac{(n+2)d}{l} - \ln\gamma + \ln k\right], \tag{G.2}$$

$$L_{G\mathrm{nd1}-G\mathrm{nd2}} = 0.2l\left(\ln\frac{2l}{d} - 1 + \frac{d}{l} - \ln\gamma + \ln k\right), \tag{G.3}$$

$$L_{V\mathrm{dd2}-G\mathrm{nd1}} = 0.2l\left(\ln\frac{2l}{nd} - 1 + \frac{nd}{l} - \ln\gamma + \ln k\right). \tag{G.4}$$

Substituting (G.1), (G.2), (G.3), and (G.4) into (41.8), the mutual inductive coupling $M_{loop}^{\mathrm{prdII}}$ between the two current loops in a pseudo-paired power distribution grid with DSDG is

$$\begin{aligned} M_{loop}^{\mathrm{prdII}} = 0.2l\Bigg[& \ln\frac{2l}{d} - 1 + \frac{d}{l} - \ln\gamma + \ln k - \ln\frac{2l}{(n+2)d} + 1 \\ & -\frac{(n+2)d}{l} + \ln\gamma - \ln k + \ln\frac{2l}{d} - 1 + \frac{d}{l} - \ln\gamma + \ln k \\ & -\ln\frac{2l}{nd} + 1 - \frac{nd}{l} + \ln\gamma - \ln k\Bigg]. \end{aligned} \tag{G.5}$$

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0

Simplifying (G.5) and considering that $\ln \gamma$ and $\ln k$ are approximately the same for different distances between the lines, M_{loop}^{prdII} is

$$\begin{aligned} M_{loop}^{\text{prdII}} &= 0.2l \left[\ln\frac{2l}{d} + \frac{d}{l} - \ln\frac{2l}{(n+2)d} - \frac{(n+2)d}{l} + \ln\frac{2l}{d} + \frac{d}{l} \right. \\ &\qquad \left. - \ln\frac{2l}{nd} - \frac{nd}{l} \right] \\ &= 0.2l \left[\ln\frac{2l \times (n+2)d \times 2l \times nd}{d \times 2l \times d \times 2l} + \frac{2d - (n+2)d - nd}{l} \right] \\ &= 0.2l \left[\ln\left(n^2 + 2n\right) - \frac{2nd}{l} \right] > 0 \quad \text{for } n \geq 1. \end{aligned} \tag{G.6}$$

Appendix H
Derivation of $R_{2\,(x,y)}$

The integral characterizing the effective impedance in a semi-uniform mesh structure consists of two separate integrals, $R_{x,y}/r = R_{1(x,y)} + R_{2(x,y)}$. A derivation of the second part of the integral is provided in this appendix where $R_2(x, y)$ is simplified to obtain a numerical solution similar to $R_1(x, y)$. Multiple numerical solutions exist for different values of k. To obtain a general solution of $R_2(x, y)$ for all possible values of k, k is expanded when approaching a positive real number n. In this appendix, the second part of the integral in (22.30) is simplified by applying well known trigonometric identities and a Taylor series expansion when $k \to n+\epsilon$ where $\epsilon \ll 1$ (i.e., when k approaches n). From (22.30), $R_{2(x,y)}$ is

$$R_{2(x,y)} = \frac{k}{\pi}\int_0^{\pi}\left(\frac{1}{\sqrt{(k+1-k\cos\beta)^2-1}} - \frac{1}{\beta\sqrt{k}}\right)d\beta. \tag{H.1}$$

Substituting $(1+\epsilon)^m \approx 1+m\epsilon$ multiple times into (H.1), $R_{2(x,y)}$ simplifies to (H.2), (H.3), (H.4), and (H.5).

$$R_{2(x,y)} = \frac{k}{\pi}\int_0^{\pi}\Big(\{(n+\epsilon+1-(n+\epsilon)\cos\beta)^2-1\}^{-1/2} - \frac{1}{\beta}(n+\epsilon)^{-1/2}\Big)d\beta. \tag{H.2}$$

$$R_{2(x,y)} = \frac{k}{\pi}\int_0^{\pi}\Bigg(\Bigg\{(n+1-n\cos\beta)^2\Big(1+\epsilon\frac{1-\cos\beta}{n+1-n\cos\beta}\Big)^2-1\Bigg\}^{-1/2} - \frac{1}{\beta\sqrt{n}}\Big(1-\epsilon\frac{1}{2n}\Big)\Bigg)d\beta. \tag{H.3}$$

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0

$$R_{2(x,y)} = \frac{k}{\pi}\int_0^{\pi}\Big((n+1-n\cos\beta)^2-1 + 2\epsilon(1-\cos\beta)(n+1-n\cos\beta)\Big)^{-1/2}d\beta - \frac{k}{\pi}\int_0^{\pi}\left(\frac{1}{\beta\sqrt{n}}-\epsilon\frac{1}{2n\sqrt{n}\beta}\right)d\beta. \tag{H.4}$$

$$R_{2(x,y)} = \frac{k}{\pi}\int_0^{\pi}((n+1-n\cos\beta)^2-1)^{-1/2}\left(1-\epsilon\frac{(1-\cos\beta)(n+1-n\cos\beta)}{(n+1-n\cos\beta)^2-1}\right)d\beta - \frac{k}{\pi}\int_0^{\pi}\left(\frac{1}{\beta\sqrt{n}}-\epsilon\frac{1}{2n\sqrt{n}\beta}\right)d\beta. \tag{H.5}$$

$R_{2(x,y)}$ is grouped into two parts, as follows,

$$R_{2(x,y)} = \frac{k}{\pi}\int_0^{\pi}\left(\left((n+1-n\cos\beta)^2-1\right)^{-1/2}-\frac{1}{\beta\sqrt{n}}\right)d\beta + \frac{k}{\pi}\int_0^{\pi}\left(-\epsilon\frac{(1-\cos\beta)(n+1-n\cos\beta)}{((n+1-n\cos\beta)^2-1)^{3/2}}+\frac{\epsilon}{2\beta n\sqrt{n}}\right)d\beta. \tag{H.6}$$

$R_{2(x,y)}$ can be numerically determined by assigning n to a constant value. For instance, when $k \to 1$ (*i.e.*, $n = 1$), the first and second parts of (H.6) are numerically determined by, respectively, assigning $n = 1$ and substituting $\epsilon = k-1$. $R_{2(x,y)}$ becomes

$$R_{2(x,y)} = -\frac{k(k-1)}{\pi}\int_0^{\pi}\left(\frac{(1-\cos\beta)(2-\cos\beta)}{((2-\cos\beta)^2-1)^{3/2}}-\frac{1}{2\beta}\right)d\beta - 0.033425k. \tag{H.7}$$

The second integral is numerically solved and the closed-form expression for $R_{2(x,y)}$ when $k \to 1$ is

$$R_{2(x,y)} = -0.033425k - 0.0629k(k-1). \tag{H.8}$$

When k approaches another constant, (H.6) is similarly determined. Closed-form approximations for $R_{1(x,y)}$ and $R_{2(x,y)}$ are listed in Table 22.1 for different values of n, where the effective resistance $R_{x,y} = R_{1(x,y)} + R_{2(x,y)}$.

Appendix I
Closed-Form Expressions for Interconnect Resistance, Capacitance, and Inductance

Closed-form expressions for the resistance, capacitance, and inductance of a line are summarized in this appendix to provide additional background on the effect of technology and certain design parameters on the interconnect impedance. The interconnect line resistance is

$$R = \frac{\rho L}{WT}, \tag{I.1}$$

where ρ, L, W, and T are, respectively, the resistivity, length, width, and thickness of the interconnect. The line-to-substrate capacitance and coupling capacitance are, respectively, [206]

$$\frac{C_s}{\varepsilon_{ox}} = \frac{W}{h} + 2.2217\left(\frac{s}{s+0.7h}\right)^{3.193} + \\ + 1.171\left(\frac{s}{s+1.51h}\right)^{0.7642}\cdot\left(\frac{T}{T+4.532h}\right)^{0.1204}, \tag{I.2}$$

and

$$\frac{C_c}{\varepsilon_{ox}} = 1.144\frac{T}{s}\left(\frac{h}{h+2.059s}\right)^{0.0944} + 0.7428\left(\frac{W}{W+1.592s}\right)^{1.144} \\ + 1.158\left(\frac{W}{W+1.874s}\right)^{0.1612}\cdot\left(\frac{h}{h+0.9801s}\right)^{1.179}, \tag{I.3}$$

where ε_{ox}, h, and s are, respectively, the oxide permittivity, distance from the interconnect to the substrate, and spacing between adjacent interconnects. Closed-form expressions for the self- and mutual inductance of a line are, respectively, [45, 631]

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0

$$L_s = \frac{\mu_0.L}{2\pi}\left[ln(\frac{2L}{W+T}) + \frac{1}{2} + \frac{0.22(W+T)}{L}\right], \tag{I.4}$$

and

$$L_m = \frac{\mu_0.L}{2\pi}\left[ln(\frac{2L}{d}) - 1 + \frac{d}{L}\right], \tag{I.5}$$

where μ_0 and d are, respectively, the magnetic permeability of free space and the center-to-center distance between two adjacent interconnects.

References

1. C. Pirtle, *Engineering the World: Stories from the First 75 Years of Texas Instruments* (Southern Methodist University Press, Dallas, 2005)
2. T.R. Reid, *The Chip: How Two Americans Invented the Microchip and Launched a Revolution* (Random House, New York, 2001)
3. L. Berlin, *Man Behind the Microchip: Robert Noyce and the Invention of Silicon Valley* (Oxford University Press, New York, 2005)
4. J.S. Kilby, Miniaturized Electronic Circuits, U.S. Patent 3,138,743, 23 June 1964
5. J.A. Hoerni, Planar silicon transistors and diodes. IRE Trans. Electron Devices **8**(2), 178, (1961)
6. D. Kahng, A historical perspective on the development of MOS transistors and related devices. IEEE Trans. Electron Devices **23**(7), 655–657 (1976)
7. R.E. Kerwin, D.L. Klein, J.C. Sarace, Method for Making MIS Structures, U.S. Patent 3,475,234, 28 Oct 1969
8. G.E. Moore, Cramming more components onto integrated circuits. Electronics **32**(8), 114–117 (1965). ISSN:2079-9292; CODEN:ELECGJ published quarterly online by MDPI
9. G.E. Moore, Progress in digital integrated electronics, in *Proceedings of the IEEE International Electron Devices Meeting*, pp. 11–13, Dec 1975
10. *International Technology Roadmap for Semiconductors*, 2007 edn. (Semiconductor Industry Association, 2012). http://public.itrs.net
11. B.T. Murphy, D.E. Haggan, W.W. Troutman, From circuit miniaturization to the scalable IC. Proc. IEEE **88**(5), 691–703 (2000)
12. J. Millman, *Microelectronics* (McGraw-Hill, New York, 1979)
13. F. Faggin, M.E. Hoff, Standard parts and custom design merge in four-chip processor kit. Electronics 112–116 (1972). ISSN:2079-9292; CODEN:ELECGJ published quarterly online by MDPI
14. J.D. Warnock et al., 22nm next-generation IBM system z microprocessor, in *Proceedings of the IEEE International Solid-State Circuits Conference*, Electronics (ISSN 2079-9292; CODEN: ELECGJ) published quarterly online by MDPI, pp. 1–3, Feb 2015
15. S. Rusu et al., A 65-nm dual-core multithreaded Xeon processor with 16-MB L3 cache. IEEE J. Solid State Circuits **42**(1), 17–25 (2007)
16. J. Chang et al., The 65-nm 16-MB shared on-die L3 cache for the dual-core Intel Xeon processor 7100 Series. IEEE J. Solid-State Circuits **42**(4), 846–852 (2007)
17. J.M. Hart et al., Implementation of a forth-generation 1.8-GHz dual-core SPARC V9 microprocessor. IEEE J. Solid-State Circuits **41**(1), 210–217 (2006)

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0

18. R. Kalla, B. Sinharoy, J.M. Tendler, IBM Power5 chip: a dual-core multithreaded processor. IEEE Micro **24**(2), 40–47 (2004)
19. Intel, *Intel Product Specifications* (2015). Available online: http://ark.intel.com
20. *The International Technology Roadmap for Semiconductors* (Semiconductor Industry Association, 2012). Available online: www.itrs.net
21. K.T. Tang, E.G. Friedman, Estimation of transient voltage fluctuations in the CMOS-based power distribution networks, in *Proceedings of the IEEE International Symposium on Circuit and Systems*, vol. 5, pp. 463–466, May 2001
22. K.T. Tang, E.G. Friedman, On-chip ΔI noise in the power distribution networks of high speed CMOS integrated circuits, in *Proceedings of the IEEE International ASIC/SOC Conference*, pp. 53–57, Sept 2000
23. K.T. Tang, E.G. Friedman, Simultaneous switching noise in on-chip CMOS power distribution networks. IEEE Trans. Very Large Scale Integr. (VLSI) Circuits **10**(4), 487–493 (2002)
24. *International Technology Roadmap for Semiconductors*, 2007 edn. (Semiconductor Industry Association, 2007). http://public.itrs.net
25. M. Benoit, S. Taylor, D. Overhauser, S. Rochel, Power distribution in high-performance design, in *Proceedings of the IEEE International Symposium on Low Power Electronics and Design*, pp. 274–278, Aug 1998
26. L.C. Tsai, A 1 GHz PA-RISC processor, in *Proceedings of the IEEE International Solid-State Circuits Conference*, pp. 322–323, Feb 2001
27. C.J. Anderson et al., Physical design of a fourth-generation POWER GHz microprocessor, in *Proceedings of the IEEE International Solid-State Circuits Conference*, pp. 232–233, Feb 2001
28. M. Popovich, E.G. Friedman, Decoupling capacitors for multi-voltage power distribution systems. IEEE Trans. Very Large Scale Integr. (VLSI) Syst. **14**(3), 217–228 (2006)
29. M. Popovich, E.G. Friedman, Impedance Characteristics of Decoupling Capacitors in Multi-Power Distribution Systems, in *Proceedings of the IEEE International Conference on Electronics, Circuits and Systems*, pp. 160–163, Dec 2004
30. A.V. Mezhiba, E.G. Friedman, *Power Distribution Networks in High Speed Integrated Circuits* (Kluwer Academic, Norwell, 2004)
31. K.T. Tang, E.G. Friedman, Delay uncertainty due to on-chip simultaneous switching noise in high performance CMOS integrated circuits, in *Proceedings of the IEEE Workshop on Signal Processing Systems*, pp. 633–642, Oct 2000
32. K.T. Tang, E.G. Friedman, Incorporating voltage fluctuations of the power distribution network into the transient analysis of CMOS logic gates. Analog Integr. Circuits Signal Process. **31**(3), 249–259 (2002)
33. M. Saint-Laurent, M. Swaminathan, Impact of power supply noise on timing in high-frequency microprocessors, in *IEEE Topical Meeting on Electrical Performance of Electronic Packaging*, pp. 261–264, Oct 2002
34. M. Saint-Laurent, M. Swaminathan, Impact of power-supply noise on timing in high-frequency microprocessors. IEEE Trans. Adv. Packag. **27**(1), 135–144 (2004)
35. A. Waizman, C.-Y. Chung, Package capacitor impact on microprocessor maximum operating frequency, in *Proceedings of the IEEE International Electronic Components and Technology Conference*, pp. 118–122, June 2001
36. A. Elkholy, T. Anand, W.-S. Choi, A. Elshazly, P.K. Hanumolu, A 3.7 mW low-noise wide-bandwidth 4.5 GHz digital fractional-N PLL using time amplifier-based TDC. IEEE J. Solid-State Circuits **50**(4), 867–881 (2015)
37. E.G. Friedman (ed.), *Clock Distribution Networks in VLSI Circuits and Systems* (IEEE Press, Piscataway, 1995)
38. E.G. Friedman (ed.), *High Performance Clock Distribution Networks* (Kluwer Academic, Norwell, 1997)
39. I.S. Kourtev, E.G. Friedman, *Timing Optimization Through Clock Skew Scheduling* (Kluwer Academic, Norwell, 2000)

40. J.P. Eckhardt, K.A. Jenkins, PLL phase error and power supply noise, in *IEEE Topical Meeting on Electrical Performance of Electronic Packaging*, pp. 73–76, Oct 1998
41. A.W. Strong, E.Y. Wu, R.-P. Vollertsen, J. Sune, G.L. Rosa, T.D. Sullivan, S.E. Rauch, *Reliability Wearout Mechanisms in Advanced CMOS Technologies* (Wiley, Hoboken, 2006)
42. L. Smith, Reliability and performance tradeoffs in the design of on-chip power delivery and interconnects, in *IEEE Topical Meeting on Electrical Performance of Electronic Packaging*, pp. 49–52, Nov 1999
43. J.D. Jackson, *Classical Electrodynamics* (Wiley, New York, 1975)
44. F.W. Grover, *Inductance Calculations: Working Formulas and Tables* (D. Van Nostrand Company, New York, 1946)
45. A.E. Ruehli, Inductance calculations in a complex integrated circuit environment. IBM J. Res. Dev. **16**(5), 470–481 (1972)
46. E.B. Rosa, The self and mutual inductance of linear conductors. Bull. Natl. Bur. Stand. **4**(2), 301–344 (1908). Government Printing Office, Washington, DC
47. E.B. Rosa, L. Cohen, Formulæ and tables for the calculation of mutual and self-inductance. Bull. Natl. Bur. Stand. **5**(1), 1–132 (1908). Government Printing Office, Washington, DC
48. E.B. Rosa, F.W. Grover, Formulæ and tables for the calculation of mutual and self-inductance. Bull. Natl. Bur. Stand. **8**(1), 1–237 (1912). Government Printing Office, Washington, DC
49. R.F. German, H.W. Ott, C.R. Paul, Effect of an image plane on printed circuit board radiation, in *Proceedings of the IEEE International Symposium on Electromagnetic Compatibility*, pp. 284–291, Aug 1990
50. T.S. Smith, C.R. Paul, Effect of grid spacing on the inductance of ground grids, in *Proceedings of the IEEE International Symposium on Electromagnetic Compatibility*, pp. 72–77, Aug 1991
51. C.R. Paul, *Introduction to Electromagnetic Compatibility* (Wiley, New York, 1992)
52. R.E. Matick, *Transmission Lines for Digital and Communication Networks* (McGraw-Hill, New York, 1969)
53. D.W. Bailey, B.J. Benschneider, Clocking design and analysis for a 600-MHz alpha microprocessor. IEEE J. Solid-State Circuits **33**(11), 1627–1633 (1998)
54. R.M. Averill et al., Chip integration methodology for the IBM S/390 G5 and G6 custom microprocessors. IBM J. Res. Dev. **43**(5/6), 681–706 (1999)
55. H.A. Wheeler, Formulas for the skin effect, in *Proceedings of the IRE*, pp. 412–424, Sept 1942
56. C.-S. Yen, Z. Fazarinc, R.L. Wheeler, Time-domain skin-effect model for transient analysis of lossy transmission lines. Proc. IEEE **70**(7), 750–757 (1982)
57. T.V. Dinh, B. Cabon, J. Chilo, New skin-effect equivalent circuit. Electron. Lett. **26**(19), 1582–1584 (1990)
58. S. Kim, D.P. Neikirk, Compact equivalent circuit model for the skin effect, in *Proceedings of the IEEE International Microwave Symposium*, pp. 1815–1818, June 1996
59. B. Krauter, S. Mehrotra, Layout based frequency dependent inductance and resistance extraction for on-chip interconnect timing analysis, in *Proceedings of the IEEE/ACM Design Automation Conference*, pp. 303–308, June 1998
60. G.V. Kopcsay, B. Krauter, D. Widiger, A. Deutsch, B.J. Rubin, H.H. Smith, A comprehensive 2-D inductance modeling approach for VLSI interconnects: frequency-dependent extraction and compact model synthesis. IEEE Trans. Very Large Scale Integr. (VLSI) Circuits **10**(6), 695–711 (2002)
61. S. Mei, C. Amin, Y.I. Ismail, Efficient model order reduction including skin effect, in *Proceedings of the IEEE/ACM Design Automation Conference*, pp. 232–237, June 2003
62. A. Deutsch et al., When are transmission-line effects important for on-chip interconnections? IEEE Trans. Microw. Theory Techn. **45**(10), 1836–1846 (1997)
63. Y.I. Ismail, E.G. Friedman, J.L. Neves, Figures of merit to characterize the importance of on-chip inductance. IEEE Trans. Very Large Scale Integr. (VLSI) Circuits **7**(4), 442–449 (1999)
64. Y.I. Ismail, E.G. Friedman, *On-Chip Inductance in High Speed Integrated Circuits* (Kluwer Academic, Norwell, 2001)
65. R. Schmitt, *Electromagnetics Explained* (Newnes—Elsevier Science, Boston, 2002)

66. B. Krauter, L.T. Pileggi, Generating sparse partial inductance matrices with guaranteed stability, in *Proceedings of the IEEE/ACM Design Automation Conference*, pp. 45–52, Nov 1995
67. Z. He, M. Celik, L.T. Pileggi, SPIE: sparse partial inductance extraction, in *Proceedings of the IEEE/ACM Design Automation Conference*, pp. 137–140, June 1997
68. A.J. Dammers, N.P. van der Meijs, Virtual screening: a step towards a sparse partial inductance matrix, in *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pp. 445–452, June 1999
69. M.W. Beattie, L. Pileggi, Efficient inductance extraction via windowing, in *Proceedings of the IEEE Design, Automation, and Test Conference in Europe*, pp. 430–436, Mar 2001
70. M. Kamon, M.J. Tsuk, J. White, FastHenry: a multipole-accelerated 3-D inductance extraction program. IEEE Trans. Microw. Theory Techn. **42**(9), 1750–1758 (1994)
71. A.V. Mezhiba, E.G. Friedman, Properties of on-chip inductive current loops, in *Proceedings of the ACM Great Lakes Symposium on Very Large Scale Integration*, pp. 12–17, Apr 2002
72. A.V. Mezhiba E.G. Friedman, Inductive characteristics of power distribution grids in high speed integrated circuits, in *Proceedings of the IEEE International Symposium on Quality Electronic Design*, pp. 316–321, Mar 2002
73. A.V. Mezhiba, E.G. Friedman, Inductive properties of high-performance power distribution grids. IEEE Trans. Very Large Scale Integr. (VLSI) Circuits **10**(6), 762–776 (2002)
74. J.M. Maxwell, *A Treatise on Electricity and Magnetism*, vol. 2, 2nd edn. (The Clarendon Press, Oxford, 1881), Part IV, Chapter XIII
75. J.J. Clement, Electromigration reliability, in *Design of High-Performance Microprocessor Circuits*, ed. by A.P. Chandrakasan, W.J. Bowhill, F. Fox (IEEE Press, New York, 2001), Chapter 20, pp. 429–448
76. I.A. Blech, H. Sello, *Mass Transport of Aluminum by Momentum Exchange with Conducting Electrons*. USAF-RADC Series, vol. 5 (United State Air Force – Rome Air Development Center, Rome, 1966), pp. 496–505
77. J.R. Black, Mass transport of aluminum by moment exchange with conducting electrons, in *Proceedings of the IEEE International Reliability Physics Symposium*, pp. 148–159, Apr 1967
78. F.M. D'Heurle, Electromigration and failure in electronics: an introduction. Proc. IEEE **59**(10), 1409–1417 (1971)
79. C.-K. Hu, K.P. Rodbell, T.D. Sullivan, K.Y. Lee, D.P. Bouldin, Electromigration and stress-induced voiding in fine Al and Al-Alloy thin-film lines. IBM J. Res. Dev. **39**(4), 465–497 (1995)
80. C. Ryu, K.-W. Kwon, A.L.S. Loke, H.Lee, T. Nogami, V.M. Dubin, R.A. Kavari, G.W. Ray, S.S. Wong, Microstructure and reliability of copper interconnects. IEEE Trans. Electron Devices **46**(6), 1113–1120 (1999)
81. M.J. Attardo, R. Rosenberg, Electromigration damage in aluminum film conductors. J. Appl. Phys. **41**(5), 2381–2386 (1970)
82. C.-K. Hu, R. Rosenberg, H.S. Rathore, D.B. Nguyen, B. Agarwala, Scaling effect on electromigration in on-chip Cu wiring, in *Proceedings on the IEEE International Conference on Interconnect Technology*, pp. 267–269, May 1999
83. R.H. Havemann, J.A. Hutchby, High-performance interconnects: an integration overview. Proc. IEEE **89**, 586–601 (2001)
84. F.G. Yost, D.E. Amos, A.D. Romig, Jr., Stress-driven diffusive voiding of aluminum conductor lines, in *Proceedings of the IEEE International Reliability Physics Symposium*, pp. 193–201, Apr 1989
85. I.A. Blech, K. L. Tai, Measurements of stress gradients generated by electromigration. Appl. Phys. Lett. **30**(8), 387–389 (1977)
86. I.A. Blech, Electromigration in thin aluminum films on titanium nitride. J. Appl. Phys. **47**(4), 1203–1208 (1976)
87. R.G. Filippi, R.A. Wachnik, H. Aochi, J.R. Lloyd, M.A. Korhonen, The effect of current density and stripe length on resistance saturation during electromigration testing. Appl. Phys. Lett. **69**(16), 2350–2352 (1996)

88. P. Børgesen, M.A. Korhonen, D.D. Brown, C.-Y. Li, H.S. Rathore, P.A. Totta, Stress evolution during stress migration and electromigration in passivated interconnect lines, in *Proceedings of the American Institute of Physics Conference*, vol. 305, pp. 231–253, June 1994
89. J.J. Clement, J.R. Lloyd, C.V. Thompson, Failure in tungsten-filled via structures, in *Proceedings of the Materials Research Society*, vol. 391, pp. 423–428, Apr 1995
90. B.N. Argarwala, M.J. Attardo, A.J. Ingraham, Dependence of electromigration-induced failure time on length and width of aluminum thin film conductors. J. Appl. Phys. **41**, 3954–3960 (1970)
91. J. Cho, C.V. Thompson, Grain size dependence of electromigration-induced failures in narrow interconnects. Appl. Phys. Lett. **54**(25), 2577–2579 (1989)
92. J.R. Black, Electromigration—a brief survey and some recent results. IEEE Trans. Electron Devices **4**2, 338–347 (1969)
93. J.J. Clement, Electromigration modeling for integrated circuit interconnect relibility analysis. IEEE Trans. Device Mater. Reliab. **1**(1), 33–42, (2001)
94. J.M. Towner, E.P. Van de Ven, Aluminum electromigration under pulsed DC conditions, in *Proceedings of the IEEE International Reliability Physics Symposium*, pp. 36–39, Apr 1983
95. J.A. Maiz, Characterization of electromigration under bidirectional (BC) and pulsed unidirectional (PDC) currents, in *Proceedings of the IEEE International Reliability Physics Symposium*, pp. 220–228, Apr 1989
96. L.M. Ting, J.S. May, W.R. Hunter, J.W. McPherson, AC electromigration characterization and modeling of multilayered interconnects, in *Proceedings of the IEEE International Reliability Physics Symposium*, pp. 311–316, Mar 1993
97. D. Edelstein et al., Full copper wiring in a sub-0.25 μm CMOS ULSI technology, in *Proceedings of the IEEE International Electron Devices Meeting*, pp. 773–776, Dec 1997
98. E.T. Ogawa, K.-D. Lee, V.A. Blaschke, P.S. Ho, Electromigration reliability issues in dual-damascene Cu interconnections. IEEE Trans. Reliab. **51**(4), 403–419 (2002)
99. S. Thrasher, C. Capasso, L. Zhao, R. Hernandez, P. Mulski, S. Rose, T. Nguyen, H. Kawasaki, Blech effect in single-inlaid Cu interconnects, in *Proceedings of the IEEE International Interconnect Technology Conference*, pp. 177–179, June 2001
100. P.C. Wang, R.G. Filippi, L.M. Gignac, Electromigration threshold in single-damascene copper interconnects with SiO_2 dielectrics, in *Proceedings of the IEEE International Interconnect Technology Conference*, pp. 263–265, June 2001
101. E.T. Ogawa, Direct observation of a critical length effect in dual-damascene Cu/oxide interconnects. Appl. Phys. Lett. **78**(18), 2652–2654 (2001)
102. S.P. Hau-Riege, Probabilistic immortality of Cu damascene interconnects. J. Appl. Phys. **91**(4), 2014–2022 (2002)
103. C.-K. Hu, L. Gignac, E. Liniger, R. Rosenberg, A. Stamper, Bimodal electromigration mechanisms in dual-damascene Cu line/via on W, in *Proceedings of the IEEE International Interconnect Technology Conference*, pp. 133–135, June 2002
104. B. Li, T.D. Sullivan, T.C. Lee, Line depletion electromigration characteristics of Cu interconnects, in *Proceedings of the IEEE International Reliability Physics Symposium*, pp. 140–145, Mar 2003
105. P. Moon, V. Chikarmane, K. Fischer, R. Grover, T.A. Ibrahim, D. Ingerly, K.J. Lee, C. Litteken, T. Mule, S. Williams, Process and electrical results for the on-die interconnect stack for Intel's 45nm process generation. Intel Technol. J. **12**(2), 87–92 (2008)
106. P. Justison, E. Ogawa, M. Gall, C. Capasso, D. Jawarani, J. Wetzel, H. Kawasaki, P.S. Ho, Electromigration in multi-level interconnects with polymeric low-*k* interlevel dielectrics, in *Proceedings of the IEEE International Interconnect Technology Conference*, pp. 202–204, June 2000
107. K.-D. Lee, X. Lu, E.T. Ogawa, H. Matsuhashi, P.S. Ho, V.A. Blaschke, R. Augur, Electromigration study of Cu/low *k* dual damascene interconnects, in *Proceedings of the IEEE International Reliability Physics Symposium*, pp. 322–326, Mar 2002
108. C.S. Hau-Riege, A.P. Marathe, V. Pham, The effect of low-*k* ILD on the electromigration reliability of Cu interconnects with different line lengths, in *Proceedings of the IEEE International Reliability Physics Symposium*, pp. 173–177, Mar 2003

109. E.T. Ogawa, K.-D. Lee, H. Matsuhashi, K.-S. Ko, P.R. Justison, A.N. Ramamurthi, A.J. Bierwag, P.S. Ho, V.A. Blaschke, R.H. Havemann, Statistics of electromigration early failures in Cu/oxide dual damascene interconnects, in *Proceedings of the IEEE International Reliability Physics Symposium*, pp. 341–349, Mar 2001
110. F.G. Yost, D.E. Amos, A.D. Romig Jr., Statistical electromigration budgeting for reliable design and verification in a 300-MHz microprocessor, in *Proceedings of the IEEE Symposium on VLSI Circuits*, pp. 115–116, 1995
111. *International Technology Roadmap for Semiconductors*, 2006 Update (Semiconductor Industry Association, 2006). http://public.itrs.net
112. R.H. Dennard, F.H. Gaensslen, V.L. Rideout, E. Bassous, A.R. LeBlanc, Design of ion-implanted MOSFET's with very small physical dimensions. IEEE J. Solid-State Circuits **SC–33**(5), 256–268 (1974)
113. K.C. Saraswat, E. Mohammadi, Effect of scaling of interconnections on the time delay of VLSI circuits. IEEE Trans. Electron Devices **ED–29**(4), 645–650 (1982)
114. H.B. Bakoglu, *Circuit, Interconnections and Packaging for VLSI* (Addison Wesley, Reading, 1990)
115. L.A. Arledge Jr., W.T. Lynch, Scaling and performance implications for power supply and other signal routing constraints imposed by I/O limitations, in *Proceedings of the IEEE Symposium on IC/Package Design Integration*, pp. 45–50, Feb 1998
116. W.S. Song, L.A. Glasser, Power distribution techniques for VLSI circuits. IEEE J. Solid-State Circuits **SC–21**(1), 150–156, (1986)
117. P. Larsson, Noise in CMOS integrated circuits di/dt. Analog Integr. Circuits Signal Process. **14**(1/2), 113–129 (1997)
118. G.A. Katopis, Delta-I noise specification for a high-performance computing machine. Proc. IEEE **73**(9), 1405–1415 (1985)
119. B.D. McCredie, W.D. Becker, Modeling, measurement, and simulation of simultaneous switching noise. IEEE Trans. Compon. Packag. Manuf. Technol. Pt. B: Adv. Packag. **19**(3), 461–472 (1996)
120. S.R. Nassif, O. Fakhouri, Technology trends in power-grid-induced noise, in *Proceedings of the Workshop on System Level Interconnect Prediction*, pp. 55–59, Apr 2002
121. *International Technology Roadmap for Semiconductors*, 1999 edn. (Semiconductor Industry Association, 1999). http://public.itrs.net
122. *International Technology Roadmap for Semiconductors*, 1997 edn. (Semiconductor Industry Association, 1997). http://public.itrs.net
123. *International Technology Roadmap for Semiconductors*, 1998 Update (Semiconductor Industry Association, 1998). http://public.itrs.net
124. *International Technology Roadmap for Semiconductors*, 2001 edn. (Semiconductor Industry Association, 2001). http://public.itrs.net
125. A.V. Mezhiba, E.G. Friedman, Inductance/area/resistance tradeoffs in high performance power distribution grids, in *Proceedings of the IEEE International Symposium on Circuit and Systems*, vol. I, pp. 101–104, May 2002
126. R.R. Tummala, E.J. Rymaszewski, A.G. Klopfenstein (eds.), *Microelectronics Packaging Handbook* (Chapman & Hall, New York, 1997)
127. D. Sylvester, H. Kaul, Future performance challenges in nanometer design, in *Proceedings of the IEEE/ACM Design Automation Conference*, pp. 3–8, June 2001
128. A.V. Mezhiba, E.G. Friedman, Scaling trends of on-chip power distribution noise, in *Proceedings of the Workshop on System Level Interconnect Prediction*, pp. 47–53, Apr 2002
129. A.V. Mezhiba, E.G. Friedman, Scaling trends of on-chip power distribution noise. IEEE Trans. Very Large Scale Integr. (VLSI) Circuits **12**(4), 386–394 (2004)
130. R. Jakushokas, E.G. Friedman, Line width optimization for interdigitated power/ground networks, in *Proceedings of the ACM Great Lakes Symposium on Very Large Scale Integration*, pp. 329–334, May 2010

131. A.V. Mezhiba, E.G. Friedman, Electrical characteristics of multi-layer power distribution grids, in *Proceedings of the IEEE International Symposium on Circuit and Systems*, vol. 5, pp. 473–476, May 2003
132. A. Deutsch et al., The importance of inductance and inductive coupling for on-chip wiring, in *IEEE Topical Meeting on Electrical Performance of Electronic Packaging*, pp. 53–56, Oct 1997
133. R. Evans, M. Tsuk, Modeling and measurement of a high-performance computer power distribution system. IEEE Trans. Compon. Packag. Manuf. Technol. Pt. B: Adv. Packag. **17**(4), 467–471 (1994)
134. D.J. Herrell, B. Beker, Modeling of power distribution systems for high-performance processors. IEEE Trans. Adv. Packag. **22**(3), 240–248 (1999)
135. T. Rahal-Arabi, G. Taylor, M. Ma, J. Jones, C. Webb, Design and validation of the core and IOs decoupling of the Pentium III and Pentium 4 processors, in *IEEE Topical Meeting on Electrical Performance of Electronic Packaging*, pp. 249–252, Oct 2002
136. L.D. Smith, R.E. Anderson, D.W. Forehand, T.J. Pelc, T. Roy, Power distribution system design methodology and capacitor selection for modern CMOS technology. IEEE Trans. Adv. Packag. **22**(3), 284–291 (1999)
137. A. Waizman, C.-Y. Chung, Resonant free power network design using extended adaptive voltage positioning (EAVP) methodology. IEEE Trans. Adv. Packag. **24**(3), 236–244, (2001)
138. I.Novak, L.M. Noujeim, V. St Cyr, N. Biunno, A. Patel, G. Korony, A. Ritter, Distributed matched bypassing for board-level power distribution networks. IEEE Trans. Adv. Packag. **25**(2), 230–242 (2002)
139. G.F. Taylor, C. Deutschle, T. Arabi, B. Owens, An approach to measuring power supply impedance of microprocessors, in *IEEE Topical Meeting on Electrical Performance of Electronic Packaging*, pp. 211–214, Oct 2001
140. R. Panda, D. Blaauw, R. Chaudhry, V. Zolotov, B. Young, R. Ramaraju, Model and analysis for combined package and on-chip power grid simulation, in *Proceedings of the IEEE International Symposium on Low Power Electronics and Design*, pp. 179–184, Aug 2000
141. A. Hasan, A. Sarangi, A. Sathe, G. Ji, High performance mobile Pentium III package development and design, in *Proceedings of the IEEE International Electronic Components and Technology Conference*, pp. 378–385, June 2002
142. I. Novak, Lossy power distribution networks with thin dielectric layers and/or thin conductive layers. IEEE Trans. Adv. Packag. **23**(3), 353–360 (2000)
143. H. Braunisch, S.N. Towle, R.D. Emery, C. Hu, G.J. Vandentop, Electrical performance of bumpless build-up layer packaging, in *Proceedings of the IEEE International Electronic Components and Technology Conference*, pp. 353–358, June 2002
144. B.W. Amick, C.R. Gauthier, D. Liu, Macro-modeling concepts for the chip electrical interface, in *Proceedings of the IEEE/ACM Design Automation Conference*, pp. 391–394, June 2002
145. D. Blaauw, R. Panda, R. Chaudhry, Design and analysis of power distribution networks, in *Design of High-Performance Microprocessor Circuits*, ed. by A.P. Chandrakasan, W.J. Bowhill, F. Fox (IEEE Press, New York, 2001), Chapter 24, pp. 499–522
146. A. Dalal, L. Lev, S. Mitra, Design of an efficient power distribution network for the UltraSPARC-I microprocessor, in *Proceedings of the IEEE International Conference on Computer Design*, pp. 118–123, Oct 1995
147. M.K. Gowan, L.L. Biro, D.B. Jackson, Power considerations in the design of the alpha 21264 microprocessor, in *Proceedings of the IEEE/ACM Design Automation Conference*, pp. 726–731, June 1998
148. L. Cao, J.P. Krusius, A new power distribution strategy for area array bonded ICs and packages of future deep sub-micron ULSI, in *Proceedings of the IEEE International Electronic Components and Technology Conference*, pp. 915–920, June 1999
149. P.E. Gronowski, W.J. Bowhill, R.P. Preston, M.K. Gowan, R.L. Allmon, High-performance microprocessor design. IEEE J. Solid-State Circuits **33**(5), 676–686 (1998)

150. D.A. Priore, Inductance on silicon for sub-micron CMOS VLSI, in *Proceedings of the IEEE Symposium on VLSI Circuits*, pp. 17–18, May 1993
151. L.-R. Zheng, H. Tenhunen, Effective power and ground distribution scheme for deep submicron high speed VLSI circuits, in *Proceedings of the IEEE International Symposium on Circuit and Systems*, vol. I, pp. 537–540, May 1999
152. B.J. Benschneider et al., A 1 GHz alpha microprocessor, in *Proceedings of the IEEE International Solid-State Circuits Conference*, pp. 86–87, Feb 2000
153. A. Dharchoudhury, Design and analysis of power distribution networks in PowerPC microprocessors, in *Proceedings of the IEEE/ACM Design Automation Conference*, pp. 738–743, June 1998
154. Y.-L. Li, T.-G. Yew, C.-Y. Chung, D.G. Figueroa, Design and performance evaluation of microprocessor packaging capacitors using integrated capacitor-via-plane model. IEEE Trans. Adv. Packag. **23**(3), 361–367 (2000)
155. S.H. Hashemi, P.A. Sandborn, D. Disko, R. Evans, The close attached capacitor: a solution to switching noise problems. IEEE Trans. Adv. Packag. **15**(6), 1056–1063 (1992)
156. B.A. Gieseke et al., A 600 MHz superscalar RISC microprocessor with out-of-order execution, in *Proceedings of the IEEE International Solid-State Circuits Conference*, pp. 176–177, Feb 1997
157. S.H. Hall, G.W. Hall, J.A. McCall, *High-Speed Digital System Design: A Handbook of Interconnect Theory and Design Practices* (Wiley, New York, 2000)
158. A. Jain et al., A 1.2 GHz alpha microprocessor with 44.8 GB/s chip pin bandwidth, in *Proceedings of the IEEE International Solid-State Circuits Conference*, pp. 240–241, Feb 2001
159. R. Heald et al., Implementation of a 3rd generation SPARC V9 64b microprocessor, in *Proceedings of the IEEE International Solid-State Circuits Conference*, pp. 412–413, Feb 2000
160. J.D. Warnock, J.M. Keaty, J. Petrovick, J.G. Clabes, C.J. Kircher, B.L. Krauter, P.J. Restle, B.A. Zoric, C.J. Anderson, The circuit and physical design of the POWER4 microprocessor. IBM J. Res. Dev. **46**(1), 27–51 (2002)
161. W.T. Lynch, L.A. Arledge Jr., Power supply distribution and other wiring issues for deep-submicron ICs, in *Proceedings of the Material Research Society Symposia*, vol. 514, pp. 11–27, Apr 1998
162. L. Zu, M. Joshi, C. Houghton, B. Loughlin, G. Vaccaro, J. Dietz, Improving microprocessor performance with flip chip package design, in *Proceedings of the IEEE symposium on IC/package design integration*, pp. 82–87, Feb 1998
163. S. Lipa, J.T. Schaffer, A.W. Glaser, P.D. Franzon, Flip-chip power distribution, in *Proceedings of the IEEE Topical Meeting on Electrical Performance of Electronic Packaging*, pp. 39–41, Oct 1998
164. P.D. Franzon, J.T. Schaffer, S. Lipa, A.W. Glaser, Issues in chip-package codesign with MCM-D/flip-chip technology, in *Proceedings of the IEEE Topical Meeting on Electrical Performance of Electronic Packaging*, pp. 88–92, Oct 1998
165. A. Sarangi, G. Ji, T. Arabi, G.F. Taylor, Design and performance evaluation of Pentium III microprocessor packaging, in *IEEE Topical Meeting on Electrical Performance of Electronic Packaging*, pp. 291–294, Oct 2001
166. R. Mahajan, R. Nair, V. Wakharkar, J. Swan, J. Tang, G. Vandentop, High performance package design for a 1 GHz microprocessor. Intel Technol. J. **6**(2), 62–75 (2002)
167. T. Kawahara, SuperCSP™. IEEE Trans. Adv. Packag. **23**(2), 215–219 (2000)
168. S.N. Towle, H. Braunisch, C. Hu, R.D. Emery, G.J. Vandentop, Bumpless build-up layer packaging, in *Proceedings of the ASME International Mechanical Engineering Congress and Exposition*, pp. 11–16, Nov 2001
169. A. Hasan, A. Sarangi, C.S. Baldwin, R.L. Sankman, G.F. Taylor, High performance package design for a 1 GHz microprocessor. IEEE Trans. Adv. Packag. **24**(4), 470–476 (2001)

170. M. Tsuk, R. Dame, D. Dvorscak, C. Houghton, J.S. Laurent, Modeling and measurement of the alpha 21364 package, in *Proceedings of the IEEE International Electronic Components and Technology Conference*, pp. 283–286, June 2001
171. P. Saxena, S. Gupta, Shield count minimization in congested regions, in *Proceedings of the ACM International Symposium on Physical Design*, pp. 78–83, Apr 2002
172. H. Su, J. Hu, S.S. Sapatnekar, S.R. Nassif, Congestion-driven codesign of power and signal networks, in *Proceedings of the IEEE/ACM Design Automation Conference*, pp. 64–69, June 2002
173. P. Saxena, S. Gupta, On integrating power and signal routing for shield count minimization in congested regions. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **22**(4), 437–445 (2003)
174. P. Restle, A.E. Ruehli, S.G. Walker, Dealing with inductance in high-speed chip design, in *Proceedings of the IEEE/ACM Design Automation Conference*, pp. 904–909, June 1999
175. P.J. Restle, A.E. Ruehli, S.G. Walker, Multi-GHz interconnect effects in microprocessors, in *Proceedings of the ACM international symposium on physical design*, pp. 93–97, Apr 2001
176. R.R. Troutman, *Latch-Up in CMOS Technology: The Problem and Its Cure* (Kluwer Academic, Boston, 1986)
177. R. Panda, S. Sundareswaran, D. Blaauw, On the interaction of power distribution network with substrate, in *Proceedings of the IEEE International Symposium on Low Power Electronics and Design*, pp. 388–393, Aug 2001
178. M. Badaroglu, G. Van der Plas, P. Wambacq, S. Donnay, G.G.E. Gielen, H.J. De Man, SWAN: high-level simulation methodology for digital substrate noise generation. IEEE Trans. Very Large Scale Integr. (VLSI) Circuits **14**(1), 23–33 (2006)
179. E. Salman, R. Jakushokas, E.G. Friedman, R.M. Secareanu, O.L. Hartin, Methodology for efficient substrate noise analysis in large scale mixed-signal circuits. IEEE Trans. Very Large Scale Integr. (VLSI) Circuits **17**(10), 1405–1418 (2009)
180. D. Ozis, T. Fiez, K. Mayaram, A comprehensive geometry-dependent macromodel for substrate noise coupling in heavily doped CMOS processes, in *Proceedings of the IEEE Custom Integrated Circuits Conference*, pp. 497–500, Sept 2002
181. H. Lan, T.W. Chen, C.O. Chui, P. Nikaeen, J.W. Kim, R.W. Dutton, Synthesized compact models and experimental verifications for substrate noise coupling in mixed-signal ICs. IEEE J. Solid-State Circuits **41**(8), 1817–1829 (2006)
182. A. Samavedam, A. Sadate, K. Mayaram, T.S. Fiez, A scalable substrate noise coupling model for design of mixed-signal IC's. IEEE J. Solid-State Circuits **35**(6), 895–904 (2000)
183. R. Jakushokas, E. Salman, E.G. Friedman, R.M. Secareanu, O.L. Hartin, C.L. Recker, Compact substrate models for efficient noise coupling and signal isolation analysis, in *Proceedings of the IEEE International Symposium on Circuit and Systems*, pp. 2346–2349, May/June 2010
184. R.M. Secareanu, S. Warner, S. Seabridge, C. Burke, T.E. Watrobski, C. Morton, W. Staub, T. Tellier, E.G. Friedman, Placement of substrate contacts to minimize substrate noise in mixed-signal integrated circuits. Analog Integr. Circuits Signal Process. **28**(3), 253–264 (2001)
185. J. Kim, W. Lee, Y. Shim, J. Shim, K. Kim, J.S. Pak, J. Kim, Chip-package hierarchical power distribution network modeling and analysis based on a segmentation method. IEEE Trans. Adv. Packag. **33**(3), 647–659 (2010)
186. P. Hazucha et al., A 233-MHz 80%–87% efficient four-phase DC – DC converter utilizing air-core inductors on package. IEEE J. Solid-State Circuits **40**(4), 838–845 (2005)
187. I. Vaisband, E.G. Friedman, Heterogeneous methodology for energy efficient distribution of on-chip power supplies. IEEE Trans. Power Electron. **28**(9), 4267–4280 (2013)
188. P.J. Restle et al., A clock distribution network for microprocessors. IEEE J. Solid-State Circuits **36**(5), 792–799, May 2001
189. I. Vaisband, E.G. Friedman, R. Ginosar, A. Kolodny, Low power clock network design. J. Low Power Electron. Appl. **1**(1), 219–246(2011)

190. M. Popovich, M. Sotman, A. Kolodny, E.G. Friedman, Effective radii of on-chip decoupling capacitors. IEEE Trans. Very Large Scale Integr. (VLSI) Circuits **16**(7), 894–907 (2008)
191. S. Kose, E.G. Friedman, Distributed on-chip power delivery. IEEE J. Emerg. Sel. Top. Circuits Syst. **2**(4), 704–713 (2012)
192. S.R. Nassif, Power grid analysis benchmarks, in *Proceedings of the IEEE/ACM Asia and South Pacific Design Automation Conference*, pp. 376–381, Jan 2008
193. I. Vaisband, E.G. Friedman, Power Network On-Chip for Scalable Power Delivery. U.S. Patent 62/042,572
194. I. Vaisband, E.G. Friedman, Power network on-chip for scalable power delivery, in *Proceedings of the Workshop on System Level Interconnect Prediction*, pp. 1–5, June 2014
195. I. Vaisband, E.G. Friedman, Dynamic power management with power network-on-chip, in *Proceeding of the IEEE International Conference on New Circuits and Systems*, pp. 225–228, June 2014
196. G.E.R. Lloyd, *Early Greek Science: Thales to Aristotle* (W. W. Norton, New York, 1974).
197. A.D. Moore, *Electrostatics and Its Applications* (Wiley, New York, 1973)
198. J.L. Heilborn, *Electricity in the 17th & 18th Centuaries: A Study in Early Modern Physics* (Dover, Mineola, 1999)
199. A. Guillemin, *Electricity and Magnetism* (Macmillam, London, 1891)
200. M. Faraday, *Experimental Researches in Electricity* (Dover, Mineola, 2004)
201. J.D. Cutnell, K.W. Johnson, *Physics*, 6th edn. (Wiley, Hoboken, 2003)
202. T.H. Lee, *The Design of CMOS Radio-Frequency Integrated Circuits*, 2nd edn. (Cambridge University Press, New York, 2004)
203. C.P. Yuan, T.N. Trick, A simple formula for the estimation of the capacitance of two-dimensional interconnects in VLSI circuits. IEEE Electron Device Lett. **3**(12), 391–393 (1982)
204. T. Sakurai, K. Tamaru, Simple formulas for two- and three-dimensional capacitance. IEEE Trans. Electron Devices **30**(2), 183–185 (1983)
205. J.-H. Chern, J. Huang, L. Arledge, P.-C. Li, P. Yang, Multilevel metal capacitance models for CAD design synthesis systems. IEEE Electron Device Lett. **13**(1), 32–34 (1992)
206. S.-C. Wong, G.-Y. Lee, D.-J. Ma, Modeling of interconnect capacitance, delay, and crosstalk in VLSI. IEEE Trans. Semicond. Manuf. **13**(1), 108–111 (2000)
207. E. Barke, Line-to-ground capacitance calculation for VLSI: a comparison. IEEE Trans. Comput. Aided Design Integr. Circuits Syst. **7**(2), 295–298 (1988)
208. N.P. van der Meijs, J.T. Fokkema, VLSI circuit reconstruction from mask topology. Integration **2**(2), 85–119 (1984)
209. T. Roy, L. Smith, J. Prymak, ESR and ESL of ceramic capacitor applied to decoupling applications, in *Proceedings of the IEEE Conference on Electrical Performance of Electronic Packaging*, pp. 213–216, Oct 1998
210. D.A. Neamen, *Semiconductor Physics and Devices: Basic Principles*, 3rd edn. (McGraw-Hill, New York, 2002)
211. Power Distribution System (PDS) Design: Using Bypass/Decoupling Capacitors. http://direct.xilinx.com/bvdocs/appnotes/xapp623.pdf.
212. W. Becker, H. Smith, T. McNamara, P. Muench, J. Eckhardt, M. McAllister, G. Katopis, S. Richter, R. Frech, E. Klink, Mid-frequency simultaneous switching noise in computer systems, in *Proceedings of the IEEE Electronic Components and Technology Conference*, pp. 676–681, May 1997
213. W.D. Becker, J. Eckhardt, R.W. Frech, G.A. Katopis, E. Klink, M.F. McAllister, T.G. McNamara, P. Muench, S.R. Richter, H. Smith, Modeling, simulation, and measurement of mid-frequency simultaneous switching noise in computer systems. IEEE Trans. Compon. Packag. Manuf. Technol. Pt. B: Adv. Packag. **21**(2), 157–163 (1998)
214. T. Zhou, T. Strach, W.D. Becker, On chip circuit model for accurate mid-frequency simultaneous switching noise prediction, in *Proceedings of the IEEE Conference on Electrical Performance of Electronic Packaging*, pp. 275–278, Oct 2005

215. S. Bobba, T. Thorp, K. Aingaran, D. Liu, IC power distribution challenges, in *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pp. 643–650, Nov 2001
216. S. Chun, M. Swaminathan, L.D. Smith, J. Srinivasan, Z. Jin, M.K. Iyer, Physics based modeling of simultaneous switching noise in high speed systems, in *Proceedings of the IEEE Electronic Components and Technology Conference*, pp. 760–768, May 2000
217. S. Chun, M. Swaminathan, L.D. Smith, J. Srinivasan, Z. Jin, M.K. Iyer, Modeling of Simultaneous Switching Noise in High Speed Systems. IEEE Trans. Adv. Packag. **24**(2), 132–142 (2001)
218. L. Smith, Simultaneous switching noise and power plane bounce for CMOS technology, in *Proceedings of the IEEE Conference on Electrical Performance of Electronic Packaging*, pp. 163–166, Oct 1999
219. F.Y. Yuan, Electromagnetic modeling and signal integrity simulations of power/ground networks in high speed digital packages and printed circuit boards, in *Proceedings of the IEEE/ACM Design Automation Conference*, pp. 421–426, June 1998
220. Z. Mu, Simulation and modeling of power and ground planes in high speed printed circuit boards, in *Proceedings of the IEEE International Symposium on Circuits and Systems*, pp. 459–462, May 2001
221. N. Na, J. Choi, S. Chun, M. Swaminathan, J. Srinivasan, Modeling and transient simulation of planes in electronic packages. IEEE Trans. Adv. Packag. **23**(3), 340–352 (2000)
222. T.-G. Yew, Y.-L. Li, C.-Y. Chung, D.G. Figueroa, Design and performance evaluation of chip capacitors on microprocessor packaging, in *Proceedings of the IEEE Conference on Electrical Performance of Electronic Packaging*, pp. 175–178, Oct 1999
223. J. Kim, B. Choi, H. Kim, W. Ryu, Y. Yun, S. Ham, S.-H. Kim, Y. Lee, J. Kim, Separated role of on-chip and on-PCB decoupling capacitors for reduction of radiated emission on printed circuit board, in *Proceedings of the IEEE International Symposium on Electromagnetic Compatibility*, pp. 531–536, Aug 2001
224. B. Garben, G.A. Katopis, W.D. Becker, Package and chip design optimization for mid-frequency power distribution decoupling, in *Proceedings of the IEEE Conference on Electrical Performance of Electronic Packaging*, pp. 245–248, Oct 2002
225. M. Xu, T.H. Hubing, J. Chen, T.P. Van Doren, J.L. Drewniak, R.E. DuBroff, Power-bus decoupling with embedded capacitance in printed circuit board design. IEEE Trans. Electromagn. Compat. **45**(1), 22–30 (2003)
226. M.I. Montrose, Analysis on loop area trace radiated emissions from decoupling capacitor placement on printed circuit boards, in *Proceedings of the IEEE International Symposium on Electromagnetic Compatibility*, pp. 423–428, Aug 1999
227. P. Muthana, M. Swaminathan, E. Engin, P. Markondeya Raj, R. Tummala, Mid frequency decoupling using embedded decoupling capacitors, in *Proceedings of the IEEE Conference on Electrical Performance of Electronic Packaging*, pp. 271–274, Oct 2005
228. O.P. Mandhana, Design oriented analysis of package power distribution system considering target impedance for high performance microprocessors, in *Proceedings of the IEEE Conference on Electrical Performance of Electronic Packaging*, pp. 273–276, Oct 2001
229. L.D. Smith, D. Hockanson, Distributed SPICE circuit model for ceramic capacitors, in *Proceedings of the IEEE Electronic Components and Technology Conference*, pp. 523–528, May/June 2001
230. P. Larsson, Resonance and damping in CMOS circuits with on-chip decoupling capacitance. IEEE Trans. Circuits Syst. I: Fundam. Theory Appl. **45**(8), 849–858 (1998)
231. L.D. Smith, R.E. Anderson, T. Roy, Chip-package resonance in core power supply structures for a high power microprocessor, in *Proceedings of the ASME International Electronic Packaging Technical Conference and Exhibition*, vol. 5, July 2001
232. C.R. Paul, Effectiveness of multiple decoupling capacitors. IEEE Trans. Electromag. Compat. **34**(2), 130–133 (1992)

233. M. Popovich, E.G. Friedman, Decoupling capacitors for power distribution systems with multiple power supplies, in *Proceedings of the IEEE EDS/CAS Activities in Western New York Conference*, p. 9, Nov 2004
234. A. Waizman, C.-Y. Chung, Extended adaptive voltage positioning (EAVP), in *Proceedings of the IEEE Conference on Electrical Performance of Electronic Packaging*, pp. 65–68, Oct 2000
235. M. Sotman, M. Popovich, A. Kolodny, E.G. Friedman, Leveraging symbiotic on-die decoupling capacitance, in *Proceedings of the IEEE Conference on Electrical Performance of Electronic Packaging*, pp. 111–114, Oct 2005
236. H.H. Chen, J.S. Neely, Interconnect and circuit modeling techniques for full-chip power supply noise analysis. IEEE Trans. Compon. Packag. Manuf. Technol. Pt. B: Adv. Packag. **21**(3), 209–215 (1998)
237. S. Bobba, I.N. Hajj, Input vector generation for maximum intrinsic decoupling capacitance of VLSI circuits, in *Proceedings of the IEEE International Symposium on Circuits and Systems*, pp. 195–198, May 2001
238. R. Panda, S. Sundareswaran, D. Blaauw, Impact of low-impedance substrate on power supply integrity. IEEE Trans. Des. Test Comput. **20**(3), 16–22, May/June 2003
239. G. Steele, D. Overhauser, S. Rochel, S.Z. Hussain, Full-chip verification methods for DSM power distribution systems, in *Proceedings of the IEEE/ACM Design Automation Conference*, pp. 744–749, June 1998
240. N.H. Pham, On-chip capacitor measurement for high performance microprocessor, in *IEEE Topical Meeting on Electrical Performance of Electronic Packaging*, pp. 65–68, Oct 1998.
241. H. Seidl et al., A fully integrated Al_2O_3 trench capacitor DRAM for Sub-100 nm technology, in *Proceedings of the IEEE International Electron Devices Meeting*, pp. 839–842, Dec 2002
242. K.V. Rao, M. Elahy, D.M. Bordelon, S.K. Banerjee, H.L. Tsai, W.F. Richardson, R.H. Womack, Trench capacitor design issues in VLSI DRAM cells, in *Proceedings of the IEEE International Electron Devices Meeting*, pp. 140–143, Dec 1986
243. W.J. Bowhill et al., Circuit implementation of a 300 MHz 64-bit second generation CMOS alpha CPU. Digit. Tech. J. **7**(1), 100–118 (1995)
244. P. Larsson, Parasitic resistance in an MOS transistor used as on-chip decoupling capacitance. IEEE J. Solid-State Circuits **32**(4), 574–576 (1997)
245. A. Hastings, *The Art of Analog Layout* (Prentice Hall, Upper Saddle River, 2001)
246. J.L. McCreary, Matching properties, and voltage and temperature dependence of MOS capacitors. IEEE J. Solid-State Circuits **16**(6), 608–616 (1981)
247. R.T. Howe, C.G. Sodini, *Microelectronics: An Integrated Approach* (Prentice Hall, Upper Saddle River, 1996)
248. C.T. Black, K.W. Guarini, Y. Zhang, H. Kim, J. Benedict, E. Sikorski, I.V. Babich, K.R. Milkove, High-capacity, self-assembled metal-oxide-semiconductor decoupling capacitors. IEEE Electron Device Lett. **25**(9), 622–624 (2004)
249. A.R. Alvarez, BiCMOS technology and applications (Kluwer Academic, Norwell, 1993)
250. A. Behr, M. Schneider, S. Filho, C. Montoro, Harmonic distortion caused by capacitors implemented with MOSFET gates. IEEE J. Solid-State Circuits **27**(10), 1470–1475 (1992)
251. S. Rusu, J. Stinson, S. Tam, J. Leung, H. Muljono, B. Cherkauer, A 1.5-GHz 130-nm itanium 2 processor with 6-MB on-die L3 cache. IEEE J. Solid-State Circuits **38**(11), 1887–1895 (2003)
252. Optimization of Metal-Metal Comb-Capacitors for RF Applications. http://www.oea.com/document/OptimizMetal.pdf
253. M.J. Deen, T.A. Fjeldly, *CMOS RF Modeling, Characterization and Applications* (World Scientific, River Edge, 2004)
254. B. Razavi, *RF Microelectronics* (Prentice Hall, Upper Saddle River, 1998)
255. R.K. Ulrich, L.W. Schaper, *Integrated Passive Component Technology* (Wiley-IEEE Press, New York, 2003)
256. S.B. Chen, C.H. Lai, A. Chin, J.C. Hsieh, J. Liu, High-density MIM capacitors using Al_2O_3 and $AlTiO_x$ dielectrics. IEEE Electron Device Lett. **23**(4), 185–187 (2002)

257. M.Y. Yang, C.H. Huang, A. Chin, C. Zhu, M.F. Li, D.-L. Kwong, High-density MIM capacitors using $AlTaO_x$ dielectrics. IEEE Electron Device Lett. **24**(5), 306–308 (2003)
258. X. Yu, C. Zhu, H. Hu, A. Chin, M.F. Li, B.J. Cho, D.-L. Kwong, P.D. Foo, M.B. Yu, A high-density MIM capacitor (13 fF/μm^2) using ALD HfO_2 dielectrics. IEEE Electron Device Lett. **24**(2), 63–65 (2003)
259. H. Hu et al., High performance ALD HfO_2–Al_2O_3 laminate MIM capacitors for RF and mixed signal IC applications, in *Proceedings of the IEEE International Electron Devices Meeting*, pp. 15.6.1–15.6.4, Dec 2003
260. S.-J. Ding et al., High-density MIM capacitor using ALD high-k HfO_2 laminate dielectrics. IEEE Electron Device Lett. **24**(12), 730–732 (2003)
261. S.-J. Kim, B.J. Cho, M.B. Yu, M.-F. Li, Y.-Z. Xiong, C. Zhu, A. Chin, D.-L. Kwong, Metal-insulator-metal RF bypass capacitor using niobium oxide (Nb_2O_5) with HfO_2–Al_2O_3 barriers. IEEE Electron Device Lett. **26**(9), 625–627 (2005)
262. Y.H. Wu, A. Chin, K.H. Shih, C.C. Wu, C.P. Liao, S.C. Pai, C.C. Chi, The fabrication of very high resistivity Si with low loss and cross talk. IEEE Electron Device Lett. **21**(9), 442–444 (2000)
263. A. Kar-Roy, C. Hu, M. Racanelli, C.A. Compton, P. Kempf, G. Jolly, P.N. Sherman, J. Zheng, Z. Zhang, A. Yin, High density metal insulator metal capacitors using PECVD nitride for mixed signal and RF circuits, in *Proceedings of the IEEE International Conference on Interconnect Technology*, pp. 245–247, May 1999
264. J.A. Babcock, S.G. Balster, A. Pinto, C. Dirnecker, P. Steinmann, R. Jumpertz, B. El-Kareh, Analog characteristics of metal-insulator-metal capacitors using PECVD nitride dielectrics. IEEE Electron Device Lett. **22**(5), 230–232 (2001)
265. P. Zurcher et al., Integration of thin film MIM capacitors and resistors into copper metallization based RF-CMOS and Bi-CMOS technologies, in *Proceedings of the IEEE International Electron Devices Meeting*, pp. 153–156, Dec 2000
266. M. Armacost, A. Augustin, P. Felsner, Y. Feng, G. Friese, J. Heidenreich, G. Hueckel, O. Prigge, K. Stein, A high reliability metal insulator metal capacitor for 0.18 μm copper technology, in *Proceedings of the IEEE International Electron Devices Meeting*, pp. 157–160, Dec 2000
267. C.H. Ng, K.W. Chew, J.X. Li, T.T. Tjoa, L.N. Goh, S.F. Chu, Characterization and comparison of two metal-insulator-metal capacitor schemes in 0.13 μm copper dual damascene metallization process for mixed-mode and RF applications, in *Proceedings of the IEEE International Electron Devices Meeting*, pp. 241–244, Dec 2002
268. N. Inoue, H. Ohtake, I. Kume, N. Furutake, T. Onodera, S. Saito, A. Tanabe, M. Tagami, M. Tada, Y. Hayashi, High performance high-k MIM capacitor with plug-in plate (PiP) for power delivery line of high-speed MPUs, in *Proceedings of the IEEE International Interconnect Technology Conference*, pp. 63–65, June 2006
269. T. Soorapanth, CMOS RF Filtering at GHz Frequency, Ph.D. Thesis, Stanford University, Stanford, 2002
270. Applications of Metal-Insulator-Metal (MIM) Capacitors, International SEMATECH, Technology Transfer No. 00083985A-ENG, Oct 2000.
271. O.E. Akcasu, High Capacitance Structures in a Semiconductor Device, U.S. Patent 5,208,725, 4 May 1993
272. B.B. Mandelbrot, *The Fractal Geometry of Nature* (Freeman, New York, 1983)
273. H. Samavati, A. Hajimiri, A.R. Shahani, G.N. Nasserbakht, and T.H. Lee, Fractal capacitors. IEEE J. Solid-State Circuits **33**(12), 2035–2041 (1998)
274. R. Aparicio A. Hajimiri, Capacity limits and matching properties of integrated capacitors. IEEE J. Solid-State Circuits **37**(3), 384–393 (2002)
275. A.C.C. Ng, M. Saran, Capacitor Structure for an Integrated Circuit, U.S. Patent 5,583,359, 10 Dec 1996
276. M. Popovich, E.G. Friedman, Decoupling capacitors for power distribution systems with multiple power supply voltages, in *Proceedings of the IEEE International SOC Conference*, pp. 331–334, Sept 2004

277. M. Popovich, E.G. Friedman, M. Sotman, A. Kolodny, R.M. Secareanu, Maximum effective distance of on-chip decoupling capacitors in power distribution grids, in *Proceedings of the ACM Great Lakes Symposium on Very Large Scale Integration*, pp. 173–179, Mar 2006
278. M. Ang, R. Salem, A. Taylor, An on-chip voltage regulator using switched decoupling capacitors, in *Proceedings of the IEEE International Solid-State Circuits Conference*, pp. 438–439, Feb 2000
279. M. Popovich, E.G. Friedman, Noise aware decoupling capacitors for multi-voltage power distribution systems, in *Proceedings of the ACM/IEEE International Symposium on Quality Electronic Design*, pp. 334–339, Mar 2005
280. L. Smith, Decoupling capacitor calculations for CMOS circuits, in *IEEE Topical Meeting on Electrical Performance of Electronic Packaging*, pp. 101–105, Nov 1994
281. H.H. Chen, S.E. Schuster, On-chip decoupling capacitor optimization for high-performance vlsi design, in *Proceedings of the IEEE International Symposium on VLSI Technology, Systems, and Applications*, pp. 99–103, May 1995
282. M.D. Pant, P. Pant, D.S. Wills, On-chip decoupling capacitor optimization using architectural level prediction. IEEE Trans. Very Large Scale Integr. (VLSI) Circuits **10**(3), 319–326 (2002)
283. H. Su, S.S. Sapatnekar, S.R. Nassif, Optimal decoupling capacitor sizing and placement for standard cell layout designs. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **22**(4), 428–436 (2003)
284. S. Zhao, K. Roy, C.-K. Koh, Decoupling capacitance allocation and its application to power-supply noise-aware floorplanning. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **21**(1), 81–92 (2002)
285. F. Moll, M. Roca, *Interconnect Noise in VLSI Circuits* (Kluwer Academic, Norwell, 2003)
286. *International Technology Roadmap for Semiconductors*, 2005 edn. (Semiconductor Industry Association, 2005). http://public.itrs.net
287. M.E. Van Valkenburg, *Network Analysis* (Prentice Hall, Upper Saddle River, 1974)
288. M. Takamiya, M. Mizuno, A 6.7 fF/μm^2 bias-independent gate capacitor (BIGCAP) with digital CMOS process and its application to the loop filter of a differential PLL. IEEE J. Solid-State Circuits **40**(3), 719–725 (2005)
289. V. Kursun, E.G. Friedman, *Multi-Voltage CMOS Circuit Design* (Wiley, Hoboken, 2006)
290. D. Lee, D. Blaauw, D. Sylvester, Gate oxide leakage current analysis and reduction for VLSI circuits. IEEE Trans. Very Large Scale Integr. (VLSI) Syst. **12**(2), 155–166 (2004)
291. M. Anis, Y. Massoud, Power design challenges in deep-submicron technology, in *Proceedings of the IEEE International Midwest Symposium on Circuits and Systems*, pp. 1510–1513, Dec 2003
292. D. Deleganes, J. Douglas, B. Kommandur, M. Patyra, Designing a 3 GHz, 130 nm, Intel Pentium 4 Processor, in *Proceedings of the IEEE Symposium on VLSI Circuits*, pp. 130–133, June 2002
293. R. McGowen, C.A. Poirier, C. Bostak, J. Ignowski, M. Millican, W.H. Parks, S. Naffziger, Power and temperature control on a 90-nm itanium family processor. IEEE J. Solid-State Circuits **41**(1), 229–237, Jan 2006
294. S. Naffziger, B. Stackhouse, T. Grutkowski, D. Josephson, J. Desai, E. Alon, M. Horowitz, The implementation of a 2-Core, multi-threaded itanium family processor. IEEE J. Solid-State Circuits **41**(1), 197–209 (2006)
295. T. Hubing, Effective strategies for choosing and locating printed circuit board decoupling capacitors, in *Proceedings of the IEEE International Symposium on Electromagnetic Compatibility*, pp. 632–637, Aug 2005
296. Mathematica 5.2, Wolfram Research, Inc.
297. M.P. Goetz, Time and frequency domain analysis of integral decoupling capacitors. IEEE Trans. Compon. Packag. Manuf. Technol. Pt. B: Adv. Packag. **19**(3), 518–522 (1996)
298. T. Murayama, K. Ogawa, H. Yamaguchi, Estimation of peak current trough CMOS VLSI circuit supply lines, in *Proceedings of the ACM Asia and South Pacific Design Automation Conference*, pp. 295–298, Jan 1999

299. M. Popovich, A.V. Mezhiba, E.G. Friedman, *Power Distribution Networks with On-Chip Decoupling Capacitors* (Springer, New York, 2008)
300. S. Pant, D. Blaauw, E. Chiprout, Power Grid Physics and Implications for CAD. IEEE Des. Test Comput. **24**(3), 246–254 (2007)
301. S. Kose, E.G. Friedman, On-chip point-of-load voltage regulator for distributed power supplies, in *Proceedings of the ACM Great Lakes Symposium on Very Large Scale Integration*, pp. 377–380, May 2010
302. S. Kose, E.G. Friedman, An area efficient fully monolithic hybrid voltage regulator, in *Proceedings of the IEEE International Symposium on Circuits and Systems*, pp. 2718–2721, May 2010
303. P. Hazucha, T. Karnik, B.A. Bloechel, C. Parsons, D. Finan, S. Borkar, Area-efficient linear regulator with ultra-fast load regulation. IEEE J. Solid-State Circuits **40**(4), 933–940 (2005)
304. M. Popovich, E.G. Friedman, M. Sotman, A. Kolodny, On-chip power distribution grids with multiple supply voltages for high-performance integrated circuits. IEEE Trans. Very Large Scale Integr. (VLSI) Circuits **16**(7), 908–921 (2008)
305. E. Salman, E.G. Friedman, R.M. Secareanu, O.L. Martin, Worst case power/ground noise estimation using an equivalent transition time for resonance. IEEE Trans. Circuits Syst. I: Regul. Pap. **56**(5), 997–1004 (2009)
306. A. Todri, M. Sadowska, F. Maire, C. Matheron, A study of decoupling capacitor effectiveness in power and ground grid networks, in *Proceedings of the IEEE International Symposium on Quality Electronic Design*, pp. 653–658, Mar 2009
307. G. Venezian, On the resistance between two points on a grid. Am. J. Phys. **62**(11), 1000–1004 (1994)
308. E. Salman, E.G. Friedman, *High Performance Integrated Circuit Design* (McGraw-Hill, New York, 2012)
309. F. Waldron, J. Slowey, A. Alderman, B. Narveson, S.C. O'Mathuna, Technology roadmapping for power supply in package (PSiP) and power supply on chip (PwrSoC), in *Proceedings of the IEEE International Applied Power Electronics Conference and Exposition*, pp. 525–532, Feb 2010
310. S. Kose, S. Tam, S. Pinzon, B. McDermott, E.G. Friedman, Active filter based hybrid on-chip DC-DC converters for point-of-load voltage regulation. IEEE Trans. Very Large Scale Integr. (VLSI) Circuits **21**(4), 680–691 (2013)
311. S. Kose, E.G. Friedman, Distributed power network co-design with on-chip power supplies and decoupling capacitors, in *Proceedings of the Workshop on System Level Interconnect Prediction*, vol. 13, June 2011
312. S. Kose, E.G. Friedman, Simultaneous co-design of distributed on-chip power supplies and decoupling capacitors, in *Proceedings of the IEEE International SOC Conference*, pp. 15–18, Sept 2010
313. S. Kose, S. Tam, S. Pinzon, B. McDermott, E.G. Friedman, An area efficient on-chip hybrid voltage regulator, in *Proceedings of the IEEE International Symposium on Quality Electronic Design*, pp. 398–403, Mar 2012
314. V. Kursun, S.G. Narendra, V.K. De, E.G. Friedman, Analysis of buck converters for on-chip integration with a dual supply voltage microprocessor. IEEE Trans. Very Large Scale Integr. (VLSI) Syst. **11**(3), 514–522 (2003)
315. V. Kursun, S.G. Narendra, V.K. De, E.G. Friedman, High input voltage step-down DC-DC converters for integration in a low voltage CMOS process, in *Proceedings of the IEEE International Symposium on Quality Electronic Design*, pp. 517–521, Mar 2004
316. M. Al-Shyoukh, H. Lee, R. Perez, A transient-enhanced low-quiescent current low-dropout regulator with buffer impedance attenuation. IEEE J. Solid-State Circuits **42**(8), 1732–1742 (2007)
317. T.Y. Man, K.N. Leung, C.Y. Leung, P.K.T. Mok, M. Chan, Development of single-transistor-control LDO based on flipped voltage follower for SoC. IEEE Trans. Circuits Syst. I: Fundam. Theory Appl. **55**(5), 1392–1401 (2008)

318. J. Guo, K.N. Leung, A 6-μW chip-area-efficient output-capacitorless LDO in 90-nm CMOS technology. IEEE J. Solid-State Circuits **45**(9), 1896–1905 (2010)
319. M. Wens, M.S.J. Steyaert, A fully integrated CMOS 800-mW fourphase semiconstant ON/OFF-time step-down converter. IEEE Trans. Power Electron. **26**(2), 326–333 (2011)
320. H. Nam, Y. Ahn, J. Roh, 5-V buck converter using 3.3-V standard CMOS process with adaptive power transistor driver increasing efficiency and maximum load capacity. IEEE Trans. Power Electron. **27**(1), 463–471, (2012)
321. L. Wang, Y. Pei, X. Yang, Y. Qin, Z. Wang, Improving light and intermediate load efficiencies of buck converters with planar nonlinear inductors and variable on time control. IEEE Trans. Power Electron. **27**(1), 342–353 (2012)
322. W. Yan, W. Li, R. Liu, A noise-shaped buck DC-DC converter with improved light-load efficiency and fast transient response. IEEE Trans. Power Electron. **26**(12), 3908–3924 (2011)
323. H. Jia, J. Lu, X. Wang, K. Padmanabhan, Z.J. Shen, Integration of a monolithic buck converter power IC and bondwire inductors with ferrite epoxy glob cores. IEEE Trans. Power Electron. **26**(6), 1627–1630 (2011)
324. Y.-H. Lee, S.-C. Huang, S.-W. Wang, W.-C. Wu, P.-C. Huang, H.-H. Ho, Y.-T. Lai, K.-H. Chen, Power-tracking embedded buck-boost converter with fast dynamic voltage scaling for the SoC system. IEEE Trans. Power Electron. **27**(3), 1271–1282 (2012)
325. Y. Ahn, H. Nam, J. Roh, A 50-MHz fully integrated low-swing buck converter using packaging inductors. IEEE Trans. Power Electron. **27**(10), 4347–4356 (2012)
326. M. Bathily, B. Allard, F. Hasbani, A 200-MHz integrated buck converter with resonant gate drivers for an RF power amplifier. IEEE Trans. Power Electron. **27**(2), 610–613 (2012)
327. Y. Ramadass, A. Fayed, A. Chandrakasan, A fully-integrated switched-capacitor step-down DC-DC converter with digital capacitance modulation in 45 nm CMOS. IEEE J. Solid-State Circuits **45**(12), 2557–2565 (2010)
328. H.-P. Le, S.R. Sanders, E. Alon, Design techniques for fully integrated switched-capacitor DC-DC converters. IEEE J. Solid-State Circuits **46**(9), 2120–2131 (2011)
329. I. Vaisband, B. Price, S. Kose, Y. Kolla, E.G. Friedman, J. Fischer, Distributed LDO regulators in a 28 nm power delivery system. Analog Integr. Circuits Signal Process. **83**(3), 295–309 (2015)
330. Y.-H. Lee, S.-Y. Peng, C.-C. Chiu, A.C.-H. Wu, K.-H. Chen, Y.-H. Lin, S.-W. Wang, T.-Y. Tsai, C.-C. Huang, C.-C. Lee, A low quiescent current asynchronous digital-LDO with PLL-modulated fast-DVS power management in 40 nm SoC for MIPS performance improvement. IEEE J. Solid-State Circuits **48**(4), 1018–1030 (2013)
331. P. Li, Design and analysis of IC power delivery, in *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pp. 664–666, Nov 2012
332. S. Lai, B. Yan, P. Li, Stability assurance and design optimization of large power delivery networks with multiple on-chip voltage regulators, in *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pp. 247–254, Nov 2012
333. S. Lai, P. Li, A fully on-chip area-efficient CMOS low-dropout regulator with load regulation. Analog Integr. Circuits Signal Process. **72**(2), 925–1030 (2012)
334. S. Lai, B. Yan, P. Li, Localized stability checking and design of IC power delivery with distributed voltage regulators. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **32**(9), 1321–1334 (2013)
335. A.J. D'Souza, R. Singh, J.R. Prabhu, G. Chowdary, A. Seedher, S. Somayajula, N.R. Nalam, L. Cimaz, S. Le Coq, P. Kallam, S. Sundar, S. Cheng, S. Tumati, W. Huang, A fully integrated power-management solution for a 65nm CMOS cellular handset chip, in *Proceedings of the IEEE International Solid-State Circuits Conference*, pp. 382–384, Feb 2011
336. J.F. Bulzacchelli et al., Dual-loop system of distributed microregulators with high DC accuracy, load response time below 500 ps, and 85-mV dropout voltage. IEEE J. Solid-State Circuits **47**(4), 863–874 (2012)
337. F. Lima, A. Geraldes, T. Marques, J.N. Ramalho, P. Casimiro, Embedded CMOS distributed voltage regulator for large core loads, in *Proceedings of the IEEE European Solid-State Circuits Conference*, pp. 521–524, Sept 2003

338. R.W. Erickson, M. Dragan, *Fundamentals of Power Electronics* (Kluwer Academic, Norwell, 2001)
339. C. O'Mathuna, N. Wang, S. Kulkarni, S. Roy, Review of integrated magnetics for power supply on chip (PwrSoC). IEEE Trans. Power Electron. **27**(1), 4799–4816 (2012)
340. V. Vorperian, Simplified analysis of PWM converters using model of PWM switch. II. Discontinuous conduction mode. IEEE Trans. Aerosp. Electron. Syst. **26**(3), 497–505 (1990)
341. V. Vorperian, Simplified analysis of PWM converters using model of PWM switch. Continuous conduction mode. IEEE Trans. Aerosp. Electron. Syst. **26**(3), 490–496 (1990)
342. F. Wei, A. Fayed, A feasibility study of high-frequency buck regulators in nanometer CMOS technologies, in *Proceedings of the IEEE Dallas Workshop on Circuits and Systems*, pp. 1–4, Oct 2009
343. Vishay, Passive components. Available online: http://www.vishay.com
344. Johanson Technology, Integrated passive components. Available online: http://www.johansontechnology.com
345. D. Lu, C.P. Wong, *Materials for Advanced Packaging* (Springer, New York, 2008)
346. *International Technology Roadmap for Semiconductors*. Available online: www.itrs.net/Links/2011ITRS/2011Tables/PIDS_2011Tables.xlsx
347. G. Palumbo, D. Pappalardo, Charge pump circuits: an overview on design strategies and topologies. IEEE Circuits Syst. Mag. **10**(1), 31–45 (2010)
348. C. Jia, H. Chen, M. Liu, C. Zhang, Z. Wang, Integrated power management circuit for piezoelectronic generator in wireless monitoring system of orthopaedic implants. IET Circuits Dev. Syst. **2**(6), 485–494 (2008)
349. C. Hong, L. Ming, H. Wenhan, C. Yi, J. Chen, Z. Chun, W. Zihua, Low-power circuits for the bidirectional wireless monitoring system of the orthopedic implants. IEEE J. Biomed. Circuits Syst. **3**(6), pp. 437–443 (2009)
350. A. Cabrini, A. Fantini, G. Torell, High-efficiency regulated charge pump for non-volatile memories, in *Proceedings of the IEEE International Conference on Electronics, Circuits and Systems*, pp. 720–723, Dec 2006
351. Q. Fan, X. Fu, P. Niu, G. Yang, T. Gao, A novel low voltage and high speed CMOS charge pump circuit, in *Proceeding of the IEEE International Conference on Signal Processing Systems*, pp. V3–389–V3–391, July 2010
352. I. Vaisband, M. Saadat, B. Murmann, A closed-loop reconfigurable switched-capacitor DC-DC converter for sub-mW energy harvesting applications. IEEE Trans. Circuits Syst. I: Regul. Pap. **62**(2), 385–394 (2015)
353. A. Shapiro, E.G. Friedman, Power efficient level shifter for 16 nm FinFET near threshold circuits. IEEE Trans. Very Large Scale Integr. (VLSI) Circuits **24**(2), 774–778 (2015)
354. A. Shapiro, E.G. Friedman, MOS Current Mode Logic Near Threshold Circuits. J. Low Power Electron. Appl. **4**(2), 138–152 (2014)
355. A. Shapiro, E.G. Friedman, Performance characteristics of 14 nm near threshold MCML circuits, in *Proceedings of the IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference*, pp. 79–80, Oct 2013
356. S. Bhunia, S. Mukhopadhyay, *Low-Power Variation-Tolerant Design in Nanometer Silicon* (Springer, New York, 2011)
357. R.G. Dreslinski, M. Wieckowski, D. Blaauw, D. Sylvester, T. Mudge, Near-threshold computing: reclaiming Moores law through energy efficient integrated circuits. Proc. IEEE **98**(2), 253–266 (2010)
358. Y. Okuma, K. Ishida, Y. Ryu, X. Zhang, P.-H. Chen, K. Watanabe, M. Takamiya, T. Sakurai, 0.5-V input digital LDO with 98.7% current efficiency and 2.7-μA quiescent current in 65nm CMOS, in *Proceedings of the IEEE Custom Integrated Circuits Conference*, pp. 1–4, Sept 2010

359. K. Hirairi, Y. Okuma, H. Fuketa, T. Yasufuku, M. Takamiya, M. Nomura, H. Shinohara, T. Sakurai, 13% power reduction in 16b integer unit in 40nm CMOS by adaptive power supply voltage control with parity-based error prediction and detection (PEPD) and fully integrated digital LDO, in *Proceedings of the IEEE International Solid-State Circuits Conference*, pp. 486–488, Feb 2012
360. M. Onouchi, K. Otsuga, Y. Igarashi, T. Ikeya, S. Morita, K. Ishibashi, K. Yanagisawa, A 1.39-V input fast-transient-response digital LDO composed of low-voltage MOS transistors in 40-nm CMOS process, in *Proceedings of the IEEE Asian Solid-State Circuits Conference*, pp. 37–40, Nov 2011
361. A. Raychowdhury, D. Somasekhar, J. Tschanz, V. De, A fully-digital phase-locked low dropout regulator in 32nm CMOS, in *Proceedings of the IEEE Symposium on VLSI Circuits*, pp. 115–116, June 2012
362. J. Gjanci, M.H. Chowdhury, A hybrid scheme for on-chip voltage regulation in system-on-a-chip (SOC). IEEE Trans. Very Large Scale Integr. (VLSI) Circuits **19**(11), 1949–1959 (2011)
363. G.A. Rincon-Mora, P.E. Allen, Optimized frequency-shaping circuit topologies for LDOs. IEEE Trans. Circuits Syst. II: Analog Digit. Signal Process. **45**(6), 703–708 (1998)
364. K.N. Leung, P. K.T. Mok, A Capacitor-free CMOS low-dropout regulator with damping-factor-control frequency compensation. IEEE J. Solid-State Circuits **38**(10), 1691–1702 (2003)
365. Y.-H. Lam, W.-H. Ki, A 0.9 V 0.35 μm adaptively biased CMOS LDO regulator with fast transient response, in *Proceedings of the IEEE International Solid-State Circuits Conference*, pp. 442–626, Feb 2008
366. P.Y. Or, K.N. Leung, An output-capacitorless low-dropout regulator with direct voltage-spike detection. IEEE J. Solid-State Circuits **45**(2), 458–466 (2010)
367. C.-H. Wu, L.-R. Chang-Chien, L.-Y. Chiou, Active filter based on-chip step-down DC-DC switching voltage regulator, in *Proceedings of the IEEE TENCON Conference*, pp. 1–6, Nov 2005
368. P.R. Sallen, E.L. Key, A practical method for designing RC active filter. IRE Trans. Circuit Theory **CT-2**, 74–85 (1955)
369. G. Daryanani, *Principles of Active Network Synthesis and Design* (Wiley, New York, 1976)
370. D.A. Johns, K. Martin, *Analog Integrated Circuit Design* (Wiley, New York, 1997)
371. Y. Ramadass, A. Fayed, B. Haroun, A. Chandrakasan, A 0.16mm^2 completely on-chip switched-capacitor DC-DC converter using digital capacitance modulation for LDO replacement in 45nm CMOS, in *Proceedings of the IEEE International Solid-State Circuits Conference*, pp. 208–209, Feb 2010
372. H.-P. Le et al., A 32 nm fully integrated reconfigurable switched-capacitor DC-DC converter delivering 0.55W/mm^2 at 81% efficiency, in *Proceedings of the IEEE International Solid-State Circuits Conference*, pp. 210–211, Feb 2010
373. G.W. den Besten, B. Nauta, Embedded 5 V-to-3.3 V voltage regulator for supplying digital IC's in 3.3 V CMOS technology. IEEE J. Solid-State Circuits **33**(7), 956–962 (1998)
374. K. Onizuka, K. Inagaki, H. Kawaguchi, M. Takamiya, T. Sakurai, Stacked-chip implementation of on-chip buck converter for distributed power supply system in SiPs. IEEE J. Solid-State Circuits **42**(11), 2404–2410, Nov. (2007)
375. T.Y. Man, P.K.T. Mok, M. Chan, A high slew-rate push-pull output amplifier for low-quiescent current low-dropout regulators with transient-response improvement. IEEE Trans. Circuits Syst. II: Express Briefs **54**(9), 755–759 (2007)
376. U.Y. Ogras, R. Marculescu, D. Marculescu, E.G. Jung, Design and management of voltage-frequency island partitioned networks-on-chip. IEEE Trans. Very Large Scale Integr. (VLSI) Circuits **17**(3), 330–341 (2009)
377. Q. Zhou, J. Shi, B. Liu, Y. Cai, Floorplanning considering IR drop in multiple supply voltages island designs. IEEE Trans. Very Large Scale Integr. (VLSI) Circuits **19**(4), 638–646 (2011)
378. E. Fayneh, E. Knoll, On-Chip Filter-Regulator, Such as One for a Microprocessor Phase Locked Loop (PLL) Supply, U.S. Patent 6,661,213, Dec 2003

379. M.G. Degrauwe, J. Rijmenants, E.A. Vittoz, H.J. de Man, Adaptive biasing CMOS amplifiers. IEEE J. Solid-State Circuits **17**(3), 522–528 (1982)
380. W.-J. Huang, S.-I. Liu, Capacitor-free low dropout regulators using nested miller compensation with active resistor and 1-bit programmable capacitor array. IET Electron. Lett. **2**(3), 306–316 (2008)
381. C.K. Chava, J. Silva-Martinez, A frequency compensation scheme for LDO voltage regulators. IEEE Trans. Circuits Syst. I: Regul. Pap. **51**(6), 1041–1050 (2004)
382. X. Fan, C. Mishra, E.S.-Sinencio, Single miller capacitor frequency compensation technique for low-power multistage amplifiers. IEEE J. Solid-State Circuits **40**(3), 584–592 (2005)
383. R.J. Milliken, J.S.-Martinez, E.S.-Sinencio, Full on-chip CMOS low-dropout voltage regulator. IEEE Trans. Circuits Syst. I: Regul. Pap. **54**(9), 1879–1890 (2007)
384. K.N. Leung, Y.S. Ng, K.Y. Yim, P.Y. Or, An adaptive current-boosting voltage buffer for low-power low dropout regulators, in *Proceeding of the IEEE Conference on Electron Devices and Solid-State Circuits*, pp. 485–488, Dec 2007
385. M. El-Nozahi, A. Amer, J. Torres, K. Entesari, E. Sanchez Sinencio, High PSR low drop-out regulator with feedforward ripple cancellation technique. IEEE J. Solid-State Circuits **45**(3), 565–577 (2010)
386. M. Ho, K.N. Leung, K.-L. Mac, A low-power fast-transient 90-nm low-dropout regulator with multiple small-gain stages. IEEE J. Solid-State Circuits **45**(11), 2466–2475 (2010)
387. G.A. Rincon-Mora P.E. Allen, A low-voltage, low quiescent current, low drop-out regulator. IEEE J. Solid-State Circuits **33**(1), 36–44 (1998)
388. T. Hattori et al., A power management scheme controlling 20 power domains for a single-chip mobile processor, in *Proceedings of the IEEE International Solid-State Circuits Conference*, pp. 542–543, Feb 2006
389. T. Dhaene, D.D. Zutter, Selection of lumped element models for coupled lossy transmission lines. IEEE Trans. Computer-Aided Des. Integr. Circuits Syst. **1**(7), 805–815 (1992)
390. Z. Toprak-Deniz et al., Distributed system of digitally controlled microregulators enabling per-core DVFS for the POWER8™microprocessor, in *Proceedings of the IEEE International Solid-State Circuits Conference*, pp. 98–99, Feb 2014
391. T. Coulot et al., Stability analysis and design procedure of multiloop linear LDO regulators via state matrix decomposition. IEEE Trans. Power Electron. **28**(11), 5352–5363 (2013)
392. S. Bin Nasir, Y. Lee, A. Raychowdhury, Modeling and analysis of system stability in a distributed power delivery network with embedded digital linear regulators, in *Proceedings of the IEEE International Symposium on Quality Electronic Design*, pp. 68–75, Mar 2014
393. B. Razavi, *Design of Analog CMOS Integrated Circuits* (McGraw-Hill, New York, 2001)
394. B.H. Calhoun, A. Chandrakasan, Ultra-dynamic voltage scaling using sub-threshold operation and local voltage dithering in 90nm CMOS, in *Proceedings of the IEEE International Solid-State Circuits Conference*, pp. 300–599, Feb 2005
395. T.D. Burd, T.A. Pering, A.J. Stratakos, R.W. Brodersen, A dynamic voltage scaled microprocessor system. IEEE J. Solid-State Circuits **35**(11), 1571–1580 (2000)
396. R. Jakushokas, M. Popovich, A.V. Mezhiba, S. Kose, E.G. Friedman, *Power Distribution Networks with On-Chip Decoupling Capacitors*, 2nd edn. (Springer, New York, 2011)
397. S. Kose, I. Vaisband, E.G. Friedman, Digitally controlled wide range pulse width modulator for on-chip power supplies, in *Proceedings of the IEEE International Symposium on Circuit and Systems*, pp. 2251–2254, May 2013
398. S. Docking, M. Sachdev, A method to derive an equation for the oscillation frequency of a ring oscillator. IEEE Trans. Circuits Syst. I: Regul. Pap. **50**(2), 259–264 (2003)
399. W. Kolodziejski, S. Kuta, J. Jasielski, Current controlled delay line elements' improvement study, in *Proceedings of the International Conference on Signals and Electronic Systems*, pp. 1–4, Sept 2012
400. I. Vaisband, M. Azhar, E.G. Friedman, S. Kose, Digitally controlled pulse width modulator for on-chip power management. IEEE Trans. Very Large Scale Integr. (VLSI) Circuits **22**(12), 2527–2534 (2014)

401. A. Djemouai, M. Sawan, M. Slamani, High performance integrated CMOS frequency to voltage converter, in *Proceedings of the International Conference on Microelectronics*, pp. 63–66, Dec 1998
402. A.M. Pappu, X. Zhang, A.V. Harrison, A.B. Apsel, Process-invariant current source design: methodology and examples. IEEE J. Solid-State Circuits **42**(10), 2293–2302 (2007)
403. X. Zhang, A.B. Apsel, A low-power, process and temperature compensated ring oscillator with addition based current source. IEEE Trans. Circuits Syst. I: Regul Pap. **58**(5), 868–878 (2011)
404. M. Anis, M.H. Aburahma, Leakage current variability in nanometer technologies, in *Proceedings of the International Workshop on System-on-Chip for Real-Time Applications*, pp. 60–63, July 2005
405. NIMO Group, Predictive Technology Model (PTM) (Arizona State University, 2008). Available Online: http://www.eas.asu.edu/~ptm
406. Y. Cao, *Predictive Technology Model for Robust Nanoelectronic Design* (Springer, New York, 2011)
407. Y. Kishiwada, S. Ueda, Y. Miyawaki, T. Matusoka, Process variation compensation with effective gate-width tuning for low-voltage cmos digital circuits, in *Proceedings of the IEEE International Meeting for Future of Electron Devices*, pp. 1–2, May 2012
408. M.S. Gupta, J.A. Rivers, P. Bose, G.-Y. Wei, D. Brooks, Tribeca: design for PVT variations with local recovery and fine-grained adaptation, in *Proceedings of the Annual IEEE/ACM International Symposium on Microarchitecture*, pp. 435–446, Dec 2009
409. N.H.E. Weste, K. Eshraghian, *Principles of CMOS VLSI Design* (Addison-Wesley, Boston, 1992)
410. B. Yan, S.X.-D. Tan, G. Chen, L. Wu, Modeling and simulation for on-chip power grid networks by locally dominant Krylov subspace method, in *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pp. 744–749, Nov 2008
411. L. Ren, J. Kim, G. Feng, B. Archambeault, J.L. Knighten, J. Drewniak, J. Fan, Frequency-dependent via inductances for accurate power distribution network modeling, in *Proceedings of the IEEE International Symposium on Electromagnetic Compatibility*, pp. 63–68, Aug 2009
412. R. Jakushokas, E.G. Friedman, Inductance model of interdigitated power and ground distribution networks. IEEE Trans. Circuits Syst.—II: Analog Digit. Signal Process. **56**(7), 585–589 (2009)
413. G. Huang, A. Naeemi, T. Zhou, D. O'Connor, A. Muszynski, B. Singh, D. Becker, J. Venuto, J.D. Meindl, Compact physical models for chip and package power and ground distribution networks for gigascale integration (GSI), in *Proceedings of the Electronic Components and Technology Conference*, pp. 646–651, May 2008
414. H.H. Chen, D.D. Ling, Power supply noise analysis methodology for deep-submicron VLSI chip design, in *Proceedings of the IEEE/ACM Design Automation Conference*, pp. 638–643, June 1997
415. M. Horowitz, R.W. Dutton, Resistance extraction from mask layout data. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **2**(3), 145–150 (1983)
416. L. Ladage, R. Leupers, Resistance extraction using a routing algorithm, in *Proceedings of the IEEE/ACM Design Automation Conference*, pp. 38–42, June 1993
417. E. Macii, M. Pedram, F. Somenzi, High-level power modeling, estimation, and optimization. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **17**(11), 1061–1079 (1998)
418. D. Brooks, T. Tiwari, M. Martonosi, Wattch: a framework for architectural-level analysis and optimization, in *Proceedings of the ACM International Symposium on Computer Architecture*, pp. 83–94, June 2000
419. A. Gstottner, T. Steinecke, M. Huemer, Activity based high level modeling of dynamic switching currents in digital IC modules, in *Proceedings of the International Zurich Symposium on Electromagnetic Compatibility*, pp. 598–601, Feb 2006
420. A. Gstottner, J. Kruppa, M. Huemer, Modeling of dynamic switching currents of digital VLSI IC modules and verification by on-chip measurement, in *Proceedings of the International Zurich Symposium on Electromagnetic Compatibility*, pp. 1–4, Sept 2007

421. F.N. Najm, A survey of power estimation techniques in VLSI circuits. IEEE Trans. Very Large Scale Integr. (VLSI) Circuits **2**(4), 446–455 (1994)
422. H. Kriplani, F.N. Najm, I.N. Hajj, Pattern independent maximum current estimation in power and ground buses of CMOS VLSI circuits: algorithms, signal correlations, and their resolution. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **14**(8), 998–1012 (1995)
423. S. Bobba, I.N. Hajj, Estimation of maximum current envelope for power bus analysis and design, in *Proceedings of the ACM International Symposium on Physical Design*, pp. 141–146, Apr 1998
424. S. Bobba, I.N. Hajj, Maximum voltage variation in the power distribution network of VLSI circuits with *RLC* models, in *Proceedings of the IEEE International Symposium on Low Power Electronics and Design*, pp. 376–381, Aug 2001
425. G. Bai, I.N. Hajj, Simultaneous switching noise and resonance analysis of on-chip power distribution network, in *Proceedings of the IEEE International Symposium on Quality Electronic Design*, pp. 163–168, Mar 2002
426. L.T. Pillage, R.A. Rohrer, C. Visweswariah, *Electronic Circuit and System Simulation Methods* (McGraw-Hill, New York, 1994)
427. G. Golub, C. Van Loan, *Matrix Computations* (Johns Hopkins University Press, Baltimore, 1989)
428. H. Li, J. Jain, V. Balakrishnan, C-K. Koh, Efficient analysis of large-scale power grids based on a compact Cholesky factorization, in *Proceedings of the IEEE International Symposium on Quality Electronic Design*, pp. 627–632, Mar 2007
429. J.M.S. Silva, J.R. Phillips, L.M. Silveira, Efficient representation and analysis of power grids, in *Proceedings of the IEEE/ACM Design Automation and Test in Europe Conference*, pp. 420–425, Mar 2008
430. Y. Zhong, M.D.F. Wong, Efficient second-order iterative methods for IR drop analysis in power grid, in *Proceedings of the IEEE/ACM Asia and South Pacific Design Automation Conference*, pp. 768–773, Jan 2007
431. Y. Zhong, M.D.F. Wong, Fast block-iterative domain decomposition algorithm for *IR* drop analysis in large power grid, in *Proceedings of the IEEE International Symposium on Quality Electronic Design*, pp. 277–283, Mar 2010
432. L.-R. Zheng, H. Tenhunen, Design and analysis of power integrity in deep submicron system-on-chip circuits. Analog Integr. Circuits Signal Process. **30**(1), 15–29 (2002)
433. L.-R. Zheng, Design, Analysis, and Integration of Mixed-Signal Systems for Signal and Power Integrity, Ph.D. Thesis, Royal Institute of Technology, Stockholm, 2001
434. M. Zhao, R. Panda, S.S. Sapatnekar, D. Blaauw, Hierarchical analysis of power distribution networks. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **21**(2), 159–168 (2002)
435. Z. Zeng, P. Li, Locality-driven parallel power grid optimization. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **28**(8), 1190–1200 (2009)
436. Z. Zeng, P. Li, Z. Feng, Parallel partitioning based on-chip power distribution network analysis using locality acceleration, in *Proceedings of the IEEE International Symposium on Quality Electronic Design*, pp. 776–781, Mar 2009
437. H. Qian, S.R. Nassif, S.S. Sapatnekar, Power grid analysis using random walks. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **24**(8), 1204–1224 (2005)
438. B. Boghrati, S. Sapatnekar, Incremental solution of power grids using random walks, in *Proceedings of the IEEE/ACM Asia and South Pacific Design Automation Conference*, pp. 757–762, Jan 2010
439. P. Feldmann, R.W. Freund, E. Acar, Power Grid Analysis Using a Flexible Conjugate Gradient Algorithm with Sparsification, Technical Report, Department of Mathematics, University of California, Davis, June 2006
440. S.R. Nassif, J.N. Kozhaya, Multi-grid methods for power grid simulation, in *Proceedings of the IEEE International Symposium on Circuit and Systems*, vol. V, pp. 457–460, May 2000
441. S.R. Nassif, J.N. Kozhaya, Fast power grid simulation, in *Proceedings of the IEEE/ACM Design Automation Conference*, pp. 156–161, June 2000

442. J.N. Kozhaya, S.R. Nassif, F.N. Najm, A multigrid-like technique for power grid analysis. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **21**(10), 1148–1160 (2002)
443. C. Zhuo, J. Hu, M. Zhao, K. Chen, Power grid analysis and optimization using algebraic multigrid. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **27**(4), 738–751 (2008)
444. H. Su, K. Gala, S.S. Sapatnekar, Fast analysis and optimization of power/ground networks, in *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pp. 477–480, Nov 2000
445. A. Odabasioglu, M. Celik, L.T. Pileggi, PRIMA: passive reduced-order interconnect macromodeling algorithm. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **17**(8), 645–654 (1998)
446. S. Zhao, K. Roy, C.-K. Koh, Estimation of inductive and resistive switching noise on power supply network in deep submicron CMOS circuits, in *Proceedings of the IEEE International Conference on Computer Design*, pp. 65–72, Oct 2000
447. S. Zhao, K. Roy, C.-K. Koh, Frequency domain analysis of switching noise on power supply network, in *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pp. 487–492, Nov 2000
448. M.D. Pant, P. Pant, D.S. Wills, On-chip decoupling capacitor optimization using architectural level current signature prediction, in *Proceedings of the IEEE International ASIC/SOC Conference*, pp. 288–292, Sept 2000
449. M.D. Pant, P. Pant, D.S. Wills, On-chip decoupling capacitor optimization using architectural level prediction, in *Proceedings of the IEEE Midwest Symposium on Circuit and Systems*, pp. 772–775, Aug 2000
450. S. Zhao, K. Roy, C.-K. Koh, Decoupling capacitance allocation for power supply noise suppression, in *Proceedings of the ACM International Symposium on Physical Design*, pp. 66–71, Apr 2001
451. S. Zhao, K. Roy, C.-K. Koh, Power supply noise aware floorplanning and decoupling capacitance placement, in *Proceedings of the IEEE International Conference on VLSI Design*, pp. 489–495, Jan 2002
452. A.R. Conn, R.A. Haring, C. Visweswariah, Noise considerations in circuit optimization, in *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pp. 220–227, Nov 1998
453. C. Viswesvariah, R.A. Haring, A.R. Conn, Noise considerations in circuit optimization. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **19**(6), 679–690 (2000)
454. H. Su, S.S. Sapatnekar, S.R. Nassif, An algorithm for optimal decoupling capacitor sizing and placement for standard cell layouts, in *Proceedings of the ACM International Symposium on Physical Design*, pp. 68–73, Apr 2002
455. L. Zlydina, Y. Yagil, 3D power grid modeling, in *Proceedings of the IEEE International Conference on Electronics, Circuits and Systems*, pp. 129–132, Dec 2004
456. J. Singh, S.S. Sapatnekar, Partition-based algorithm for power grid design using locality. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **25**(4), 664–677 (2006)
457. K. Lee, A. Barber, Modeling and analysis of multichip module power supply planes. IEEE Trans. Compon. Packag. Manuf. Technol. Pt. B: Adv. Packag. **18**(4), 628–639 (1995)
458. S. Kose, E.G. Friedman, Fast algorithms for IR voltage drop analysis exploiting locality, in *Proceedings of the IEEE/ACM Design Automation Conference*, pp. 996–1001, June 2011
459. S. Kose, E.G Friedman, Efficient algorithms for fast IR drop analysis exploiting locality. Integr. VLSI J. **45**(2), 149–161 (2012).
460. Y. Ogasahara, M. Hashimoto, T. Kanamoto, T. Onoye, Measurement of supply noise suppression by substrate and deep N-well in 90nm process, in *Proceedings of the IEEE Asian Solid-State Circuits Conference*, pp. 397–400, Nov 2008
461. E. Wong, J.R. Minz, S.K. Lim, Decoupling-capacitor planning and sizing for noise and leakage reduction. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **26**(11), 2023–2034 (2007)
462. J. Rommes, W.H.A. Schilders, Efficient methods for large resistor networks. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **29**(1), 28–39 (2010)

463. R. Helinski, J. Plusquellic, Measuring power distribution system resistance variations. IEEE Trans. Semicond. Manuf. **21**(3), 444–453 (2008)
464. A.K. Chandra, P. Raghavan, W.L. Ruzzo, R. Smolensky, The electrical resistance of a graph captures its commute and cover times, in *Proceedings of the Annual ACM Symposium on Theory of Computing*, pp. 574–586, May 1989
465. D.J. Klein, M. Randić, Resistance distance. J. Math. Chem. **12**, 81–95 (1993)
466. P. Barooah, J.P. Hespanha, Graph effective resistance and distributed control: spectral properties and applications, in *Proceedings of the IEEE Conference on Decision and Control*, pp. 3479–3485, Dec 2006
467. C.R. Paul, *Analysis of Linear Circuits* (McGraw-Hill, New York, 1989)
468. M. Abramowitz, I.A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables* (Dover, Mineola, 1972)
469. E. Chiprout, Fast flip-chip power grid analysis via locality and grid shells, in *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pp. 485–488, Nov 2004
470. P. Gupta, A.B. Kahng, Efficient design and analysis of robust power distribution meshes, in *Proceedings of the IEEE International Conference on VLSI Design*, pp. 337–342, Jan 2006
471. K. Shakeri, J.D. Meindl, Compact physical IR-drop models for chip/package co-design of gigascale integration (GSI). IEEE Trans. Electron Devices **52**(6), 1087–1096 (2005)
472. S. Kose, E.G. Friedman, Fast algorithms for power grid analysis based on effective resistance, in *Proceedings of the IEEE International Symposium on Circuits and Systems*, pp. 3661–3664, May 2010
473. H. Qian, S.S. Sapatnekar, Hierarchical random-walk algorithms for power grid analysis, in *Proceedings of the IEEE/ACM Asia and South Pacific Design Automation Conference*, pp. 499–504, Jan 2004
474. H. Qian, S.R. Nassif, S.S. Sapatnekar, Random walks in a supply network, in *Proceedings of the IEEE/ACM Design Automation Conference*, pp. 93–98, June 2003
475. E. Alon, M. Horowitz, Integrated regulation for energy-efficient digital circuits. IEEE J. Solid-State Circuits **43**(8), 1795–1807 (2008)
476. Z. Zeng, X. Ye, Z. Feng, P. Li, Tradeoff analysis and optimization of power delivery networks with on-chip voltage regulation, in *Proceedings of the IEEE/ACM Design Automation Conference*, pp. 831–836, June 2010
477. J.D. van Wyk, F.C. Lee, On a future of power electronics. IEEE J. Emerg. Sel. Top. Power Electron. **1**(2), 59–72 (2013)
478. R.A. Rutenbar, G.G.E. Gielen, J. Roychowdhury, Hierarchical modeling, optimization, and synthesis for system-level analog and RF designs. Proc. IEEE **95**(3), 640–669 (2007)
479. S. Bin Nasir, S. Gangopadhyay, A. Raychowdhury, 5.6 A 0.13μm fully digital low-dropout regulator with adaptive control and reduced dynamic stability for ultra-wide dynamic range, in *Proceedings of the IEEE International Solid-State Circuits Conference*, pp. 1–3, Feb 2015
480. A. Maity, A. Patra, Design and analysis of an adaptively biased low drop-out regulator using enhanced current mirror buffer. IEEE Trans. Power Electron. **31**(3), 2324–2336 (2015)
481. Y.-C. Chu, L.-R. Chang-Chien, Digitally controlled low-dropout regulator with fast-transient and autotuning algorithms. IEEE Trans. Power Electron. **28**(9), 4308–4317 (2013)
482. C.-H. Wu, L.-R. Chang-Chien, Design of the output-capacitorless low-dropout regulator for nano-second transient response. IET Power Electron. **5**(8), 1551–1559 (2012)
483. S. Lim, A.Q. Huang, Low-dropout (LDO) regulator output impedance analysis and transient performance enhancement circuit, in *Proceedings of the IEEE International Applied Power Electronics Conference and Exposition*, pp. 1875–1878, Feb 2010
484. C.-Y. Hsieh, C.-Y. Yang, K.-H. Chen, A low-dropout regulator with smooth peak current control topology for overcurrent protection. IEEE Trans. Power Electron. **25**(6), 1386–1394 (2010)
485. C.-H. Lin, K.-H. Chen, H.-W. Huang, Low-dropout regulators with adaptive reference control and dynamic push-pull techniques for enhancing transient performance. IEEE Trans. Power Electron. **24**(4), 1016–1022 (2009)

486. I. Vaisband, E.G. Friedman, Stability of distributed power delivery systems with multiple parallel on-chip LDO regulators. IEEE Trans. Power Electron. **31**(8), 5626–5634 (2015)
487. J.H. Mulligan Jr., The effect of pole and zero locations on the transient response of linear dynamic systems. Proc. Inst. Radio Eng. **37**(5), 516–529 (1949)
488. J.J. Kelly, M.S. Ghausi, J.H. Mulligan Jr., On the analysis of composite lumped distributed systems. Elsevier J. Solid-State Electron. **284**(3), 170–192 (1967)
489. A. Riccobono, E. Santi, A novel passivity-based stability criterion (PBSC) for switching converter DC distribution systems, in *Proceedings of the IEEE International Applied Power Electronics Conference and Exposition*, pp. 2560–2567, Feb 2012
490. J.C. West, J. Potts, A simple connection between closed-loop transient response and open-loop frequency response. Proc. IEE – Pt. II: Power Eng. **100**(75), 201–212 (1953)
491. J. Wagner, G. Stolovitzky, Stability and time-delay modeling of negative feedback loops. Proc. IEEE **96**(8), 1398–1410 (2008)
492. J.E. Colgate, The Control of Dynamically Interacting Systems, Ph.D. Thesis, Massachusetts Institute of Technology, Aug 1988
493. J.L. Wyatt, L.O. Chua Jr., J. Gannett, I. Goknar, D. Green, Energy concepts in the state-space theory of nonlinear n-ports: part I-passivity. IEEE Trans. Circuits Syst. **28**(1), 48–61 (1981)
494. O. Brune, Synthesis of a Finite Two-Terminal Network Whose Driving-Point Impedance Is a Prescribed Function of Frequency, Ph.D. Thesis, Massachusetts Institute of Technology, Aug 1931
495. N. Viswanathan et al., The ISPD-2011 routability-driven placement contest and benchmark suite, in *Proceedings of the ACM International Symposium on Physical Design*, pp. 141–146, Mar 2011
496. A. Brooke, D. Kendrick, A. Meeraus, *GAMS: A User's Guide* (The Scientific Press, Redwood, 1992)
497. G.A. Rincon-Mora, Current Efficient, Low Voltage, Low Dropout Regulators, Ph.D. Thesis, Georgia Institute of Technology, Nov 1996
498. C. Tirumurti, S. Kundu, S. Sur-Kolay, Y.-S. Chang, A modeling approach for addressing power supply switching noise related failures of integrated circuits, in *Proceedings of the IEEE Design, Automation, and Test Conference in Europe*, vol. 2, pp. 1078–1083, Feb 2004
499. L.H. Chen, M. Marek-Sadowska, F. Brewer, Buffer delay change in the presence of power and ground noise. IEEE Trans. Very Large Scale Integr. (VLSI) Circuits **11**(3), 461–473 (2003)
500. X. Lai, J. Roychowdhury, Fast, accurate prediction of PLL jitter induced by power grid noise, in *Proceedings of the IEEE Custom Integrated Circuits Conference*, pp. 121–124 (2004)
501. P. Heydari, Analysis of the PLL jitter due to power/ground and substrate noise. IEEE Trans. Circ. Syst. I: Regul. Pap. **51**(12), 2404–2416 (2004)
502. P. Larsson, Measurements and analysis of PLL jitter caused by digital switching noise. IEEE J. Solid-State Circuits **36**(7), 1113–1119 (2001)
503. M. Alioto, G. Palumbo, Impact of supply voltage variations on full adder delay: analysis and comparison. IEEE Trans. Very Large Scale Integr. (VLSI) Circuits **14**(12), 1322–1335 (2006)
504. P.G. Doyle, J.L. Snell, *Random Walks and Electrical Networks* (Mathematical Association of America, Washington, DC, 1984)
505. J.N. Kozhaya, S.R. Nassif, F.N. Najm, Multigrid-like technique for power grid analysis, in *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pp. 480–487, Nov 2001
506. M.K. Tavana, M.H. Hajkazemi, D. Pathak, I. Savidis, H. Homayoun, ElasticCore: enabling dynamic heterogeneity with joint core and voltage/frequency scaling, in *Proceedings of the IEEE/ACM Design Automation Conference*, pp. 1–6, June 2015
507. B. Amelifard, M. Pedram, Optimal design of the power-delivery network for multiple voltage-island system-on-chips. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **28**(6), 888–900 (2009)
508. I. Vaisband E.G. Friedman, Energy efficient adaptive clustering of on-chip power delivery systems. Integr. VLSI J. **48**, 1–9 (2015)

509. I. Vaisband, E.G. Friedman, Computationally efficient clustering of power supplies in heterogeneous real time systems, in *Proceedings of the IEEE International Symposium on Circuit and Systems*, pp. 1628–1631, June 2014
510. Y. Kim, P. Li, An ultra-low voltage digitally controlled low-dropout regulator with digital background calibration, in *Proceedings of the IEEE International Symposium on Quality Electronic Design*, pp. 151–158, Mar 2012
511. Y. Xiong, S. Sun, H. Jia, P. Shea, Z.J. Shen, New physical insights on power MOSFET switching losses. IEEE Trans. Power Electron. **24**(2), 525–531 (2009)
512. C.P. Robert, G. Casella, *Monte Carlo Statistical Methods* (Springer, New York, 1999)
513. S. Venkatesan et al., A high performance 1.8 V 0.20 μm CMOS technology with copper metalization, in *Proceedings of the IEEE International Electron Devices Meeting*, pp. 769–772, Dec 1997
514. *International Technology Roadmap for Semiconductors*, 2000 Update (Semiconductor Industry Association, 2000). http://public.itrs.net
515. A.V. Mezhiba, E.G. Friedman, Variation of inductance with frequency in high performance power distribution grids, in *Proceedings of the IEEE International ASIC/SOC Conference*, pp. 421–425, Sept 2002
516. A.V. Mezhiba, E.G. Friedman, Frequency characteristics of high speed power distribution grids. Analog Integr. Circuits Signal Process. **35**(2/3), 207–214 (2003)
517. W.K. Henson, N. Yang, S. Kubicek, E.M. Vogel, J.J. Wortman, K. De Meyer, A. Naem, Analysis of leakage currents and impact on off-state power consumption for CMOS technology in the 100-nm regime. IEEE Trans. Electron Devices **47**(7), 1393–1400 (2000)
518. Y. Taur, CMOS design near the limit of scaling. IBM J. Res. Dev. **46**(2/3), 213–221 (2002)
519. M. Powell, S.H. Yang, B. Falsafi, K. Roy, T.N. Vijaykumar, Gated-Vdd: a circuit technique to reduce leakage in deep-submicron cache memories, in *Proceedings of the IEEE International Symposium on Low Power Electronics and Design*, pp. 90–95, Jan 2000
520. Z. Hu, A. Buyuktosunoglu, V. Srinivasan, V. Zyuban, H. Jacobson, P. Bose, Microarchitectural techniques for power gating of execution units, in *Proceedings of the IEEE International Symposium on Low Power Electronics and Design*, pp. 32–37, Aug 2004
521. S. Mutoh, T. Douseki, Y. Matsuya, T. Aoki, S. Shigematsu, J. Yamada, 1-V power supply high-speed digital circuit technology with multithreshold-voltage CMOS. IEEE J. Solid-State Circuits **30**(8), 847–854 (1995)
522. J. Rabaey, M. Pedram, *Low Power Design Methodologies* (Kluwer Academic, Norwell, 1996)
523. N.S. Kim, T. Austin, D. Baauw, T. Mudge, K. Flautner, J.S. Hu, M.J. Irwin, M. Kandemir, V. Narayanan, Leakage current: Moore's law meets static power. IEEE Trans. Comput. **36**(12), 68–75 (2003)
524. K. Usami, T. Shirai, T. Hashida, H. Masuda, S. Takeda, M. Nakata, N. Seki, H. Amano, M. Namiki, M. Imai, M. Kondo, H. Nakamura, Design and implementation of fine-grain power gating with ground bounce suppression, in *Proceedings of the IEEE International Conference on VLSI Design*, pp. 381–386, Jan 2009
525. S. Kim, S.V. Kosonocky, D.R. Knebel, K. Stawiasz, M.C. Papaefthymiou, A multi-mode power gating structure for low-voltage deep-submicron CMOS ICs. IEEE Trans. Circuits Syst. II: Express Briefs **54**(7), 586–590 (2007)
526. A. Valentian, E. Beigne, Automatic gate biasing of an SCCMOS power switch achieving maximum leakage reduction and lowering leakage current variability. IEEE J. Solid-State Circuits **43**(7), 1688–1698, (2008)
527. H. Tabkhi, G. Schirner, Application-guided power gating reducing register file static power. IEEE Trans. Very Large Scale Integr. (VLSI) Circuits **22**(12), 2513–2526 (2014)
528. H. Xu, R. Vemuri, W.-B. Jone, Dynamic characteristics of power gating during mode transition. IEEE Trans. Very Large Scale Integr. (VLSI) Circuits **19**(2), 237–249 (2011)
529. M.B. Henry, L. Nazhandali, NEMS-based functional unit power-gating: design, analysis, and optimization. IEEE Trans. Circ. Syst. I: Regul. Pap. **60**(2), 290–302 (2013)

530. A. Ramalingam, B. Zhang, D.Z. Pan, A. Devgan, Sleep transistor sizing using timing criticality and temporal currents, in *Proceedings of the IEEE/ACM Asia and South Pacific Design Automation Conference*, vol. 2, pp. 1094–1097, Jan 2005
531. M. Kazemi, E. Ipek, E. Friedman, Energy efficient nonvolatile flip flop with subnanosecond data backup time for fine grain power gating. IEEE Trans. Circuits Syst. II: Express Briefs **62**(12), 1154–1158 (2015)
532. R. Senthinathan, J.L. Prince, Simultaneous switching ground noise calculation for packaged CMOS devices. IEEE J. Solid-State Circuits **26**(11), 1724–1728 (1991)
533. A. Vaidyanath, B. Thoroddsen, J.L. Prince, Effect of CMOS driver loading conditions on simultaneous switching noise. IEEE Trans. Compon. Packag. Manuf. Technol. Pt. B: Adv. Packag. **17**(4), 480–485 (1994)
534. S.R. Vemuru, Accurate simultaneous switching noise estimation including velocity-saturation effects. IEEE Trans. Compon. Packag. Manuf. Technol. Pt. B: Adv. Packag. **19**(2), 344–349 (1996)
535. S.-J. Jou, W.-C. Cheng, Y.-T. Lin, Simultaneous switching noise analysis and low-bounce buffer design. IEE Proc. Circuits Devices Syst. **148**(6), 303–311 (2001)
536. H.-R. Cha, O.-K. Kwon, A new analytic model of simultaneous switching noise in CMOS systems, in *Proceedings of the IEEE Electronic Components and Technology Conference*, pp. 615–621, May 1998
537. S.R. Vemuru, Effects of simultaneous switching noise on the tapered buffer design. IEEE Trans. Very Large Scale Integr. (VLSI) Syst. **5**(3), 290–300 (1997)
538. P. Heydari, M. Pedram, Ground bounce in digital VLSI circuits. IEEE Trans. Very Large Scale Integr. (VLSI) Syst. **11**(2), 180–193 (2003)
539. T.J. Gabara, Ground bounce and reduction techniques, in *Proceedings of the IEEE ASIC Conference*, pp. T13–2/1–2, Sept 1991
540. A. Vittal, H. Ha, F. Brewer, M. Marek-Sadowska, Clock skew optimization for ground bounce control, in *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pp. 395–399, Nov 1996
541. M.D. Pant, P. Pant, D.S. Wills, V. Tiwari, An architectural solution for the inductive noise problem due to clock-gating, in *Proceedings of the IEEE International Symposium on Low Power Electronics and Design*, pp. 255–257, Aug 1999
542. J. Oh, M. Pedram, Multi-pad power/ground network design for uniform distribution of ground bounce, in *Proceedings of the ACM/IEEE Design Automation Conference*, pp. 287–290, June 1998
543. H.H. Chen, Minimizing chip-level simultaneous switching noise for high-performance microprocessor design, in *Proceedings of the IEEE International Symposium on Circuit and Systems*, vol. IV, pp. 544–547, May 1996
544. A. Zenteno, V.H. Champac, M. Renovell, F. Azais, Analysis and attenuation proposal in ground bounce, in *Proceedings of the IEEE Asian Test Symposium*, pp. 460–463, Nov 2004
545. M. Badaroglu, P. Wambacq, G. Van der Plas, S. Donnay, G.G.E. Gielen, H.J. De Man, Digital ground bounce reduction by supply current shaping and clock frequency modulation. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **24**(1), pp. –76 (2005)
546. I. Catt, Crosstalk (noise) in digital systems. IEEE Trans. Electron. Comput. **16**(6), 743–763 (1967)
547. A. Vittal, M. Sadowska, Crosstalk reduction for VLSI. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **16**(3), 290–298 (1997)
548. J. Zhang, E.G. Friedman, Crosstalk modeling for coupled *RLC* interconnects with application to shield insertion. IEEE Trans. Very Large Scale Integr. (VLSI) Circuits **14**(6), 641–646 (2006)
549. J. Zhang, E.G. Friedman, Effects of shield insertion on reducing crosstalk noise between coupled interconnects, in *Proceedings of the IEEE International Symposium on Circuit and Systems*, vol. 2, pp. 529–532, May 2004

550. R. Arunachalam, E. Acar, S.R. Nassif, Optimal shielding/spacing metrics for low power design, in *Proceedings of the IEEE computer society annual symposium on VLSI*, pp. 167–172, Feb 2003
551. M.R. Becer, D. Blaauw, V. Zolotov, R. Panda, I.N. Hajj, Analysis of noise avoidance techniques in DSM interconnects using a complete crosstalk noise model, in *Proceedings of the IEEE Design, Automation, and Test Conference in Europe*, pp. 456–463, Mar 2002
552. J. Zhang, E.G. Friedman, Mutual inductance modeling for multiple *RLC* interconnects with application to shield insertion, in *Proceedings of the IEEE International SOC Conference*, pp. 344–347, Sept 2004
553. A. Roy, J. Xu, M.H. Chowdhury, Analysis of the impacts of signal slew and skew on the behavior of coupled *RLC* interconnects for different switching patterns. IEEE Trans. Very Large Scale Integr. (VLSI) Circuits **18**(2), 338–342 (2010)
554. H. Kaul, D. Sylvester, D. Blaauw, Active shields: a new approach to shielding global wires, in *Proceedings of the ACM/IEEE Great Lakes Symposium on VLSI*, pp. 112–117 (2002)
555. H. Kaul, D. Sylvester, D. Blaauw, Active shielding of RLC global interconnect, in *Proceedings of the ACM/IEEE International Workshop on Timing Issues in the Specification and Synthesis of Digital Systems*, pp. 98–104, Dec 2002
556. M. Ghoneima, Y. Ismail, Formal derivation of optimal active shielding for low-power on-chip buses. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **5**(5), 821–836 (2006)
557. A. Kabbani, A.J. Al-Khalili, Estimation of Ground Bounce Effects on CMOS Circuits. IEEE Trans. Compon. Packag. Technol. **22**(2), 316–325 (1999)
558. P. Larsson, di/dt noise in CMOS integrated circuits. Analog Integr. Circuits Signal Process. **14**(2), 113–129 (1997)
559. X. Huang, Y. Cao, D. Sylvester, S. Lin, T.-J. King, C. Hu, RLC signal integrity analysis of high-speed global interconnects, in *Proceedings of the IEEE International Electron Devices Meeting*, pp. 731–734, Dec 2000
560. S. Kose, E. Salman, E.G. Friedman, Shielding methodologies in the presence of power/ground noise, in *Proceedings of the IEEE International Symposium on Circuit and Systems*, pp. 2277–2280, May 2009
561. S. Kose, E. Salman, E.G. Friedman, Shielding methodologies in the presence of power/ground noise. IEEE Trans. Very Large Scale Integr. (VLSI) Circuits **19**(8), 1458–1468 (2011)
562. J. Cong, L. He, C.-K. Koh, Z. Pan, Interconnect sizing and spacing with consideration of coupling capacitance. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **20**(9), 1164–1169 (2001)
563. V. Adler, E.G. Friedman, Repeater design to reduce delay and power in resistive interconnect. IEEE Trans. Circuits Syst. II: Analog Digit. Signal Process. **CAS II-45**(5), 607–616 (1998)
564. M. Ghoneima, Y. Ismail, Optimum positioning of interleaved repeaters in bidirectional buses. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **24**(3), 461–469 (2005)
565. Alphaworks Group, IBM Electromagnetic Field Solver Suite Tools, IBM. Available Online: http://www.alphaworks.ibm.com/tech/eip.
566. P. Bai et al., A 65 nm logic technology featuring 35 nm gate lengths, enhanced channel strain, 8 Cu interconnect layers, low-k ILD and 0.57 μm^2 SRAM cell, in *Proceedings of the IEEE International Electron Devices Meeting*, pp. 57–60, Dec 2004
567. K. Mistry et al., A 45 nm logic technology with high-k + metal gate transistors, strained silicon, 9 Cu interconnect layers, 193nm dry patterning, and 100% Pb-free packaging, in *Proceedings of the IEEE International Electron Devices Meeting*, pp. 247–250, Dec 2007
568. S. Natarajan et al., A 32 nm logic technology featuring 2nd-generation high-k + metal-gate transistors, enhanced channel strain and 0.171 μm^2 SRAM cell size in a 291 Mb array, in *Proceedings of the IEEE International Electron Devices Meeting*, pp. 1–3, Dec 2008
569. *The International Technology Roadmap for Semiconductors* (Semiconductor Industry Association, California, 2007)
570. M.R. Becer et al., Postroute gate sizing for crosstalk noise reduction. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **23**(12), 1670–1677 (2004)

571. R. Weerasekera, L.R. Zheng, D. Pamunuwa, H. Tenhunen, Crosstalk immune interconnect driver design, in *Proceedings of the IEEE International SOC Conference*, pp. 139–142, Nov 2004
572. N. Hanchate, N. Ranganathan, Simultaneous interconnect delay and crosstalk noise optimization through gate sizing using game theory. IEEE Trans. Comput. **55**(8), 1011–1023 (2006)
573. T. Sakurai, Closed-form expressions for interconnection delay, coupling, and crosstalk in VLSI's. IEEE Trans. Electron Devices **40**(1), 118–124 (1993)
574. K.T. Tang, E.G. Friedman, Delay and noise estimation of CMOS logic gates driving coupled resistive-capacitive interconnections. Integr. VLSI J. **29**(2), 131–165 (2000)
575. F.W. Grover, *Inductance Calculations: Working Formulas and Tables* (Dover, Mineola, 1962)
576. J. Wallis, *Opera Mathematica* (Oxonii, Leon: Lichfield Academiæ Typographi, 1656)
577. B. Kleveland, X. Qi, L. Madden, T. Furusawa, R.W. Dutton, M.A. Horowitz, S.S. Wong, High-frequency characterization of on-chip digital interconnects. IEEE J. Solid-State Circuits **37**(6), 716–725 (2002)
578. K.-H. Erhard, F.M. Johannes, R. Dachauer, Topology optimization techniques for power/ground networks in VLSI, in *Proceedings of the IEEE/ACM European Design Automation Conference*, pp. 362–367, Sept 1992
579. J. Fu, X. Wu, X. Hong, Y. Cai, PG2000: a CAD tool for power/ground network design, optimization and verification based on standard cell VLSIs, in *Proceedings of the IEEE international conference on communications, circuits and systems*, vol. 2, pp. 1424–1428, June 2002
580. S.X.D. Tan, C.J.R. Shi, J.-C. Lee, Reliability-constrained area optimization of VLSI power/ground networks via sequence of linear programmings. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **22**(12), 1678–1684 (2003)
581. K. Wang, M. Marek-Sadowska, On-chip power-supply network optimization using multigrid-based technique. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **24**(3), 407–417 (2005)
582. D. Khalil, Y.I. Ismail, Approximate frequency response models for RLC power grids, in *Proceedings of the IEEE International Symposium on Circuits and Systems*, pp. 3784–3787, May 2007
583. N. Srivastava, X. Qi, K. Banerjee, Impact of on-chip inductance on power distribution network design for nanometer scale integrated circuits, in *Proceedings of the IEEE International Symposium on Quality Electronic Design*, pp. 346–351, Mar 2005
584. R. Jakushokas, E.G. Friedman, Methodology for multi-layer interdigitated power and ground network design, in *Proceedings of the IEEE International Symposium on Circuits and Systems*, pp. 3208–3211, May/June 2010
585. K. Mistry et al., A 45nm logic technology with high-k + metal gate transistors, strained silicon, 9 Cu interconnect layers, 193nm dry patterning, and 100% Pb-free packaging, in *Proceedings of the IEEE International Electron Devices Meeting*, pp. 247–250, Dec 2007
586. S. Natarajan et al., A 32nm logic technology featuring 2nd-generation high-k + metal gate transistors, enhanced channel strain and 0.171 μm^2 SRAM cell size in a 291Mb array, in *Proceedings of the IEEE International Electron Devices Meeting*, pp. 1–3, Dec 2008
587. M.L. Doelz, E.T. Heald, D.L. Martin, Binary data transmission techniques for linear systems, in *Proceedings of the Institute of Radio Engineers*, vol. 45, pp. 656–661, May 1957
588. S. Galli, O. Logvinov, Recent developments in the standardization of power line communications within the IEEE. IEEE Commun. Mag. **46**(7), 64–71 (2008)
589. K.L. Wong, T. Rahal-Arabi, M. Ma, G. Taylor, Enhancing microprocessor immunity to power supply noise with clock-data compensation. IEEE J. Solid-State Circuits **41**(4), 749–758 (2006)
590. M. Bazes, Two novel full complementary self-biased CMOS differential amplifiers. IEEE J. Solid-State Circuits **26**(2), 165–168 (1991)
591. P. Cauvet, S. Bernard, M. Renovell, System-in-package, a combination of challenges and solutions, in *Proceedings of the IEEE VLSI Test Symposium*, pp. 193–199, May 2007

592. I. Savidis, B. Vaisband, E.G. Friedman, Experimental analysis of thermal coupling in 3-D integrated circuits. IEEE Trans. Very Large Scale Integr. (VLSI) Circuits **23**(10), 2077–2089 (2015)
593. B. Vaisband, I. Savidis, E.G. Friedman, Thermal conduction path analysis in 3-D ICs, in *Proceedings of the IEEE International Symposium on Circuit and Systems*, pp. 594–597, June 2014
594. A.P. Chandrakasan, R.W. Brodersen, *Low-Power CMOS Design* (Wiley-IEEE Press, New York, 1998)
595. C. Piguet, *Low-Power Processors and Systems on Chips* (CRC Press, Boca Raton, 2005)
596. A.P. Chandrakasan, M. Potkonjak, J. Rabaey, R.W. Brodersen, HYPER-LP: a system for power minimization using architectural transformations, in *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pp. 300–303, Nov 1992
597. A.P. Chandrakasan, M. Potkonjak, R. Mehra, J. Rabaey, R.W. Brodersen, Optimizing power using transformations. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **14**(1), 12–31 (1995)
598. A.P. Chandrakasan, S. Sheng, R.W. Brodersen, Low-power CMOS digital design. IEEE J. Solid-State Circuits **27**(4), 473–484 (1992)
599. T. Kuroda et al., A high-speed low-power 0.3 μm CMOS gate array with variable threshold voltage (VT) scheme, in *Proceedings of the IEEE Custom Integrated Circuits Conference*, pp. 53–56, May 1996
600. V. Kursun, E.G. Friedman, Domino logic with variable threshold keeper. IEEE Trans. Very Large Scale Integr. (VLSI) Syst. **11**(6), 1080–1093 (2003)
601. V. Kursun, E.G. Friedman, Sleep switch dual threshold voltage domino logic with reduced standby leakage current. IEEE Trans. Very Large Scale Integr. (VLSI) Syst. **12**(5), 485–496 (2004)
602. K. Usami, T. Ishikawa, M. Kanazawa, H. Kotani, Low-power design technique for ASIC's by partially reducing supply voltage, in *Proceedings of the IEEE International ASIC Conference*, pp. 301–304, Sept 1996
603. D. Marculescu, Power efficient processors using multiple supply voltages, in *Proceedings of the Workshop on Compilers and Operating Systems for Low Power*, Oct 2000
604. J.-M. Chang, M. Pedram, Energy minimization using multiple supply voltages, in *Proceedings of the IEEE International Symposium on Low Power Electronics and Design*, pp. 157–162, Aug 1996
605. J.-M. Chang, M. Pedram, Energy minimization using multiple supply voltages. IEEE Trans. Very Large Scale Integr. (VLSI) Syst. **5**(4), 436–443 (1997)
606. R.I. Bahar, H. Cho, G.D. Hachtel, E. Macii, F. Somenzi, An application of ADD-based timing analysis to combinational low power Re-synthesis, in *Proceedings of the ACM/IEEE International Workshop on Low Power Design*, pp. 39–44, Apr 1994
607. V. Kursun, R.M. Secareanu, E.G. Friedman, CMOS voltage interface circuit for low power systems, in *Proceedings of the IEEE International Symposium on Circuits and Systems*, pp. 3667–3670, May 2002
608. K. Usami, M. Igarashi, F. Minami, T. Ishikawa, M. Kanzawa, M. Ichida, K. Nogami, Automated low-power technique exploiting multiple supply voltages applied to a media processor. IEEE J. Solid-State Circuits **33**(3), 463–472 (1998)
609. K. Usami, M. Horowitz, Clustered voltage scaling technique for low-power design, in *Proceedings of the IEEE International Symposium on Low Power Electronics and Design*, pp. 3–8, Apr 1995
610. K. Usami et al., Automated low-power technique exploiting multiple supply voltages applied to a media processor, in *Proceedings of the IEEE Custom Integrated Circuit Conference*, pp. 131–134, May 1997
611. V. Kursun, S.G. Narendra, V.K. De, E.G. Friedman, Low-voltage-swing monolithic DC – DC conversion. IEEE Trans. Circuits Syst. II: Express Briefs **51**(5), 241–248 (2004)
612. V. Kursun, V.K. De, E.G. Friedman, S.G. Narendra, Monolithic voltage conversion in low-voltage CMOS technologies. Microelectron. J. **36**(9), 863–867 (2005)

613. R.K. Krishnamurthy, A. Alvandpour, V.K. De, S. Borkar, High-performance and low-power challenges for sub-70 nm microprocessor circuits, in *Proceedings of the IEEE Custom Integrated Circuit Conference*, pp. 125–128, May 2002
614. S.H. Kulkarni, D. Sylvester, High performance level conversion for dual V_{dd} design. IEEE Trans. Very Large Scale Integr. (VLSI) Syst. **12**(9), 926–936 (2004)
615. M. Igarashi, K. Usami, K. Nogami, F. Minami, Y. Kawasaki, T. Aoki, M. Takano, S. Sonoda, M. Ichida, N. Hatanaka, A low-power design method using multiple supply voltages, in *Proceedings of the IEEE International Symposium on Low Power Electronics and Design*, pp. 36–41, Aug 1997
616. J.-S. Wang, S.-J. Shieh, J.-C. Wang, C.-W. Yeh, Design of standard cells used in low-power ASIC's exploiting the multiple-supply-voltage scheme, in *Proceedings of the IEEE International ASIC Conference*, pp. 119–123, Sept 1998
617. M. Hamada, Y. Ootaguro, T. Kuroda, Utilizing surplus timing for power reduction, in *Proceedings of the IEEE Conference on Custom Integrated Circuits*, pp. 89–92, May 2001
618. T. Sakurai, A.R. Newton, Alpha-power law MOSFET model and its application to CMOS inverter delay and other formulas. IEEE J. Solid-State Circuits **25**(2), 584–594 (1990)
619. W. Hung, Total power optimization through simultaneously multiple-V_{dd} multiple-V_{TH} assignment and device sizing with stack forcing, in *Proceedings of the IEEE International Symposium on Low Power Electronics and Design*, pp. 144–149, Aug 2004
620. S.K. Mathew, M.A. Anders, B. Bloechel, T. Nguyen, R.K. Krishnamurthy, S. Borkar, A 4-GHz 300-mW 64-bit integer execution ALU with dual supply voltages in 90-nm CMOS. IEEE J. Solid-State Circuits **40**(1), 44–51 (2005)
621. D. Nguyen, A. Davare, M. Orshansky, D. Chinnery, B. Thompson, K. Keutzer, Minimization of dynamic and static power through joint assignment of threshold voltages and sizing optimization, in *Proceedings of the IEEE International Symposium on Low Power Electronics and Design*, pp. 158–163, Aug 2003
622. M. Takahashi et al., A 60-mW MPEG4 video codec using clustered voltage scaling with variable supply-voltage scheme. IEEE J. Solid-State Circuits **33**(11), 1772–1780, Nov 1998
623. K. Zhang, U. Bhattacharya, Z. Chen, F. Hamzaoglu, D. Murray, N. Vallepalli, Y. Wang, B. Zheng, M. Bohr, A 3-GHz 70-Mb SRAM in 65-nm CMOS technology with integrated column-based dynamic power supply. IEEE J. Solid-State Circuits **41**(1), 146–151 (2006)
624. S. Raje, M. Sarrafzadeh, Variable voltage scheduling, in *Proceedings of the ACM International Symposium on Low Power Design*, pp. 9–14, Apr 1995
625. A.V. Mezhiba, E.G. Friedman, Impedance characteristics of power distribution grids in nanoscale integrated circuits. IEEE Trans. Very Large Scale Integr. (VLSI) Syst. **12**(11), 1148–1155 (2004)
626. M. Popovich, E.G. Friedman, M. Sotman, A. Kolodny, On-chip power distribution grids with multiple supply voltages for high-performance integrated circuits, in *Proceedings of the ACM/IEEE Great Lakes Symposium on VLSI*, pp. 2–7, Apr 2005
627. J. Mermet, W. Nebel, *Low Power Design in Deep Submicron Electronics* (Kluwer Academic, Norwell, 1997)
628. A.P. Chandrakasan, W.J. Bowhill, F. Fox, *Design of High-Performance Microprocessor Circuits* (Wiley-IEEE Press, New York, 2000)
629. L.D. Smith, Packaging and power distribution design considerations for a sun microsystems desktop workstation, in *Proceedings of the IEEE Conference on Electrical Performance of Electronic Packaging*, pp. 19–22, Oct 1997
630. M. Popovich, E.G. Friedman, Noise coupling in multi-voltage power distribution systems with decoupling capacitors, in *Proceedings of the IEEE International Symposium on Circuits and Systems*, pp. 620–623, May 2005
631. G. Zhong, C.K. Koh, Exact closed form formula for partial mutual inductance of on-chip interconnects, in *Proceedings of the IEEE International Conference on Computer Design*, pp. 428–433, Sept 2002

Index

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0