High Performance Power Distribution Networks with On-Chip Decoupling Capacitors for Nanoscale Integrated Circuits

by

Mikhail Popovich

Submitted in Partial Fulfillment
of the
Requirements for the Degree
Doctor of Philosophy

Supervised by
Professor Eby G. Friedman

Department of Electrical and Computer Engineering
The College
School of Engineering and Applied Sciences

University of Rochester
Rochester, New York

2007

It has become appallingly obvious that our technology has exceeded our humanity.

— Albert Einstein

Dedication

This work is dedicated to my parents, Mr. Evgeniy Antonovich and Mrs. Lyudmila Mikhailovna, my wife Oksana, and my daughter Elizabeth Michelle.

Curriculum Vitae

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. degree in electrical and computer engineering from the University of Rochester, Rochester, NY in 2002, where he is completing the Ph.D. degree in electrical engineering.

He was an intern at Freescale Semiconductor Corporation, Tempe, AZ, in the summer 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 the Eastman Kodak Company, Rochester, NY. He has authored a book and 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. His research interests are in the areas of on-chip noise, signal integrity, and interconnect design including on-chip inductive effects, optimization of power distribution networks, and the design 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.

Publications

Book

1. M. Popovich, A. V. Mezhiba, and E. G. Friedman, *Power Distribution Networks with On-Chip Decoupling Capacitors*, Springer Publishing Company (in press).

Journal Papers

2. M. Popovich, R. M. Secareanu, E. G. Friedman, and O. L. Hartin, "Efficient Placement of Distributed On-Chip Decoupling Capacitors in Nanoscale ICs," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* (in submission).

3. M. Popovich, M. Sotman, A. Kolodny, and E. G. Friedman, "Effective Radii of On-Chip Decoupling Capacitors," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* (in submission).

4. M. Popovich, E. G. Friedman, M. Sotman, and A. Kolodny, "On-Chip Power Distribution Grids with Multiple Supply Voltages for High Performance Integrated Circuits," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* (in submission).

5. M. Popovich and E. G. Friedman, "Decoupling Capacitors for Multi-Voltage Power Distribution Systems," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, Vol. 14, No. 3, pp. 217–228, March 2006.

Conference Papers

6. M. Popovich, R. M. Secareanu, E. G. Friedman, and O. L. Hartin "Distributed On-Chip Decoupling Capacitors," *ACM International Conference on Nano-Networks* (in submission).

7. M. Popovich, R. M. Secareanu, E. G. Friedman, and O. L. Hartin, "Efficient Placement of Distributed On-Chip Decoupling Capacitors in Nanoscale ICs," *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, November 2007 (in press).

8. M. Sotman, A. Kolodny, M. Popovich, and E. G. Friedman, "On-Die Decoupling Capacitance: Frequency Domain Analysis of Activity Radius," *Proceedings of the IEEE International Symposium on Circuits and Systems*, pp. 489–492, May 2006.

9. M. Popovich, E. G. Friedman, M. Sotman, A. Kolodny, and R. M. Secareanu, "Maximum Effective Distance of On-Chip Decoupling Capacitors in Power Distribution Grids," *Proceedings of the ACM/IEEE Great Lakes Symposium on VLSI*, pp. 173–179, April/May 2006.

10. M. Sotman, M. Popovich, A. Kolodny, and E. G. Friedman, "Leveraging Symbiotic On-Die Decoupling Capacitance," *Proceedings of the IEEE Conference on Electrical Performance of Electronic Packaging*, pp. 111–114, October 2005.

11. M. Popovich, E. G. Friedman, R. M. Secareanu, and O. L. Hartin, "On-Chip Power Noise Reduction Techniques in High Performance SoC-Based Integrated Circuits," *Proceedings of the IEEE International SOC Conference*, pp. 309–312, September 2005.

12. M. Popovich and E. G. Friedman, "Noise Coupling in Multi-Voltage Power Distribution Systems with Decoupling Capacitors," *Proceedings of the IEEE International Symposium on Circuits and Systems*, pp. 620–623, May 2005.

13. M. Popovich, E. G. Friedman, M. Sotman, and A. Kolodny, "On-Chip Power Distribution Grids with Multiple Supply Voltages for High Performance Integrated Circuits," *Proceedings of the ACM/IEEE Great Lakes Symposium on VLSI*, pp. 2–7, April 2005 **(received the best student paper award for GLSVLSI 2005)**.

14. M. Popovich and E. G. Friedman, "Noise Aware Decoupling Capacitors for Multi-Voltage Power Distribution Systems," *Proceedings of the IEEE International Symposium on Quality Electronic Design*, pp. 334–339, March 2005.

15. M. Popovich and E. G. Friedman, "Impedance Characteristics of Decoupling Capacitors in Multi-Power Distribution Systems," *Proceedings of the IEEE International Conference on Electronics, Circuits, and Systems*, pp. 160–163, December 2004.

16. M. Popovich and E. G. Friedman, "Decoupling Capacitors for Power Distribution Systems with Multiple Power Supplies," *Proceedings of the 28th Annual IEEE EDS/CAS Activities in Western New York Conference*, p. 9, November 2004.

17. M. Popovich and E. G. Friedman, "Decoupling Capacitors for Power Distribution Systems with Multiple Power Supply Voltages," *Proceedings of the IEEE International SOC Conference*, pp. 331–334, September 2004.

18. M. Margala, R. Alonzo, G.-Q. Chen, B. J. Jasionowski, K. Kraft, M. Lay, J. Lindner, M. Popovich, and J. Suss, "Low-Voltage Power-Efficient Adder Design," *Proceedings of the IEEE Midwest Symposium on Circuits and Systems*, pp. 461–464, August 2002.

Patent Disclosures

19. M. Popovich, R. M. Secareanu, E. G. Friedman, and O. L. Hartin, "Distributed On-Chip Decoupling Capacitor Network," US patent pending.

20. M. Popovich and E. G. Friedman, "Method and Design Flow for Effective Placement of On-Chip Decoupling Capacitors Determined by Maximum Effective Radii," US patent pending **(received 2007 GRC Inventor Recognition Award from the Semiconductor Research Corporation)**.

21. M. Popovich, E. G. Friedman, R. M. Secareanu, and O. L. Hartin, "Method and Apparatus to Reduce Noise Fluctuations in On-Chip Power Distribution Networks," US patent pending.

22. R. M. Secareanu, M. Popovich, W. Parmon, and O. L. Hartin, "Design of a Distributed On-Chip Decoupling Capacitor Network – Correlation between a Distributed Capacitor and Circuit Demands," Freescale defensive patent, August 11, 2006, http://www.ip.com/pubview/IPCOM000138959D.

Acknowledgments

The unforgettable memory and experience that I have had at the University of Rochester will greatly embellish the rest of my life. It is impossible to acknowledge all those people who have made my life in Rochester joyful and meaningful. If your name is not listed, rest assured that my gratitude is not less than for those listed below.

First of all, I would like to express my deepest appreciation to my academic advisor, Professor Eby G. Friedman, for his generous mentorship in my academic and personal growth. He was the one who gave me a chance to perform this research and believed in my success. His professionalism, commitment, and kindness have brought the best out of me. Thank you for making my research experience fruitful and highly enjoyable.

I thank Professors Paul Ampadu, Michael Huang, and Daniel Stefankovic for serving on my proposal and defense committees and for their valuable advice and comments regarding this dissertation. I would also like to thank Professor David

H. Albonesi for his support. I would express my appreciation to the University of Rochester and the Department of Electrical and Computer Engineering for providing a unique environment which encourages high quality research. I would also like to thank all of the department secretaries and computer staff for their active support. Special thanks to Prof. Avinoam Kolodny and Michael Sotman from Technion – Israel Institute of Technology for their collaboration and support. Their generous feedback has greatly improved my work. I am grateful to Olin L. Hartin, Marie Burnham, and Radu M. Secareanu for giving me the opportunity of a summer internship in Freescale Semiconductor Corporation and for their collaboration and suggestions.

I would like to thank those previous and current members in the High Performance VLSI/IC Design and Analysis Laboratory: Andrey Mezhiba, Dimitris Velenis, Volkan Kursun, Boris Andreev, Weize Xu, Magdy El-Moursy, Junmou Zhang, Guoqing Chen, Vasilis Pavlidis, Jonathan Rosenfeld, Emre Salman, Renatas Jakushokas, and Ioannis Savidis for their help, advise, and companionship. I would also like to thank RuthAnn Williams for administrative support and for giving a spark to routine life in the lab.

Last, but not least, I want to thank my wife Oksana and my daughter Elizabeth for all of the support that they give me and happiness they bring to my life. I am sorry for all holidays and weekends that I missed working on my dissertation. I am deeply indebted to my parents for supporting me in every stage of my life. I owe you much of what I have become. My debt to you can never be paid. My gratitude is also

extended to relatives and friends in Russia for their understanding and encouragement throughout my life.

This work is supported in part by the Semiconductor Research Corporation under Contract Nos. 2003-TJ-1068 and 2004-TJ-1207, the National Science Foundation under Contract Nos. CCR-0304574 and CCF-0541206, grants from the New York State Office of Science, Technology & Academic Research to the Center for Advanced Technology in Electronic Imaging Systems, and by grants from Intel Corporation, Eastman Kodak Corporation, Manhattan Routing, Intrinsix Corporation, and Freescale Semiconductor Corporation.

Abstract

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 on-chip power distribution systems have significantly increased. The higher switching speed of a greater number of smaller transistors produces faster and larger current transients in the power distribution network. These conditions place strict requirements on the on-chip power distribution network to ensure the integrity of the on-chip power supply.

To manage the problem of high power dissipation, multiple on-chip power supply voltages have become commonplace in nanoscale integrated circuits. On-chip power distribution grids with multiple power supply voltages and multiple grounds are presented in this dissertation. The impedance characteristics of the power distribution grids with multiple power supply voltages and multiple grounds are described. The proposed power distribution grid structures are shown to outperform traditional power distribution grids with multiple power supply voltages and a single ground.

Decoupling capacitors are widely used to manage power supply noise. Conventional approaches for placing on-chip decoupling capacitors is shown to be ineffective in nanoscale integrated circuits. A design methodology for placing on-chip decoupling capacitors based on the maximum effective radii is described in this dissertation. Techniques to estimate the minimum required on-chip decoupling capacitance are presented. A methodology for designing decoupling capacitors for power distribution systems with multiple power supply voltages is also described.

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 in proximity of the switching circuits. A system of distributed on-chip decoupling capacitors is shown to be a good compromise, providing the required charge drawn by the load while satisfying existing technology constraints. The research presented in this dissertation provides specific methodologies, techniques, and strategies for designing robust on-chip power distribution networks with on-chip decoupling capacitors for application to high performance nanoscale integrated circuits.

Contents

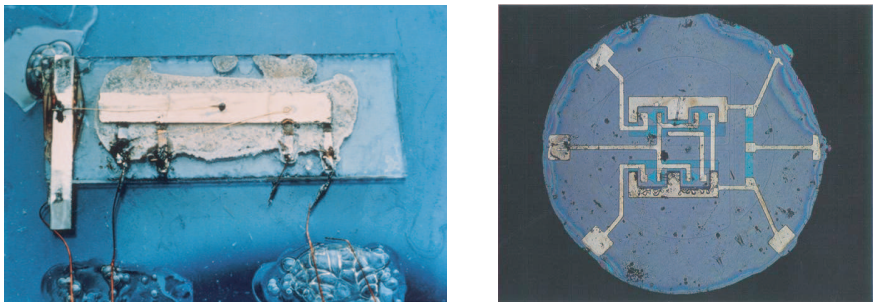
List of Tables

List of Figures

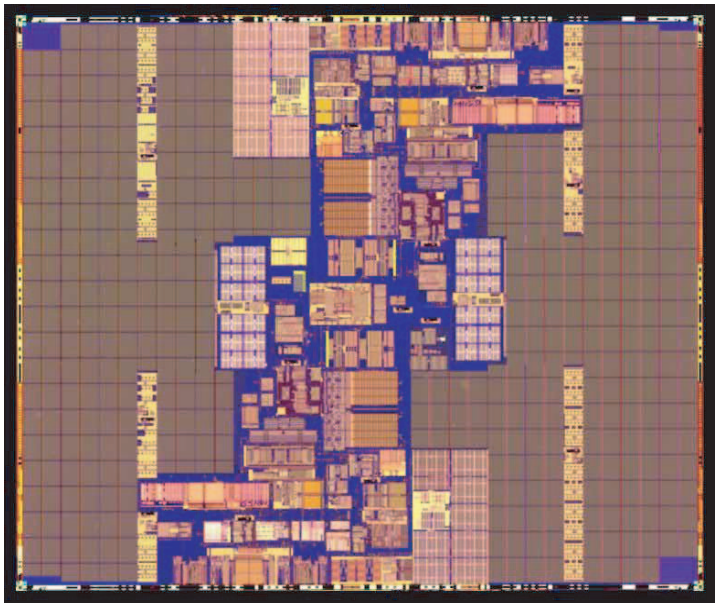
Chapter 1

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 recent high performance dual core Montecito microprocessor (Intel Corporation, 2005) are depicted in Fig. 1.1. In 1960, Jean Hoerni invented the planar process [5]. Later, in 1960, Dawon Kahng and Martin Atalla



(a) (b)



(c)

Figure 1.1: Microphotographs of the first integrated circuit (IC) and first monolithic IC along with a high performance, high complexity IC (the die size is not to scale). (a) The first IC (Texas Instruments, 1958), (b) the first monolithic IC (Fairchild Semiconductor, 1959), and (c) the high performance dual core Montecito microprocessor (Intel Corporation, 2005).

demonstrated the first silicon based 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.

With the on-going miniaturization of integrated circuit feature size, the design of the power and ground distribution networks has become a challenging task. These challenges arise from shorter transition times, lower noise margins, higher currents, and increased current densities. Furthermore, the power supply voltage has decreased in order to lower dynamic power dissipation. A greater number of transistors increases the total current drawn from the power delivery network. Simultaneously, the higher switching speed of these greater number of smaller transistors produces faster and larger current transients in the power distribution network [8].

The high average currents produce large ohmic IR voltage drops [9] and the fast current transients cause large inductive $L\frac{dI}{dt}$ voltage drops [10] (ΔI noise) in the power distribution networks. The power distribution networks are designed to minimize these voltage drops, maintaining the power and ground supply voltages at the terminals of the current load within specified noise margins. If the power supply voltage drops too much, the performance and functionality of the circuit will

be severely compromised. Furthermore, excessive overshoot of the supply voltage can affect circuit reliability [11].

The goal of this chapter is to introduce the problem of power delivery in integrated circuits, discuss the deleterious effects of power distribution noise, and provide guidance and perspective to the rest of this dissertation. Fundamental issues in the design of the power and ground distribution networks in modern high performance, high complexity ICs are described in Section 1.1. The adverse effects of power supply noise on circuit operation are discussed in Section 1.2. Finally, the overall structure of the dissertation and a description of each chapter are outlined in Section 1.3.

1.1 The Problem of Power Delivery

The problem of power delivery is illustrated in Fig. 1.2, where a circuit model of a power distribution system is shown [12]. A power distribution system consists of a power supply, a current load, and interconnect lines connecting the supply to the load. The power supply is modeled as an ideal voltage source providing nominal power and ground voltage levels (V_{dd} and V_{gnd}). The current load is modeled as a variable current source $I_{load}(t)$, representing a transistor or circuit module. Note that the power and ground lines have a finite parasitic resistance R_g and R_p and inductance L_g and L_p. As a result of the non-ideal interconnect lines connecting the supply and the load, resistive voltage drops $\Delta V_R = IR$ and inductive voltage drops $\Delta V_L = L\frac{dI}{dt}$

develop across the parasitic interconnect impedances, as the load draws current from the power distribution system. The voltage levels at the load terminals, therefore, deviate from the nominal power supply levels, dropping to $V_{dd} - IR_p - L\frac{dI}{dt}$ at the power terminal and rising to $V_{gnd} + IR_g + L\frac{dI}{dt}$ at the ground terminal, as shown in Fig. 1.2.

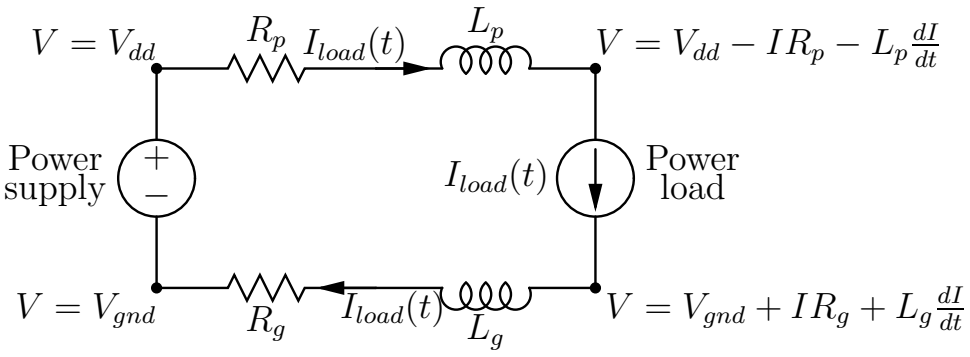


Figure 1.2: Circuit model of a power delivery system. A power and ground distribution system consists of a power supply, current load, and non-ideal interconnect lines.

These fluctuations in the supply voltages are referred to as power supply noise [13]. Power supply noise adversely affects circuit operation through several mechanisms, as described in Section 1.2. The power distribution noise at the terminals of the load should be maintained within the maximum allowed voltage fluctuations (noise margins) to ensure correct operation of the overall system. The power distribution system should therefore be carefully designed, supplying sufficient current to each transistor while satisfying target noise margins.

To maintain the supply voltage within specified noise margins, the output impedance of the power distribution network should be low as seen from the power and ground terminals of the circuit. IC technologies are expected to scale for at least another decade [8]. As a result, the average and transient currents drawn from the power delivery network will continue to rise. Simultaneously, the power supply voltage will be scaled to manage on-chip power consumption. The target impedance of a power distribution network in high speed, high complexity ICs such as microprocessors will therefore continue to drop, reaching an inconceivable level of 250 $\mu\Omega$ by the year 2017 [8], as depicted in Fig. 1.3.

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 [14]. At high frequencies, however, only 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 [15]. On-chip decoupling capacitors, however, reduce the self-resonant frequency of a power distribution system, resulting in high amplitude power supply voltage fluctuations at the resonant frequencies. A system of on-chip decoupling capacitors should therefore be carefully designed to provide a low impedance, resonant-free power distribution network 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 [11].

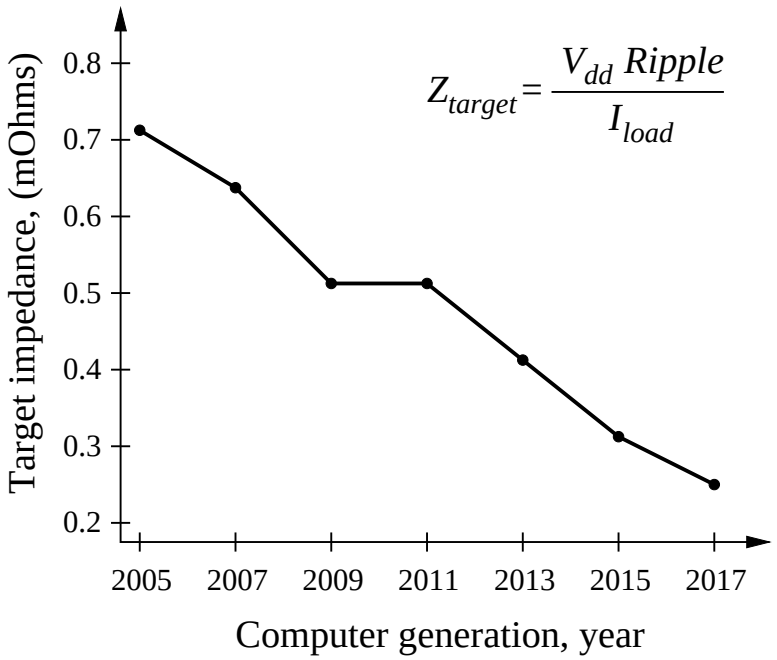


Figure 1.3: Projections of the target impedance of a power distribution system. The target impedance will continue to drop for future technology generations at an aggressive rate of 1.25 X per computer generation [8].

1.2 Detrimental Effects of Power Distribution Noise

Power distribution noise adversely affects the operation of an IC through several mechanisms. The propagation delay of the on-chip signals depends upon the power supply level during a signal transition. If the rail-to-rail power voltage is reduced as a result of power supply variations, the gate-to-source voltage of the NMOS and PMOS transistors decreases, lowering the output current of the transistors. The signal delay therefore increases as compared to the delay under a nominal power supply voltage. Conversely, a higher supply voltage will shorten the propagation delay. The power

noise thereby results in increased delay uncertainty of the clock signals and an increase in the maximum delay of the data paths [16], [17]. Power supply noise can therefore limit the maximum operating frequency of an integrated circuit [18], [19], [20].

Power supply noise can also result in on-chip clock jitter. A phase-locked loop (PLL) with a voltage controlled oscillator (VCO) is typically utilized to generate the on-chip clock signal in high performance microprocessors. PLLs and VCOs are highly susceptible to power supply variations, resulting in phase deviations in the on-chip clock signal, as illustrated in Fig. 1.4. These phase deviations are typically referred to as clock jitter [21], [22]. Cycle-to-cycle jitter (*or* random deviations) refers to independent deviations from the ideal phase at different edges of a clock signal. Peak-to-peak jitter (*or* systematic variations) refers to variations of the on-chip clock phase as compared to the system clock signal. Clock jitter contributes directly to the delay uncertainty of the clock signals and degrades the synchronization among the different clock domains, compromising overall system performance [23]. It is therefore important to provide a low noise power supply voltage to the on-chip clock generation and distribution circuitry to ensure fault-free operation of the overall system [24].

In digital ICs with single-ended signaling, the power and ground supply networks serve as a reference voltage for the on-chip data signals. Variations in the supply voltage therefore create a discrepancy between the power and ground voltage levels at the interface between the transmitting and receiving circuits. Power noise induced

uncertainty in these reference voltages degrades the noise margins of the on-chip signals, jeopardizing the functionality of a system.

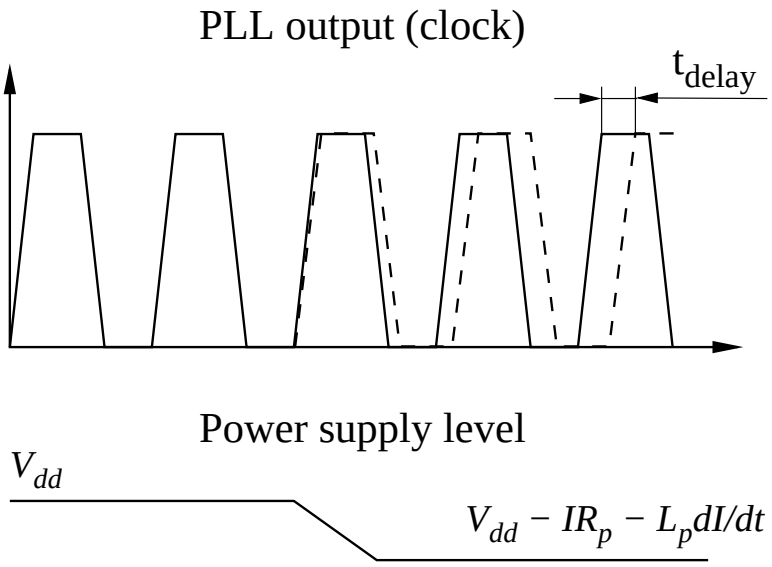


Figure 1.4: A phase error (or jitter) in the on-chip clock signal due to power supply noise. The original clock signal is depicted by a solid line and the delayed clock signal is shown by a dashed line.

In contemporary CMOS technologies, the thickness of the gate silicon dioxide has been dramatically reduced to several molecular layers to increase the current drive of the transistors. The maximum power supply voltage is limited by the maximum electric field within the gate oxide layer [25]. Variations in the power supply can increase the voltage across the gate oxide layer above the nominal power supply, degrading the long term reliability of the transistors [26]. Overshoots of the power supply voltage should therefore be limited to avoid significant degradation in reliability.

1.3 Dissertation Outline

Methodologies for designing power distribution grids in high performance nanoscale ICs is the primary topic of this dissertation. The related topic of placing on-chip decoupling capacitors to mitigate power distribution noise in nanometer ICs is also discussed. Design techniques and expressions to determine the location and magnitude of the on-chip decoupling capacitors are developed. Several power distribution schemes with multiple supply voltages are presented, resulting in reduced power distribution noise.

Decoupling capacitance is introduced in Chapter 2. A historical perspective of capacitance is provided. The decoupling capacitor is shown to be analogous to a reservoir of charge. A hydraulic analogy of a hierarchical placement of decoupling capacitors is introduced. It is demonstrated that the impedance of a power distribution system can be maintained below target specifications 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.

Systems with multiple power supply voltages are described in Chapter 3. 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 optimum number and magnitude of the multiple power supplies, maximizing any savings in power.

On-chip power distribution grids with multiple power supply voltages are discussed in Chapter 4. A power distribution grid with multiple power supplies and multiple grounds is presented. It is shown that 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.

On-chip power noise reduction techniques in high performance ICs are the primary subject of Chapter 5. 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.

Decoupling capacitors for power distribution systems with multiple power supply voltages is the topic of the following chapter – Chapter 6. With the introduction of a second power supply, the noise at one power supply can propagate to the other power supply, producing power and signal integrity problems in the overall system. The interaction 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 investigated. Design techniques to cancel and shift the antiresonant spikes out of the range of the operational frequencies are also presented.

On-chip decoupling capacitors have traditionally been allocated into the available white space on a die. The efficacy of the on-chip decoupling capacitors 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 Chapter 7. The maximum effective radii of an on-chip decoupling capacitor as determined by the target impedance (during discharge) and the charge time are developed. 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

in the proximity of the switching circuits. Moreover, a large decoupling capacitor exhibits a distributed 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 Chapter 8. Design techniques to estimate the location and magnitude of a system of distributed decoupling capacitors are presented. Various tradeoffs in the design of a system of distributed on-chip decoupling capacitors are also investigated.

In Chapter 9, the research described in the dissertation is summarized. Directions for future research are suggested in Chapter 10. A multi-layer model of an on-chip power distribution grid needs to be developed to accurately analyze power noise and signal integrity in high complexity ICs. Chip-package co-design methodologies will be developed to accurately analyze power and signal integrity in nanoscale ICs. On-chip decoupling capacitors in mixed-signal and RF ICs can dramatically worsen substrate noise coupling. A design methodology for placing on-chip decoupling capacitors in mixed-signal ICs and systems-on-chip will be required. Techniques for placing on-chip decoupling capacitors in 3-D ICs, significantly reducing the overall on-chip decoupling capacitance, are also proposed as future research.

Chapter 2

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 very deep submicrometer (VDSM) 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 Chapter 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 $L\frac{dI}{dt}$ voltage drops (ΔI noise) [27].

Decoupling capacitors are often utilized to manage this power supply noise. Decoupling capacitors 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 described 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 Section 2.1, the concept of a decoupling capacitance is introduced and an historical retrospective is described. A practical model of a decoupling capacitor is also introduced. In Section 2.2, the impedance of a power distribution system with decoupling capacitors is presented. Target specifications of the impedance 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 Section 2.3. Different types of on-chip decoupling capacitors are qualitatively analyzed in Section 2.4.

The advantages and disadvantages of several types of widely used on-chip decoupling capacitors are also discussed in Section 2.4. Three strategies for allocating on-chip decoupling capacitors are described in Section 2.5. Finally, some conclusions are offered in Section 2.6.

2.1 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 dissertation, the use of decoupling capacitors to mitigate power supply noise is investigated. The concept of a decoupling capacitor is introduced in this section. A historical retrospective is presented in Section 2.1.1. A description of a decoupling capacitor as a reservoir of charge is discussed in Section 2.1.2. Decoupling capacitors are shown to be an effective way to provide the required charge to a switching current load within a short period of time. A practical model of a decoupling capacitor is presented in Section 2.1.3.

2.1.1 Historical Retrospective

About 600 BC, Thales of Miletus recorded that the ancient Greeks could generate sparks by rubbing balls of amber on spindles [28]. This is the triboelectric effect [29],

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. 2.1 [30]. 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 very similar capacitor in January 1746 [31]. 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 [32]. 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 term condensor was coined by Alessandro Volta in 1782 (derived from the Italian *condensatore*), with reference to the ability of a device to store a higher density of electric charge than a normal isolated conductor [32].

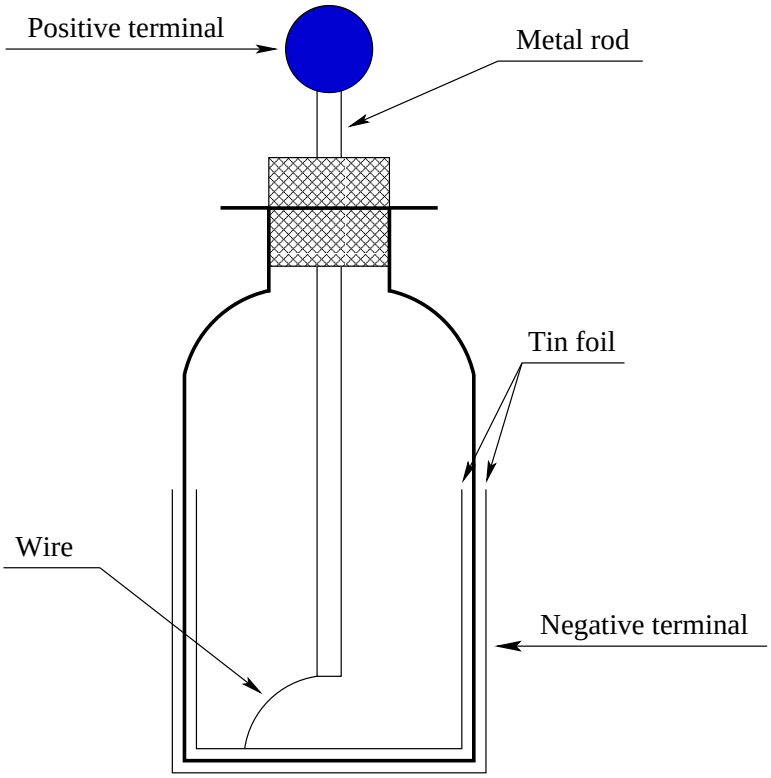


Figure 2.1: Leyden jar originally developed by Ewald Georg von Kleist in 1745 and independently invented by Pieter van Musschenbroek in 1746. The charge is stored on the glass between two tin foils (capacitor plates) [30].

2.1.2 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,

$$C = \frac{Q}{V}. \tag{2.1}$$

The capacitance is proportional to the surface area of the conducting plate and inversely proportional to the distance between the plates [33]. The capacitance is also proportional to the permittivity of the dielectric substance that separates the plates. The capacitance of a parallel-plate capacitor is

$$C \approx \frac{\epsilon A}{d}, \tag{2.2}$$

where ϵ is the permittivity of the dielectric, A is the area of the plates, and d is the spacing between the plates. Equation (2.2) 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 [34], as shown in Fig. 2.2. Accurate closed-form expressions have been developed by numerically fitting a model that describes parallel lines above the plane or between two parallel planes [35], [36], [37], [38], [39], [40].

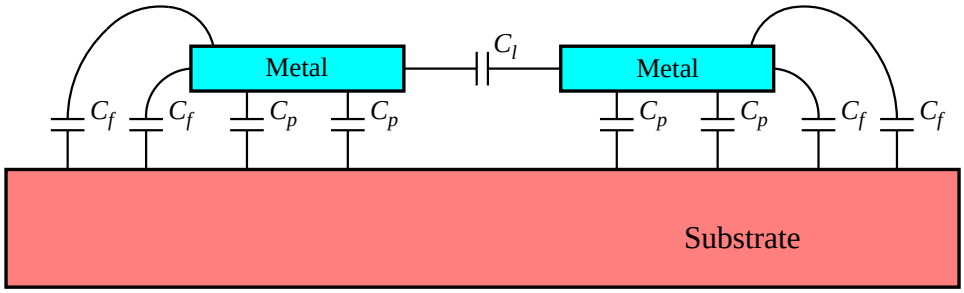


Figure 2.2: Capacitance of two metal lines placed over a substrate. Three primary components compose the total capacitance of the on-chip metal interconnects. C_l denotes the lateral flux (side) capacitance, C_f denotes the fringe capacitance, and C_p denotes the parallel plate capacitance.

As opposite charges accumulate on the plates of a capacitor across an insulator, a voltage develops across the capacitor due to the electric field formed by the opposite charges. 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

$$E_{stored} = \frac{1}{2}CV^2 = \frac{1}{2}\frac{Q^2}{C} = \frac{1}{2}VQ. \tag{2.3}$$

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. 2.3. 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 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.

2.1.3 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 [41]. 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

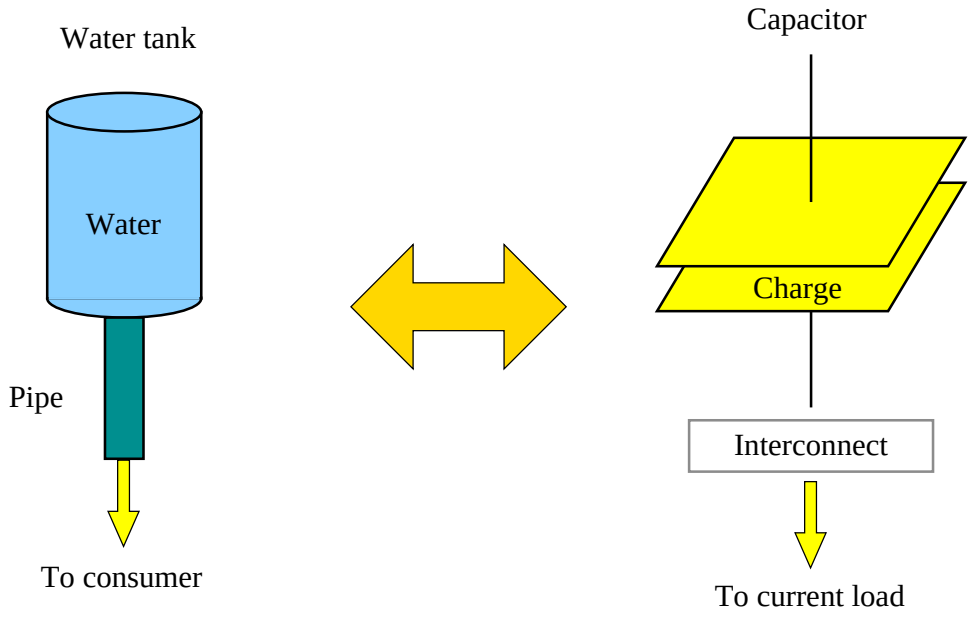


Figure 2.3: Hydraulic model of a decoupling capacitor as a reservoir of charge. Similar to water stored in a water tank and connected to the consumer through a system of pipes, charge at the decoupling capacitor is stored between the conductive plates connected to the current load through a hierarchical interconnect system.

(with a slope of 20 dB/dec in a logarithmic scale), as shown in Fig. 2.4. 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 [42]. 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 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 example on-chip metal-oxide-semiconductor (MOS) decoupling capacitor are illustrated in Fig. 2.5. 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. 2.5(b) [43].

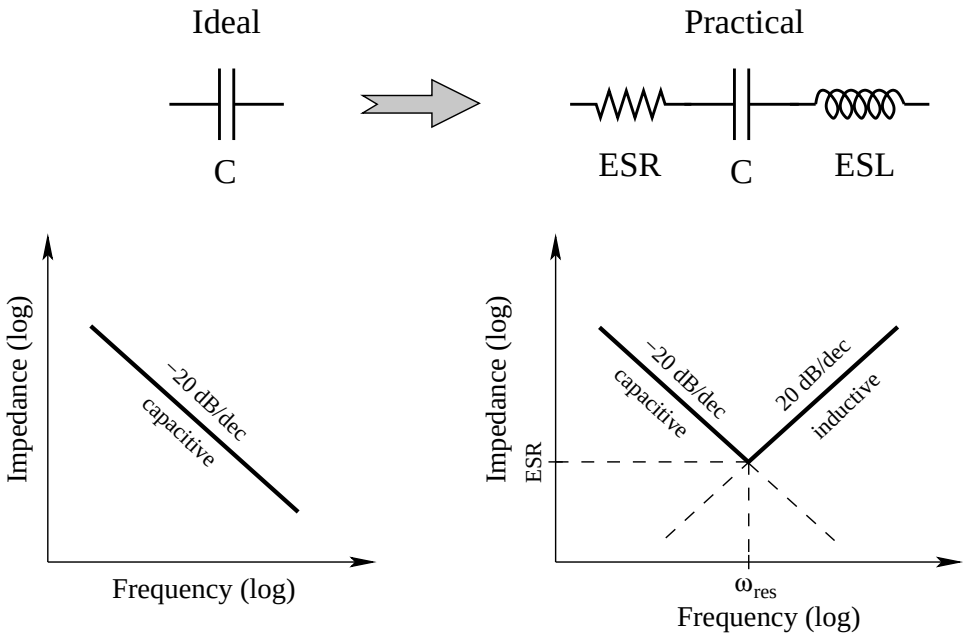
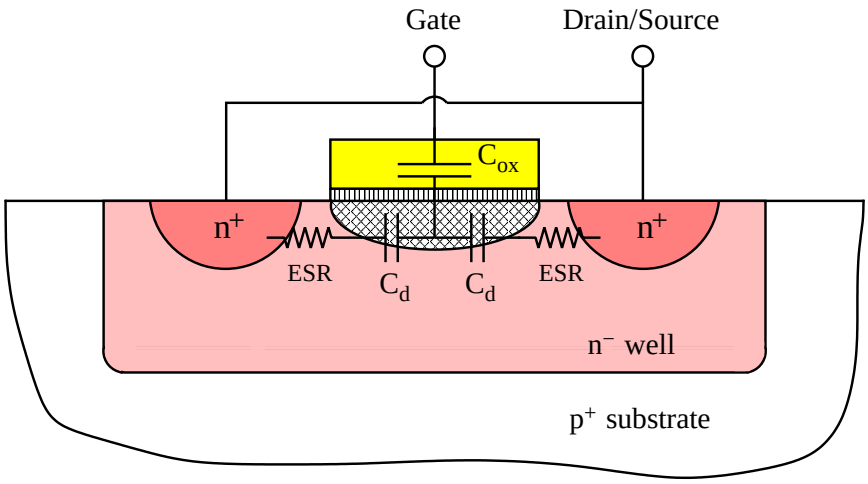
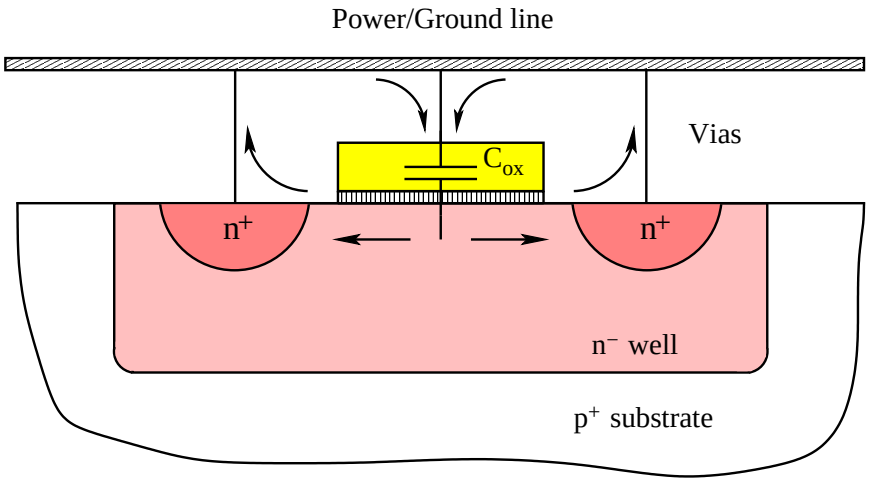


Figure 2.4: Practical model of a decoupling capacitor. The impedance of a practical decoupling capacitor decreases linearly with frequency, reaching the minimum at a resonant frequency. Beyond the resonant frequency, the impedance of the decoupling capacitor increases linearly with frequency due to the ESL. The minimum impedance is determined by the ESR of the decoupling capacitor.



(a) ESR of a 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 [44].



(b) ESL of a 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.

Figure 2.5: Physical structure of an on-chip MOS decoupling capacitor.

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. 2.4). 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 [45]. 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 two for each doubling in the number of capacitors [46].

2.2 Impedance of Power Distribution System with Decoupling Capacitors

As described in Section 2.1.2, 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 Section 2.2.1, the target impedance of a power distribution system is introduced. 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. Antiresonance phenomenon is presented in Section 2.2.2. A hydraulic analogy of a system of decoupling capacitors is described in Section 2.2.3. The analogy is drawn between a water supply system and the hierarchical placement of decoupling capacitors at different levels of a power delivery network.

2.2.1 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 [47]. A circuit network representing the impedance of a power distribution system as seen from the terminals of the current load is shown in Fig. 2.6.

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 [48], [49], [50]. The maximum tolerable impedance of a power

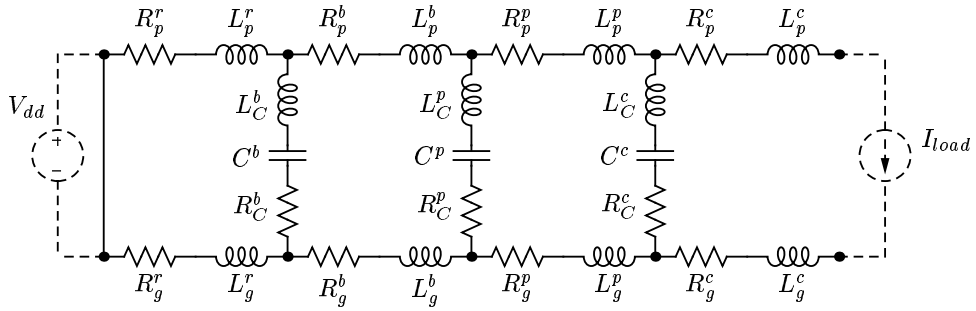


Figure 2.6: A circuit network representing the impedance of a power distribution system with decoupling capacitors as seen from the terminals of the current load. The ESR and ESL of the decoupling capacitors are also included. Subscript p denotes the power paths and subscript g denotes the ground path. Superscripts r, b, p, and c refer to the voltage regulator, board, package, and on-chip power delivery networks, respectively.

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.

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. 2.6, depends upon the impedance of the comprising circuit elements. A power distribution system with no decoupling capacitors is shown in Fig. 2.7. The power source and load are connected by interconnect with resistive and

inductive parasitic impedances. The magnitude of the impedance of this network is

$$|Z_{tot}(\omega)| = |R_{tot} + j\omega L_{tot}|\,, \tag{2.4}$$

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

$$R_{tot} = R^p_{tot} + R^g_{tot}, \tag{2.5}$$

$$R^p_{tot} = R^r_p + R^b_p + R^p_p + R^c_p, \tag{2.6}$$

$$R^g_{tot} = R^r_g + R^b_g + R^p_g + R^c_g, \tag{2.7}$$

$$L_{tot} = L^p_{tot} + L^g_{tot}, \tag{2.8}$$

$$L^p_{tot} = L^r_p + L^b_p + L^p_p + L^c_p, \tag{2.9}$$

$$L^g_{tot} = L^r_g + L^b_g + L^p_g + L^c_g. \tag{2.10}$$

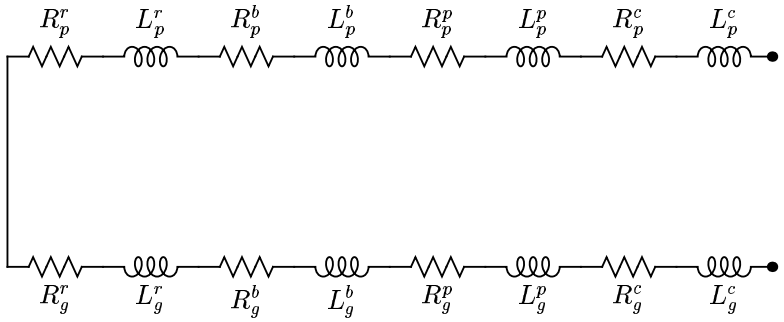


Figure 2.7: A circuit network representing the impedance of a power distribution system without decoupling capacitors.

The variation of the impedance with frequency is illustrated in Fig. 2.8. To satisfy a specification at low frequency, the resistance of the power delivery network should be sufficiently low, $R_{tot} < Z_{target}$. Above the frequency $f_{L_{tot}} = \frac{1}{2\pi}\frac{R_{tot}}{L_{tot}}$, however, the impedance of the power delivery network is dominated by the inductive reactance $j\omega L_{tot}$ and increases linearly with frequency, exceeding the target impedance at the frequency $f_{max} = \frac{1}{2\pi}\frac{Z_{target}}{L_{tot}}$.

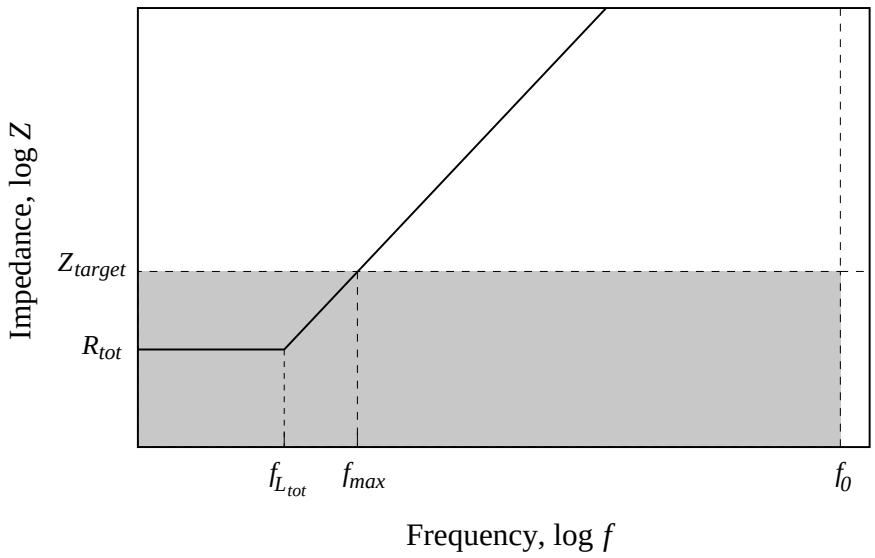


Figure 2.8: Impedance of a power distribution system without decoupling capacitors. The shaded area denotes the target impedance specifications of the overall power distribution system.

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 [51], [52], [53], [54], [55]. The inductance of the power distribution system is mainly determined by the board and package

interconnects [56], [57], [58]. 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 which is difficult to lower [59].

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 [60], [61], [62], [63]. 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 [64], [65]. Several stages of decoupling capacitors are typically utilized to maintain the output impedance of a power distribution system below a target impedance [46], [66], as described in Section 2.2.3.

2.2.2 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 [67], [68] or by an on-chip decoupling capacitor and the parasitic inductance of the package (*i.e.*, chip-package resonance) [69], [70], 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. 2.9. 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. 2.10. Observe that the impedance of the power delivery network is reduced by a factor of two 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

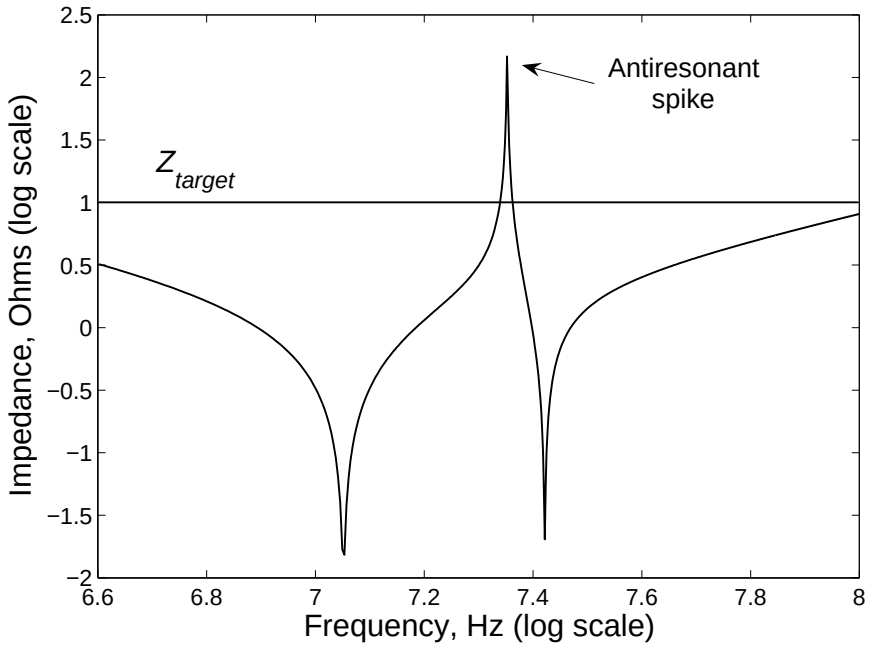


Figure 2.9: Antiresonance of the output impedance of a power distribution network. Antiresonance results in a distinctive peak, exceeding the target impedance specification.

impedance specification [46].

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 [15]. The antiresonance phenomenon for different capacitive values is illustrated in Fig. 2.11. The antiresonance of parallel decoupling capacitors can be explained as 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. 2.11). 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.

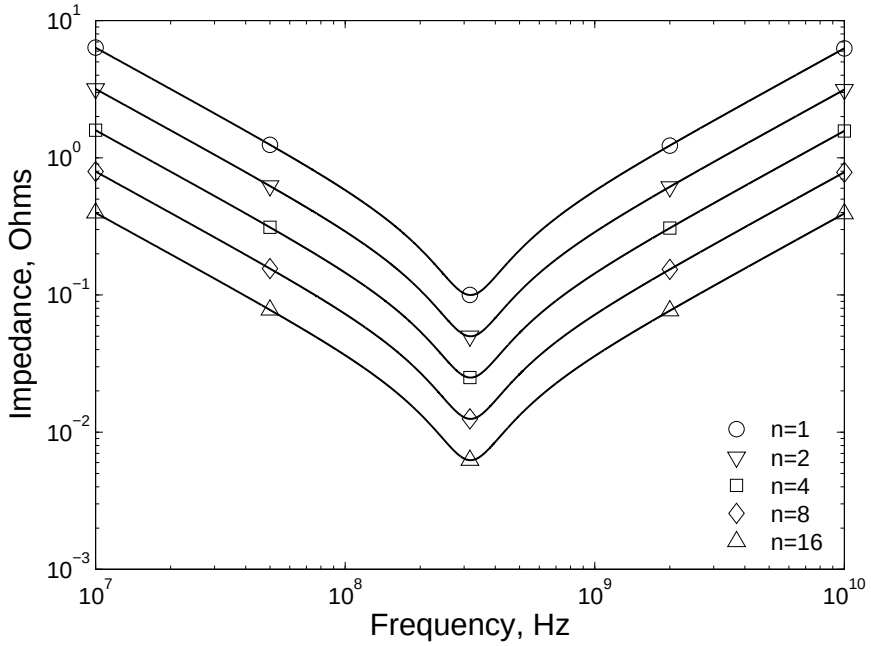


Figure 2.10: Impedance of a power distribution system with n identical decoupling capacitors connected in parallel. The ESR of each decoupling capacitor is $R = 0.1\,\Omega$, the ESL is $L = 100\,\text{pH}$, and the capacitance is $C = 1\,\text{nF}$. The impedance of a power distribution system is reduced by a factor of two as the number of capacitors is doubled.

The magnitude of the antiresonant spike can be effectively reduced by lowering the parasitic inductance of the decoupling capacitors. For instance, as discussed in [46], 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,

$$Q = \frac{L}{R}. \tag{2.11}$$

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. 2.12. In the case of two capacitors with distinctive nominal values ($C_1 \gg C_2$), a significant 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}, as shown in Fig. 2.12(a). 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. 2.12(b). As described in [71], the high frequency impedance of two parallel decoupling capacitors

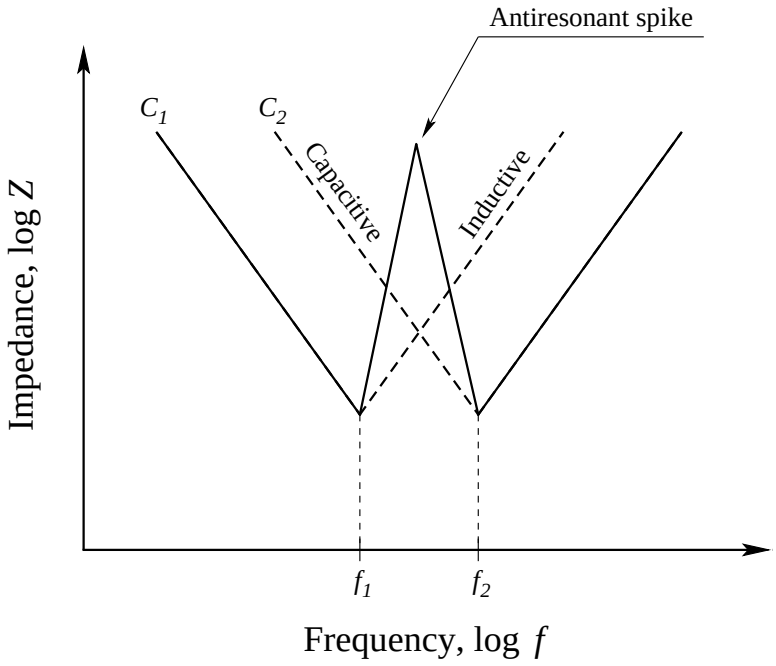
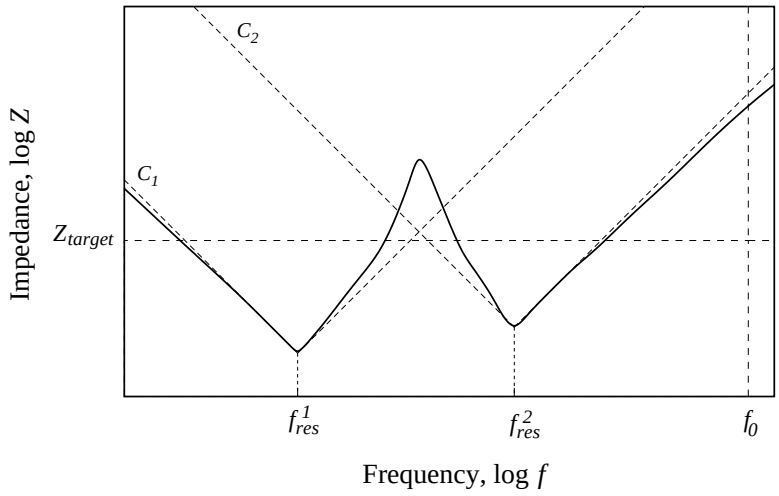
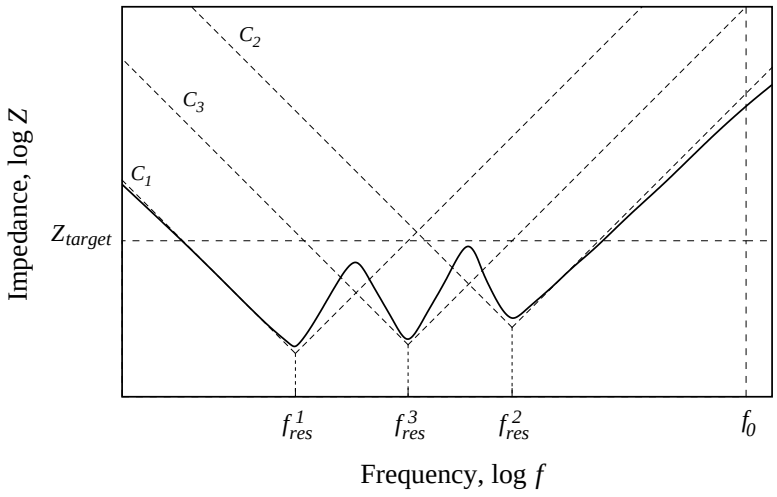


Figure 2.11: Antiresonance of parallel capacitors, $C_1 > C_2$, $L_1 = L_2$, and $R_1 = R_2$. A parallel LC tank is formed in the frequency range from f_1 to f_2. The total impedance drastically increases in the frequency range from f_1 to f_2 (the solid line), producing an antiresonant spike.

is only reduced by a factor of two (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) [72].



(a) Impedance of a power distribution system with two decoupling capacitors, $C_1 \gg C_2$



(b) Impedance of a power distribution system with three decoupling capacitors, $C_1 > C_3 > C_2$

Figure 2.12: 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.

2.2.3 Hydraulic Analogy of Hierarchical Placement of Decoupling Capacitors

As discussed in Section 2.1.2, 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 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 [11].

The hydraulic analogy of the hierarchical placement of decoupling capacitors is shown in Fig. 2.13. 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 an 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.

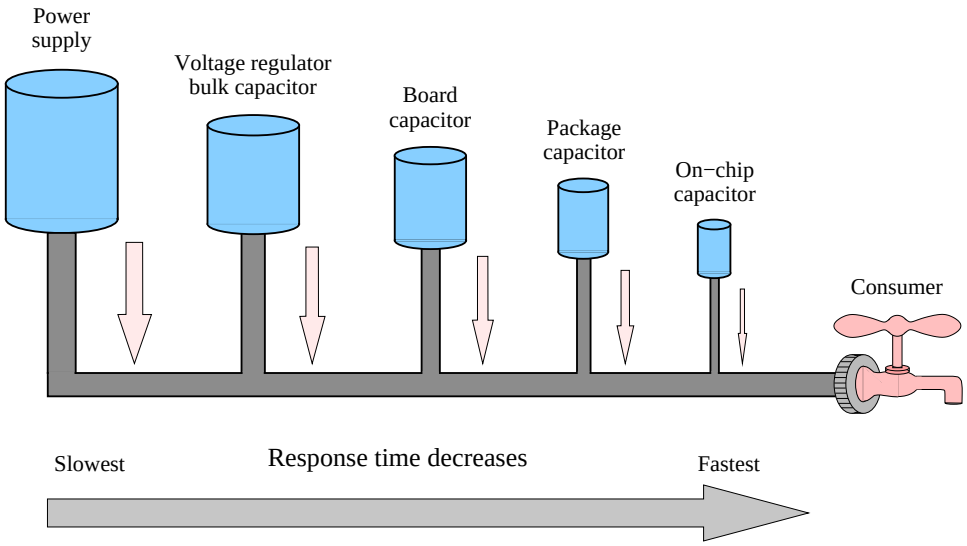


Figure 2.13: Hydraulic analogy of the hierarchical placement of decoupling capacitors. The decoupling capacitors are represented by the water tanks. The response time is proportional to the size of the capacitor and inversely proportional to the distance from a capacitor to the consumer. The on-chip decoupling capacitor has the shortest response time (located closer to the consumer), but is capable of providing the smallest amount of charge.

A power supply is typically treated as an infinite amount of charge. Due to large 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 closer 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 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. 2.6) are illustrated in Fig. 2.14. 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).

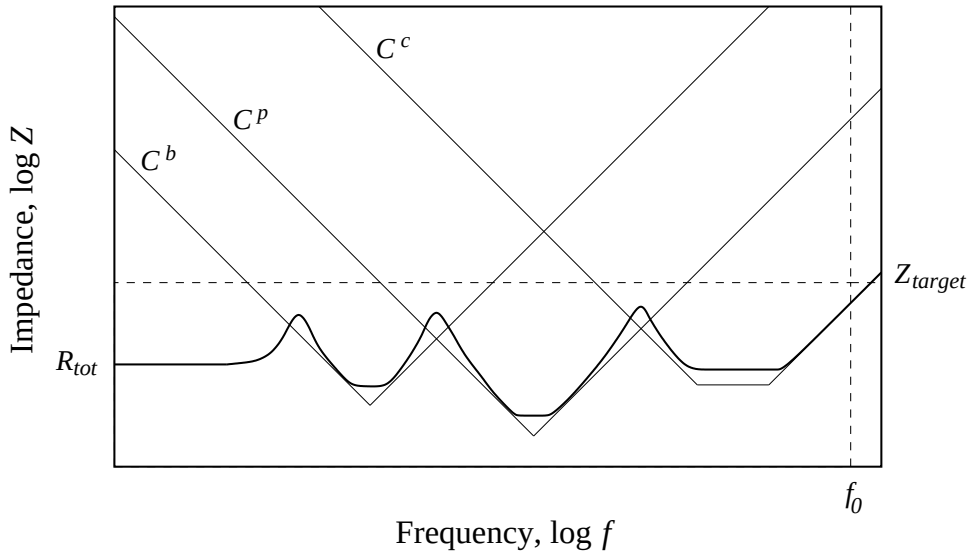


Figure 2.14: Impedance of a power distribution system with board, package, and on-chip decoupling capacitances. The overall impedance is shown with a black line. The impedance of a power distribution system with three levels of decoupling capacitors is maintained below the target impedance (dashed line) over the frequency range of interest. The impedance characteristics of the decoupling capacitors are shown by the thin solid lines.

Fully Compensated System

A special case in the impedance of an RLC circuit formed by the decoupling capacitor and the parasitic inductance of the power/ground (P/G) lines is achieved

when the zeros of the tank circuit impedance cancel the poles, making the impedance purely resistive and independent of frequency,

$$R_L = R_C = R_0 = \sqrt{\frac{L}{C}}, \tag{2.12}$$

$$\frac{L}{R_L} = R_C C, \tag{2.13}$$

where R_L and R_C are the parasitic resistance of the P/G lines and the ESR of the decoupling capacitor, respectively. In this case, the impedance of the RLC tank is fully compensated. Equations (2.12) and (2.13) 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 [73], [74]. 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

$C_{total} = C^b + C^p + C^c$ is larger than the total decoupling capacitance of a single-tier solution, where C^b, C^p, and C^c are the board, package, and on-chip decoupling capacitances, respectively. 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 [11].

2.3 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 Section 2.3.1. The intentional decoupling capacitance is reviewed in Section 2.3.2.

2.3.1 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 [75].

One component of the intrinsic capacitance is the parasitic capacitance of the interconnect lines. Three types of intrinsic interconnect capacitances are illustrated in Fig. 2.15. 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. 2.15).

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. 2.16, 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, as illustrated in Fig. 2.16(a). 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. 2.16(b).

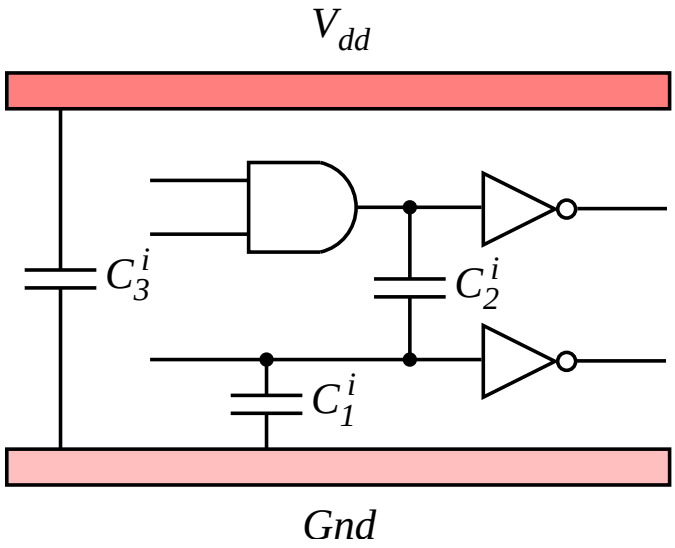


Figure 2.15: Intrinsic decoupling capacitance of the interconnect lines. C_1^i denotes the capacitance between the signal line and the power/ground line. C_2^i denotes the capacitance between signal lines. C_3^i denotes the capacitance between the power and ground lines.

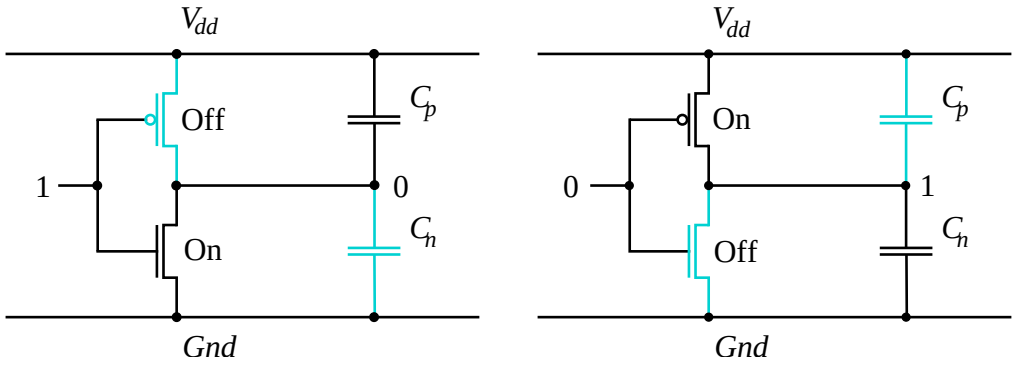


Figure 2.16: Intrinsic decoupling capacitance of a non-switching circuit. (a) Inverter input is high. (b) Inverter input is low.

Depending upon the total capacitance $(C_p + C_n)$ and the switching factor SF, the decoupling capacitance from the non-switching circuits is [76]

$$C_{circuit} = \frac{P}{V_{dd}^2 f} \frac{(1 - SF)}{SF}, \tag{2.14}$$

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 $R_{PMOS}C_n$ or $R_{NMOS}C_p$ and usually varies in a $0.18\,\mu$m CMOS technology from about 50 ps to 250 ps [76].

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 [77].

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 while the P-type wells are connected to the ground supply 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. 2.17. 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 $(R_p + R_n + R_{contact})C_{nw}$ for an N-well capacitor is typically in the range of 250 ps to 500 ps in a 0.18 μm CMOS technology. The parasitic capacitance of the wells usually dominates the intrinsic decoupling capacitance of ICs fabricated in an epitaxial CMOS process [78], [79]. The overall intrinsic on-chip decoupling capacitance consists of several components and is

$$C_{intrinsic} = C_{inter} + C_{pn} + C_{well} + C_{load} + C_{gs} + C_{gb}, \tag{2.15}$$

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 [80], [81]. The overall intrinsic decoupling capacitance of an IC can also be determined experimentally. In [82], 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 [70].

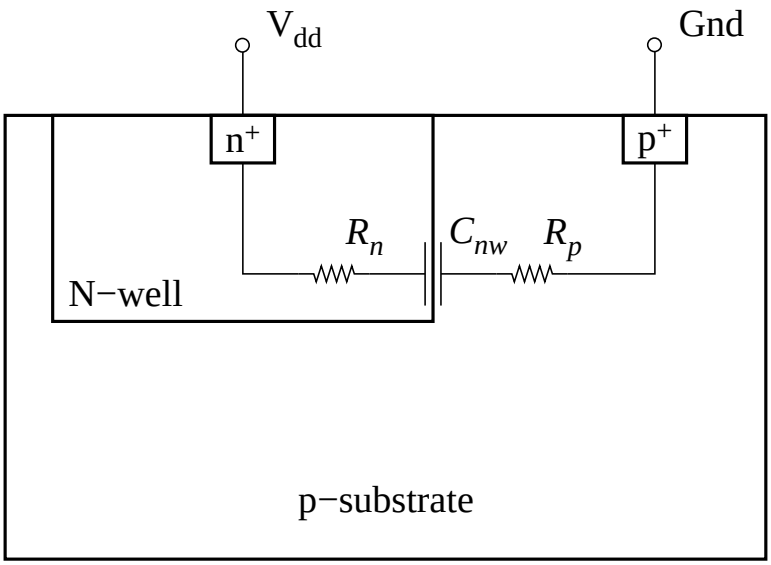


Figure 2.17: N-well junction intrinsic decoupling capacitance. The capacitor C_{nw} is formed by the reverse-biased p-n junction between the N-well and the p-substrate.

2.3.2 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 [83], [84].

A 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. 2.18. An optional fuse (or control gate) is typically provided to disconnect the thin oxide capacitor from the rest of the circuits in the undesirable case of a short circuit due to process defects. As the size and shape of MOS capacitors vary, the R_nC_{ox} time constant typically ranges from 40 ps to 200 ps. Depending upon the switching speed of the circuit, typical on-chip MOS decoupling capacitors are effective for RC time constants below 200 ps [76]. Using more than 20% of the overall die area for intentional on-chip decoupling capacitance is common in modern high performance, high complexity ICs, such as microprocessors [85], [86].

A 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 a distributed RC circuit, as shown in Fig. 2.19. The impedance of the distributed RC structure shown in Fig. 2.19 is frequency dependent, $Z(\omega) = R(\omega) + \frac{1}{j\omega C(\omega)}$. Both the resistance $R(\omega)$

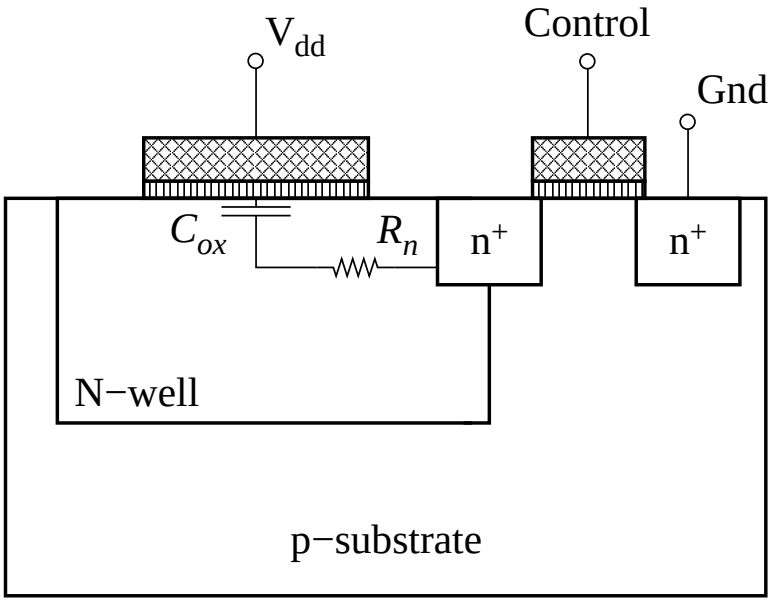


Figure 2.18: Thin oxide MOS decoupling capacitor.

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 [87]. 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 [87].

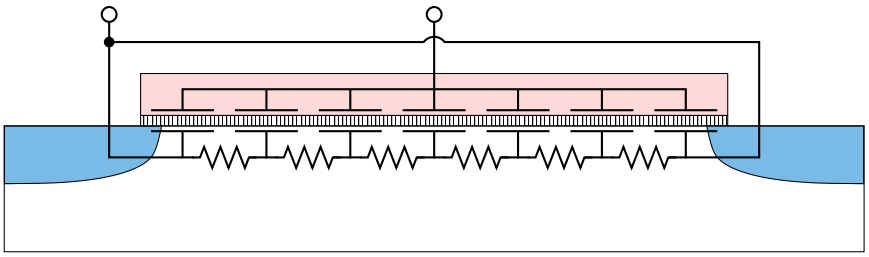


Figure 2.19: Equivalent RC model of a MOS decoupling capacitor.

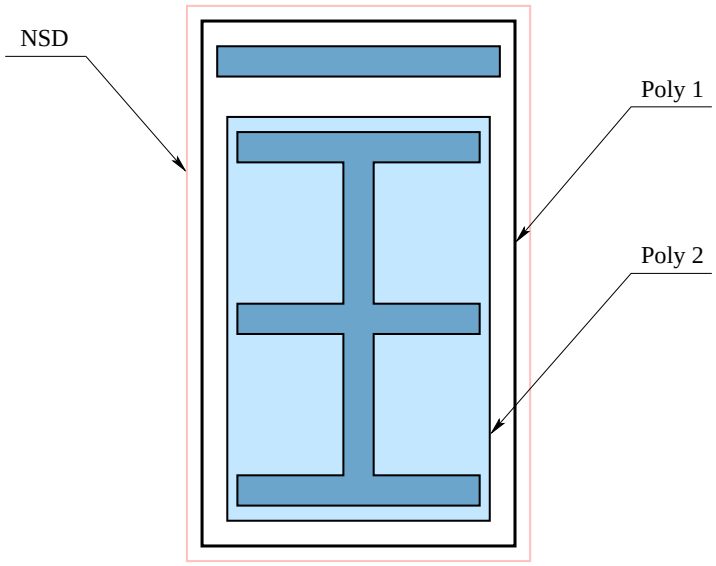
Long channel transistors, however, are more area efficient. In transistors with a minimum length channel, the source and body contacts dominate the transistor area, while the MOS capacitor stack occupies a relatively small fraction of the total area. For longer channels, the area of the MOS capacitor increases while the area overhead of the source/drain contacts remain constant, increasing the capacitance per total area [11]. A tradeoff therefore exists between the area efficiency and the ESR of the MOS decoupling capacitor. Transistors with a channel length twelve times greater than the minimum length are a good compromise [87]. 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.

2.4 Types of On-Chip Decoupling Capacitors

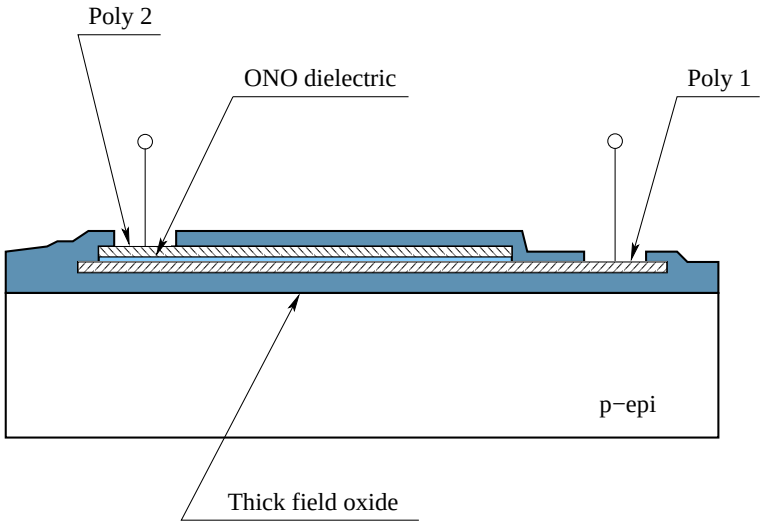
Multiple on-chip capacitors are utilized in ICs to satisfy various design requirements. Four types of widely utilized on-chip decoupling capacitors are the subject of this section. Polysilicon-insulator-polysilicon (PIP) capacitors are presented in Section 2.4.1. Three types of MOS decoupling capacitors, accumulation, depletion, and inversion, are described in Section 2.4.2. Metal-insulator-metal (MIM) decoupling capacitors are reviewed in Section 2.4.3. In Section 2.4.4, lateral flux decoupling capacitors are described. The design and performance characteristics of the different on-chip decoupling capacitors are compared in Section 2.4.5.

2.4.1 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 to 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 [88], as illustrated in Fig. 2.20. 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



(a)



(b)

Figure 2.20: Layout (a) and cross section (b) of a 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.

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 [88].

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. 2.20(b)). Observe from Fig. 2.20 that the PIP capacitors normally 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/volt [88]. The temperature coefficient of a PIP capacitor also depends on voltage modulation effects and is typically less than 250 ppm/°C [89].

2.4.2 MOS Capacitors

A MOS capacitor consists of a metal-oxide-semiconductor structure, as illustrated in Fig. 2.21. 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 to 10 Ω-cm (with a doping of 10^{15} cm^{-3}).

The capacitance of a MOS capacitor depends upon the voltage applied to the gate with respect to the body. The dependence of the capacitance upon the voltage

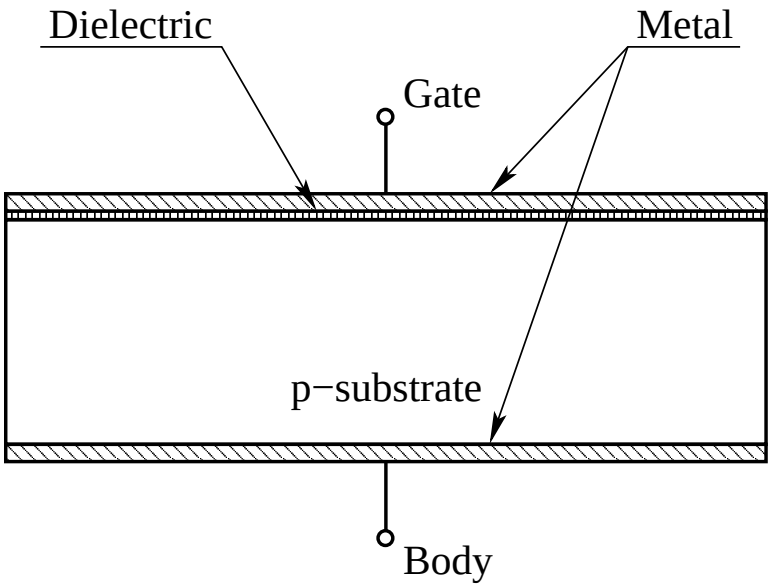


Figure 2.21: The structure of an n-type MOS capacitor.

across a MOS capacitor (a capacitance versus voltage (CV) diagram) is plotted in Fig. 2.22. Depending upon the gate-to-body potential V_{gb}, three regions of operation are distinguished in the CV diagram of a 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 two voltages (see Fig. 2.22). 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.

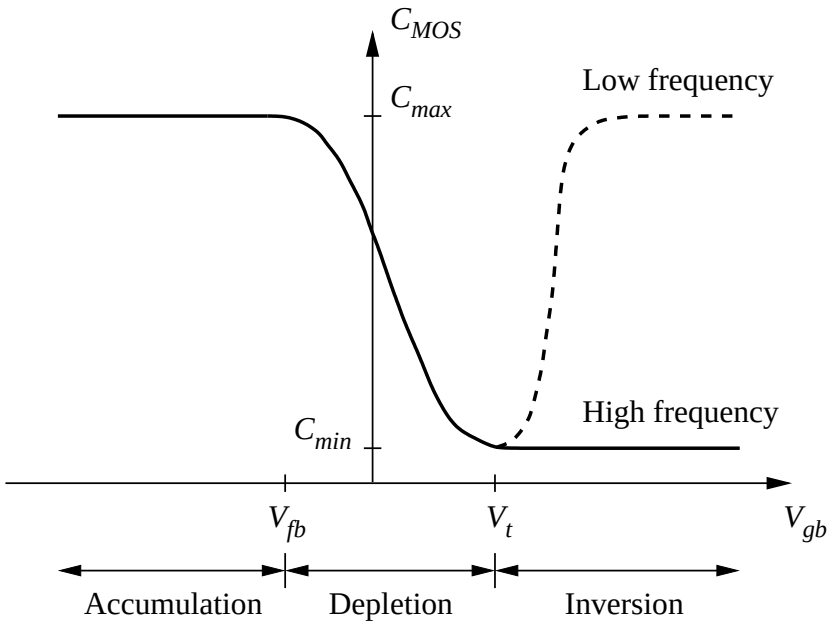


Figure 2.22: Capacitance versus gate voltage (CV) diagram of an n-type MOS capacitor. The flat band voltage V_{fb} separates the accumulation region from the depletion region. The threshold voltage V_t separates the depletion region from the inversion region.

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 a MOS capacitor operating in accumulation is shown in Fig. 2.23. 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

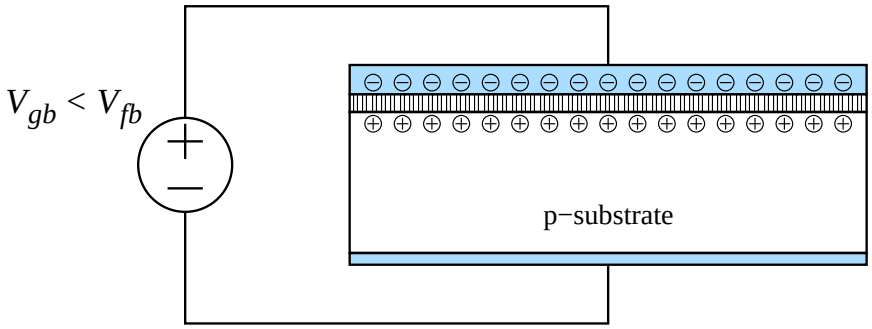


Figure 2.23: Charge distribution in an NMOS capacitor operating in accumulation ($V_{gb} < V_{fb}$).

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. 2.24. The capacitance of a MOS capacitor operating in accumulation achieves the maximum value and is

$$C_{MOS_{accum}} = C_{max} = A\,C_{ox} = A\,\frac{\epsilon_{ox}}{t_{ox}}, \tag{2.16}$$

where A is the area of the gate electrode, ϵ_{ox} is the permittivity of the oxide, and t_{ox} is the oxide thickness.

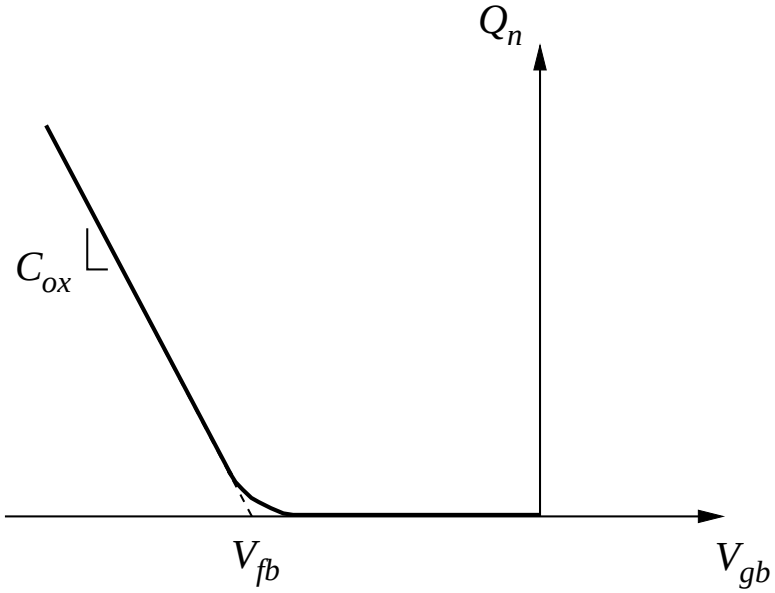


Figure 2.24: Accumulation charge density as a function of the applied gate voltage. The capacitance per unit area C_{ox} is determined by the slope of the line.

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. 2.25.

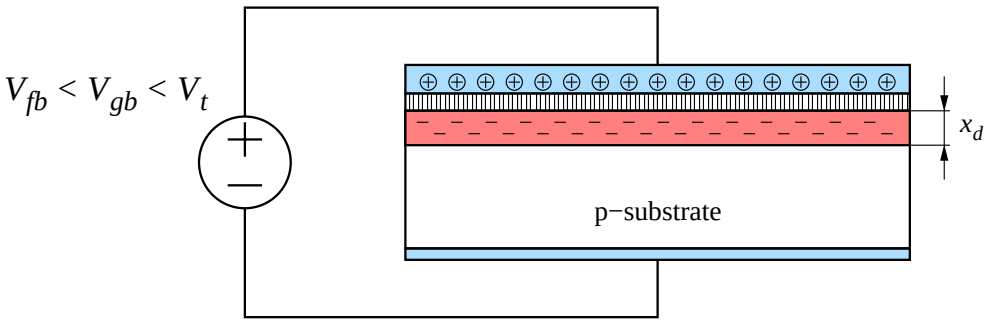


Figure 2.25: Charge distribution in an NMOS capacitor operating in depletion ($V_{fb} < V_{gb} < V_t$). Under this bias condition, all of the mobile positive carriers (holes) are pushed away, depleting the surface of the semiconductor, resulting in a negative space charge with thickness x_d.

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

$$C_d(V_{gb}) = \frac{\epsilon_{Si}}{x_d(V_{gb})}, \tag{2.17}$$

where ϵ_{Si} is the permittivity of the silicon and x_d is the thickness of the depletion layer (space charge). Observe from Fig. 2.25 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

$$C_{MOS_{deplet}} = A\,\frac{C_{ox}\,C_d}{C_{ox} + C_d}. \tag{2.18}$$

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 a MOS capacitor operating in inversion is depicted in Fig. 2.26.

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

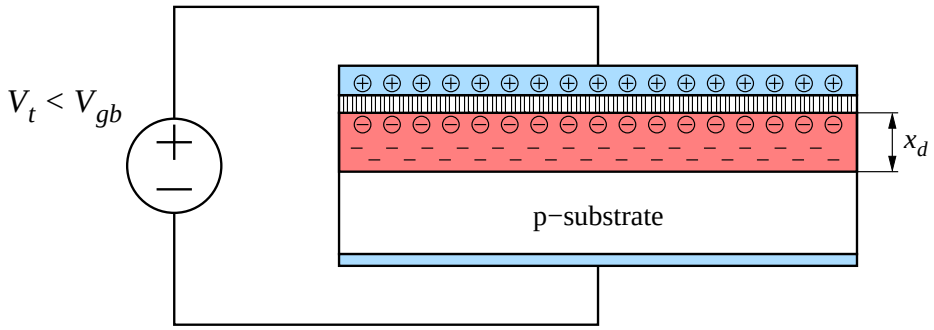


Figure 2.26: Charge distribution of an NMOS capacitor operating in inversion ($V_t < V_{gb}$). Under this bias condition, a negative charge is accumulated at the semiconductor surface, inverting the conductivity of the semiconductor surface (from p-type to n-type).

oxide layer. If the gate voltage approaches the threshold voltage, the depleted layer capacitance per unit area C_d^{min} reaches a minimum [90]. In this case, the overall MOS capacitance reaches the minimum value and is

$$C_{MOS_{inv}} = C_{MOS}^{min} = A\,\frac{C_{ox}\,C_d^{min}}{C_{ox} + C_d^{min}}, \tag{2.19}$$

where

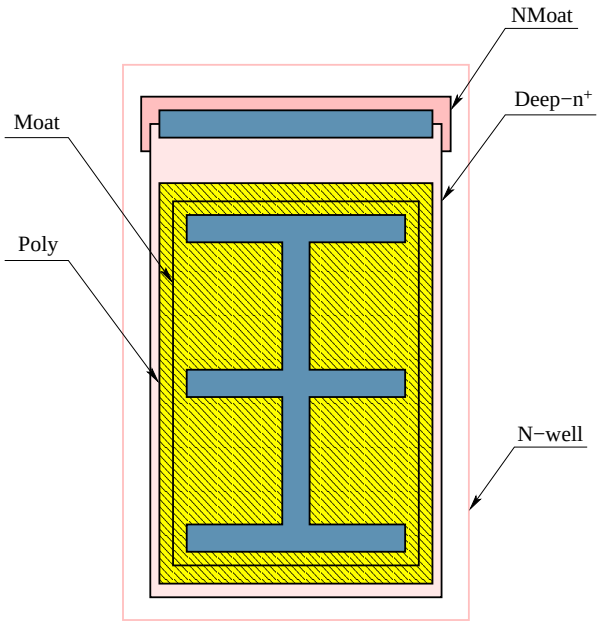
$$C_d^{min} = \frac{\epsilon_{Si}}{x_d^{max}}. \tag{2.20}$$

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

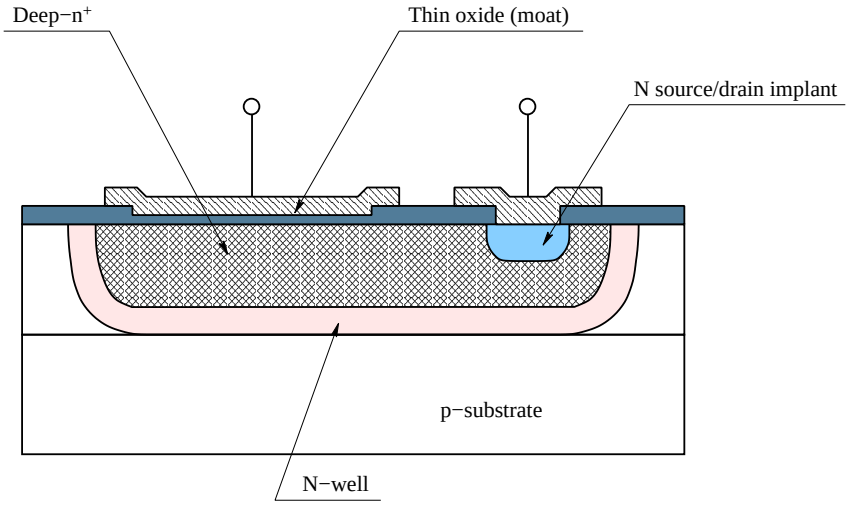
reaches the maximum value. At high frequencies, however, the generation rate is not sufficiently high, prohibiting the formation of the electron charge at the Si/SiO_2 interface. In this case, the thickness of the silicon depletion layer reaches the maximum. Hence, the inversion layer capacitance reaches the minimum.

A 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 ($25\,\mu$m or less) [88]. If the source and drain diffusions are omitted, the backgate contact typically runs entirely around the gate.

A layout and cross section of a MOS capacitor formed in a BiCMOS process are illustrated in Fig. 2.27. 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 $100\,\Omega/\square$), resulting in a substantial parasitic resistance of the lower plate. The heavily concentrated n-type doping thickens the gate oxide by 10% to 30% through dopant-enhanced oxidation, resulting in higher working voltages but 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.



(a)



(b)

Figure 2.27: Layout (a) and cross section (b) of a deep-n$^+$ MOS capacitor constructed in a BiCMOS process.

Regardless of how a 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 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 a MOS capacitor should therefore be connected to the driven node (with the lower impedance). Swapping the two electrodes of a 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 [91], providing a cost effective on-chip decoupling capacitance. MOS capacitors result in relatively high matching: the gate oxide capacitance is typically controlled with less than 5% error [89]. MOS capacitors, however, are non-linear devices that exhibit strong voltage dependence (more than 100 ppm/volt [92]) 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.

A 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 [93]. The MOS capacitance operating in accumulation is almost independent of frequency. Moreover, MOS decoupling capacitors operating in accumulation result in an approximately 15 X reduction in leakage current as compared to MOS decoupling capacitors operating in inversion [94]. MOS decoupling capacitors operating in accumulation should therefore be the primary form of MOS decoupling capacitors in modern high performance ICs.

2.4.3 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. 2.28. 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 [95], the parasitic capacitance to the substrate is less than 2% of the total capacitance.

Historically, MIM capacitors have been widely used in RF and mixed-signal ICs due to low leakage, high linearity, low process variations (high accuracy), and low

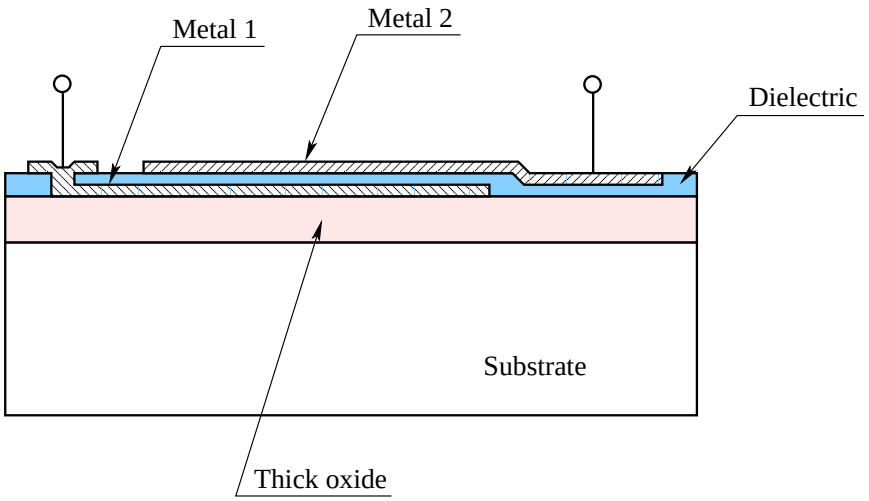


Figure 2.28: Cross section of a MIM capacitor. A thick oxide (SiO_2) layer is typically deposited on the substrate to reduce the parasitic capacitance to the substrate.

temperature variations [96], [97], [98] 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 to $10\,\mathrm{fF}/\mu\mathrm{m}^2$) have been fabricated using Al_2O_3 and $AlTiO_x$ dielectrics [99], $AlTaO_x$ [100], and HfO_2 dielectric using atomic layer deposition (ALD) [101]. A higher capacitance density ($13\,\mathrm{fF}/\mu\mathrm{m}^2$) is achieved using laminate ALD HfO_2–Al_2O_3 dielectrics [102],

[103]. 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 [104]. A capacitance density of $17\,\mathrm{fF}/\mu\mathrm{m}^2$ is achieved using 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 ($\leq 400\,^{\circ}\mathrm{C}$) back-end high complexity digital process is therefore a challenging problem [105]. This problem can be overcome by utilizing MIM capacitors with plasma enhanced chemical vapor deposition (PECVD) nitride dielectrics [106], [107]. 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 [108], [109], [110]. In [111], 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 [112]. 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 $t_{ox} = 1\,\text{nm}$, a capacitance density of $34.5\,\text{fF}/\mu\text{m}^2$ has been achieved [113].

2.4.4 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. 2.29. The two terminals of the capacitor are shown in light grey and dark grey. 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 [114], as illustrated in Fig. 2.30. Note that a higher capacitance density is achieved by using a lateral flux together with a vertical flux (parallel plate structure).

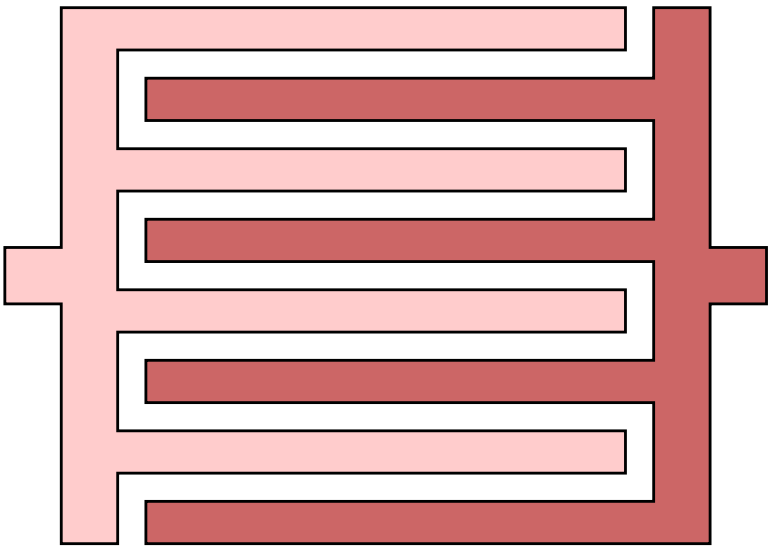


Figure 2.29: A simplified structure of an interdigitated lateral flux capacitor (top view). Two terminals of the capacitor are shown in light grey and dark grey.

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 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. 2.31. 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.

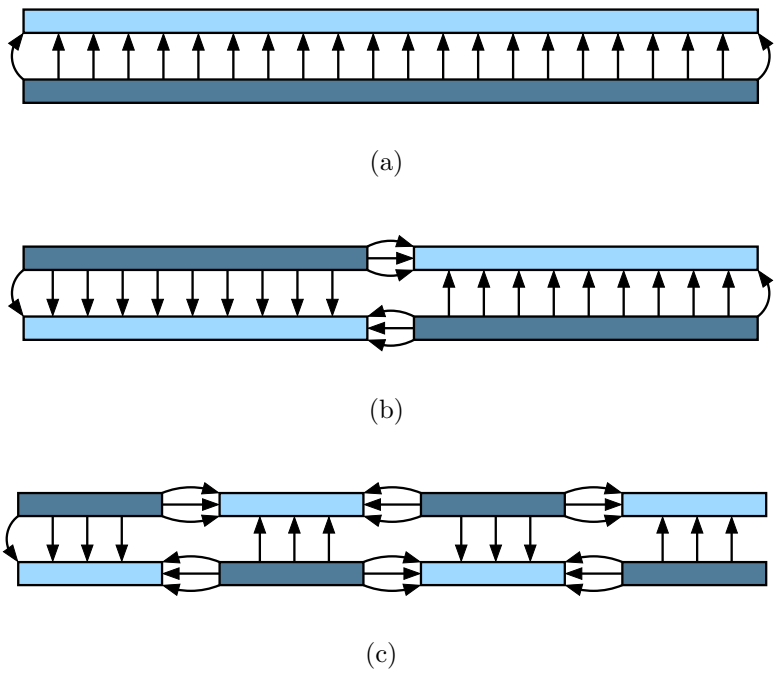


Figure 2.30: Vertical flux versus lateral flux. (a) A standard parallel plate structure, (b) divided by two cross-connected metal layers, and (c) divided by four cross-connected metal layers.

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 [115]. 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 [116]. It has been demonstrated that in certain cases, the effective capacitance of fractal capacitors can be increased by more than ten times.

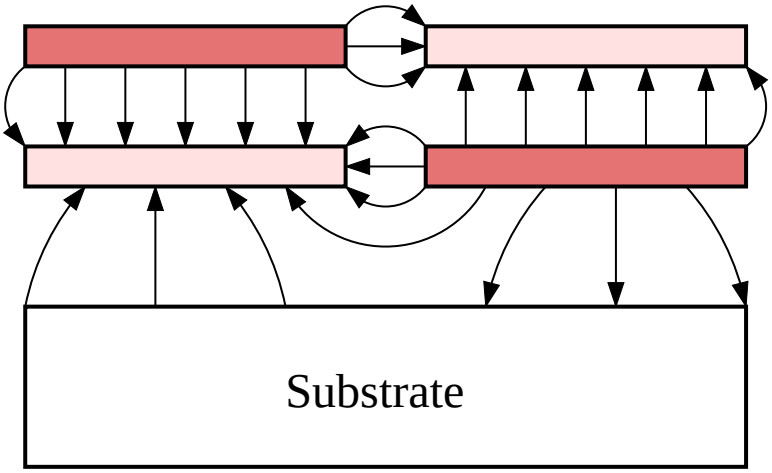


Figure 2.31: Reduction of the bottom plate parasitic capacitance through flux stealing. Shades of grey denote the two terminals of the capacitor.

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 [115]. 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 [116].

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%) [116]. 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. 2.32. 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 overlapping 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 [117]. 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.

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 [118]. A VPP capacitor consists of metal slabs connected vertically using multiple vias between the vertical plates. This structure fully exploits the lateral scaling trends as compared to fractal structures [119].

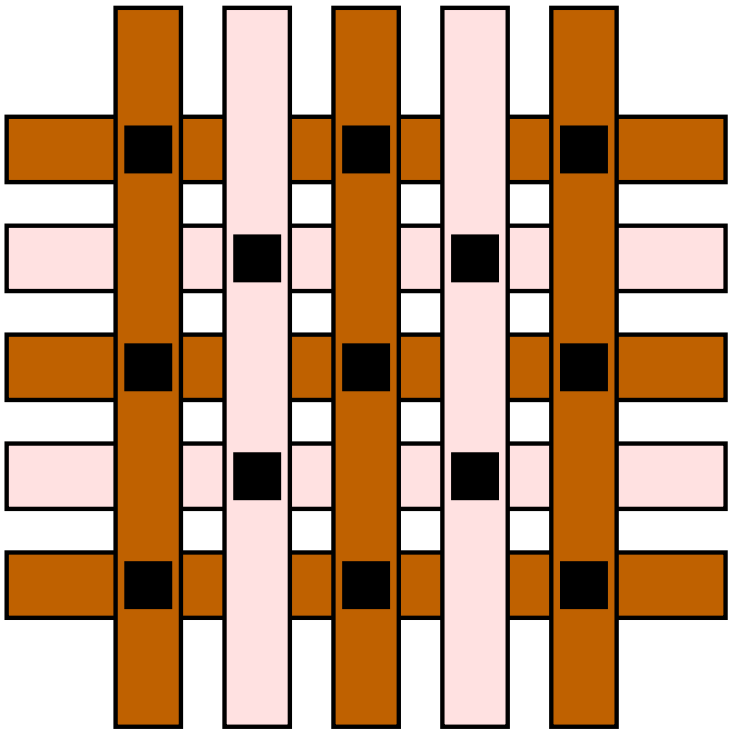


Figure 2.32: Woven capacitor. The two terminals of the capacitor are shown in light grey and dark grey. The vias are illustrated by the black colored squares.

2.4.5 Comparison of On-Chip Decoupling Capacitors

On-chip decoupling capacitors can be implemented in ICs in a number of ways. The primary characteristics of four common types of on-chip decoupling capacitors, discussed in Sections 2.4.1 – 2.4.4, are listed in Table 2.1. Note that typical MIM capacitors provide a lower capacitance density ($1\,\mathrm{fF}/\mu\mathrm{m}^2 - 10\,\mathrm{fF}/\mu\mathrm{m}^2$) than MOS capacitors. Recently, a higher capacitance density ($13\,\mathrm{fF}/\mu\mathrm{m}^2$) of MIM capacitors has

been achieved using laminate ALD HfO_2–Al_2O_3 dielectrics [102], [103]. A capacitance density of $34.5\,fF/\mu m^2$ has been reported in [113] for a MIM capacitor with a dielectric thickness of 1 nm.

Table 2.1: 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 ($fF/\mu m^2$)	1 – 5	10 – 20	1 – 30	10 – 20
Bottom plate capacitance (%)	5 – 10	20 – 30	2 – 5	1 – 5
Linearity (ppm/volt)	50 – 150	300 – 500	10 – 50	50 – 100
Quality factor	5 – 15	1 – 10	50 – 150	10 – 50
Parasitic resistance ($m\Omega$)	500 – 2000	1000 – 10000	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

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 [120]. 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 [14].

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 [121]. 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 [119]. 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 2.1 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. 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 [108].

Note that PIP capacitors typically require additional process steps, adding extra cost. From the information listed in Table 2.1, 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.

2.5 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. 2.33. 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.

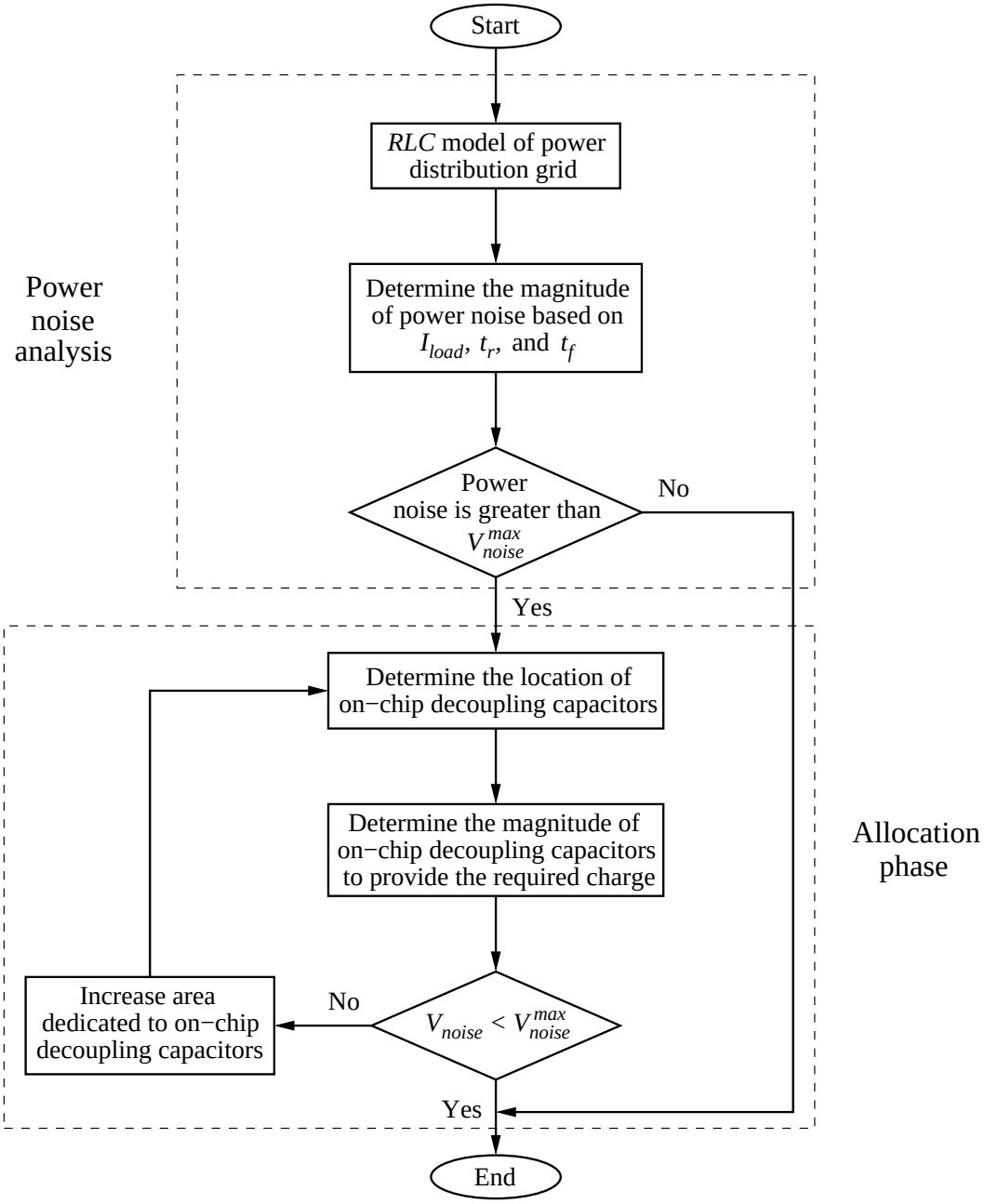


Figure 2.33: Flow chart for allocating on-chip decoupling capacitors.

In some cases, large functional blocks should be partitioned, permitting the allocation 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 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 Section 2.5.1. An allocation strategy based on an excessive noise amplitude is described in Section 2.5.2. An allocation strategy based on excessive charge is discussed in Section 2.5.3.

2.5.1 Charge-Based Allocation Methodology

One of the first approaches is based on the average power current drawn by a circuit block [122]. 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, *i.e.*, to release charge $\delta Q_i = \frac{I_i^{avg}}{f_{clk}}$ as the power voltage level varies by a noise margin δV_{dd},

$$C_i^{dec} = \frac{\delta Q_i}{\delta V_{dd}} = \frac{I_i^{avg}}{f_{clk}\,\delta V_{dd}}, \tag{2.21}$$

where f_{clk} is the clock frequency.

The rationale behind the approach represented by (2.21) 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 (2.21) 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 (2.21). 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 (2.21) insufficient. This inconsistency is largely responsible for the unrealistic dependence of the decoupling capacitance as determined by (2.21) 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 [123], [124], [125]. The characteristics of the power distribution network based on (2.21) are simulated 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 the microprocessor functional units are estimated based on the average power current obtained with 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.

2.5.2 Allocation Strategy Based on the Excessive Noise Amplitude

More aggressive capacitance budgeting is proposed in [126], [127] to amend the allocation strategy described by (2.21). 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 decoupling capacitance is allocated to those blocks where the power noise target specifications have already been achieved. Alternatively, 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 decoupling capacitance is

$$C_{dec} = \frac{V_{noise} - \delta V_{dd}}{V_{noise}} \frac{\delta Q}{\delta V_{dd}}, \tag{2.22}$$

where δQ is the charge drawn from the power distribution system by the current load during a single clock period.

The rationale behind (2.22) is that in order to reduce the power noise from V_{noise} to δV_{dd} (*i.e.*, by a factor of $\frac{V_{noise}}{\delta V_{dd}}$), the capacitance C_{dec} should supply a $1 - \frac{\delta V_{dd}}{V_{noise}}$ share of the total current. Consequently, the same share of charge as the power voltage is decreased by δV_{dd}, making $C_{dec}\,\delta V_{dd} = \frac{V_{noise} - \delta V_{dd}}{V_{noise}}\,\delta Q$. 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 decoupling effect of the on-chip intrinsic decoupling capacitance and the off-chip decoupling capacitors [11].

The efficacy of a capacitance allocation methodology based on (2.22) has been tested on five MCNC benchmark circuits [128]. For a $0.25\,\mu$m CMOS technology, the proposed methodology requires, on average, 28% lower overall decoupling capacitance as compared to the more conservative allocation methodology based on (2.21) [122]. A noise aware floorplanning methodology based on this allocation strategy has also been developed [128]. The noise aware floorplanning methodology results, 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.

2.5.3 Allocation Strategy Based on Excessive Charge

The allocation strategy presented in Section 2.5.2 can be further refined. Note that (2.22) 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 proposed in [129], [130] as a more accurate metric characterizing the severity of the noise margin violation. According to this approach, the metric of the ground supply quality at node j is

$$M_j = \int_0^T \max\left[\left(V_j^{gnd}(t) - \delta V\right), 0\right]\, dt, \tag{2.23}$$

or, assuming a single peak noise violates the noise margin between times t_1 and t_2,

$$M_j = \int_{t_1}^{t_2} \left(V_j^{gnd}(t) - \delta V\right)\, dt, \tag{2.24}$$

where $V_j^{gnd}(t)$ is the ground voltage at node j of the power distribution grid.

Worst case switching patterns are used to calculate (2.23) and (2.24). This metric is illustrated in Fig. 2.34. The value of the integral in (2.24) 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,

$$M = \sum_j M_j. \tag{2.25}$$

This metric becomes zero when the power noise margins are satisfied at all times throughout the circuit.

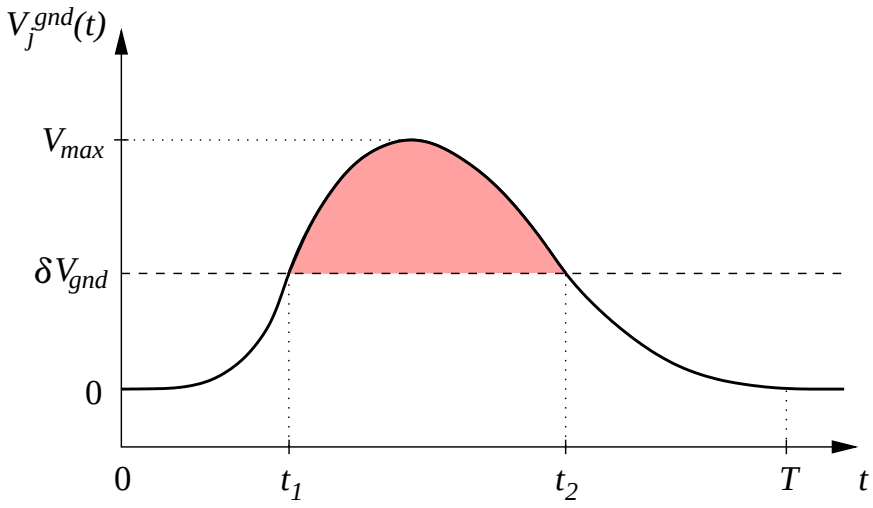


Figure 2.34: Variation of ground supply voltage with time. The integral of the excess of the ground voltage deviation over the noise margin δV_{gnd} (the shaded area) is used as a quality metric to guide the process of allocating the decoupling capacitors.

Application of (2.23) and (2.25) to the decoupling capacitance allocation process requires a more complex procedure as compared to (2.21) and (2.22). Note that utilizing (2.23) 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 (2.25) 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 (2.25). The integral formulation expressed by (2.23) is, fortunately, amenable to efficient optimization algorithms. The primary motivation for the original integral formulation of the 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 (2.25) in application-specific ICs has been demonstrated in [131], [132]. 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 rows of cells. 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 proposed methodology results in a significant reduction in the number of circuit nodes exhibiting noise margin violations and a significant reduction in the maximum power supply noise.

2.6 Chapter 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
- 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

Chapter 3

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 (SoC) and systems-in-package (SiP) technologies, the problem of heat removal has worsened further. 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 personal digital assistants (PDA), wireless communications, and imaging systems that demand high speed computation and complex functionality while dissipating as little power as possible [133]. 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 integrated circuit.

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 [134]. Reducing the supply voltage, however, increases the circuit delay. Chandrakasan *et al.* demonstrated in [135] 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 [136].

Power consumption can also be reduced by scaling the threshold voltage V_{th} while simultaneously reducing V_{dd} [137]. This approach, however, results in significantly increased standby leakage current. To limit the leakage current during sleep mode, several techniques have been proposed, such as multi-threshold voltage CMOS [138], [139], variable threshold voltage schemes [140], [141], and circuits with an additional transistor behaving as a sleep switch [142]. These techniques, however, require additional process steps and/or additional circuitry to control the substrate bias or switch off portions of the circuit [141].

The total power dissipation can also be reduced by utilizing multiple power supply voltages [143], [144]. 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 [139]. 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 Section 3.1. Challenges to ICs with multiple supply voltages are discussed in Section 3.2. Choosing the optimum number and magnitude of the multi-voltage power supplies is discussed in Section 3.3. Some conclusions are offered in Section 3.4.

3.1 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 power supply voltage V_{dd}^{L} is provided to those gates to reduce power. Note that in practical applications, the number of critical paths is only a small portion of the total number of paths in a circuit. Excess slack exists in the majority of paths within a circuit. Determining those gates with excessive time slack is therefore an important and complex task [139]. 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 [145], [146]. Multi-voltage low power techniques are reviewed in this section. A low power technique with multiple power supply voltages is presented in Section 3.1.1. Clustered voltage scaling (CVS) is presented in Section 3.1.2. Extended clustered voltage scaling (ECVS) is discussed in Section 3.1.3.

3.1.1 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. 3.1. 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 all of the gates along the non-critical delay paths is replaced by a lower supply voltage V_{dd}^L is illustrated in Fig. 3.2. 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 [147].

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

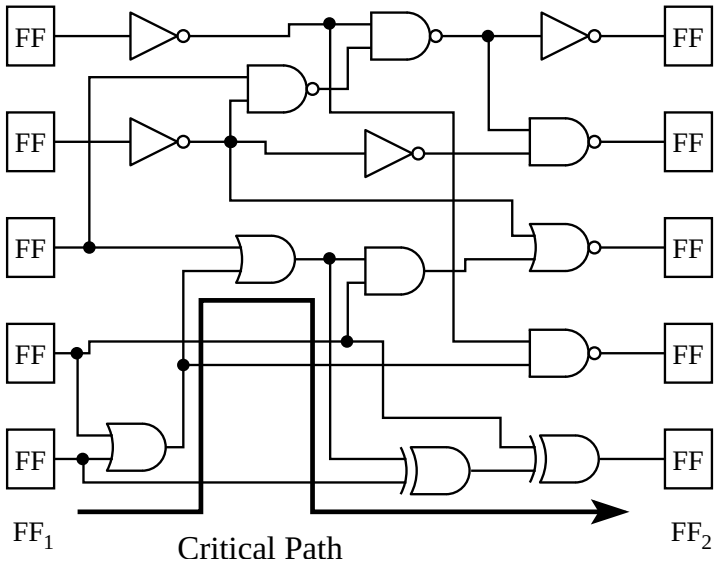


Figure 3.1: An example single supply voltage circuit.

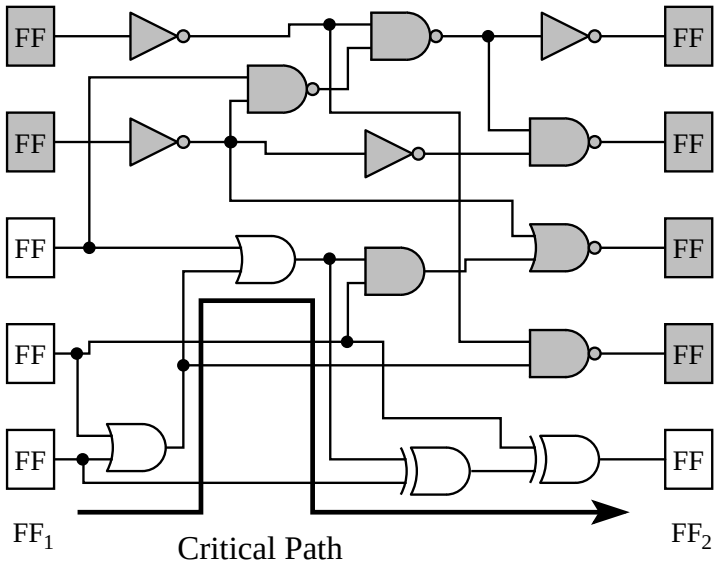


Figure 3.2: An example dual supply voltage circuit. The gates operating at a lower power supply voltage V_{dd}^{L} (located off the critical delay path) are shaded.

sufficiently high to turn off the PMOS device, as shown in Fig. 3.3. The PMOS device in the high voltage gate is therefore weakly "ON," 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.

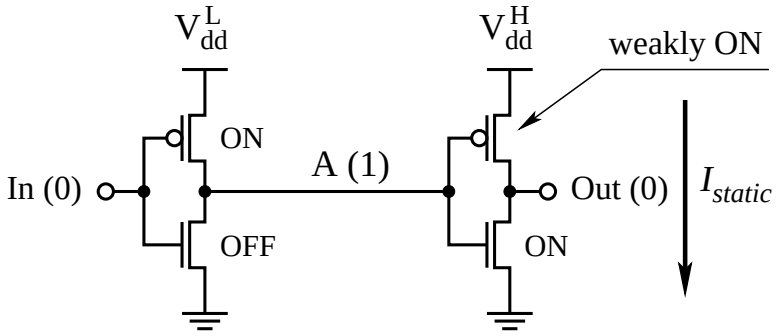


Figure 3.3: Static current as a result of a direct connection between the V_{dd}^L gate and the V_{dd}^H gate.

Level converters are typically inserted at node A to remove the static current path [148]. A simple level converter circuit is illustrated in Fig. 3.4. 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 the number of level converters are minimized while satisfying the system-level timing constraints [149].

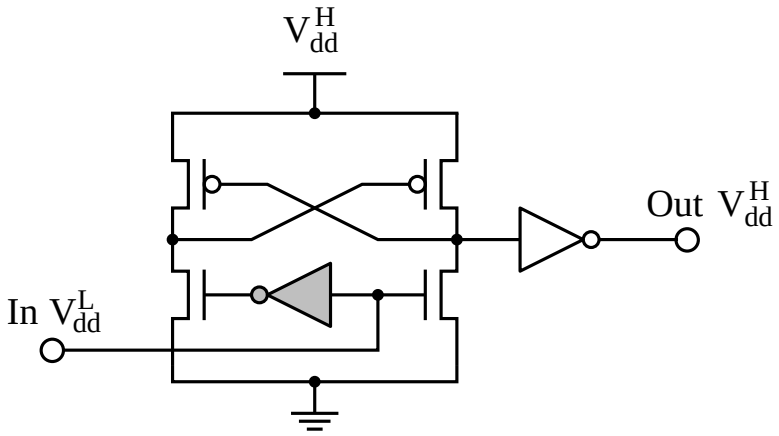


Figure 3.4: Level converter circuit. The inverter operating at the reduced power supply voltage V_{dd}^{L} is shown in grey.

3.1.2 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, proposed in [150], results in a circuit structure with a greatly reduced number of level converters, as shown in Fig. 3.5.

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. 3.5 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.

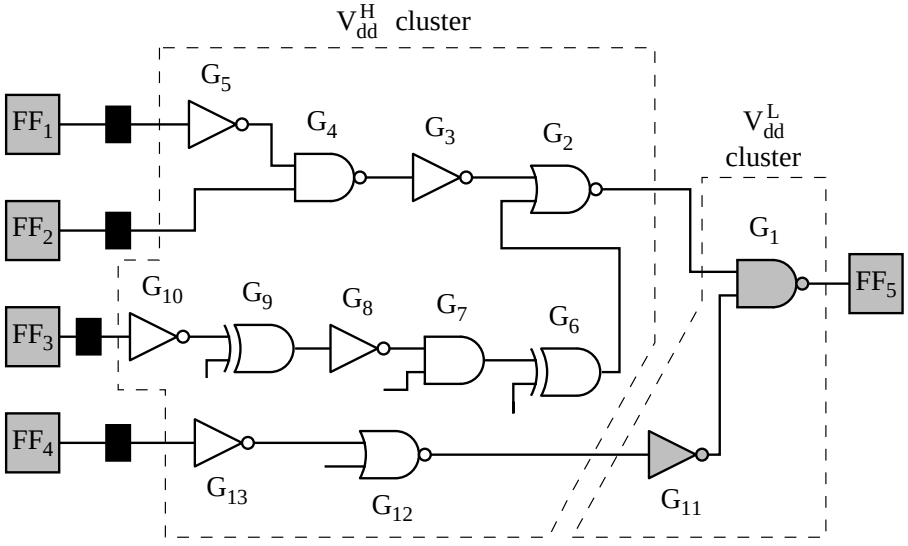


Figure 3.5: A dual power supply voltage circuit with the clustered voltage scaling (CVS) technique [150]. The gates operating at a lower supply voltage are shaded. The level converters are shown as black rectangles.

3.1.3 Extended Clustered Voltage Scaling (ECVS)

The number of gates with a lower power supply voltage can be increased by optimizing the insertion points of the level converters, further reducing overall power.

As an example, in the CVS structure shown in Fig. 3.5, 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 Section 3.1.2, 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. 3.6. Note that the structure shown in Fig. 3.6 is obtained from the CVS network by relaxing any limitations on the insertion positions of the level converters. Such a technique is often referred to as the extended clustered voltage scaling technique [149], [151].

3.2 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

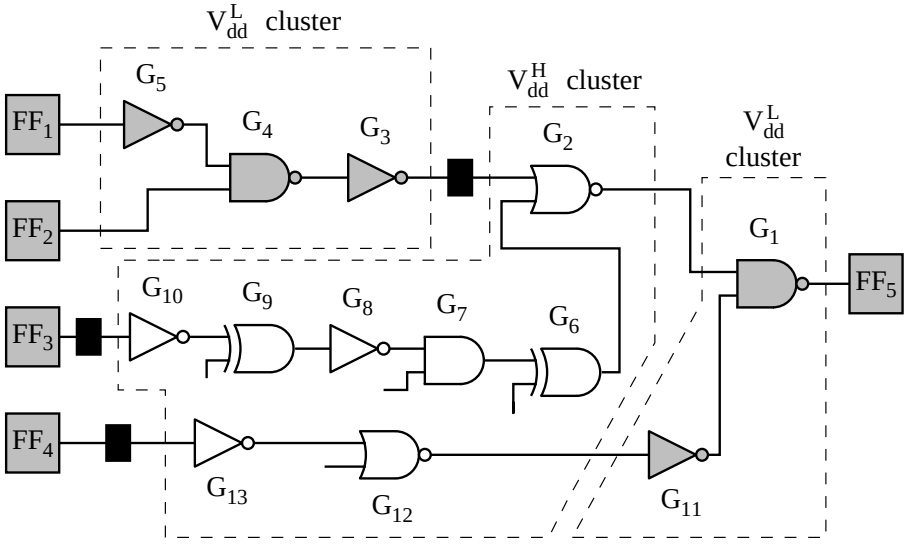


Figure 3.6: A dual power supply voltage circuit with the extended clustered voltage scaling (ECVS) technique [149]. The gates operating at a lower supply voltage are shaded. The level converters are shown as black rectangles.

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 introduced in Section 3.2.1. Power penalties are presented in Section 3.2.2. The additional design complexity associated with level converters and integrated DC–DC voltage converters is discussed in Section 3.2.3. Several placement and routing strategies are described in Section 3.2.4.

3.2.1 Die Area

As described in Section 3.1, 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.

3.2.2 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 Section 3.3. 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 [152], [153]. 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 [154].

3.2.3 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 [148], [155]. Standard logic gates with embedded level

conversion as reported in [155] 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 input/output (I/O) pads dedicated to power delivery, and reduce fabrication costs [156]. 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 [157]. 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.

3.2.4 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 [149], as shown in Fig. 3.7. 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 proposed in [158] is illustrated in Fig. 3.8. 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 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

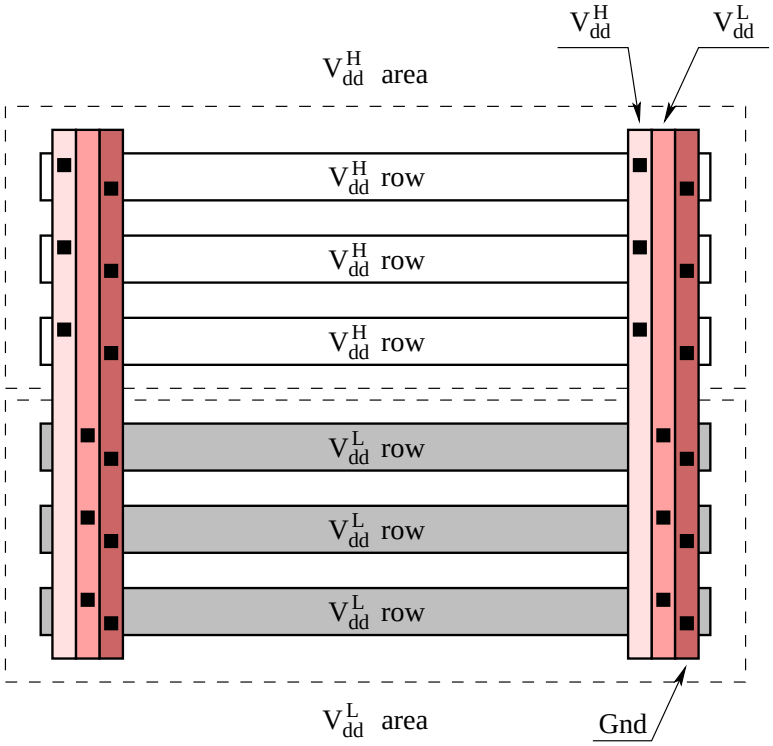


Figure 3.7: Layout of an area-by-area architecture with a dual power supply voltage. In this architecture, the V_{dd}^L cells are placed in one area, while the V_{dd}^H cells are separately placed in a different area.

row-by-row 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.

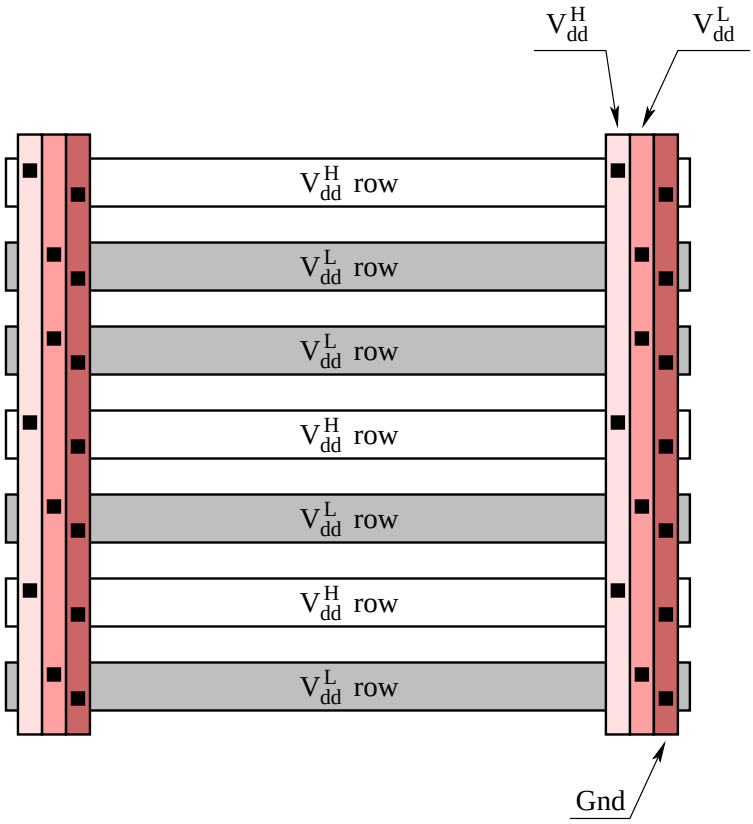


Figure 3.8: Layout of a row-by-row architecture with a dual power supply voltage. In this architecture, the V_{dd}^{L} cells and V_{dd}^{H} cells are placed in different rows. Each row consists of only V_{dd}^{L} cells *or* V_{dd}^{H} cells.

In-Row Architecture

An improved row-by-row layout architecture is presented in [159]. This architecture is based on a modified cell library [159]. 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. 3.9. 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 increasing the overall area (a 2.7% area overhead as compared to the original cell) [159]. 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 [159]. 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.

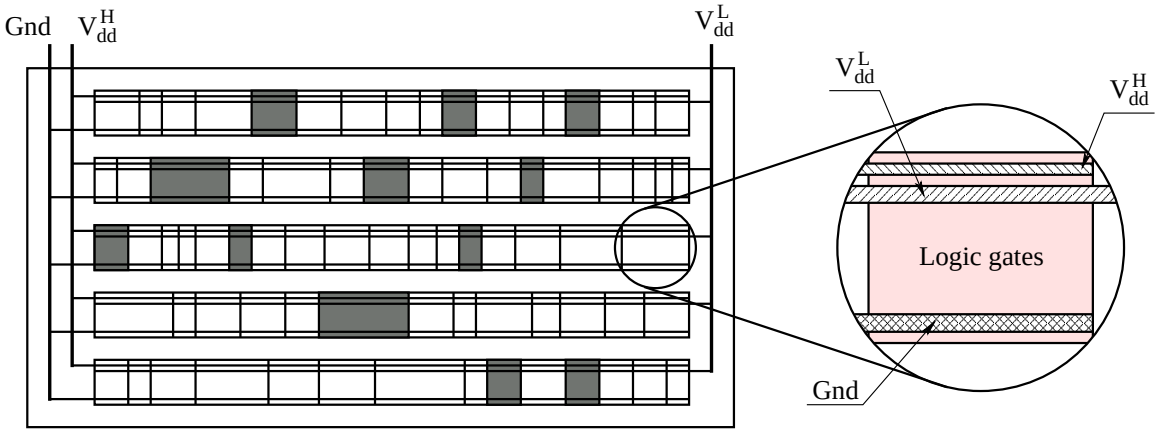


Figure 3.9: In-row dual power supply voltage scheme. This architecture is based on a modified cell library with two power rails and one ground rail in each cell. The V_{dd}^{H} cells are shown in grey and the V_{dd}^{L} cells are white.

3.3 Optimum Number and Magnitude of Available Power Supply Voltages

In low power techniques with multiple power supply voltages, any 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. 3.10. Observe from Fig. 3.10 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. 3.10). If more than the optimum number of power supplies are used, the savings in power becomes smaller, as depicted in Fig. 3.10 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 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.

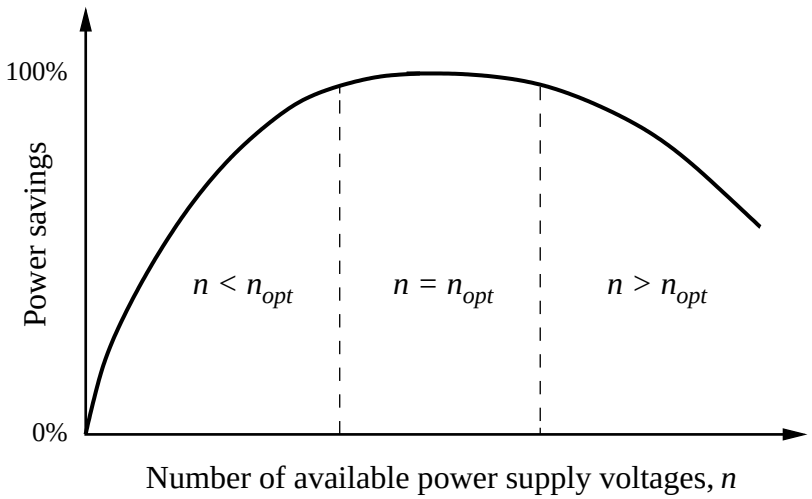


Figure 3.10: Trend in power reduction with multi-voltage scheme as a function of the number of available supply voltages.

In systems with multiple power supply voltages (where $V_1 > V_2 > \cdots > V_n$), the power dissipation is [160]

$$P_n = f \left\{ \left(C_1 - \sum_{i=2}^{n} C_i \right) V_1^2 + \sum_{i=2}^{n} C_i V_i^2 \right\}, \tag{3.1}$$

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

$$K_{V_{dd}} \equiv \frac{P_n}{P_1} = 1 - \sum_{i=2}^{n} \left[\left(\frac{C_i}{C_1} \right) \left\{ 1 - \left(\frac{V_i}{V_1} \right)^2 \right\} \right]. \tag{3.2}$$

Since delay is proportional to the total capacitance, $\frac{C_i}{C_1}$ is

$$\frac{C_i}{C_1} = \frac{\int_0^1 p(t)\, t_i\, dt}{\int_0^1 p(t)\, t\, dt}, \tag{3.3}$$

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

$$t_i = \begin{cases} \dfrac{t_{i,0}}{t_{i,0} - t_{i+1,0}}(t - t_{i+1,0}) & : \quad t_{i+1,0} \le t \le t_{i,0} \\ \dfrac{t_{i,0}}{t_{i-1,0} - t_{i,0}}(t_{i-1,0} - t) & : \quad t_{i,0} \le t \le t_{i-1,0}, \end{cases} \tag{3.4}$$

where $t_{i,0}$ is

$$t_{i,0} = \left(\frac{V_1}{V_i} \right) \left(\frac{V_i - V_{th}}{V_1 - V_{th}} \right)^{\alpha}, \tag{3.5}$$

V_{th} is the threshold voltage, and α is the velocity saturation index [161]. Note that

$t_{n+1,0} = 0$. $K_{V_{dd}}$ can be determined from (3.1) – (3.5) for a specific $p(t)$, V_1, V_i, and V_{th}.

For a lambda-shaped normalized path delay distribution function $p(t)$ (see Fig. 3.11) as determined from post-layout static timing analysis, approximate rules of thumb for determining the optimum magnitude of power supply voltages have been determined by Hamada *et al.* [160],

$$\text{for } \{V_1, V_2\} \qquad \frac{V_2}{V_1} = 0.5 + 0.5\frac{V_{th}}{V_1}, \tag{3.6}$$

$$\text{for } \{V_1, V_2, V_3\} \qquad \frac{V_2}{V_1} = \frac{V_3}{V_2} = 0.6 + 0.4\frac{V_{th}}{V_1}, \tag{3.7}$$

$$\text{for } \{V_1, V_2, V_3, V_4\} \qquad \frac{V_2}{V_1} = \frac{V_3}{V_2} = \frac{V_4}{V_3} = 0.7 + 0.3\frac{V_{th}}{V_1}. \tag{3.8}$$

Criteria (3.6) – (3.8) 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. 3.10). As determined in [160], 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 $\frac{V_{th}}{V_{dd}}$ increases.

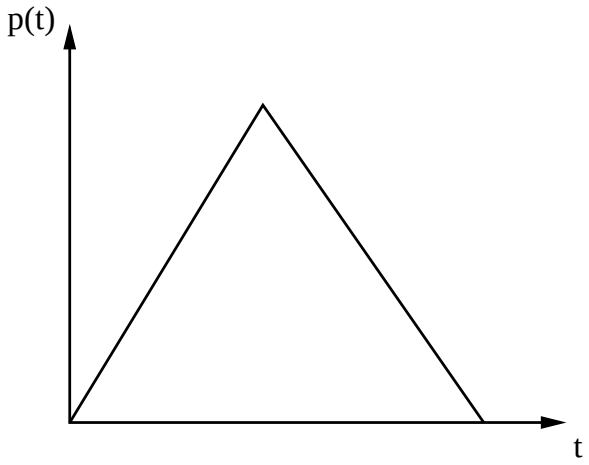


Figure 3.11: A lambda-shaped normalized path delay distribution function.

A rule of thumb for two power supply voltages has been evaluated by simulations in [149]. For $V_{dd}^H = 3.3$ volts, a V_{dd}^L of 1.9 volts has been estimated, exhibiting good agreement with (3.6). 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. 3.12. Observe from Fig. 3.12 that the minimum overall power is achieved at $V_{dd}^L = 1.9$ volts.

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

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. 3.12.

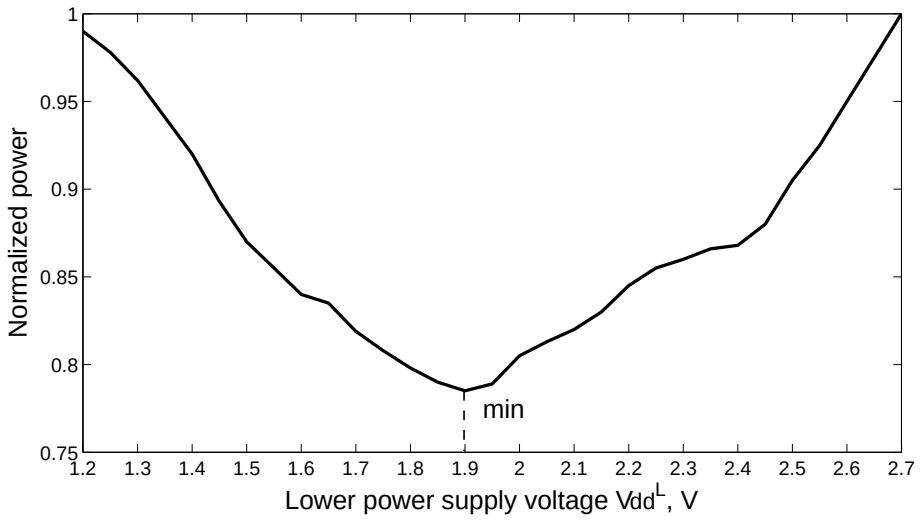


Figure 3.12: Dependence of the total power of a dual power supply system on a lower power supply voltage V_{dd}^L [149]. The original high power supply voltage $V_{dd}^H = 3.3$ volts.

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 [139], [162], [163]. 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.* [164] demonstrated power reductions approaching 32% on average (57% maximum) for the ISCAS85 benchmark circuits. CVS with variable supply voltage schemes has been presented in [165]. 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 [166], 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 in high complexity ICs. Multiple low power techniques should therefore be utilized to achieve significant power savings in modern nanoscale ICs.

3.4 Chapter Summary

The discussion of multiple on-chip power supply systems and different low power 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 4

On-Chip Power Distribution Grids with Multiple Supply Voltages for High Performance Integrated Circuits

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 [8]. The higher

currents produce large IR voltage drops. Fast current transients lead to large $L\frac{dI}{dt}$ inductive voltage drops (ΔI noise) within the power distribution networks.

The lower voltage of the power supply level can be described as

$$V_{load} = V_{dd} - IR - L\frac{dI}{dt}, \tag{4.1}$$

where V_{load} is the voltage level seen from 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 $\pm 5\%$ 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. To manage the problem of high power dissipation, multiple on-chip power supply voltages have become commonplace [150]. This strategy has the advantage of permitting those 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 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 investigated in the area of high level synthesis for low power [146], [167]. 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 [139].

Power distribution networks in high performance ICs are commonly structured as a multi-layer grid [11]. 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. 4.1, where three layers of interconnect are depicted with the power lines shown in dark grey and the ground lines shown in light grey.

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. The chapter is organized as follows. Existing work on power distribution grids and related power distribution

systems with multiple supply voltages is reviewed in Section 4.1. The structure of a power distribution grid and the simulation setup are reviewed in Section 4.2. The structure of a power distribution grid with dual supply voltages and dual grounds (DSDG) is discussed in Section 4.3. Interdigitated power distribution grids with DSDG are described in Section 4.4. Paired power distribution grids with DSDG are analyzed in Section 4.5. Simulation results are presented in Section 4.6. Circuit design implications are discussed in Section 4.7. Some specific conclusions are summarized in Section 4.8.

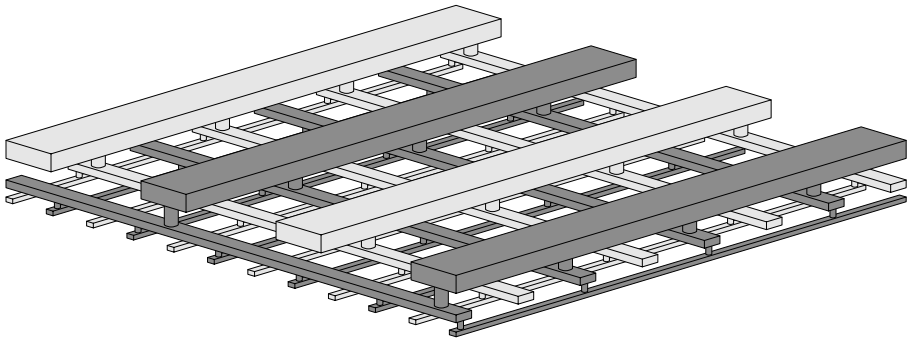


Figure 4.1: A multi-layer on-chip power distribution grid [168]. The ground lines are light grey, the power lines are dark grey. The signal lines are not shown.

4.1 Background

On-chip power distribution grids have traditionally been analyzed as purely resistive networks [169]. 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 is shown to produce 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 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 [170], 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 [171] 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 proposed 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 [158]. The original circuit is partitioned into two subcircuits by conventional layout methods. Another technique,

presented in [159], 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 [149]. 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 proposed in this chapter.

4.2 Simulation Setup

The inductance extraction program FastHenry [172] is used to analyze the inductive properties of the 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 magneto-quasistatic 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 $(1.72\,\mu\Omega \cdot \mathrm{cm})^{-1}$. A line thickness of $1\,\mu$m 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. 4.2. 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 five minutes on a Sun Blade 100 workstation.

4.3 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 [11], [14], [120], [149], [158], [159]. 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

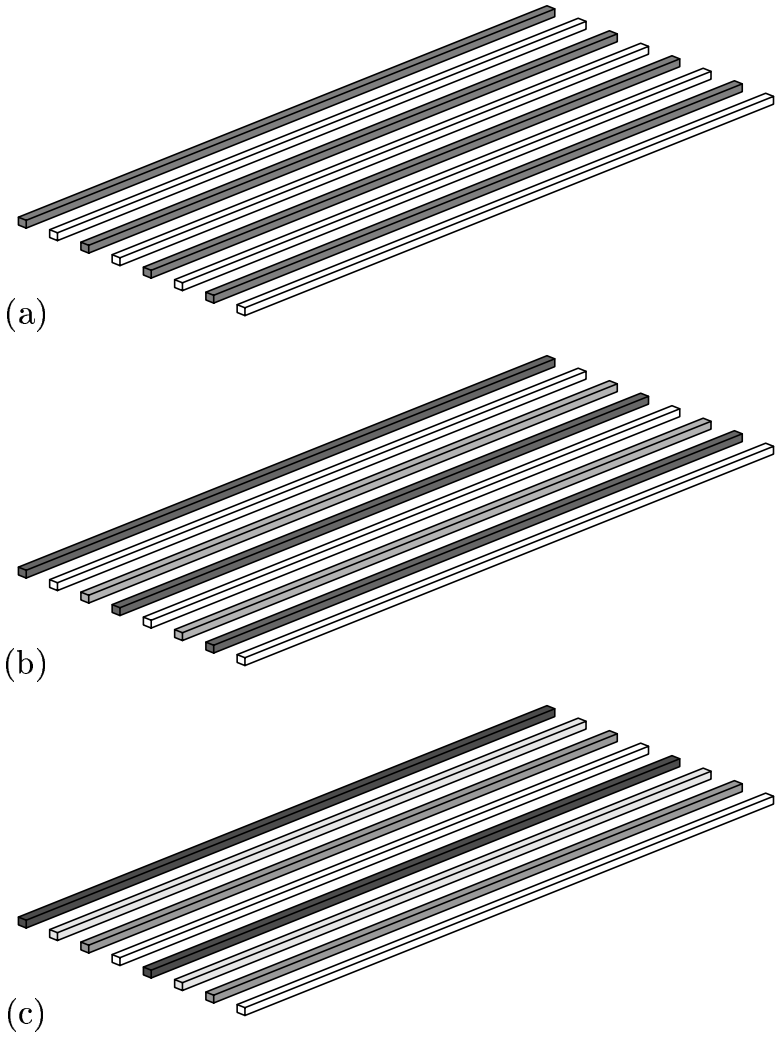


Figure 4.2: Interdigitated power distribution grids under investigation. In all of the power distribution structures, the power lines are interdigitated with the ground lines. (a) A reference power distribution grid with a single supply voltage and a single ground (SSSG). The power lines are grey colored and the ground lines are white colored, (b) a power distribution grid with DSSG. The power lines are light and dark grey colored and the ground lines are white colored, (c) the proposed power distribution grid with DSDG. The power lines are shown in black and dark grey colors and the ground lines are shown in white and light grey colors.

result, voltage fluctuations are significantly increased. To address this problem, an on-chip power distribution scheme with DSDG is proposed. 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. 4.3.

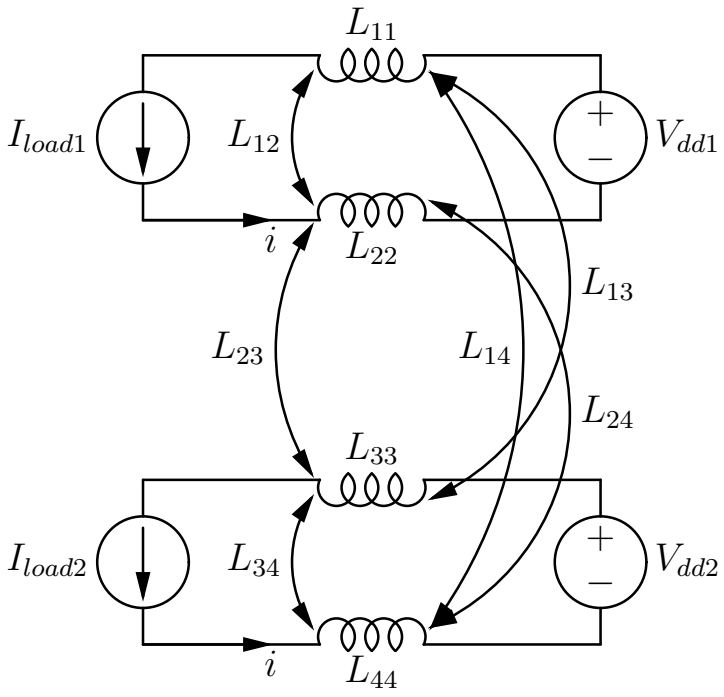


Figure 4.3: Circuit diagram of the mutual inductive coupling of the proposed power distribution grid. L_{11} and L_{33} denote the partial self-inductances of the power lines and L_{22} and L_{44} denote the partial self-inductances of the ground lines, respectively.

The loop inductance of the current loop formed by the two parallel paths is

$$L_{loop} = L_{pp} + L_{gg} - 2M, \tag{4.2}$$

where L_{pp} and L_{gg} are the partial self-inductances 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 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 [173], the mutual inductance between two parallel straight lines of equal length is

$$M_{loop} = 0.2l \left(\ln\frac{2l}{d} - 1 + \frac{d}{l} - \ln\gamma + \ln k \right) \mu\text{H}, \tag{4.3}$$

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 [11].

Analogous to inductive coupling between two parallel loop segments as described in [174], the mutual loop inductance of the two power distribution grids with DSDG

is

$$M_{loop} = L_{13} - L_{14} + L_{24} - L_{23}. \tag{4.4}$$

Note that the two negative signs before the mutual inductance components in (4.4) correspond to the current in the power and ground paths flowing in opposite directions. Also note that since the mutual inductance M in (4.2) is negative, the 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. 4.2(c), the distance between the lines d_{12} is the same as an offset between the two loops d_{23}, as illustrated in Fig. 4.4. In this case, assuming $d_{12} = d_{23} = d$, from (4.3), M_{loop} between the two grids is approximately

$$M_{loop} = 0.2l \ln\frac{3}{4}\ \mu\text{H}. \tag{4.5}$$

Thus, the 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,

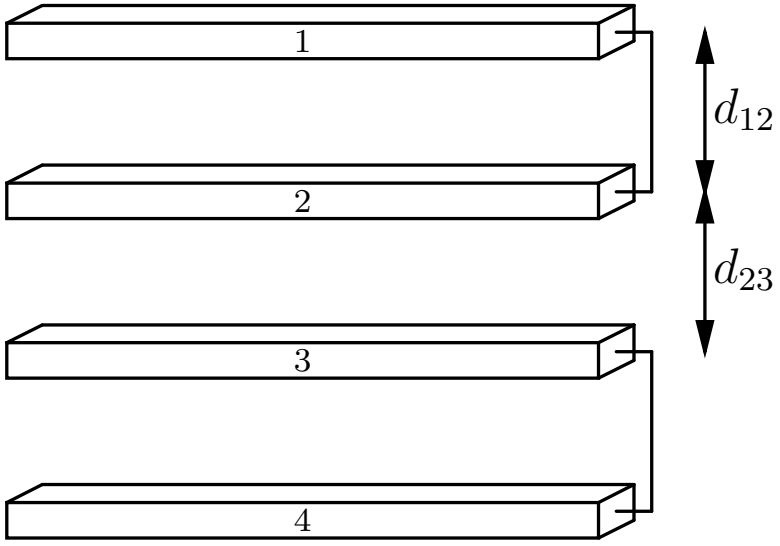


Figure 4.4: Physical structure of an interdigitated power distribution grid with DSDG. The proposed power delivery scheme consists of two independent power delivery networks.

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

$$L_{||} = \frac{L_{pp}^1 L_{pp}^2 - M^2}{L_{pp}^1 + L_{pp}^2 - 2M}, \tag{4.6}$$

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 $L\frac{dI}{dt}$ voltage fluctuations.

4.4 Interdigitated Grids with DSDG

As shown in Section 4.3, 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 (4.2) includes two terms. One term accounts for the increase (or decrease) in the mutual coupling between the power and ground paths in the 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

$$M = M' + M_{loop}, \tag{4.7}$$

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 [175]. 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 Section 4.4.1. The loop inductance in the second type of power distribution grids is discussed in Section 4.4.2.

4.4.1 Type I Interdigitated Grids with DSDG

In the first type of interdigitated power distribution grids, 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. 4.5. 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 (4.4), the mutual inductive coupling of two current loops in fully interdigitated grids with DSDG is

$$M_{loop}^{intI} = L_{Vdd1-Vdd2} - L_{Vdd1-Gnd2} + L_{Gnd1-Gnd2} - L_{Vdd2-Gnd1}, \tag{4.8}$$

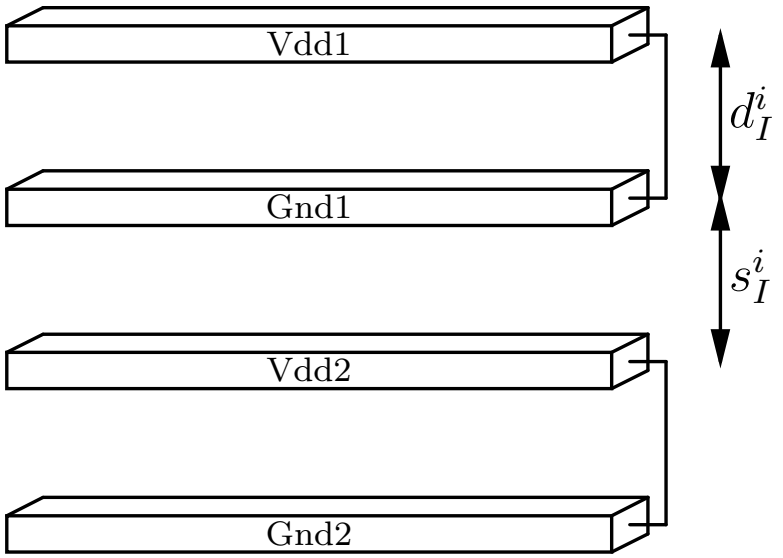


Figure 4.5: Physical structure of a fully interdigitated power distribution grid with DSDG. The distance between the lines making the loops d_I^i is equal to the separation between the two loops s_I^i.

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,

$$|L_{Vdd1-Gnd2}| + |L_{Vdd2-Gnd1}| > |L_{Vdd1-Vdd2}| + |L_{Gnd1-Gnd2}|. \tag{4.9}$$

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 (4.8), M_{loop}^{intI} between the two grids is determined by (4.5). 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 A.

4.4.2 Type II Interdigitated Grids with DSDG

In the second type of interdigitated power distribution grids, 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. 4.6. 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 (4.8). 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 (4.8), the mutual inductive coupling between the two networks M_{loop}^{intII}

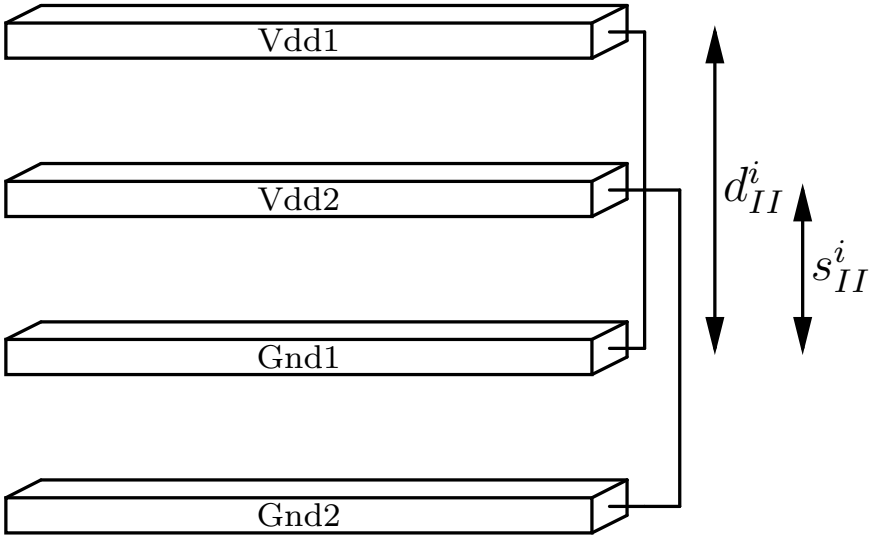


Figure 4.6: Physical structure of a pseudo-interdigitated power distribution grid with DSDG. The distance between the lines making the loops d^i_{II} is two times greater than the separation between the lines.

is

$$M^{intII}_{loop} = 0.2l\left(\ln 3 - \frac{2d}{l}\right), \tag{4.10}$$

where d is the distance between the two adjacent lines. Observe that M^{intII}_{loop} 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 B.

In modern high performance ICs, the inductive component of the power distribution noise has become comparable to the resistive noise [13]. In future nanoscale ICs, the inductive $L\frac{dI}{dt}$ 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 (4.5) to (4.10), 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 (4.7) 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. 4.7. 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.

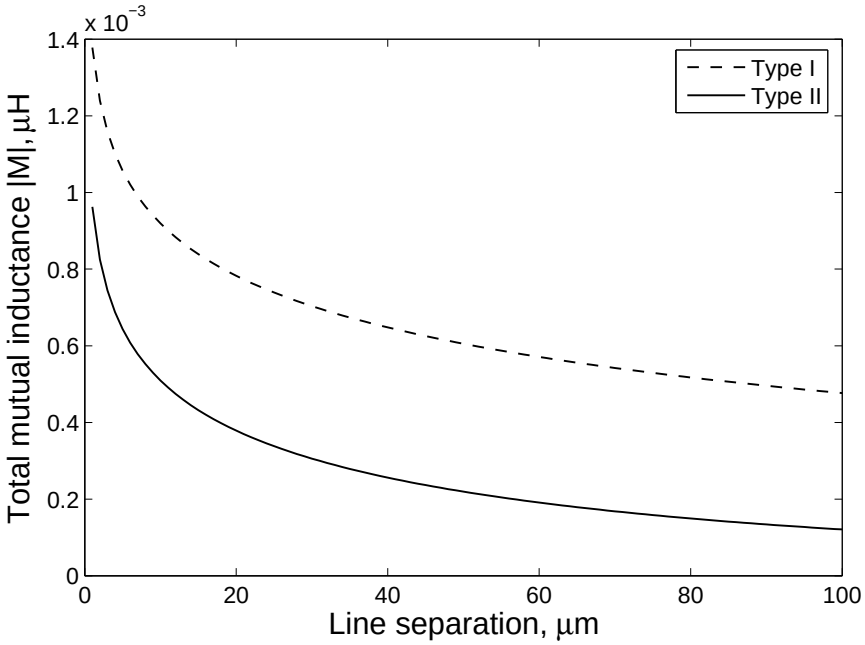


Figure 4.7: Total mutual inductance of interdigitated power distribution grids with DSDG as a function of line separation. The length of the lines is 1000 μm.

4.5 Paired Grids with DSDG

Another type of power distribution grid with alternating power and grounds lines is paired power distribution grids [11], [117]. 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 Section 4.3, 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 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 grids is described in Section 4.5.1. The loop inductance in the second type of power distribution grids is discussed in Section 4.5.2.

4.5.1 Type I Paired Grids with DSDG

In the first type of paired power distribution grids 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. 4.8. 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.

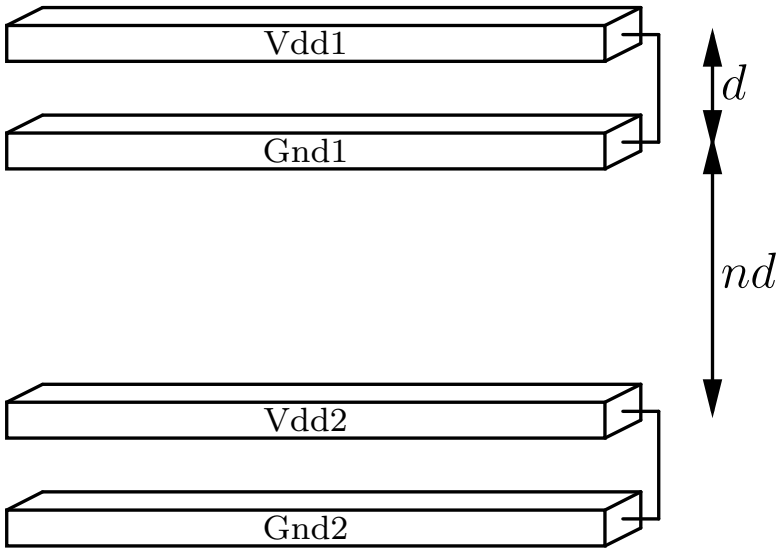


Figure 4.8: Physical structure of a fully paired power distribution grid with DSDG. In such a grid, each pair is composed of power and ground lines for a particular voltage domain. The separation between the pairs is n times larger than the distance between the lines making up the loop d.

Similar to the mutual inductance between the two loops in interdigitated power distribution grids as discussed in Section 4.4, the mutual inductive coupling of the two current loops in fully paired grids with DSDG is determined by (4.8). 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 inductances between the power and ground lines for the different voltage domains into (4.8), the mutual inductive coupling between

the two networks M_{loop}^{prdI} is

$$M_{loop}^{prdI} = 0.2l \ln \left[\frac{(n+2)n}{(n+1)^2} \right]. \tag{4.11}$$

A derivation of the mutual coupling between the two current loops in fully paired power distribution grids with DSDG is presented in Appendix C. 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 (4.3).

4.5.2 Type II Paired Grids with DSDG

In the second type of paired power distribution grids 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. 4.9. 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.

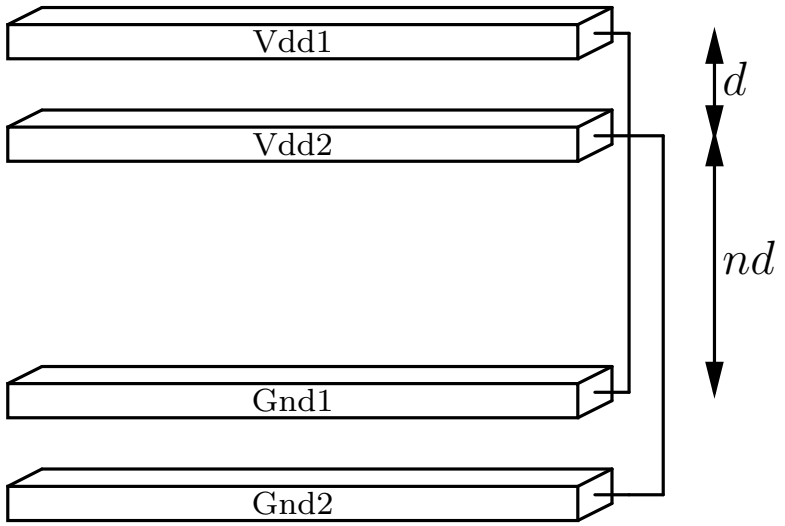


Figure 4.9: Physical structure of a pseudo-paired power distribution grid with DSDG. In such a grid, each pair is composed of power or ground lines from the two voltage domains. The separation between the pairs is n times larger than the distance between the lines making up the loop d. The effective distance between the power and ground lines in a particular power delivery network is $(n+1)d$.

As discussed in Section 4.5.1, the mutual inductive coupling between the two power delivery networks in pseudo-paired grids with DSDG is determined by (4.8). 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 inductances between the power and ground lines in the two different voltage domains into (4.8),

the mutual inductive coupling between the two networks M_{loop}^{prdII} is

$$M_{loop}^{prdII} = 0.2l \left[\ln\left(n^2 + 2n\right) - \frac{2nd}{l} \right]. \tag{4.12}$$

A derivation of the mutual coupling between the two current loops in pseudo-paired power distribution grids with DSDG is provided in Appendix D. 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,

$$M'_{prdII} = 0.2l \left[\ln\frac{2l}{(n+1)d} - 1 + \frac{(n+1)d}{l} - \ln\gamma + \ln k \right]. \tag{4.13}$$

Note that M'_{prdII} decreases with n, approaching zero for large n.

Comparing Fig. 4.8 to Fig. 4.9, note that the line separation inside each pair in the pseudo-paired power distribution grids is n times greater than the line separation between the power and ground lines making up the 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 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 (4.7) for two types of paired power distribution grids with DSDG is plotted in Fig. 4.10. 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. 4.10), 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.

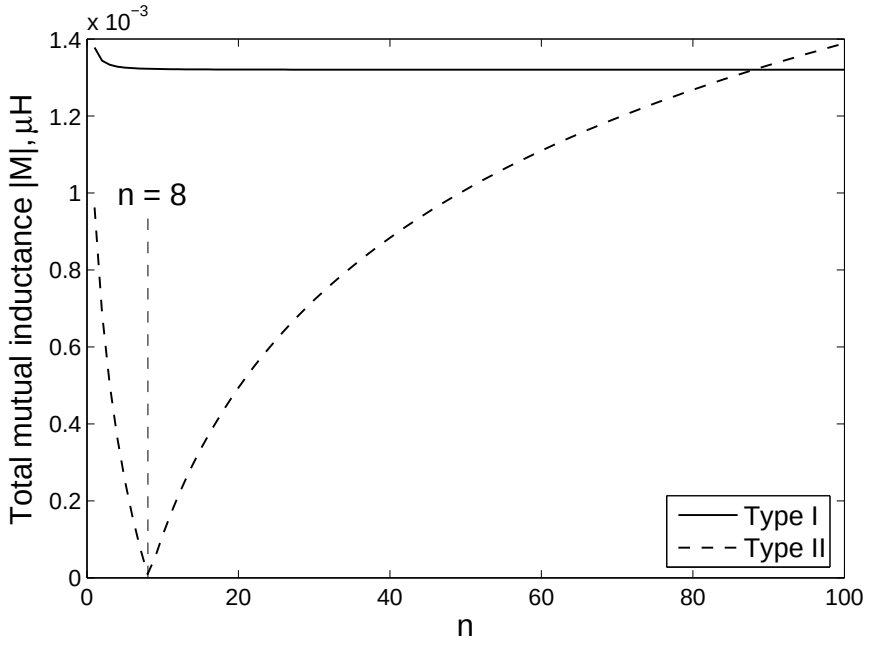


Figure 4.10: Total mutual inductance of paired power distribution grids with DSDG as a function of the ratio of the distance between the pairs to the line separation inside each pair (n). The length of the lines is $1000\,\mu$m and the line separation inside each pair d is $1\,\mu$m. Note that the total mutual inductance in pseudo-paired power distribution grids becomes zero at $n = 8$.

4.6 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 μm 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 GHZ and 100 GHz are listed in Table 4.1. The resistance and inductance for the power distribution grids with DSSG operating at 1 GHz and 100 GHz are listed in Table 4.2. Note that in the case of DSSG, only interdigitated grids can be implemented. 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 implemented. 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 Table 4.3 and 4.4, respectively. The results listed in Tables 4.1 to 4.4 are discussed in Sections 4.6.1 to 4.6.4.

Table 4.1: Impedance characteristics of power distribution grids with SSSG

Line cross section (μm × μm)	1 GHz				100 GHz			
	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 4.2: Impedance characteristics of interdigitated power distribution grids with DSSG

Line cross section (μm × μm)	R_{pp}, R_{gg} (Ω)	L^{*}_{pp}, L^{*}_{gg} (nH)	L^{*}_{pg} (nH)	k^{*}	L^{**}_{pp}, L^{**}_{gg} (nH)	L^{**}_{pg} (nH)	k^{**}
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, * denotes coupling between $V_{dd1}(V_{dd2})$ and Gnd, ** denotes coupling between V_{dd1} and V_{dd2}

The performance of interdigitated power distribution grids is quantitatively compared to the power noise of a conventional power distribution scheme with DSSG in Section 4.6.1. The maximum voltage drop from V_{dd} for paired power distribution grids is evaluated in Section 4.6.2. Both types of power distribution grids are compared

Table 4.3: Impedance characteristics of interdigitated power distribution grids with DSDG

Grid type	Line cross section (μm × μm)	R_{pp}, R_{gg} (Ω)	L^{*}_{pp}, L^{*}_{gg} (nH)	L^{*}_{pg} (nH)	k^{*}	L^{**}_{pp}, L^{**}_{gg} (nH)	L^{**}_{pg} (nH)	k^{**}	$L^{\dagger}_{pp}$, $L^{\dagger}_{gg}$ (nH)	$L^{\dagger}_{pg}$ (nH)	$k^{\dagger}$
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, * denotes coupling between V_{dd1} (V_{dd2}) and Gnd_1 (Gnd_2), ** denotes coupling between V_{dd1} (Gnd_1) and V_{dd2} (Gnd_2), † denotes coupling between Gnd_1 and V_{dd2}

Table 4.4: Impedance characteristics of paired power distribution grids with DSDG

Grid type	Line cross section (μm × μm)	R_{pp}, R_{gg} (Ω)	L^{*}_{pp}, L^{*}_{gg} (nH)	L^{*}_{pg} (nH)	k^{*}	L^{**}_{pp}, L^{**}_{gg} (nH)	L^{**}_{pg} (nH)	k^{**}	$L^{\dagger}_{pp}$, $L^{\dagger}_{gg}$ (nH)	$L^{\dagger}_{pg}$ (nH)	$k^{\dagger}$	$L^{\ddagger}_{pp}$, $L^{\ddagger}_{gg}$ (nH)	$L^{\ddagger}_{pg}$ (nH)	$k^{\ddagger}$
Type I	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
Type II	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, * denotes coupling between V_{dd1} and Gnd_1 in Type I grids (V_{dd1} – V_{dd2} in Type II grids), ** denotes coupling between V_{dd1} and V_{dd2} in Type I grids (V_{dd1} – Gnd_1 in Type II grids), † denotes coupling between V_{dd1} and Gnd_2 in Type I and Type II, ‡ denotes coupling between Gnd_1 and V_{dd2} in Type I and Type II

to the reference power distribution grid with SSSG. Power distribution schemes with decoupling capacitors are compared in Section 4.6.3. The dependence of the power noise on the switching frequency of the current loads is discussed in Section 4.6.4.

4.6.1 Interdigitated Power Distribution Grids without Decoupling Capacitors

The maximum voltage drop for four interdigitated power distribution grids without decoupling capacitors is depicted in Fig. 4.11. 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 Section 4.3, 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 improvement in noise reduction is 16.5%, which is achieved for a 8 μm wide line, and

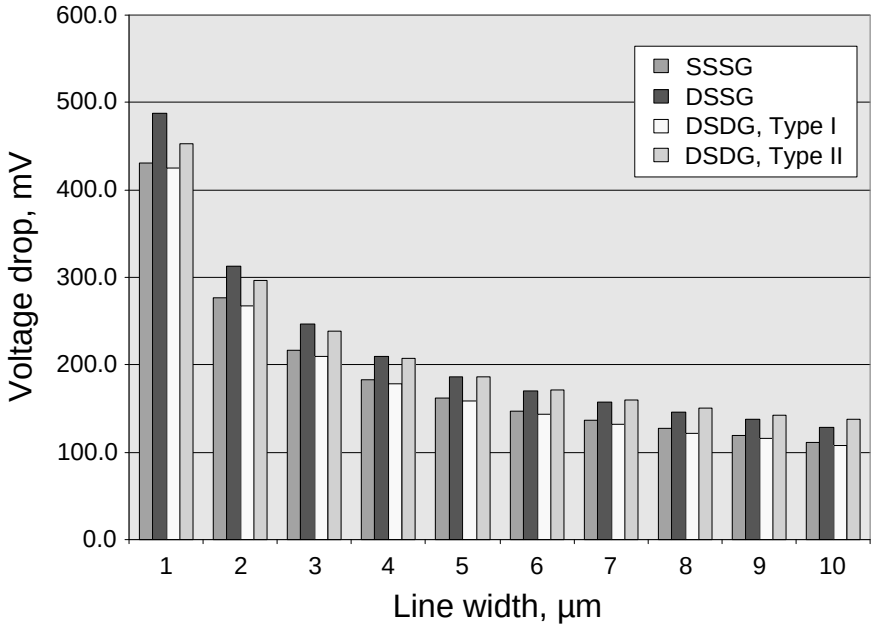


Figure 4.11: Maximum voltage drop for the four interdigitated power distribution grids under investigation. No decoupling capacitors are added.

7.1%, which is achieved for a 1 μm wide line, for fully- and pseudo-interdigitated grids with DSDG, respectively. 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. 4.11, 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 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 [117]. 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 (4.4), 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. 4.7 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.

4.6.2 Paired Power Distribution Grids without Decoupling Capacitors

The maximum voltage drop for three paired power distribution grids without decoupling capacitors is depicted in Fig. 4.12. 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. 4.12, the ratio of the separation between the pairs to the distance between the lines in each pair (n) is eighty. Also note from Fig. 4.10 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. 4.10), 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 (4.2) becomes positive). As n further increases, the power and ground paths within the two voltage domains become strongly coupled, increasing the loop inductance.

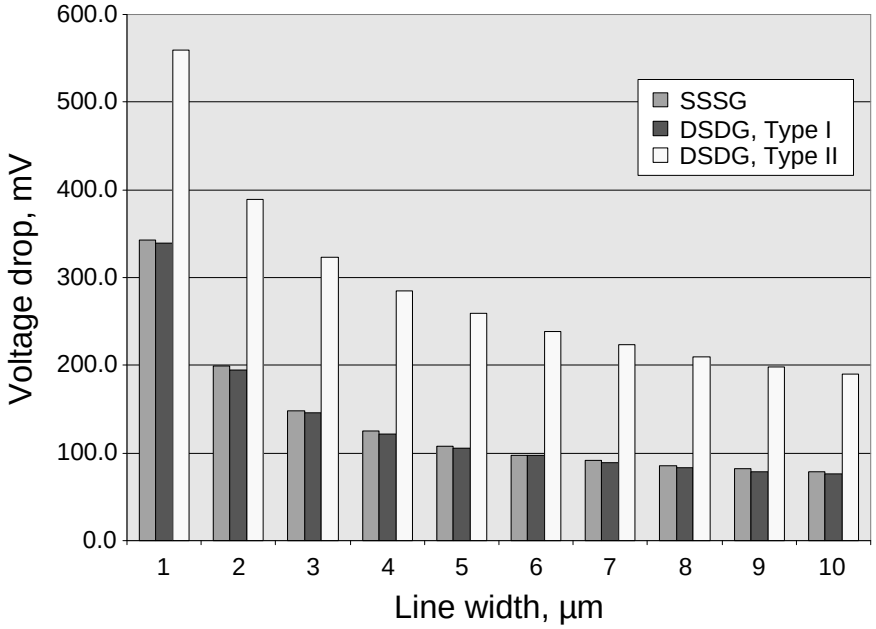


Figure 4.12: Maximum voltage drop for the three paired power distribution grids under investigation. No decoupling capacitors are added.

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. 4.13. Note in Fig. 4.13 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

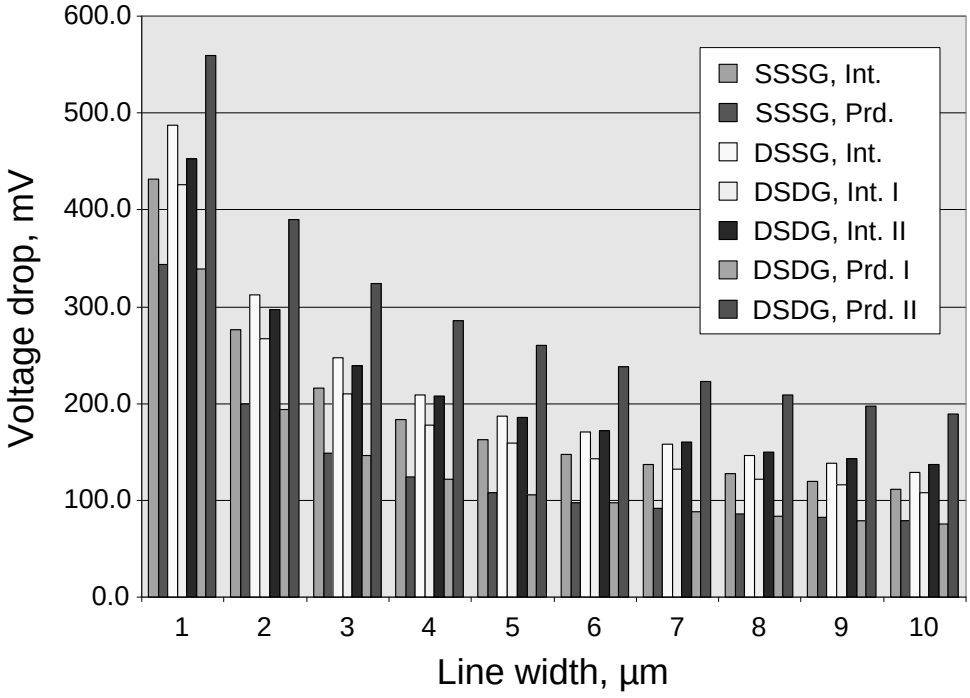


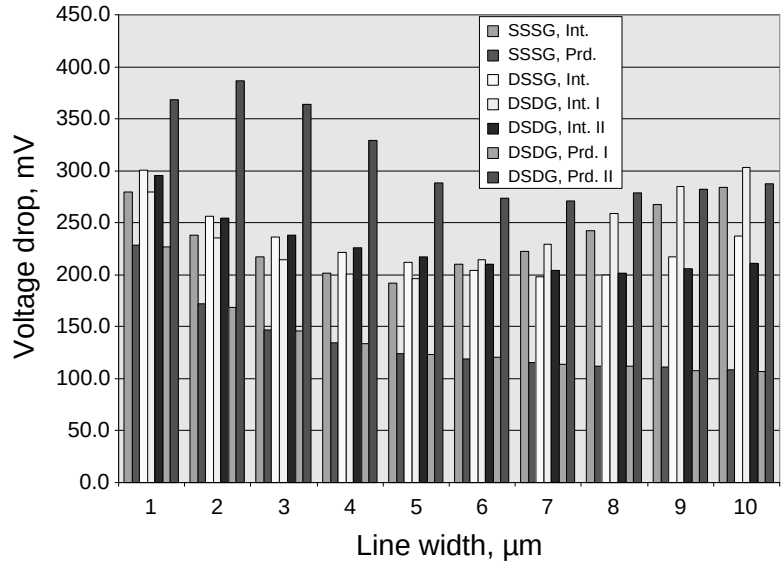
Figure 4.13: Maximum voltage drop for interdigitated and paired power distribution grids under investigation. No decoupling capacitors are added.

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

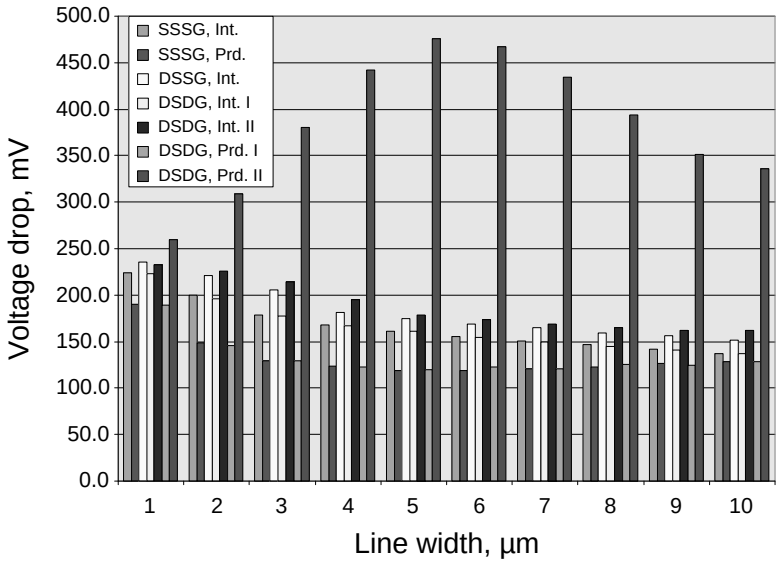
and fully paired power distribution grids with DSDG can also be a better alternative than a power distribution grid with SSSG.

4.6.3 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 [11]. The maximum voltage drop of seven power distribution schemes with decoupling capacitors operating at 1 GHz is shown in Fig. 4.14. 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 useful (located inside the effective radius of an on-chip decoupling capacitor [121]). 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. 4.14, the maximum voltage drop decreases as the lines become wider. The maximum voltage drop of the proposed 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



(a)



(b)

Figure 4.14: 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.

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. 4.13 to that shown in Fig. 4.14, 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 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. 4.14(a). 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 proposed 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 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 pF 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. 4.14(b) 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 [14], [120].

4.6.4 Dependence of Power Noise on the Switching Frequency of the Current Loads

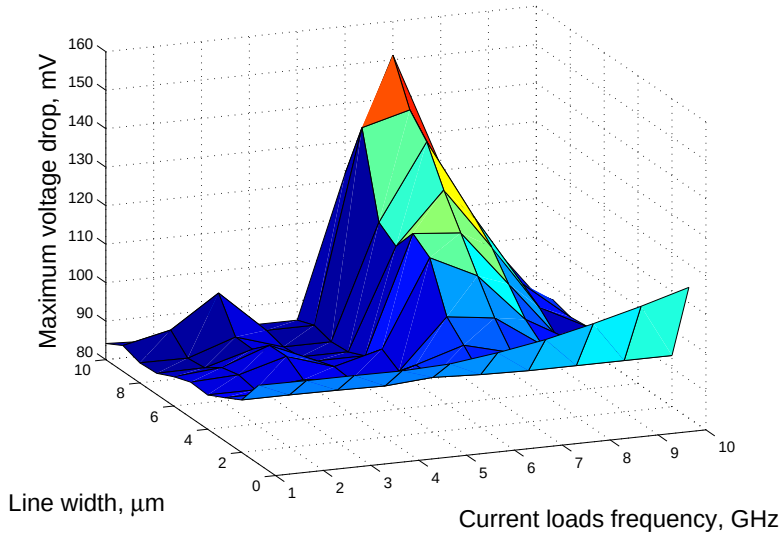
To model the dependence of the power noise on the switching frequency, the power grids are stimulated with triangular current sources with a 50 mA amplitude, 20 ps rise times, and 30 ps fall times. The switching frequency of each current source varies from 1 GHz to 10 GHz to capture the resonances in each power grid. For each grid structure, the width of the line is varied from 1 μm 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. 4.15. 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 lowered, whereas the inductance is slightly increased, decreasing the damping of the entire system. When the switching frequency of a current load approaches the

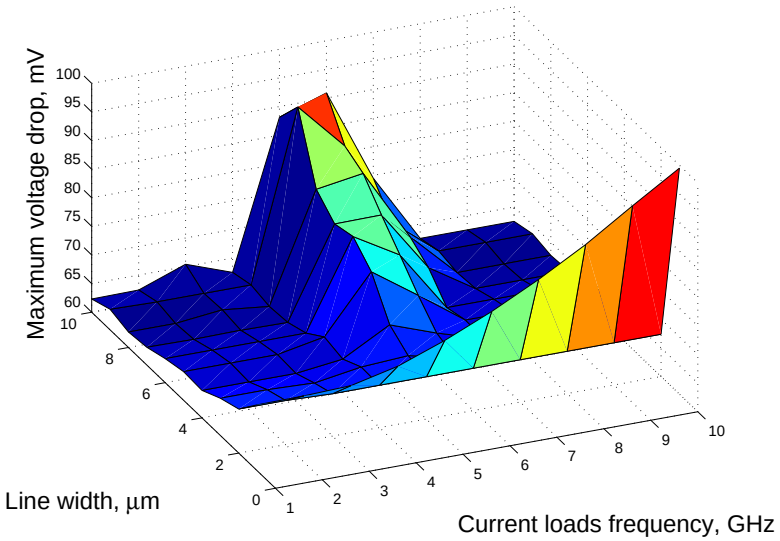
self-resonant frequency of the power grid, the voltage drop due to the RLC system increases (due to resonances). As the width of the lines increases, the system becomes more underdamped, resulting in a sharper resonant peak. The amplitude of the resonant peak increases rapidly as the system becomes less damped. The maximum voltage drop occurs between 6 GHz and 7 GHz for a power grid with a 20 pF decoupling capacitance, as shown in Fig. 4.15(a).

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 [121].

By increasing the magnitude of the decoupling capacitor, the overall power noise can be further reduced, as shown in Fig. 4.15(b). 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 Figs. 4.15(a) to 4.15(b), note that the resonant peak shifts in frequency from approximately 6 GHz to 7 GHz for a 20 pF decoupling capacitance to 5 GHz to 6 GHz for a 30 pF decoupling



(a)



(b)

Figure 4.15: 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.

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 $1\,\mu$m 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.

4.7 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 [11]. It is expected that in future nanoscale ICs, the inductive $L\frac{dI}{dt}$ voltage drop will dominate the resistive IR voltage

drop, becoming the primary component of the overall power noise [13]. As shown previously, the performance of the proposed 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 proposed power distribution grids with DSDG will increase in the future.

As discussed in Section 4.6, 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. 4.10, 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 4.1 – 4.4, 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 [8], the performance of power distribution grids with DSDG is expected to improve by reducing power distribution noise.

4.8 Chapter 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 a 20 pF added 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 the fully paired grid, the resulting 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

- It is noted that 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 5

On-chip Power Noise Reduction Techniques in High Performance Integrated Circuits

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 VLSI circuits [85], [86]. As the operating frequencies increase, the average on-chip current required to charge and discharge these capacitances also increases, while the switching time decreases. 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) [176], 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 VDSM 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 VDSM 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 proposed to analyze power and ground bounce and the effect of SSN on the performance of VLSI 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 [177]. This work suffers from the

assumption that the switching currents of the output drivers are modeled as a triangular shape. In [178], 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 [179], [180], [181]. While most prior research has concentrated on the case where all of the drivers switch simultaneously, the authors in [180] 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 [182]. Recently, 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 [27]. In [183], 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 proposed to reduce SSN. In [184], 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 [185]. An architectural approach for reducing inductive noise caused by clock gating through gradual activation/deactivation units has been introduced in [186]. In [187], 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 [183], [188]. Recently, various methods for reducing ground bounce have been introduced, such as bounce pre-generator circuits [189], supply current shaping, and clock frequency modulation [190].

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. The chapter is organized as follows. Ground noise reduction through the addition of a noise-free on-chip ground is described in Section 5.1. The efficiency of the technique as a function of the physical parameters of the system is investigated in Section 5.2. Some specific conclusions are summarized in Section 5.3.

5.1 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. 5.1. Traditionally, noisy digital circuits share the power and ground supply with noise sensitive analog circuits (see Chapter 4). 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 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. 5.2. 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 L^3_c and L^3_p, respectively. Alternatively, the capacitor is tuned in resonance with the parasitic inductances at a frequency that produces the greatest noise reduction. The

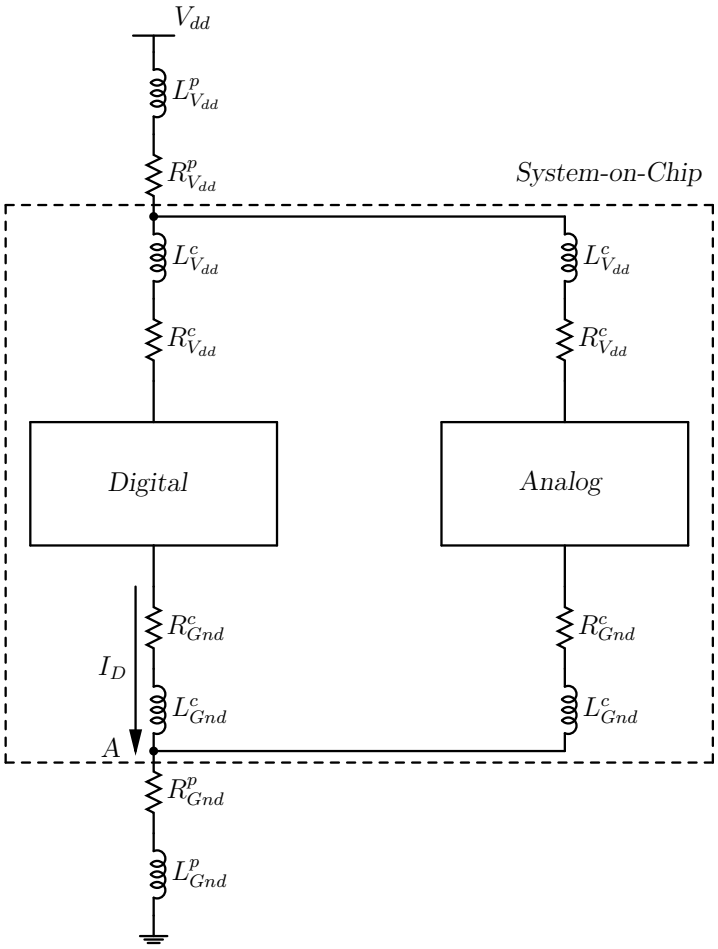


Figure 5.1: An equivalent circuit for analyzing ground bounce in an SoC. The power distribution network is modeled as a series resistance and inductance. The superscripts p and c denote the parasitic resistance and inductance of the package and on-chip power delivery systems, respectively. The subscript V_{dd} denotes the power supply voltage and the superscript G_{nd} denotes the ground.

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.

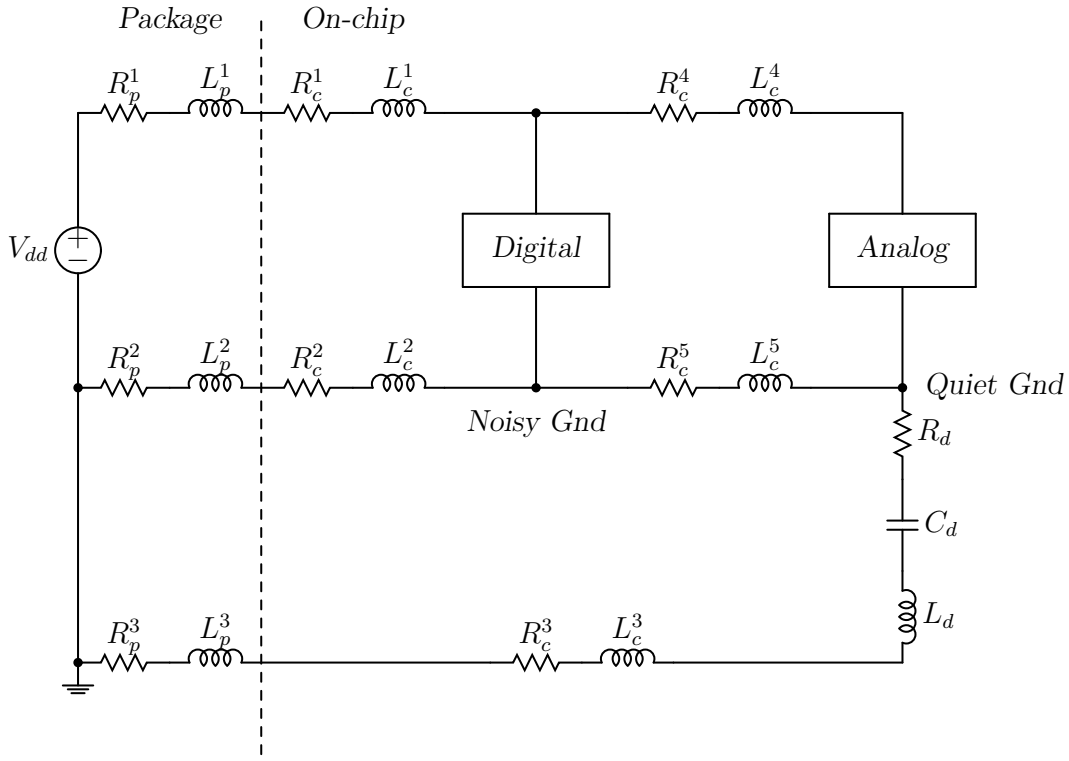


Figure 5.2: Ground bounce reduction technique. The effective series resistance and effective series inductance of the decoupling capacitor are modeled by R_d and L_d, respectively. R_c^5 and L_c^5 represent the physical separation between the noisy and noise sensitive blocks. The impedance of the additional on-chip ground is modeled by R_c^3 and L_c^3, respectively.

5.2 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. 5.3. 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, 50 ps rise time, and 200 ps fall time is utilized to estimate the reduction in ground noise.

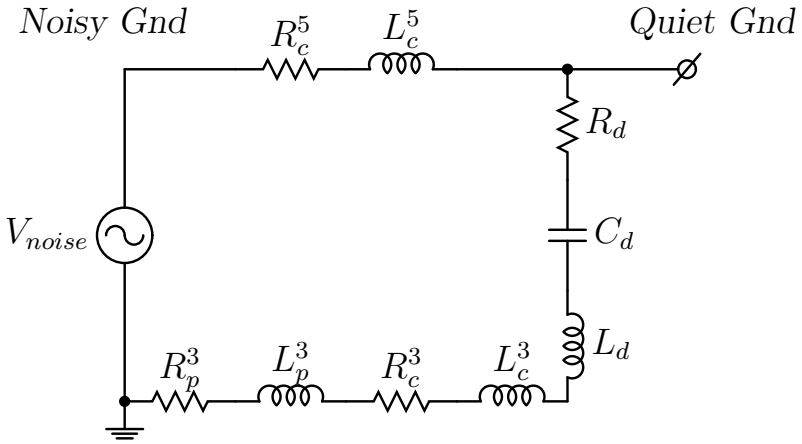


Figure 5.3: Simplified circuit of the ground bounce reduction technique. The ground bounce due to simultaneously switching the digital circuits is modeled by a voltage source. The *Noisy Gnd* denotes an on-chip ground for the simultaneously switching digital circuits. The *Quiet Gnd* denotes a low noise ground for the noise sensitive circuits.

The dependence of the noise reduction technique on the physical separation between noisy and noise sensitive circuits is presented in Section 5.2.1. The sensitivity of this technique to frequency and capacitance variations is discussed in Section 5.2.2. The dependence of ground noise on the impedance of an additional on-chip ground path is analyzed in Section 5.2.3.

5.2.1 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 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 5.1.

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

Table 5.1: 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

$V_{noise} = 100\,\text{mV}$, $f = 1\,\text{GHz}$, $R_p^3 = 10\,\text{m}\Omega$
$L_p^3 = 100\,\text{pH}$, $R_c^3 = 100\,\text{m}\Omega$, $L_c^3 = 100\,\text{fH}$, $R_d = 10\,\text{m}\Omega$
$L_d = 10\,\text{fH}$, $C_d^{Sin} = 253\,\text{pF}$, $C_d^{Triang} = 63\,\text{pF}$

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.

5.2.2 Frequency and Capacitance Variations

To determine the sensitivity of the ground bounce reduction technique on frequency and capacitance variations, the frequency is varied by ±50% from the resonant frequency and the capacitor is varied by ±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 in Figs. 5.4 and 5.5, respectively.

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. 5.5, the reduction in ground bounce is almost

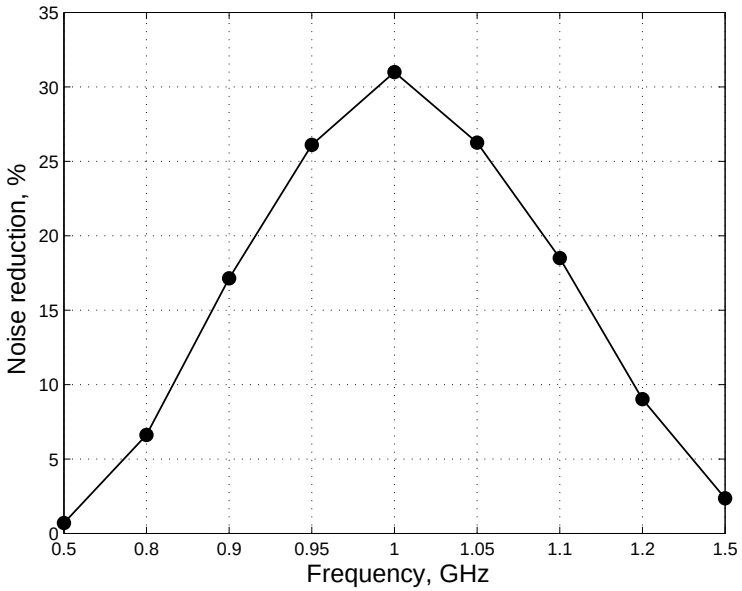


Figure 5.4: Ground bounce reduction as a function of noise frequency. The reduction in noise drops linearly as the frequency varies from the target resonant frequency. The ground noise is modeled as a sinusoidal voltage source.

insensitive to capacitance variations. The efficiency of the technique drops by about 4% as the capacitance is varied by $\pm 10\%$.

5.2.3 Impedance of an Additional Ground Path

As described in Section 5.1, 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

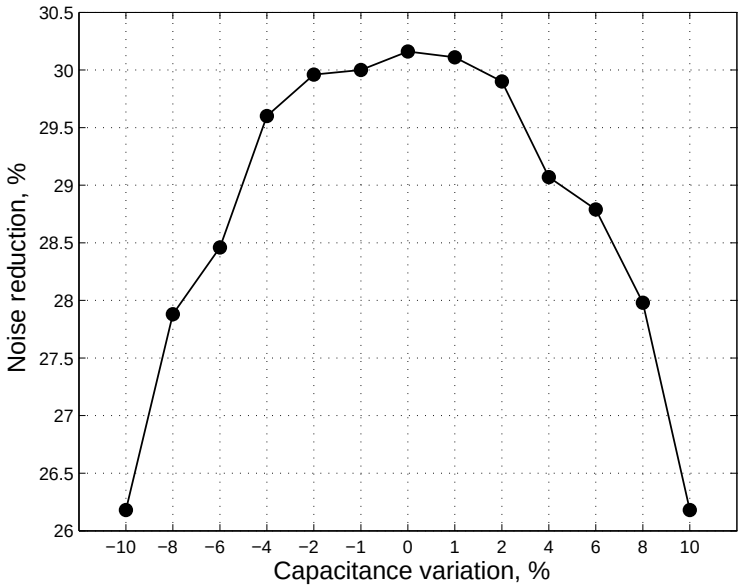


Figure 5.5: The reduction in ground bounce as a function of capacitance variations. The reduction in ground bounce is almost insensitive to capacitance variations. The ground bounce is modeled as a sinusoidal voltage source.

Section 5.2.1, placing noisy and noise sensitive blocks more distant 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 5.2.

Note from Table 5.2 that by reducing the parasitic resistance of an on-chip low noise ground, the ground bounce can be significantly lowered. Noise reductions of

Table 5.2: Ground bounce reduction for different values of parasitic resistance of the on-chip low noise ground

R_c^3 ($\mathrm{m}\Omega$)	V_{quiet} (mV)		Noise Reduction (%)	
	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

$V_{noise} = 100\,\mathrm{mV}$, $f = 1\,\mathrm{GHz}$, $R_p^3 = 10\,\mathrm{m}\Omega$, $L_p^3 = 100\,\mathrm{pH}$
$L_c^3 = 100\,\mathrm{fH}$, $R_c^5 = 80\,\mathrm{m}\Omega$, $L_c^5 = 40\,\mathrm{fH}$, $R_d = 10\,\mathrm{m}\Omega$
$L_d = 10\,\mathrm{fH}$, $C_d^{Sin} = 253\,\mathrm{pF}$, $C_d^{Triang} = 63\,\mathrm{pF}$

about 68% and 22% are demonstrated for sinusoidal and triangular noise sources, respectively. The results listed in Table 5.2 are determined for an average resistance and inductance of the on-chip power distribution ground of five unit lengths (see Table 5.1). Thus, the ground bounce can be further reduced if the analog and digital circuits are placed at a greater distance 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 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.

5.3 Chapter 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 proposed 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 a single frequency and random ground noise, respectively
- 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 6

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 [135], [137], [191]. The increased delay can be compensated by changing the critical paths with behavioral transformations such as parallelization or pipelining [192]. 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 [150]. 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 investigated in the area of high level synthesis for low power [146], [167]. 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.

The design of the power distribution system has become an increasingly difficult challenge in modern CMOS circuits [11]. 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 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 [46]

$$Z_{target} = \frac{V_{dd} \times \zeta}{I} = \frac{1.8 \text{ volts} \times 10\%}{100 \text{ ampers}} \approx 0.002 \text{ ohms}, \tag{6.1}$$

where V_{dd} is the power supply voltage, ζ is the allowed ripple voltage, and I is the

current. With general scaling theory [193], 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 five per computer generation [194]. The target impedance must be satisfied not only at DC, but also at all frequencies where current transients exist [49]. Several major components of a power delivery system are used to satisfy a target impedance over a broad frequency range. A voltage regulator module (VRM) 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,

$$|K_V| \leq \frac{\chi V_{dd1}}{V_{dd2}}, \tag{6.2}$$

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, (6.2) 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.

A power distribution network is a complex multi-level system. 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 investigated. 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. Case studies are also presented in the chapter to quantitatively illustrate this methodology for designing a system of decoupling capacitors.

The chapter is organized as follows. The impedance of a power distribution system with multiple supply voltages is described in Section 6.1. A case study of the dependence of the impedance on the power distribution system parameters is presented in Section 6.2. The voltage transfer function of a power distribution system with multiple supply voltages is discussed in Section 6.3. Case studies examining the dependence of the magnitude of the voltage transfer function on the parameters of the power distribution system are illustrated in Section 6.4. Some specific conclusions are summarized in Section 6.5.

6.1 Impedance of a Power Distribution System with Multiple Supply Voltages

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. The impedance of a power distribution system with multiple power supplies is described in Section 6.1.1. The antiresonance of capacitors connected in parallel is addressed in Section 6.1.2. The dependence of the impedance on the power distribution system is investigated in Section 6.1.3.

6.1.1 Impedance of a Power Distribution System

A model of the impedance of a power distribution system with two supply voltages is shown in Fig. 6.1. 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. 6.1 can be modeled as

$$Z = \frac{Z_1 Z_{12} + Z_1 Z_2}{Z_1 + Z_{12} + Z_2}. \tag{6.3}$$

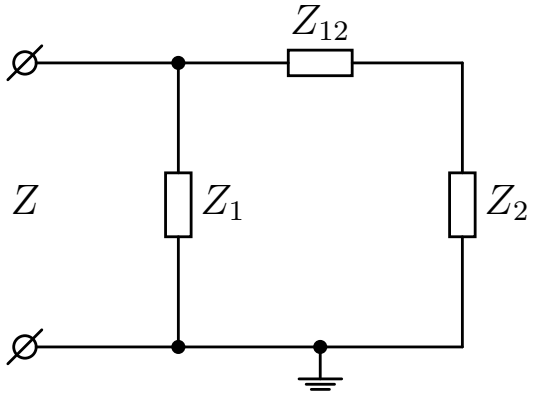


Figure 6.1: Impedance of power distribution system with two supply voltages seen from the load of the power supply V_{dd1}.

Decoupling capacitors have traditionally been modeled as a series RLC network [47]. 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. 6.2.

In this case, the impedance of the power distribution network is

$$Z = \frac{a_4 s^4 + a_3 s^3 + a_2 s^2 + a_1 s + a_0}{b_3 s^3 + b_2 s^2 + b_1 s}, \tag{6.4}$$

where

$$a_4 = L_1(L_{12} + L_2), \tag{6.5}$$

$$a_3 = R_1 L_{12} + R_{12} L_1 + R_1 L_2 + R_2 L_1, \tag{6.6}$$

$$a_2 = R_1 R_{12} + R_1 R_2 + \frac{L_1}{C_{12}} + \frac{L_{12}}{C_1} + \frac{L_1}{C_2} + \frac{L_2}{C_1}, \tag{6.7}$$

$$a_1 = \frac{R_1}{C_2} + \frac{R_2}{C_1} + \frac{R_1}{C_{12}} + \frac{R_{12}}{C_1}, \tag{6.8}$$

$$a_0 = \frac{C_{12} + C_2}{C_1 C_{12} C_2}, \tag{6.9}$$

$$b_3 = L_1 + L_{12} + L_2, \tag{6.10}$$

$$b_2 = R_1 + R_{12} + R_2, \tag{6.11}$$

$$b_1 = \frac{1}{C_1} + \frac{1}{C_{12}} + \frac{1}{C_2}, \tag{6.12}$$

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. 6.3. The minimum power distribution system impedance is limited by the ESR of the

decoupling capacitors. For on-chip applications, the ESR includes the parasitic 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 re-

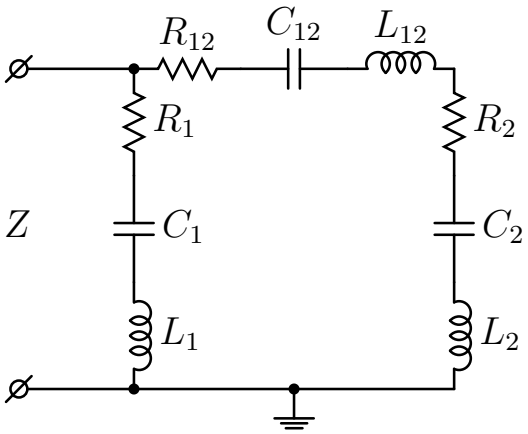


Figure 6.2: Impedance of power distribution system with two supply voltages and the decoupling capacitors represented as series RLC networks.

sistance of the power delivery system. Conversely, for printed circuit board (PCB) 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 (6.1), multiple decoupling capacitors are placed at different levels of the power grid hierarchy [46].

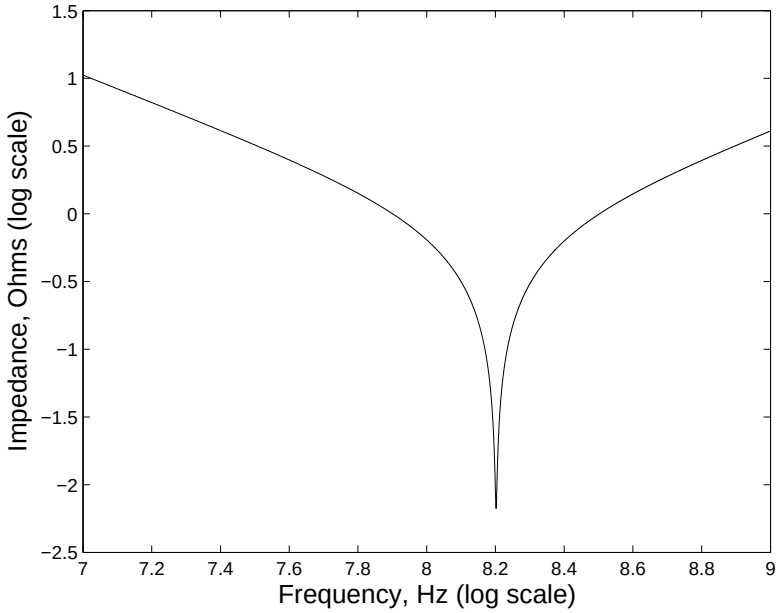


Figure 6.3: Frequency dependence of the impedance of a power distribution system with dual supply voltages, $R_1 = R_{12} = R_2 = 10\,\mathrm{m}\Omega$, $C_1 = C_{12} = C_2 = 1\,\mathrm{nF}$, and $L_1 = L_{12} = L_2 = 1\,\mathrm{nH}$. Since all of the parameters of a power distribution system are identical, 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.

As described in [67], 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

$$Z = \frac{a_4 s^4 + a_2 s^2 + a_0}{b_3 s^3 + b_1 s}, \tag{6.13}$$

where

$$a_4 = L_1(L_{12} + L_2), \tag{6.14}$$

$$a_2 = \frac{L_1}{C_{12}} + \frac{L_{12}}{C_1} + \frac{L_1}{C_2} + \frac{L_2}{C_1}, \tag{6.15}$$

$$a_0 = \frac{C_{12} + C_2}{C_1 C_{12} C_2}, \tag{6.16}$$

$$b_3 = L_1 + L_{12} + L_2, \tag{6.17}$$

$$b_1 = \frac{1}{C_1} + \frac{1}{C_{12}} + \frac{1}{C_2}. \tag{6.18}$$

6.1.2 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. 6.2 are equal, the impedance of the power distribution system can be described as a series *RLC* circuit. Expression (6.13) 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 (6.13) 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. 6.4, 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* [46].

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. 6.4. For instance, if the parasitic inductance of C_1 is greater than the 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 [15].

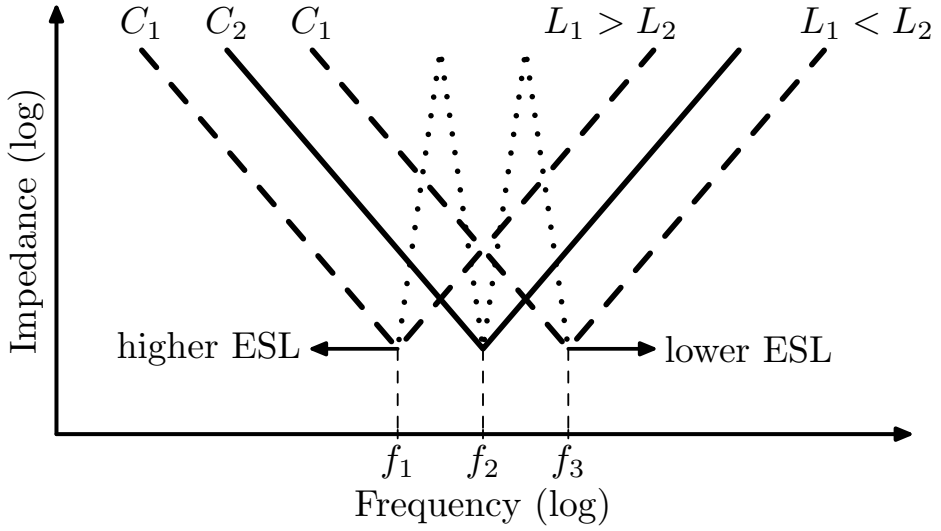


Figure 6.4: Antiresonance of the two capacitors connected in parallel, $C_2 = C_1$. Two antiresonant spikes appear between frequencies f_1 and f_2 and f_2 and f_3 (dotted lines).

6.1.3 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. 6.2, 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, the 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. 6.2 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 6.5.

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. 6.6.

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

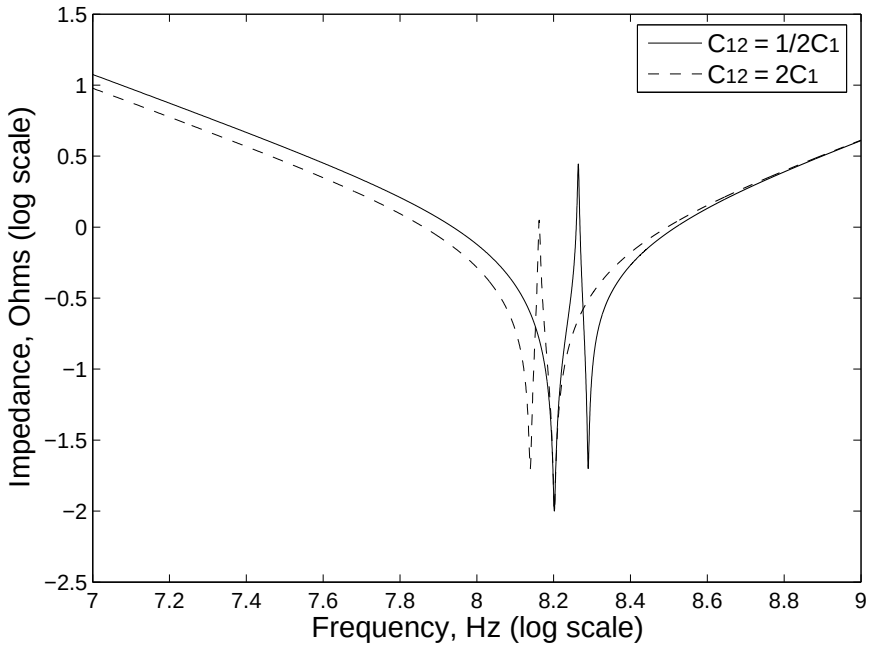


Figure 6.5: Antiresonance of a power distribution system with dual power supply voltages, $R_1 = R_{12} = R_2 = 10\,\mathrm{m}\Omega$, $C_1 = C_2 = 1\,\mathrm{nF}$, and $L_1 = L_{12} = L_2 = 1\,\mathrm{nH}$. Depending upon the ratio of C_{12} to C_1, the antiresonance appears before or after the resonant frequency of the system (the impedance minimum).

a power distribution system on the ESL is discussed below.

To determine the location of the antiresonant spikes, the roots of the denominator of (6.13) are evaluated. One pole is located at $\omega = 0$. Two other poles are located at frequencies,

$$\omega = \pm\sqrt{\frac{C_2 + C_1C_2/C_{12} + C_1}{C_1C_2(L_1 + L_{12} + L_2)}}. \tag{6.19}$$

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 (6.19) 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.

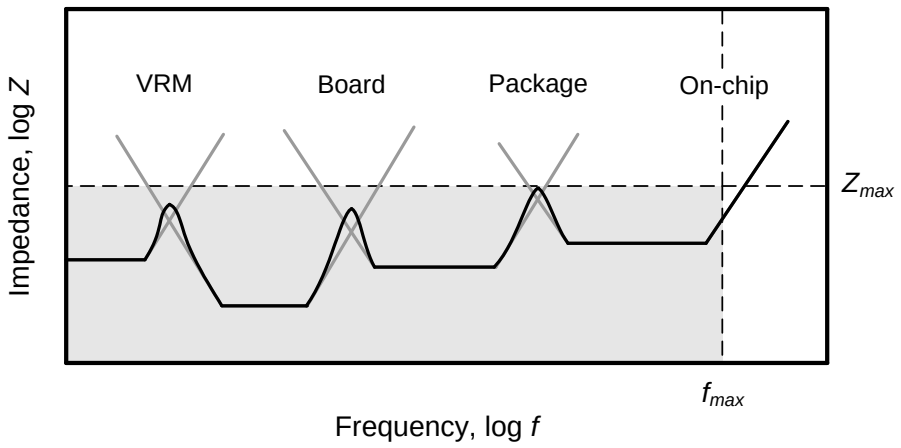
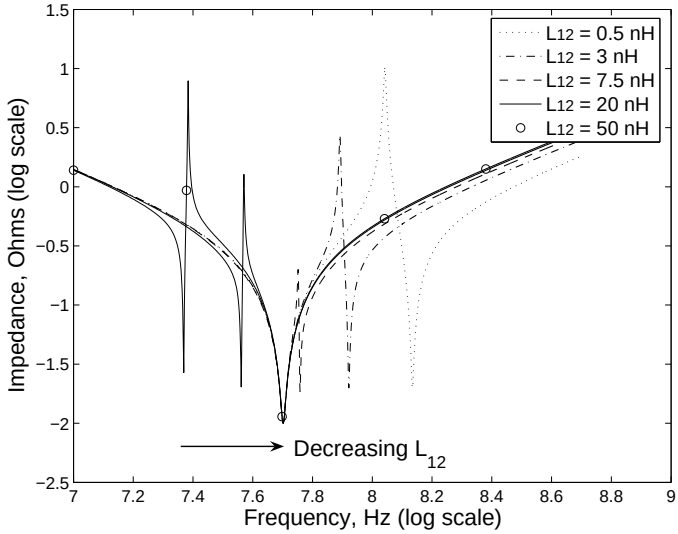
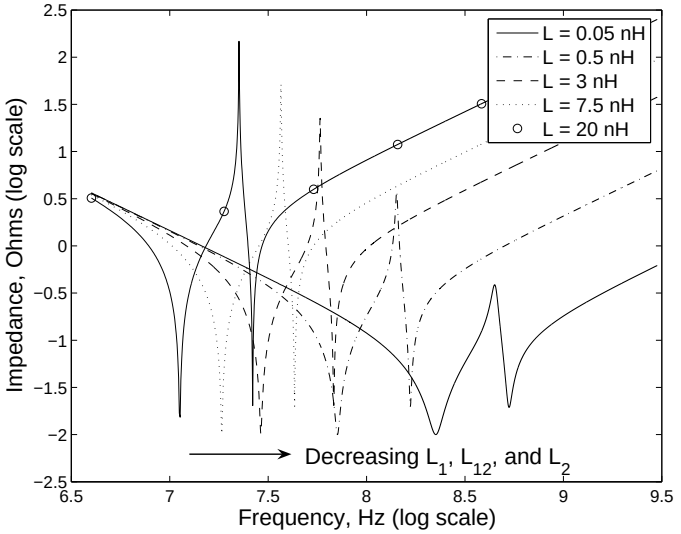


Figure 6.6: Impedance of the power distribution system as a function of frequency. Decoupling capacitors are placed at different hierarchical levels to shift an antiresonant spike above the maximum operating frequency of the system.

The dependence of the power distribution system impedance on the ESL of C_{12} is shown in Fig. 6.7(a). 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. 6.7(b).



(a) $R_1 = R_{12} = R_2 = 10\,\mathrm{m\Omega}$, $C_1 = C_2 = 10\,\mathrm{nF}$, $C_{12} = 1\,\mathrm{nF}$, and $L_1 = L_2 = 1\,\mathrm{nH}$.



(b) $R_1 = R_{12} = R_2 = 10\,\mathrm{m\Omega}$, $C_1 = C_2 = 10\,\mathrm{nF}$, $C_{12} = 1\,\mathrm{nF}$, and $L_1 = L_{12} = L_2 = L$.

Figure 6.7: 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 L_1, L_{12}, and L_2, respectively.

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. 6.7(b), 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 (6.4), the poles and zeros do not cancel, affecting the circuit behavior. 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 [117].

6.2 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 Section 6.1. An on-chip power distribution system is assumed in this example. The total budget of 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 0.1 ohms and 1 nH, respectively. The target impedance is 0.4 ohms.

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. 6.8. According to (6.19), 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

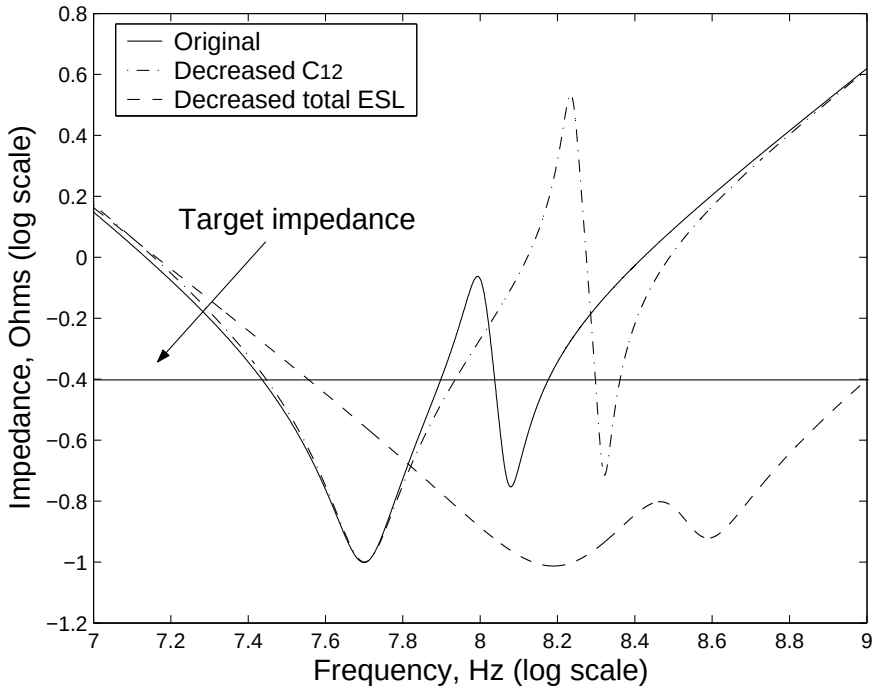


Figure 6.8: The impedance of a power distribution system with dual power supply voltages as a function of frequency, $R_1 = R_{12} = R_2 = 100\,\mathrm{m}\Omega$, $C_1 = C_2 = 10\,\mathrm{nF}$, $C_{12} = 1\,\mathrm{nF}$, and $L_1 = L_{12} = L_2 = 1\,\mathrm{nH}$. The impedance of the example power distribution 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 $0.3\,\mathrm{nF}$ (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 $40\,\mathrm{MHz}$ to $1\,\mathrm{GHZ}$ (the dashed line).

wide frequency range, from approximately 40 MHz to 1 GHz. Three different tradeoff scenarios similar to the case study illustrated in Fig. 6.8 are summarized in Table 6.1. The design parameters for each scenario represent typical values of board, package, and on-chip power distribution systems with decoupling capacitors, as shown in Fig. 6.9. The minimum and maximum frequencies denote the frequency range in

Table 6.1: 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: $R_1 = R_{12} = R_2 = 10\,\text{m}\Omega$, $L_1 = L_{12} = L_2 = 1\,\text{pH}$, $C_{12} = 1\,\text{nF}$, $C_1 = C_2 = 4\,\text{nF}$

which the impedance of a power delivery network seen from the load of V_{dd1} does not exceed the target level of $400\,\text{m}\Omega$. 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 $890\,\text{MHz}$ to $7.9\,\text{GHz}$).

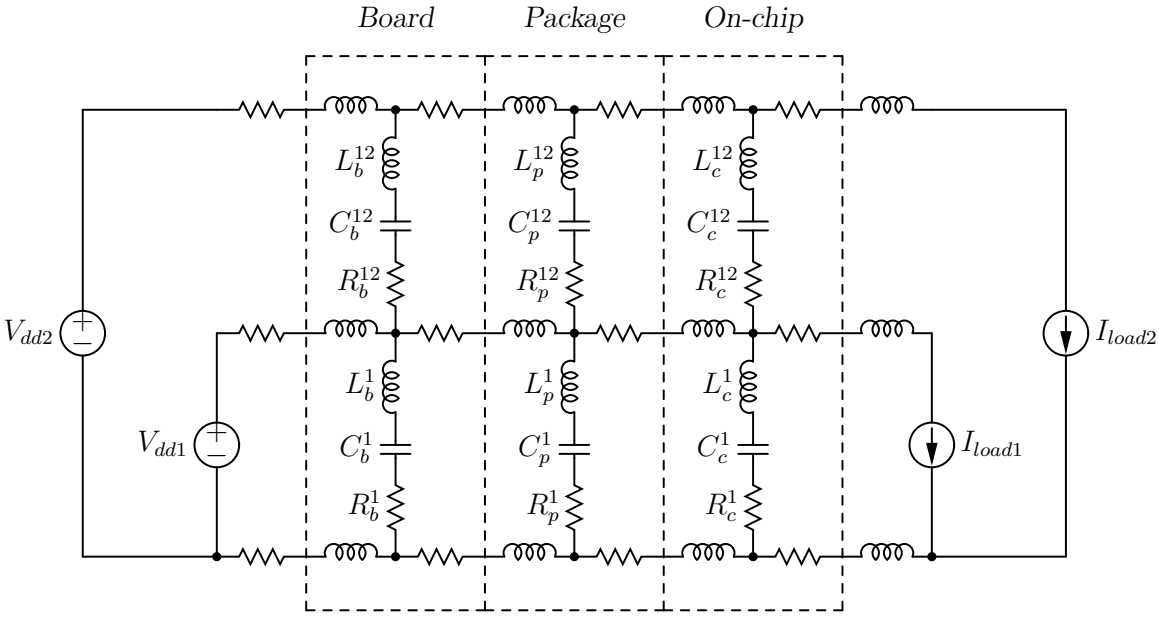


Figure 6.9: Hierarchical model of a power distribution system with dual supply voltages and a single ground. The decoupling capacitors are represented by the series connected resistance, capacitance, and inductance. For simplicity, the decoupling capacitors placed between V_{dd2} and ground are not illustrated. Subscripts b, p, and c denote the board, package, and on-chip power delivery systems, respectively. 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}.

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 (6.13), note that by increasing the decoupling capacitance, the overall impedance of a power distribution system with multiple power supply voltages can be significantly decreased.

6.3 Voltage Transfer Function of Power Distribution System with Multiple Supply Voltages

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 [120]. 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 [195], [196]. The voltage transfer function of a power distribution system with dual power supplies is described in Section 6.3.1. The dependence of the magnitude of the voltage transfer function on certain parameters of the power distribution system is described in subsection 6.3.2.

6.3.1 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. 6.10. 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 R_1, R_{12}, R_2 and L_1, L_{12}, L_2, respectively.

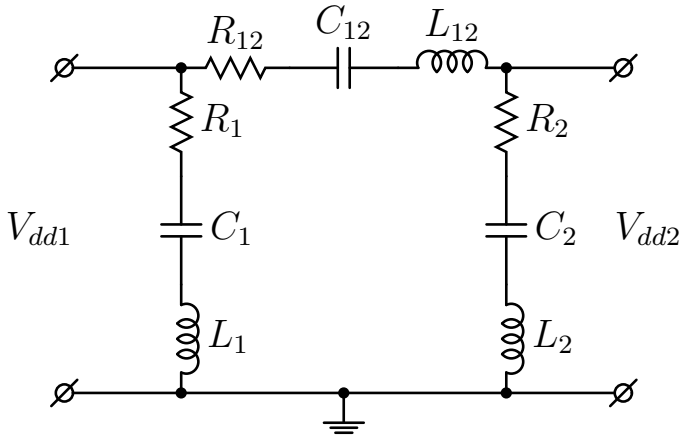


Figure 6.10: Voltage transfer function of a power distribution network with two supply voltages and the decoupling capacitors represented as series RLC networks.

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

$$K_V = \frac{a_2 s^2 + a_1 s + a_0}{b_2 s^2 + b_1 s + b_0}, \tag{6.20}$$

where

$$a_2 = L_2 C_2, \tag{6.21}$$

$$a_1 = R_2 C_2, \tag{6.22}$$

$$a_0 = C_{12}, \tag{6.23}$$

$$b_2 = C_{12} C_2 (L_{12} + L_2), \tag{6.24}$$

$$b_1 = C_{12} C_2 (R_{12} + R_2), \tag{6.25}$$

$$b_0 = C_{12} + C_2. \tag{6.26}$$

Rearranging, (6.20) can be written as

$$K_V = \frac{1}{\frac{a_2 s^2 + a_1 s + a_0}{b_2 s^2 + b_1 s + b_0} + 1}, \tag{6.27}$$

where

$$a_2 = L_{12} C_{12} C_2, \tag{6.28}$$

$$a_1 = R_{12} C_{12} C_2, \tag{6.29}$$

$$a_0 = C_2, \tag{6.30}$$

$$b_2 = L_2 C_{12} C_2, \tag{6.31}$$

$$b_1 = R_2 C_{12} C_2, \tag{6.32}$$

$$b_0 = C_{12}. \tag{6.33}$$

Equations (6.20) and (6.27) are valid only for non-zero frequency, *i.e.,* for $s > 0$. Note from (6.20) 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.

6.3.2 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 [156]. 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. 6.11. Reducing the ESR of a decoupling capacitor decreases the magnitude and range of the operating frequency of the transfer function. Note

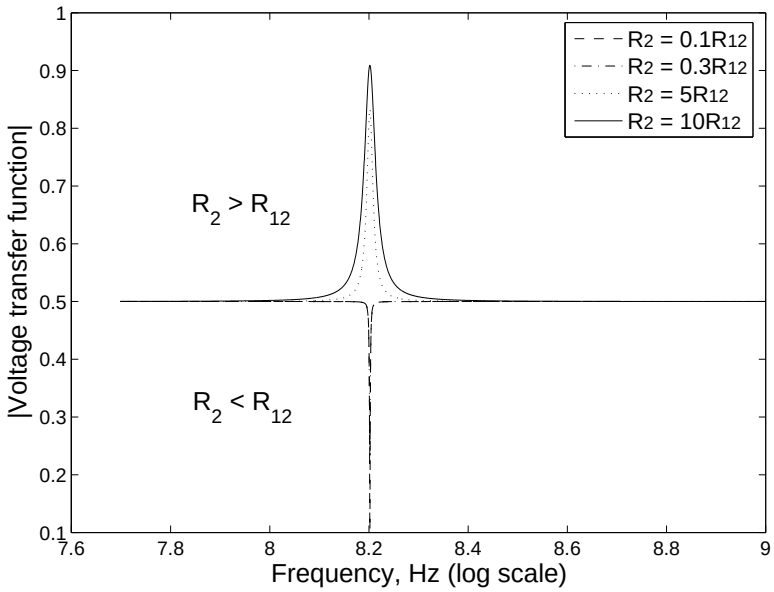


Figure 6.11: Dependence of the magnitude of the voltage transfer function on frequency of a dual V_{dd} power distribution system for different values of ESR of the decoupling capacitors, $R_{12} = 10\,\mathrm{m\Omega}$, $C_{12} = C_2 = 1\,\mathrm{nF}$, and $L_{12} = L_2 = 1\,\mathrm{nH}$.

that to maintain $|K_V|$ below or equal to 0.5, the following inequality has to be satisfied,

$$R_2 \leq R_{12}. \tag{6.34}$$

This behavior can be explained as follows. From (6.27), to maintain $|K_V|$ below or equal to 0.5,

$$\frac{L_{12}C_{12}C_2s^2 + R_{12}C_{12}C_2s + C_2}{L_2C_{12}C_2s^2 + R_2C_{12}C_2s + C_{12}} + 1 \geq 2. \tag{6.35}$$

For equal decoupling capacitors and parasitic inductances, (6.35) leads directly to

(6.34). Generally, to maintain $|K_V|$ below or equal to 0.5,

$$L_2C_2C_3s^2 + R_2C_2C_3s + C_3 \geq L_3C_2C_3s^2 + R_3C_2C_3s + C_2. \tag{6.36}$$

From (6.36), 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 (6.36).

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 (6.20), should be analyzed. To produce an overshoot-free response, the roots of the characteristic equation must be real, yielding

$$R_{12} + R_2 \geq 2\sqrt{\frac{(L_{12} + L_2)(C_{12} + C_2)}{C_{12}C_2}}. \tag{6.37}$$

In the case where $R_{12} = R_2 = R$, $L_{12} = L_2 = L$, and $C_{12} = C_2 = C$, (6.37) reduces to the well known formula [197],

$$R \geq 2\sqrt{\frac{L}{C}}. \tag{6.38}$$

The dependence of the magnitude of the voltage transfer function on the ESL of a power distribution system is shown in Fig. 6.12. For the power distribution system

parameters listed in Fig. 6.12, 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.

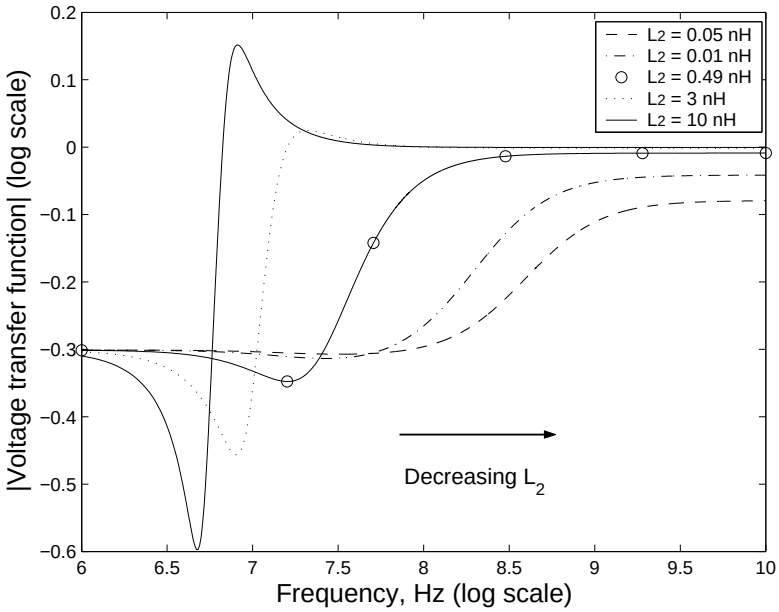


Figure 6.12: Frequency dependence of the voltage transfer function of a dual V_{dd} power distribution system for different values of ESL of the decoupling capacitors, $R_{12} = R_2 = 100\,\mathrm{m}\Omega$, $C_{12} = C_2 = 100\,\mathrm{nF}$, and $L_{12} = 10\,\mathrm{pH}$.

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 (6.37) 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. 6.12, 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 (6.37) is strict and produces an overshoot-free voltage response. In most applications, if small overshoots (about 1%) are permitted, (6.37) is less strict, permitting the parameters of a power distribution network to vary over a wider range.

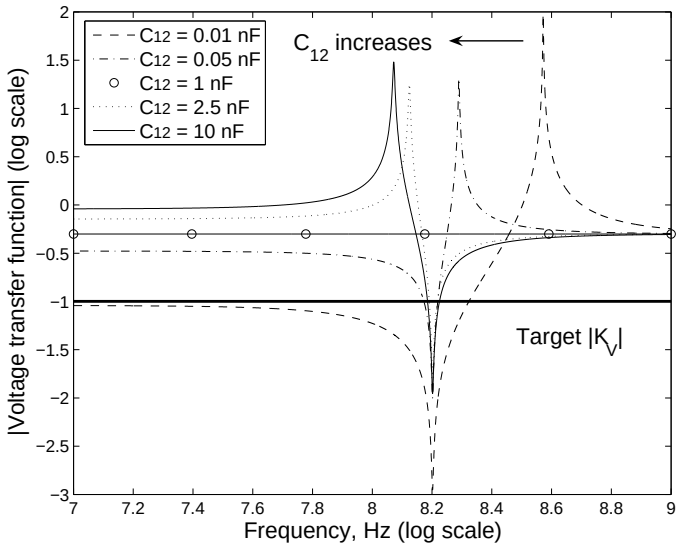
For the parameters listed in Fig. 6.12, the minimum overshoot-free voltage response equals 0.5. It is often necessary to maintain an extremely low magnitude voltage transfer function over a specific frequency range. This behavior can be achieved by varying one of the three design parameters (ESR, ESL 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 (heretofore 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

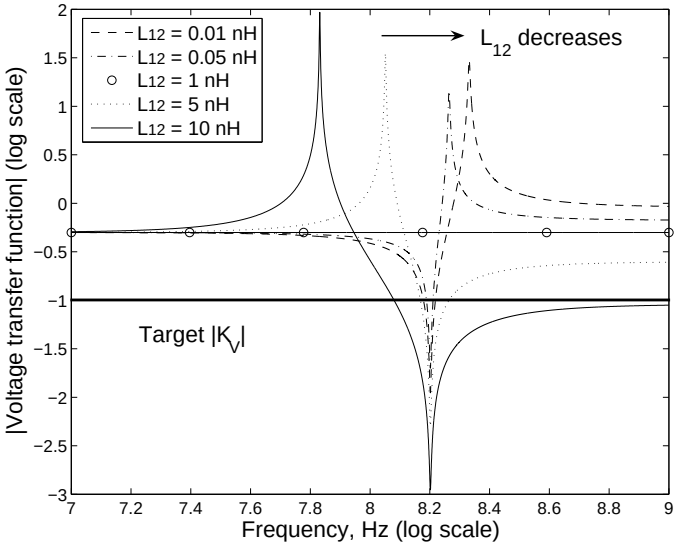
below 0.1 from the self-resonant frequency (or break frequency) of the decoupling capacitor to the maximum operating frequency. From (6.27), for frequencies smaller than the break frequency, the magnitude of the voltage transfer function is approximately $\frac{C_{12}}{C_2}$. For frequencies greater 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 (6.37), and the range of operating frequency is divided by the break frequency into two ranges. This phenomenon is illustrated in Figs. 6.13(a) and 6.13(b).

6.4 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 Section 6.3. 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 [198]. 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 ($C_1 = 100$ nF), high voltage power supply ($C_2 = 40$ nF), and



(a) $R_{12} = R_2 = 10\,\mathrm{m\Omega}$, $C_2 = 1\,\mathrm{nF}$, and $L_{12} = L_2 = 1\,\mathrm{nH}$.



(b) $R_{12} = R_2 = 10\,\mathrm{m\Omega}$, $C_{12} = C_2 = 1\,\mathrm{nF}$, and $L_2 = 1\,\mathrm{nH}$.

Figure 6.13: 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 R_{12}, R_2 and L_{12}, L_2, respectively.

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 0.1 ohms and 1 nH, respectively.

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 (6.37) 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.

6.4.1 Overshoot-Free Magnitude of a Voltage Transfer Function

For typical values of an example power distribution system, (6.37) is not satisfied and the response of the voltage transfer function produces an overshoot as shown in Fig. 6.14. 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 ohms 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. 6.14. As described in Section 6.3.2, 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.

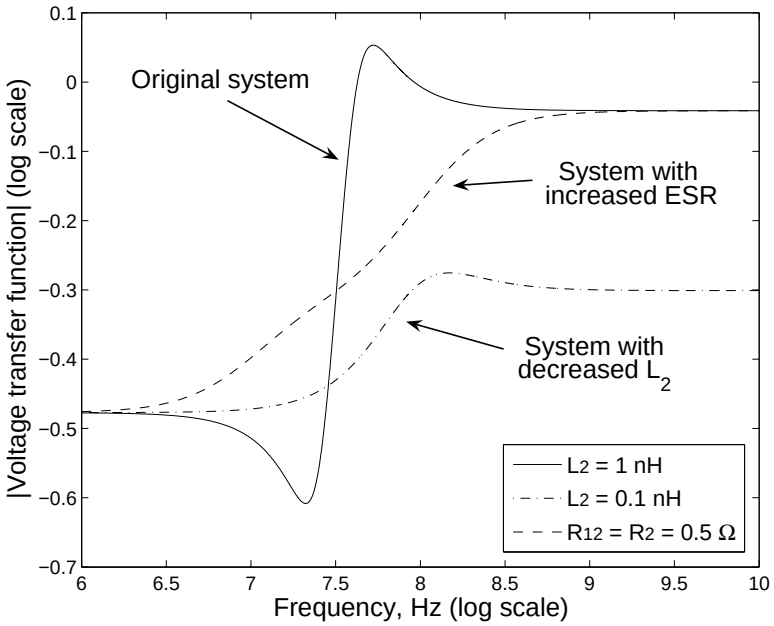


Figure 6.14: Dependence of the magnitude of the voltage transfer function of a dual V_{dd} power distribution system on frequency for different values of the ESR and ESL of the decoupling capacitors, $R_{12} = R_2 = 0.1\,\Omega$, $C_{12} = 20\,\text{nF}$, $C_2 = 40\,\text{nF}$, and $L_{12} = L_2 = 1\,\text{nH}$. 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).

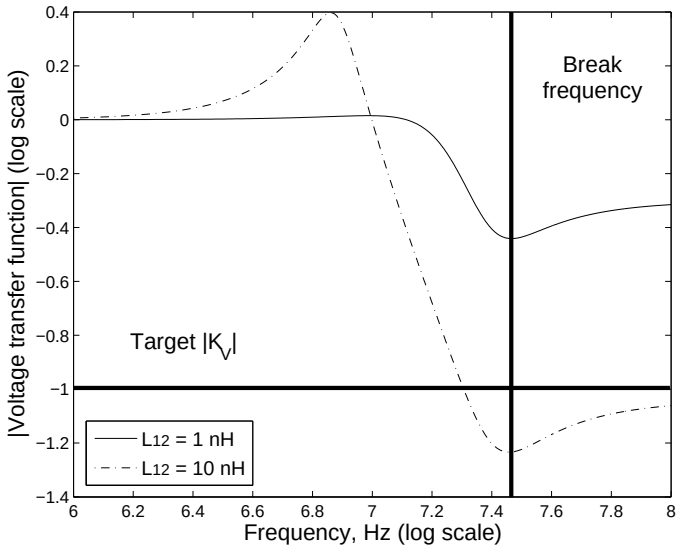
In general, a design methodology for producing an overshoot-free response of a power distribution system with dual power supply voltages is as follows. Based on the available decoupling capacitance for each power supply, the value of the decoupling capacitor placed between the two power supplies is determined 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 (6.37). 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 (6.37) is satisfied. In certain cases, the total budgeted decoupling capacitance should be increased to satisfy (6.37).

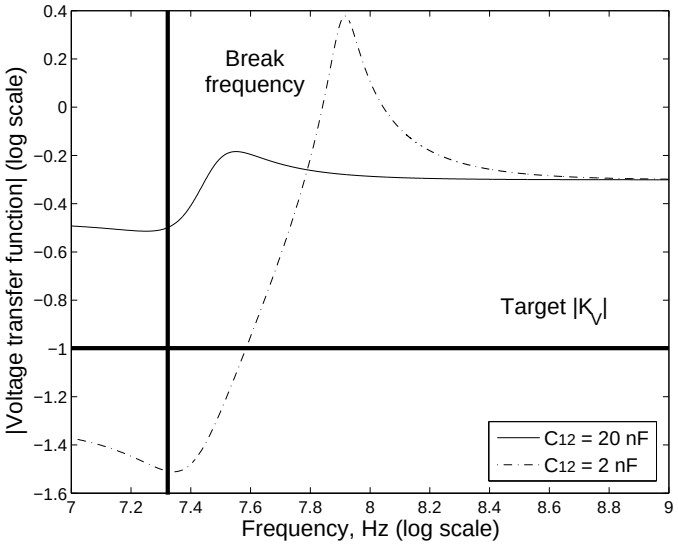
6.4.2 Tradeoff Between the Magnitude and Frequency Range

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 Section 6.3.2. 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 operating frequency. For the example power distribution system, as shown in Fig. 6.15(a), 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. 6.15(b). If it is necessary to completely decouple the two power



(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 $L_{12} = L_2 = 1\,\text{nH}$.

Figure 6.15: 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 R_{12} and R_2 and L_{12} and L_2, respectively.

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. 6.15(a). 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. 6.15(b).

Three different tradeoff scenarios similar to the case study shown in Fig. 6.14 are summarized in Table 6.2. The design parameters for each scenario represent typical values of board, package, and on-chip decoupling capacitors, as shown in Fig. 6.9. 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

Table 6.2: 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}: $R_{12} = R_2 = 10\,\text{m}\Omega$, $L_{12} = L_2 = 100\,\text{fH}$, $C_{12} = 2\,\text{nF}$, $C_2 = 40\,\text{nF}$

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 6.2, 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 described in Section 6.4.1, a design methodology to trade off the magnitude of the 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).

6.5 Chapter 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 7

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 [15]. Since the inductance scales slowly [13], 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 [196].

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 ($L\frac{dI}{dt}$ 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. 7.1.

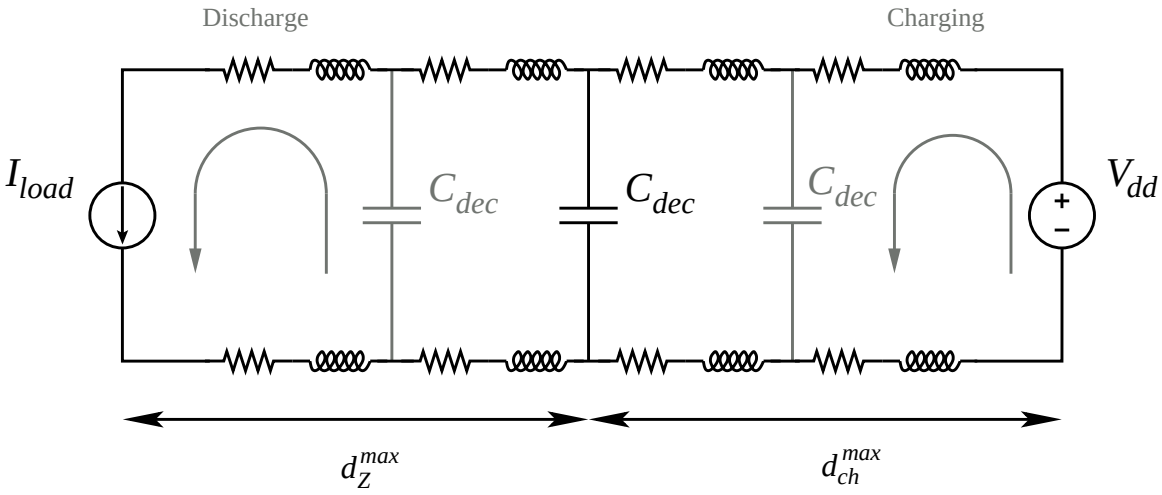


Figure 7.1: Placement of an on-chip decoupling capacitor based on the maximum effective distance. To be effective, a decoupling capacitor should be placed close to the current load during discharge. During the charging phase, however, the decoupling capacitor should be placed close to the power supply 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_{ch}^{max} during charging.

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 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_{ch}^{max}, see Fig. 7.1). 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_{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.

The chapter is organized as follows. Existing work on placing on-chip decoupling capacitors is reviewed in Section 7.1. The effective radius of an on-chip decoupling capacitor as determined by the target impedance is presented in Section 7.2. Design techniques to estimate the minimum magnitude of the required on-chip decoupling capacitance are discussed in Section 7.3. The effective radius of an on-chip decoupling capacitor based on the charge time is determined in Section 7.4. A design methodology for placing on-chip decoupling capacitors based on the maximum effective radii is presented in Section 7.5. A model of an on-chip power distribution network is developed in Section 7.6. Simulation results for typical values of on-chip parasitic resistances and inductances are presented in Section 7.7. Some circuit design implications are discussed in Section 7.8. Finally, some specific conclusions are summarized in Section 7.9.

7.1 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 [122], 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. Simplified criteria have been developed, however, significantly overestimating the required decoupling capacitance. A hierarchical placement of decoupling capacitors has been presented by Smith *et al.* in [46]. The authors of [46] 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 [75]. The concept of an effective series resistance and an effective series inductance of each decoupling capacitor is also introduced. 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 to 20 MHz. Package decoupling capacitors should therefore be utilized in the frequency range from several megahertz to several hundred megahertz [46]. In modern high performance ICs operating at several gigahertz, only decoupling capacitors placed on-chip are effective.

Two types of on-chip decoupling capacitances can be described. An intrinsic decoupling capacitance (or symbiotic capacitance) is comprised of transistors, interconnect, and well-to-substrate capacitances [75]. Since the activity factor in digital circuits is typically low (10% to 30%), the intrinsic on-chip decoupling capacitance in a particular cycle is provided by the non-switching circuits. In contrast to the intrinsic capacitance, an intentional on-chip decoupling capacitance is often added. The intentional on-chip decoupling capacitance is typically an order of magnitude greater than the existing intrinsic capacitance. The intentional on-chip decoupling capacitance is therefore assumed in this chapter to model all of the on-chip decoupling capacitance.

The optimal placement of on-chip decoupling capacitors has been discussed in [199]. The power noise is analyzed assuming an RLC network model, representing a multilayer 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 proposed in [199]. The resulting budget of on-chip decoupling capacitance is therefore significantly overestimated. Another technique for placing on-chip decoupling capacitors has been described in [125]. The decoupling capacitors are placed based on activity signatures determined from microarchitectural simulations. The proposed 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 proposed in [132]. The problem is formulated as a nonlinear optimization and solved using a sensitivity-based quadratic programming solver. The proposed 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 [128], the problem of on-chip decoupling capacitor allocation is investigated. The proposed 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 [128] 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 [121], 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 described 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 the maximum effective distance as determined by the target impedance and charge time is presented in this chapter.

7.2 Effective Radius of On-Chip Decoupling Capacitor Based on Target Impedance

Neglecting the parasitic capacitance [200], 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 [117]. The target impedance of the metal line of a particular length is therefore

$$Z(\omega) = Z'(\omega) \times d, \tag{7.1}$$

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} of (6.1) [15] into (7.1), the maximum effective radius d_Z^{max} between the decoupling capacitor and the current load is

$$d_Z^{max} = \frac{Z_{target}}{Z'(\omega)} = \frac{V_{dd} \times Ripple}{I \times \sqrt{r^2 + \omega^2 l^2}}, \tag{7.2}$$

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. 7.2 [8]. 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 Section 7.7. 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.

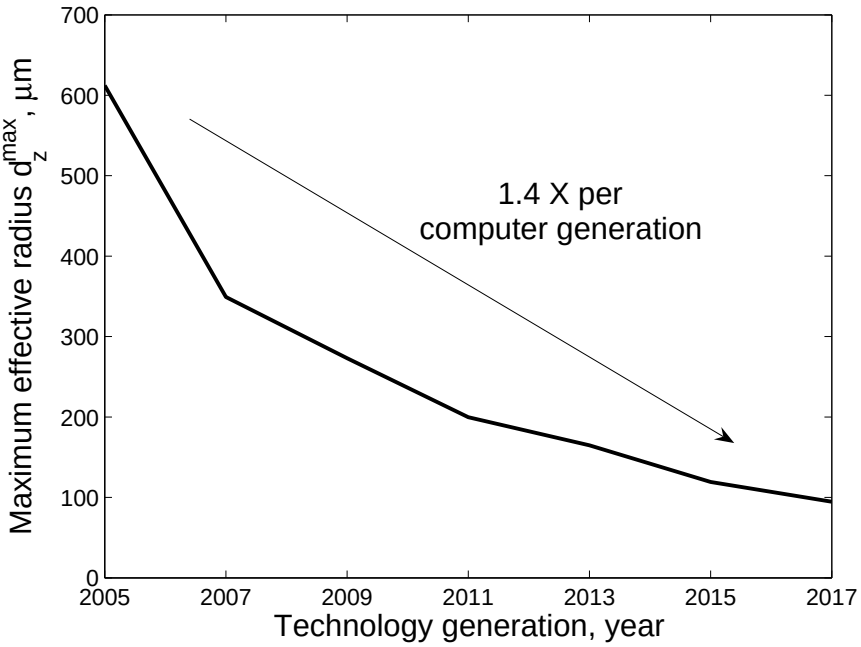


Figure 7.2: Projection of the maximum effective radius as determined by the target impedance d_Z^{max} for future technology generations: $I_{max} = 10\,\text{mA}$, $V_{dd} = 1\,\text{V}$, and $Ripple = 0.1$. Global on-chip interconnects are assumed, modeling the highly optimistic scenario. The maximum 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).

7.3 Estimation of Required On-Chip Decoupling Capacitance

Once the specific location of an on-chip decoupling capacitor is determined as described in Section 7.2, the minimum required magnitude of on-chip decoupling capacitance should be determined, providing the expected current demands. Design

expressions for determining the required magnitude of the on-chip decoupling capacitors based on the dominant power noise are presented in this section. A conventional approach with dominant resistive noise is described in Section 7.3.1. Techniques for determining the magnitude of on-chip decoupling capacitors in the case of dominant inductive noise are developed in Section 7.3.2. The critical length of the P/G paths connecting the decoupling capacitor and the current load is presented in Section 7.3.3.

7.3.1 Dominant Resistive Noise

To estimate the on-chip decoupling capacitance required to support a specific local current demand, for simplicity and without loss of generality, 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. 7.3. The on-chip power distribution network is modeled as a series RL circuit. To qualitatively illustrate the proposed 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. 7.3). Since the required charge is provided by an on-chip decoupling capacitor, the voltage across

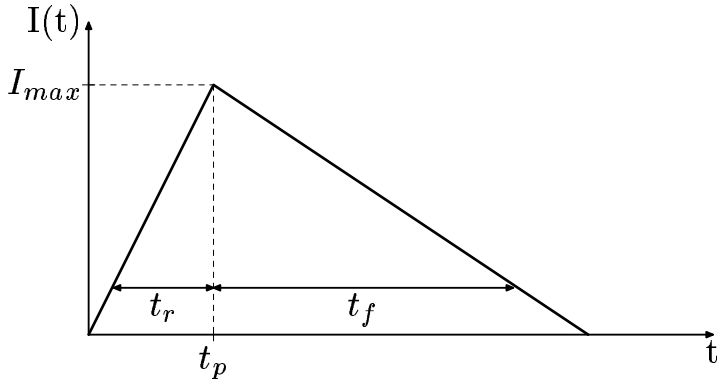


Figure 7.3: Linear approximation of the current demand of a power distribution network by a current source. The magnitude of the current source reaches the maximum current I_{max} at peak time t_p. t_r and t_f denote the rise and fall time of the current load, respectively.

the capacitor during discharge drops below the initial power supply voltage. The required charge during the entire switching event is thus[1]

$$Q_{dis}^{f} = \frac{I_{max} \times (t_r + t_f)}{2} = C_{dec} \times (V_{dd} - V_C^f), \tag{7.3}$$

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_C^f 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.

[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.

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 [201]) to guarantee fault-free operation. Thus,

$$V_C^f \equiv V_{load}^f \geq 0.9\, V_{dd}. \tag{7.4}$$

Substituting (7.4) into (7.3) and solving for C_{dec}, the minimum on-chip decoupling capacitance required to support the current demand during a switching event is

$$C_{dec}^f \geq \frac{I_{max} \times (t_r + t_f)}{0.2\, V_{dd}}, \tag{7.5}$$

where C_{dec}^f is the decoupling capacitance required to support the current demand during the entire switching event.

7.3.2 Dominant Inductive Noise

Note that (7.5) 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 ($IR \gg L\frac{dI}{dt}$). 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 $L\frac{dI}{dt}$ 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 ($L\frac{dI}{dt} \gg IR$), as shown in Fig. 7.4. 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.

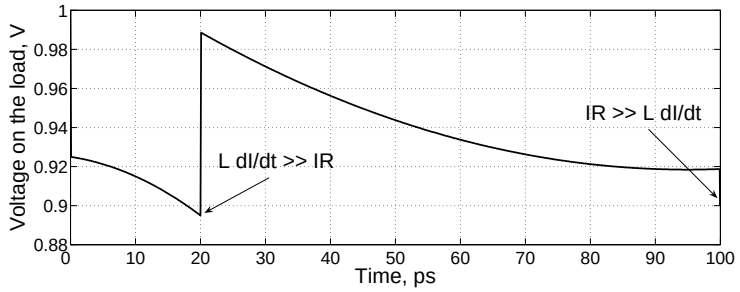
The charge Q_{dis}^r required to support the current demand during the rise time of the current load is equal to the area of the triangle formed by I_{max} and t_r. The required charge is provided by the on-chip decoupling capacitor. The voltage across the decoupling capacitor drops below the power supply level by ΔV_C^r. The required charge during t_r is[2]

$$Q_{dis}^r = \frac{I_{max} \times t_r}{2} = C_{dec} \times \Delta V_C^r, \tag{7.6}$$

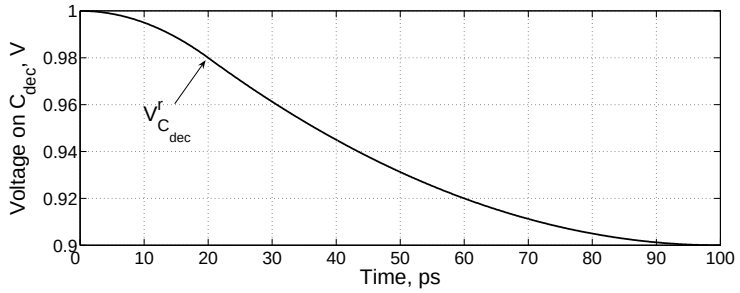
where Q_{dis}^r is the charge drawn by the current load during t_r and ΔV_C^r is the voltage drop across the decoupling capacitor at t_p. From (7.6),

$$\Delta V_C^r = \frac{I_{max} \times t_r}{2\, C_{dec}}. \tag{7.7}$$

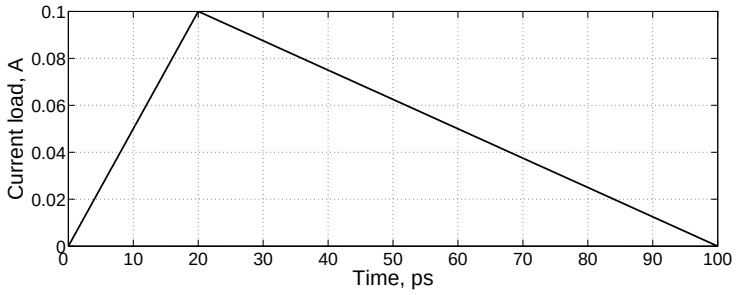
[2]In the general case with a given current profile, the required charge can be estimated as the integral of $I_{load}(t)$ from 0 to t_r.



(a)



(b)



(c)

Figure 7.4: Power distribution noise during discharge of an on-chip decoupling capacitor: $I_{max} = 100\,\text{mA}$, $V_{dd} = 1\,\text{V}$, $t_r = 20\,\text{ps}$, $t_f = 80\,\text{ps}$, $R = 100\,\text{m}\Omega$, $L = 15\,\text{pH}$, and $C_{dec} = 50\,\text{pF}$. 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 resulting voltage drop on the power terminal of a current load is therefore larger than the maximum tolerable noise.

By time t_p, the voltage drop as seen from the current load is the sum of the ohmic IR drop, the inductive $L\frac{dI}{dt}$ 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

$$V_{load}^r = V_{dd} - I \times R - L\frac{dI}{dt} - \Delta V_C^r, \tag{7.8}$$

where R and L are the parasitic resistance and inductance of the P/G lines, respectively. Linearly approximating the current load, dI is assumed equal to I_{max} and dt to t_r. Note that the last term in (7.8) accounts for the voltage drop ΔV_C^r across the decoupling capacitor during the rise time of the current at the load.

Assuming that $V_{load}^r \geq 0.9\,V_{dd}$, substituting (7.7) into (7.8), and solving for C_{dec}, the minimum on-chip decoupling capacitance to support the current demand during t_r is

$$C_{dec}^r \geq \frac{I_{max} \times t_r}{2\left(0.1\,V_{dd} - I \times R - L\frac{dI}{dt}\right)}. \tag{7.9}$$

Note that if $L\frac{dI}{dt} \gg IR$, 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 (7.5), (7.9) 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.

7.3.3 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},

$$\frac{I_{max} \times t_r}{\left(0.1\,V_{dd} - I\,r\,d_{crit} - l\,d_{crit}\,\dfrac{dI}{dt}\right)} = \frac{I_{max} \times (t_r + t_f)}{0.1\,V_{dd}}. \tag{7.10}$$

Solving (7.10) for d_{crit},

$$d_{crit} = \frac{0.1\,V_{dd}\left(1 - \dfrac{t_r}{t_r + t_f}\right)}{I\,r + l\,\dfrac{dI}{dt}}. \tag{7.11}$$

For a single line connecting a current load to a decoupling capacitor, the minimum required on-chip decoupling capacitor is determined by (7.5) for lines shorter than d_{crit} and by (7.9) for lines longer than d_{crit}, as illustrated in Fig. 7.5. Note that for a line length equal to d_{crit}, (7.5) and (7.9) result in the same required capacitance. Also note that the maximum length of a single line is determined by (7.2). 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 (7.5) or (7.9) is utilized to estimate the required on-chip decoupling capacitance.

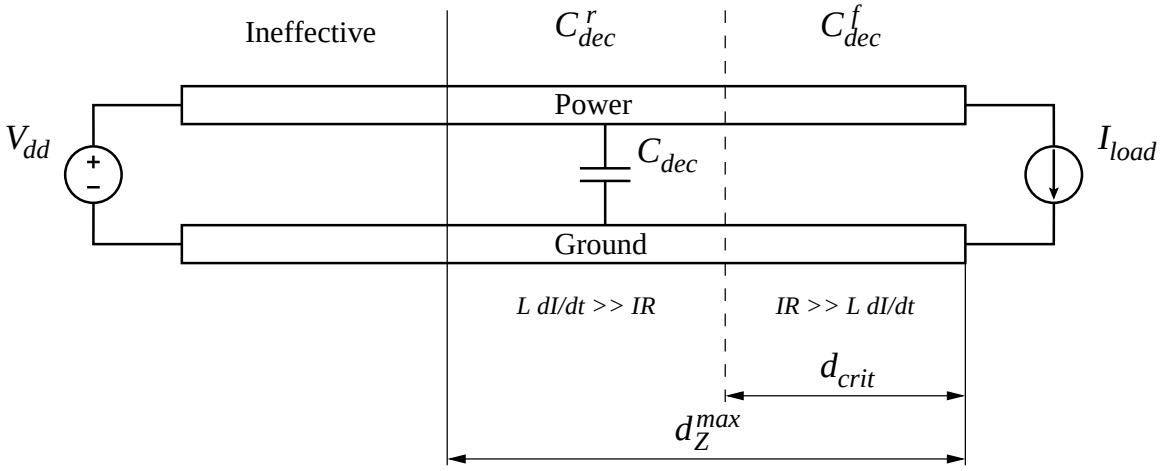


Figure 7.5: Critical line length of an interconnect between a decoupling capacitor and a current load. The minimum required on-chip decoupling capacitance is determined by (7.5) for lines shorter than d_{crit} and by (7.9) for lines longer than d_{crit}. The decoupling capacitor is ineffective beyond the maximum effective radius as determined by the target impedance d_Z^{max}.

The dependence of the critical line length d_{crit} on the rise time t_r of the current load as determined by (7.11) is depicted in Fig. 7.6. From Fig. 7.6, 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.

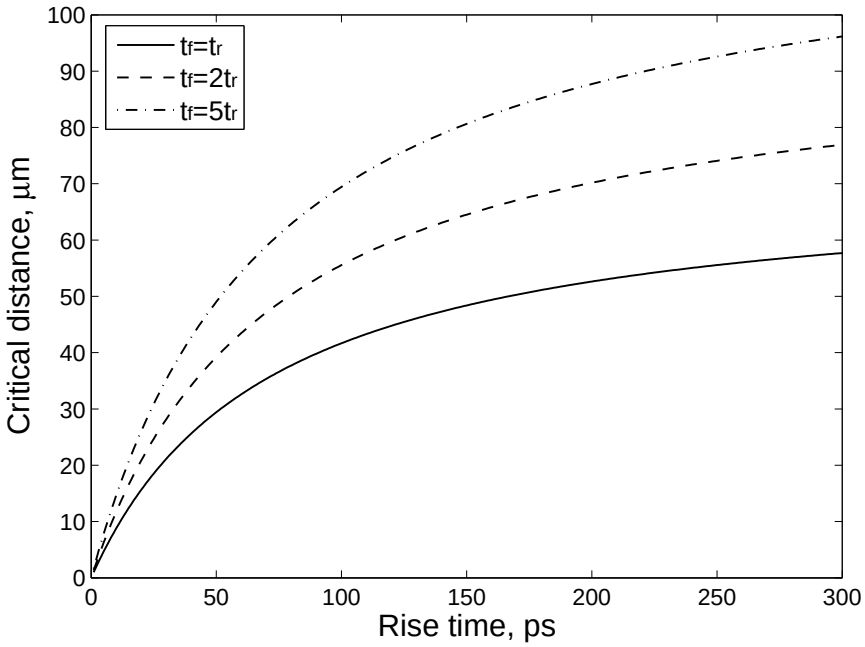
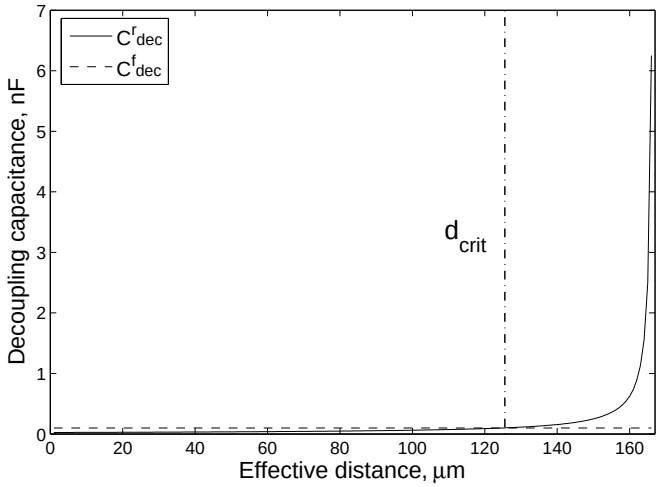
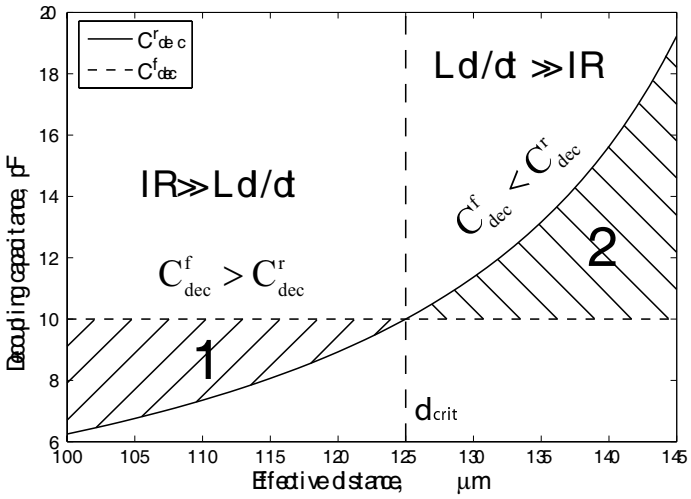


Figure 7.6: Dependence of the critical line length d_{crit} on the rise time of the current load: $I_{max} = 0.1\,\mathrm{A}$, $V_{dd} = 1\,\mathrm{V}$, $r = 0.007\,\Omega/\mu\mathrm{m}$, and $l = 0.5\,\mathrm{pH}/\mu\mathrm{m}$. Note that d_{crit} is determined by $\dfrac{t_r}{t_f}$, increasing with larger t_f. The critical line length will shrink in future nanometer technologies as transition times become shorter.

Observe in Fig. 7.5 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. 7.7. For the example parameters shown in Fig. 7.7, the critical line length is $125\,\mu$m. 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 current load increases,



(a)



(b)

Figure 7.7: Design space for determining minimum required on-chip decoupling capacitance: $I_{max} = 50\,\text{mA}$, $V_{dd} = 1\,\text{V}$, $r = 0.007\,\Omega/\mu\text{m}$, $l = 0.5\,\text{pH}/\mu\text{m}$, $t_r = 100\,\text{ps}$, and $t_f = 300\,\text{ps}$. a) The design space for determining the minimum required on-chip decoupling capacitance is broken into two regions by d_{crit}. b) The design space around d_{crit}. For the example parameters, the critical line length is $125\,\mu\text{m}$. 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.

as shown in Fig. 7.7(a). 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_{dec}^{f} is constant and greater than C_{dec}^{r} (see region 1 in Fig. 7.7(b)). If $L\frac{dI}{dt}$ noise dominates IR noise (the line length is greater than d_{crit}), the required on-chip decoupling capacitance C_{dec}^{r} increases substantially with line length and is greater than C_{dec}^{f} (see region 2 in Fig. 7.7(b)). 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_{dec}^{r} and the distance between the decoupling capacitor and the current load. As C_{dec}^{r} is placed closer to the current load, the required capacitance can be significantly reduced.

7.4 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 Section 7.2, an on-chip decoupling capacitor should be placed in close proximity to the power supply (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. 7.8. 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.

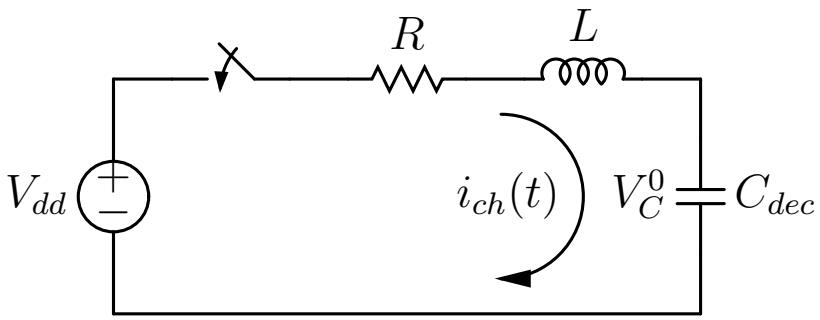


Figure 7.8: Circuit charging an on-chip decoupling capacitor. The parasitic impedance of the power distribution system connecting the decoupling capacitor to the power supply is modeled by a series RL circuit.

For the circuit shown in Fig. 7.8, the KVL equation for the current in the circuit is [202]

$$L\,\frac{di_{ch}}{dt} + R\,i_{ch} + \frac{1}{C_{dec}}\int i_{ch}\,dt = V_{dd}. \tag{7.12}$$

Differentiating (7.12),

$$L\,\frac{d^2 i_{ch}}{dt^2} + R\,\frac{di_{ch}}{dt} + \frac{1}{C_{dec}}\,i_{ch} = 0. \tag{7.13}$$

Equation (7.13) is a second order linear differential equation with the characteristic equation,

$$s^2 + \frac{R}{L}\,s + \frac{1}{LC_{dec}} = 0. \tag{7.14}$$

The general solution of (7.13) is

$$i_{ch}(t) = K_1\,e^{s_1 t} + K_2\,e^{s_2 t}, \tag{7.15}$$

where s_1 and s_2 are the roots of (7.14),

$$s_{1,2} = -\frac{R}{2L} \pm \sqrt{\left(\frac{R}{2L}\right)^2 - \frac{1}{LC_{dec}}}. \tag{7.16}$$

Note that (7.15) represents the solution of (7.13) as long as the system is over-damped. The damping factor is therefore greater than one, *i.e.,*

$$\left(\frac{R}{L}\right)^2 > \frac{4}{LC}. \tag{7.17}$$

For a single line, from (7.17), the critical line length resulting in an overdamped system is

$$D > \frac{4\,l}{r^2\,C_{dec}}, \tag{7.18}$$

where C_{dec} is the on-chip decoupling capacitance, and l and r are the per length inductance and resistance, respectively. Inequality (7.18) 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 (7.15) is therefore a general solution of (7.13) for a scaled CMOS technology.

Initial conditions are applied to determine the arbitrary constants K_1 and K_2 in (7.15). The current charging the decoupling capacitor during the charging phase is

$$\begin{aligned} i_{ch}(t) = & \frac{I_{max}\,(t_r + t_f)}{4\,L\,C_{dec}\sqrt{\left(\frac{R}{2L}\right)^2 - \frac{1}{LC_{dec}}}} \\ & \times \left\{ \exp\left[\left(-\frac{R}{2L} + \sqrt{\left(\frac{R}{2L}\right)^2 - \frac{1}{LC_{dec}}} \right) t \right] \right. \\ & \left. - \exp\left[\left(-\frac{R}{2L} - \sqrt{\left(\frac{R}{2L}\right)^2 - \frac{1}{LC_{dec}}} \right) t \right] \right\}. \end{aligned} \tag{7.19}$$

The voltage across the decoupling capacitor during the charging phase can be determined by integrating (7.19) from zero to the charge time,

$$V_C(t) = \frac{1}{C_{dec}} \int_0^{t_{ch}} i_{ch}(t)\, dt, \tag{7.20}$$

where t_{ch} is the charge time, and $V_C(t)$ and $i_{ch}(t)$ are the voltage across the decoupling capacitor and the current flowing through the decoupling capacitor during the charging phase, respectively. Substituting (7.19) into (7.20) and integrating from zero to t_{ch}, the voltage across the decoupling capacitor during the charging phase is

$$\begin{aligned} V_{C_{dec}}(t_{ch}) = {} & \frac{I_{max}\,(t_r + t_f)}{4\,C_{dec}^2\,L\sqrt{\left(\dfrac{R}{2L}\right)^2 - \dfrac{1}{LC_{dec}}}} \\ & \times \left\{ \frac{\exp\left[\left(-\dfrac{R}{2L} + \sqrt{\left(\dfrac{R}{2L}\right)^2 - \dfrac{1}{LC_{dec}}}\right) t_{ch}\right] - 1}{-\dfrac{R}{2L} + \sqrt{\left(\dfrac{R}{2L}\right)^2 - \dfrac{1}{LC_{dec}}}} \right. \\ & \left. + \frac{1 - \exp\left[\left(-\dfrac{R}{2L} - \sqrt{\left(\dfrac{R}{2L}\right)^2 - \dfrac{1}{LC_{dec}}}\right) t_{ch}\right]}{-\dfrac{R}{2L} - \sqrt{\left(\dfrac{R}{2L}\right)^2 - \dfrac{1}{LC_{dec}}}} \right\}. \end{aligned} \tag{7.21}$$

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 effective radius of an on-chip decoupling capacitor during the charging phase. Thus from (7.21), 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. 7.9. 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.

7.5 Design Methodology for Placing On-Chip Decoupling Capacitors

An overall design methodology for placing on-chip decoupling capacitors based on the maximum effective radii is illustrated in Fig. 7.10. The maximum effective radius

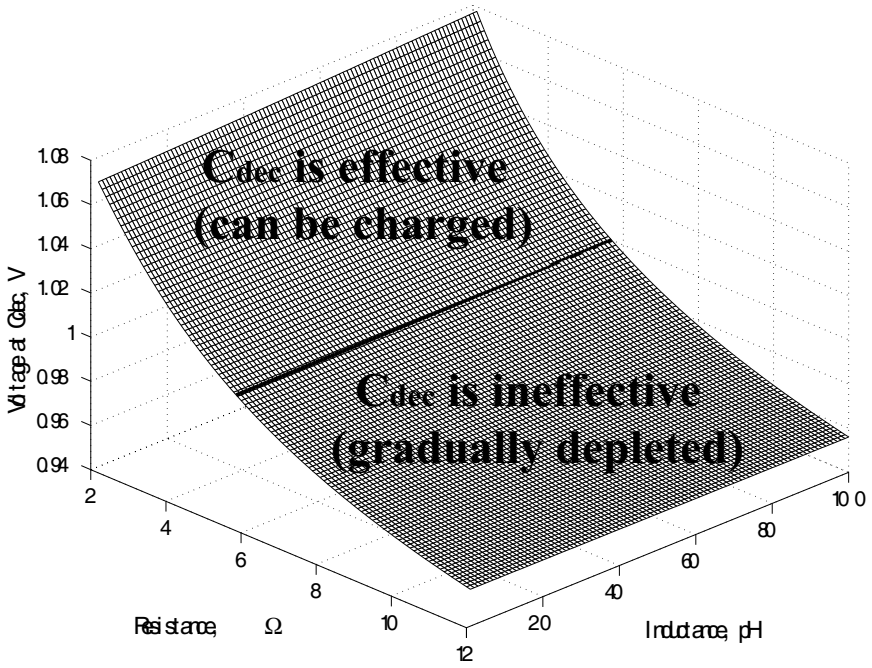


Figure 7.9: Design space for determining the maximum tolerable parasitic resistance and inductance of a power distribution grid: $I_{max} = 100\,\text{mA}$, $t_r = 100\,\text{ps}$, $t_f = 300\,\text{ps}$, $C_{dec} = 100\,\text{pF}$, $V_{dd} = 1\,\text{volt}$, and $t_{ch} = 400\,\text{ps}$. For a target charge time, the maximum resistance and inductance result in 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 results in a voltage greater than the power supply means that the charge on the decoupling capacitor can be restored within t_{ch}.

based on the target impedance is determined from (7.2) 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, (7.5) is used to determine the required on-chip decoupling capacitance. If $L\dfrac{dI}{dt}$ noise dominates, the on-chip decoupling capacitance is determined by (7.9). In the case of a single line connecting

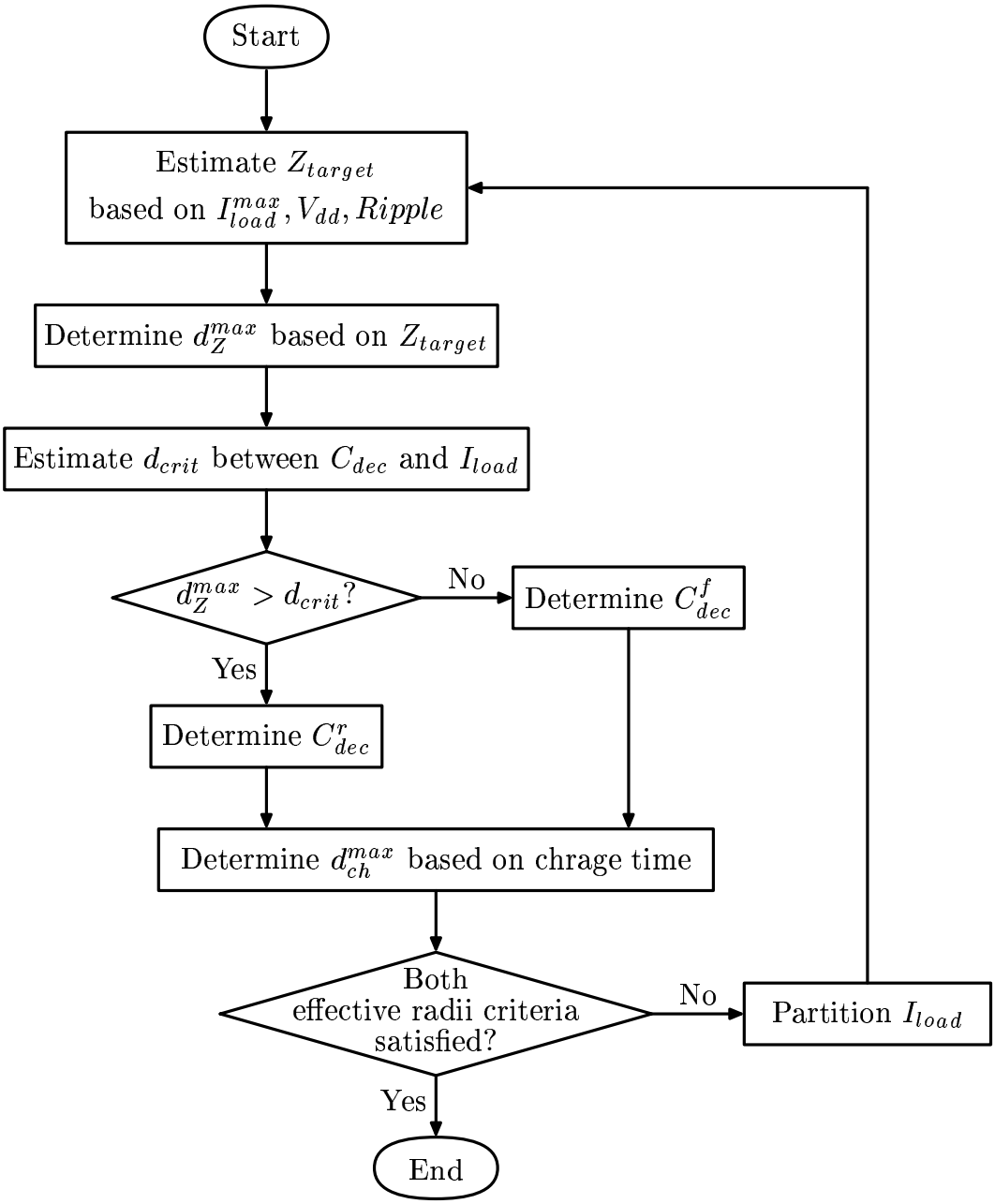


Figure 7.10: Design flow for placing on-chip decoupling capacitors based on the maximum effective radii.

a decoupling capacitor to a current load, the critical wire length is determined by (7.11).

The maximum effective distance based on the charge time is determined from (7.21). Note that (7.21) 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. 7.11. 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.

7.6 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.

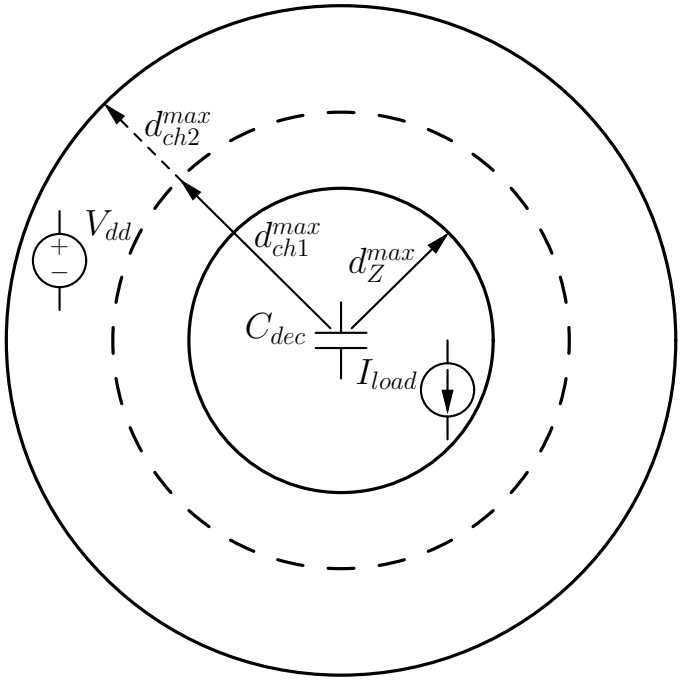


Figure 7.11: The effective radii of an on-chip decoupling capacitor. The on-chip decoupling capacitor is placed such that both the current load and the power supply are located inside the effective radius. The maximum effective radius as determined by the target impedance d_Z^{max} does not depend on 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}.

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 a 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 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 hundreds 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 inches by 1.5 inches, 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 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. 7.12. A finer mesh is used

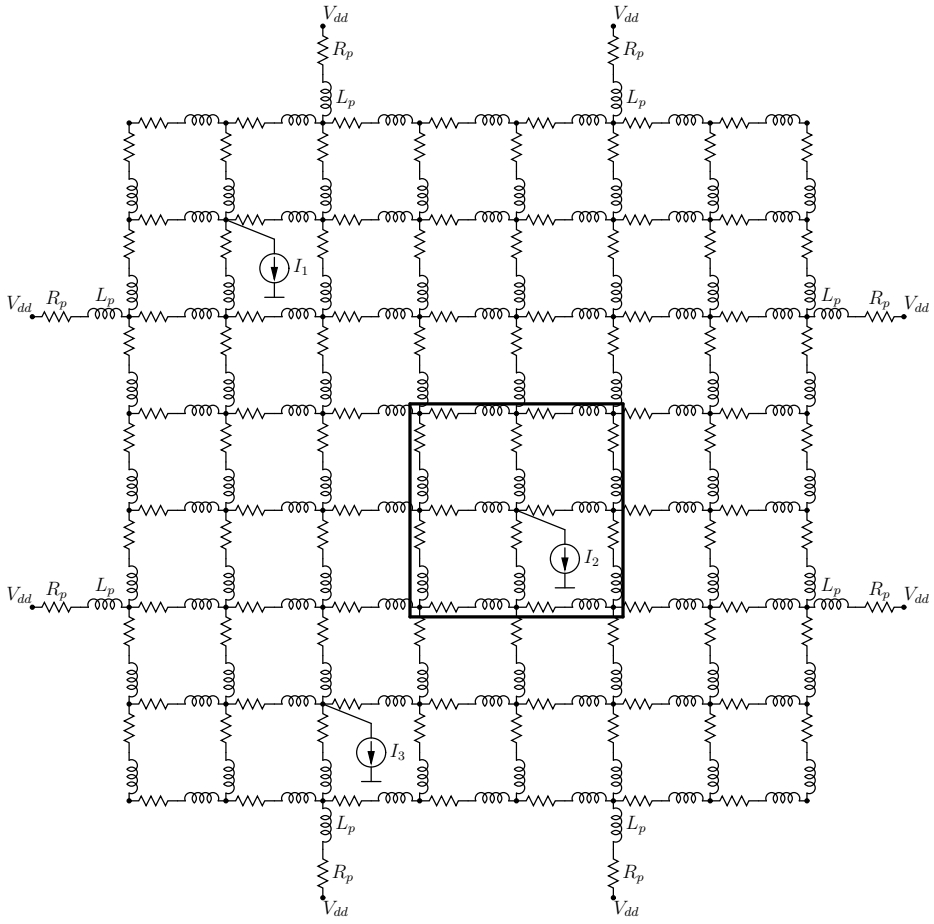


Figure 7.12: Model of a power distribution network. The on-chip power delivery system is modeled as a distributed RL mesh with seven by seven equal segments. The current loads are modeled as current sources with equivalent magnitude and switching activities. R_p and L_p denote the parasitic resistance and inductance of the package, respectively. The rectangle denotes a "hot" spot – the area where the power supply voltage drops below the minimum tolerable level.

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. Disagreeing with Fig. 7.11, the overall effective radius is actually shaped more like a diamond, as illustrated in Fig. 7.13.

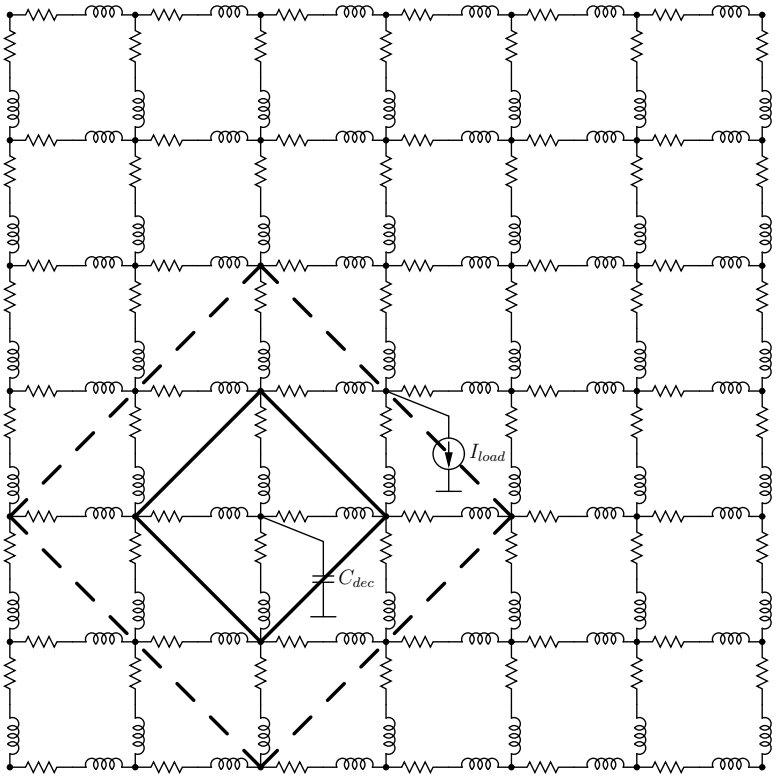


Figure 7.13: Effective radii of an on-chip decoupling capacitor. For a power distribution system modeled as a distributed RL mesh, the maximum effective radius is the Manhattan distance between two points. The overall effective radius is therefore shaped like a diamond.

In modern high performance ICs, up to 3000 I/O pins can be necessary [8]. 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 inches by 1.5 inches inside a flip-chip package, the distance between two adjacent power or ground pads is about $1300\,\mu$m. 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.

7.7 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). 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 (four closest pins of a flip-chip package) are also included in the model. The proposed 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 7.1. These three scenarios listed in Table 7.1 represent typical values of the parasitic resistance and inductance of the top, intermediate, and bottom layers of on-chip interconnects in a 90 nm CMOS technology [8]. 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 (7.11). Hence, $IR \gg L\dfrac{dI}{dt}$, and the required on-chip decoupling capacitance is determined by (7.5). 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 (7.9).

Table 7.1: 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	45454	2.3	0.2
Bottom	0.1	0.1	0.01	50000	99.8	0
	0.1	0.1	0.1	∞	0	0
	0.1	0.1	1	∞	0	0
$V_{dd} = 1\,\mathrm{V}$, $V_{ripple} = 100\,\mathrm{mV}$, $t_r = 100\,\mathrm{ps}$, $t_f = 300\,\mathrm{ps}$, $t_{ch} = 400\,\mathrm{ps}$						

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 7.2. From (7.11), for the parameters listed in Table 7.2, the critical voltage drop is 75 mV. If the voltage fluctuations at the current load do not exceed the critical voltage, $IR \gg L\dfrac{dI}{dt}$ and the required on-chip decoupling capacitance is determined by (7.5). 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 inductive $L\frac{dI}{dt}$ noise dominates and the required on-chip decoupling capacitance is determined by (7.9).

Table 7.2: 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	–

$V_{dd} = 1$ V, $V_{ripple} = 100$ mV, $t_r = 100$ ps,
$t_f = 300$ ps, $t_{ch} = 400$ ps, cell size is 32.5 μm × 32.5 μm

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. 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_{dec}^{r} as determined by (7.9) 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 (7.9) 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 $L\dfrac{dI}{dt}$ drop).

The effective radii listed in Table 7.1 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 7.1 to Table 7.2, 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 a 10 mA and 100 mA current load, respectively. 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.

7.8 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 Chapter 8. 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 proposed methodology for placing and sizing on-chip decoupling capacitors results in a greately reduced budgeted on-

chip decoupling capacitance as compared to a uniform (or blind) placement of on-chip decoupling capacitors into any available white space [128].

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. 7.14. Each decoupling capacitor provides 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).

7.9 Chapter 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

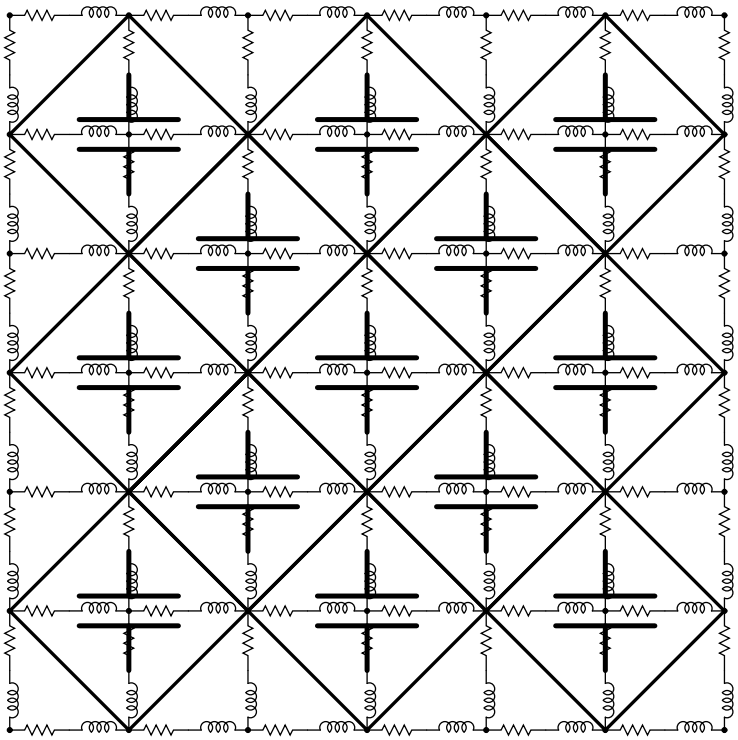


Figure 7.14: A schematic example allocation of on-chip decoupling capacitors across an IC. Similar current loads are assumed to be uniformly distributed on the die. Each on-chip decoupling capacitor provides sufficient charge to the current load(s) within the maximum effective radius.

- 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 $L\frac{dI}{dt}$ 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 8

Efficient Placement of Distributed On-Chip Decoupling Capacitors

Decoupling capacitors are widely used to manage power supply noise [199] and are an effective way to reduce the impedance of power delivery systems operating at high frequencies [14], [15]. 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 [13], 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 [196].

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 [121], 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 [121]. 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 Chapter 7, a system of *distributed* on-chip decoupling capacitors is proposed. 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.

The chapter is organized as follows. Technology limitations in nanoscale integrated circuits are reviewed in Section 8.1. The problem of placing on-chip decoupling capacitors in nanoscale ICs while satisfying technology constraints is formulated in Section 8.2. The design of a distributed on-chip decoupling capacitor network is presented in Section 8.3. Various design tradeoffs are discussed in Section 8.4. A design methodology for placing distributed on-chip decoupling capacitors is presented in Section 8.5. Related simulation results for typical values of on-chip parasitic resistances are discussed in Section 8.6. Some specific conclusions are summarized in Section 8.7.

8.1 Technology Constraints

On-chip decoupling capacitors have traditionally been designed as standard gate oxide CMOS capacitors [203]. As technology scales, leakage current through the gate oxide of an on-chip decoupling capacitor has greatly increased [139], [204], [205]. Moreover, in modern high performance ICs, a large portion (up to 40%) of the circuit area is occupied by the on-chip decoupling capacitance [206], [207]. 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 [208].

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 $10\,\mathrm{fF}/\mu\mathrm{m}^2$ to $30\,\mathrm{fF}/\mu\mathrm{m}^2$ [99], [102], [113]. 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. 8.1.

Observe from Fig. 8.1 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.

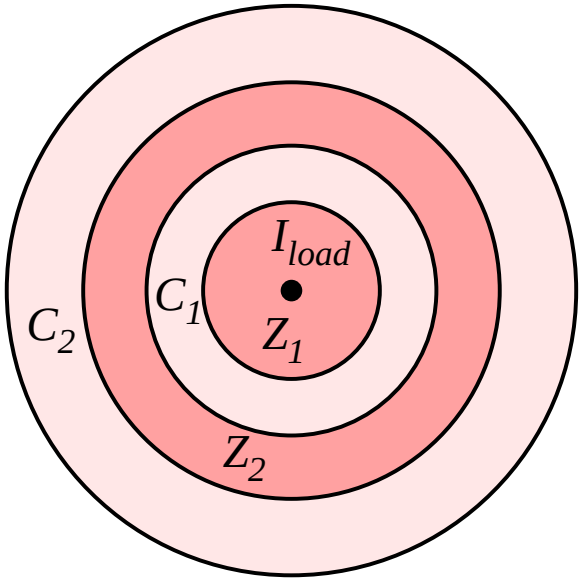


Figure 8.1: Fundamental limits of on-chip interconnections. Two levels of a distributed on-chip decoupling capacitance are allocated around a current load. The interconnect impedance is inversely proportional to the fraction of metal area dedicated to the interconnect level, decreasing as the decoupling 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 in dark grey and light grey, respectively.

8.2 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 [131], [128], as shown in Fig. 8.2. In this way, decoupling

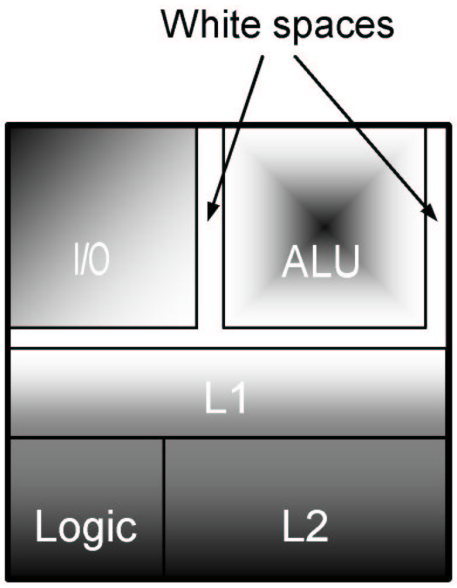


Figure 8.2: Placement of on-chip decoupling capacitors using a conventional approach. Decoupling capacitors are allocated into the white space (those areas not occupied by the circuits elements) available on the die using an unsystematic or *ad hoc* approach. As a result, the power supply voltage drops below the minimum tolerable level for remote blocks (shown in dark grey). Low noise regions are light grey.

capacitors are often placed at a significant distance from the current load. Conventional approaches for placing on-chip decoupling capacitors result 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. 8.3. 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.

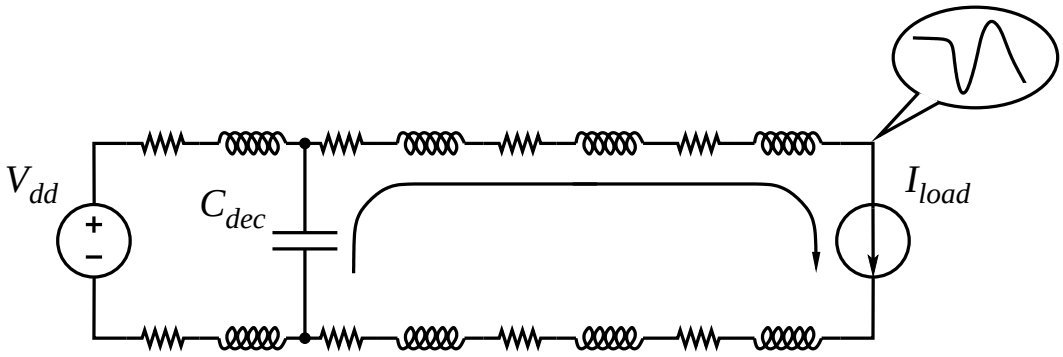


Figure 8.3: A conventional on-chip decoupling capacitor. Typically, a large decoupling capacitor is placed farther from the current load due to physical limitations. Current flowing through the long power/ground lines results in large voltage fluctuations across the terminals of the current load.

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) [61]. In this case, a lumped model of a decoupling capacitor provides sufficient accuracy [209].

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. 8.4.

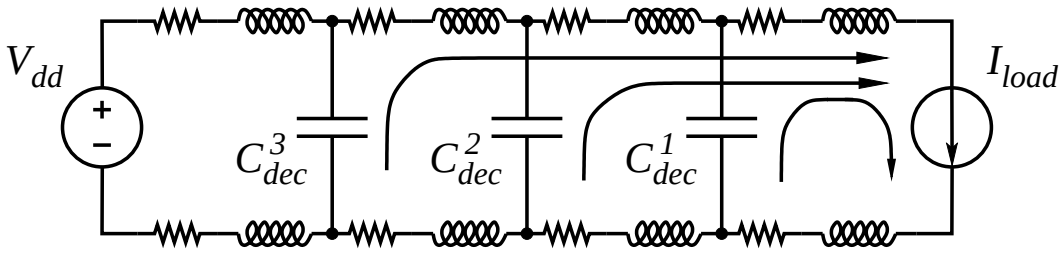


Figure 8.4: A network of distributed on-chip decoupling capacitors. The magnitude of the decoupling capacitors is based on the impedance of the interconnect segment connecting a specific capacitor to a current load. Each decoupling capacitor is designed to only provide charge during a specific time interval.

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 the proposed 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 [14], [46], the proposed 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.

8.3 Design of a Distributed On-Chip Decoupling Capacitor Network

As described in Section 8.2, 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 proposed technique is illustrated in Fig. 8.5. 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 proposed 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).

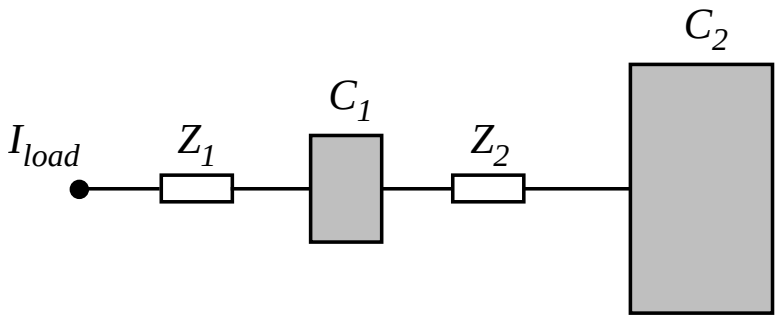


Figure 8.5: A physical model of the proposed system of distributed on-chip decoupling capacitors. Two capacitors are assumed to provide the required charge drawn by the load. Z_1 and Z_2 denote the impedance of the metal lines connecting C_1 to the current load and C_2 to C_1, respectively.

A circuit model of the proposed system of distributed on-chip decoupling capacitors is shown in Fig. 8.6. 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 current load is 10% of the power supply voltage.

Note from Fig. 8.6 that since the charge drawn by the current load is provided by the on-chip decoupling capacitors, the voltage across the capacitors during discharge

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.

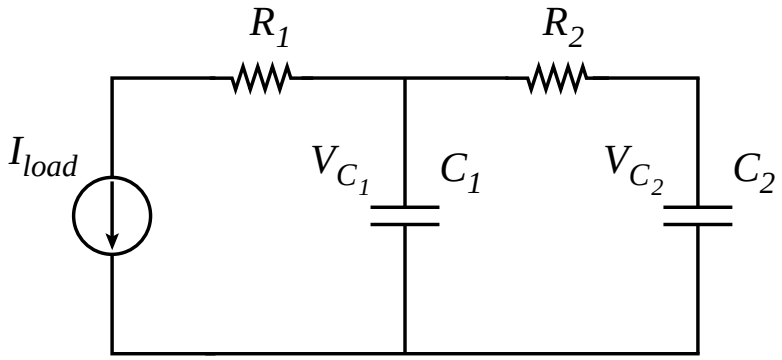


Figure 8.6: A circuit model of an on-chip distributed decoupling capacitor network. The impedance of the metal lines is modeled as R_1 and R_2, respectively.

The voltage across the decoupling capacitors at the end of the switching cycle ($t = t_r$) can be determined from Kirchhoff's laws [202]. Writing KVL and KCL equations for each of the loops (see Fig. 8.6), the system of differential equations describing the voltage across C_1 and C_2 at t_r is

$$\frac{dV_{C_1}}{dt} = \frac{V_{C_2} - V_{C_1}}{R_2 C_1} - \frac{I_{load}}{C_1}, \tag{8.1}$$

$$\frac{dV_{C_2}}{dt} = \frac{V_{C_1} - V_{C_2}}{R_2 C_2}. \tag{8.2}$$

Simultaneously solving (8.1) and (8.2) and applying the initial conditions, the voltage across C_1 and C_2 at the end of the switching activity is

$$V_{C_1}|_{t=t_r} = \frac{1}{2(C_1+C_2)^3 t_r}\Bigg[2C_1^3 t_r + C_1^2 t_r(6C_2 - I_{max}t_r) \\ - C_2^2 t_r\left(2C_2\left(I_{max}R_2 - 1\right) + I_{max}t_r\right) \\ + 2C_1C_2\left(C_2^2\left(1 - e^{-\frac{(C_1+C_2)t_r}{C_1C_2R_2}}\right)I_{max}R_2^2 \\ + C_2\left(3 - I_{max}R_2\right)t_r - I_{max}t_r^2\right)\Bigg], \tag{8.3}$$

$$V_{C_2}|_{t=t_r} = \frac{1}{2(C_1+C_2)^3 t_r}\Bigg[2C_1^3 t_r + C_2^2 t_r\left(2C_2 - I_{max}t_r\right) \\ + 2C_1C_2t_r\left(C_2\left(3 + I_{max}R_2\right) - I_{max}t_r\right) \\ + C_1^2\left(2C_2^2\left(e^{-\frac{(C_1+C_2)t_r}{C_1C_2R_2}} - 1\right)I_{max}R_2^2 \\ + 2C_2\left(3 + I_{max}R_2\right)t_r - I_{max}t_r^2\right)\Bigg], \tag{8.4}$$

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

$$V_{C_1} = V_{load} + I_{max}R_1, \tag{8.5}$$

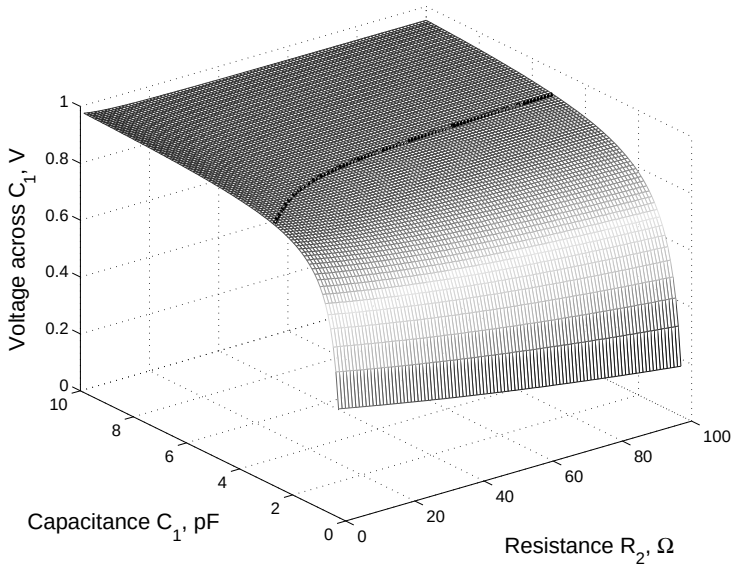
where V_{load} is the voltage across the terminals of a current load. Assuming $V_{load} \geq 0.9V_{dd}$ and $V_{C_1}^{max} = V_{dd}$ (meaning that C_1 is infinitely large), the upper bound for R_1 is

$$R_1^{max} = \frac{V_{dd}(1-\alpha)}{I_{max}}, \tag{8.6}$$

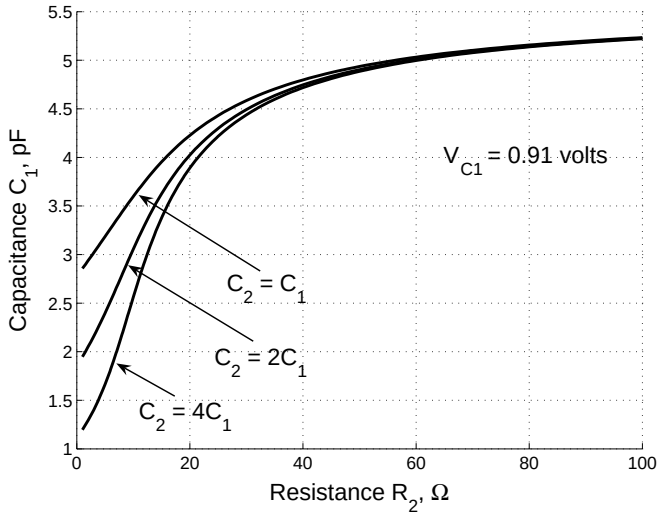
where α is the ratio of the minimum tolerable voltage across the terminals of a current load to the power supply voltage ($\alpha = 0.9$ in this chapter). If $R_1 > R_1^{max}$, no solution exists for providing sufficient charge drawn by the load. In this case, the circuit block should be partitioned, 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 (8.3) and (8.4), 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. 8.7.

Observe from Fig. 8.7 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. 8.7(a)). 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



(a)



(b)

Figure 8.7: Voltage across C_1 during discharge as a function of C_1 and R_2: $I_{max} = 0.01\,\text{mA}$, $V_{dd} = 1$ volt, and $t_r = 100\,\text{ps}$. a) Assuming $C_1 = C_2$ and $R_1 = 10\,\Omega$, the minimum tolerable voltage across C_1, resulting in $V_{load} \geq 0.9V_{dd}$, is 0.91 volts (shown as a black equipotential line). b) The design space for determining C_1 and R_2 resulting in the voltage across C_1 equal to 0.91 volts.

capacitance therefore results in a smaller voltage drop. From Fig. 8.6, 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. 8.7(b).

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 $V_{load} = 0.9V_{dd}$. The maximum magnitude of the current load I_{max} is 0.01 A, the rise time t_r is 100 ps, and the power supply voltage V_{dd} is one volt. Note that the voltage across C_2 after discharge as determined by (8.4) 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 volts).

The system of equations to determine the parameters of an on-chip distributed decoupling capacitor network as depicted in Fig. 8.6 is

$$V_{load} = V_{C_1} - I_{max} R_1, \tag{8.7}$$

$$V_{C_1} = f(C_1, C_2, R_2), \tag{8.8}$$

$$V_{C_2} = f(C_1, C_2, R_2), \tag{8.9}$$

$$\frac{I_{max} t_r}{2} = C_1 \left(V_{dd} - V_{C_1}\right) + C_2 \left(V_{dd} - V_{C_2}\right), \tag{8.10}$$

where V_{C_1} and V_{C_2} are the voltage across C_1 and C_2 and determined by (8.3) and (8.4), respectively. Equation (8.10) 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 (8.7) for V_{C_1} and substituting into (8.8), C_1, C_2, and R_2 are determined from (8.8), (8.9), and (8.10) for a specified $V_{C_2}(t_r)$, as discussed in the following section.

8.4 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 Section 8.3, in a system composed of two decoupling capacitors (see Fig. 8.6) with known R_1; R_2, C_1, and C_2 are determined from the system of equations, (8.7)–(8.10). 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 equations (8.7)–(8.10) is solved numerically [210].

Various tradeoff scenarios are discussed in this section. The dependence of the system parameters on R_1 is presented in Section 8.4.1. The design of a distributed on-chip decoupling capacitor network with the minimum magnitude of C_1 is discussed in Section 8.4.2. 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 Section 8.4.3. The minimum total budgeted on-chip decoupling capacitance is also determined in this section.

8.4.1 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 8.1. 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 8.1, the minimum magnitude of R_2 is four ohms, as determined from numerical simulations.

Table 8.1: Dependence of the parameters of a distributed on-chip decoupling capacitor network on R_1

R_1 (Ω)	$R_2 = 5\,(\Omega)$		$R_2 = 10\,(\Omega)$	
	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
$V_{dd} = 1\,\text{V}$, $V_{load} = 0.9\,\text{V}$, $t_r = 100\,\text{ps}$, and $I_{max} = 0.01\,\text{A}$				

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 8.1). 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.

8.4.2 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 on-chip decoupling capacitors as a function of R_1 under the constraint of a minimum C_1 are listed in Table 8.2. Note that V_{C_2} denotes the voltage across C_2 after discharge.

Table 8.2: 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				$V_{C_2} = 0.95$ volt	
(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	–	–

$V_{dd} = 1$ V, $V_{load} = 0.9$ V, $t_r = 100$ ps,
$I_{max} = 0.01$ A, and $C_1 = 1$ pF

Note that two scenarios are considered in Table 8.2 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 negative, 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 [211].

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.

8.4.3 Minimum Total Budgeted On-Chip Decoupling Capacitance

As discussed in Sections 8.4.1 and 8.4.2, 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 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 [208]. 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. 8.8. 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. 8.8 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

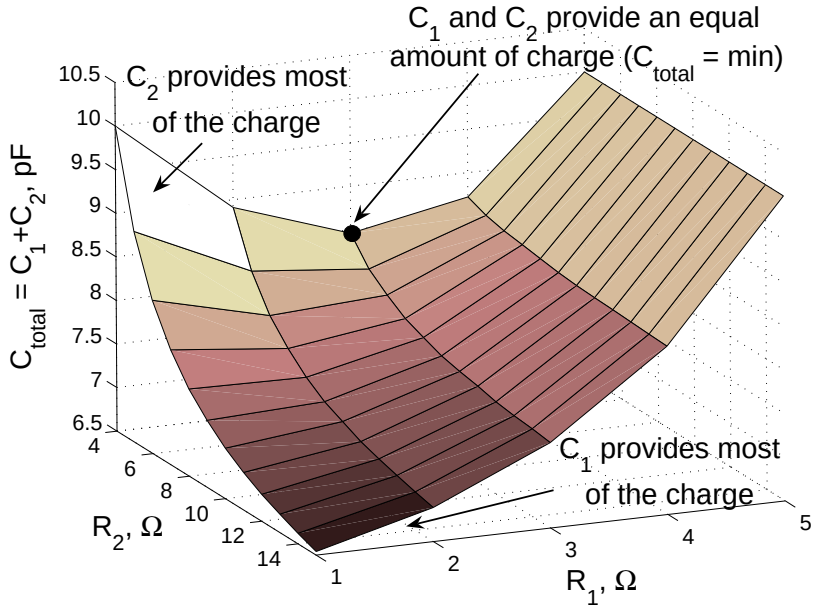


Figure 8.8: The total budgeted on-chip decoupling capacitance as a function of the parasitic resistance of the metal lines, R_1 and R_2: $I_{max} = 10\,\text{mA}$, $V_{dd} = 1\,\text{volt}$, $V_{load} = 0.9\,\text{volt}$, and $t_r = 100\,\text{ps}$. In the system of distributed on-chip decoupling capacitors, an optimal ratio $\frac{R_2}{R_1}$ exists, resulting in the minimum total budgeted on-chip decoupling capacitance.

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. 8.8 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 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. 8.9. In contrast to the results depicted in Fig. 8.8, 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. 8.9), the total budgeted on-chip decoupling capacitance increases rapidly, becoming impractically large.

Comparing Fig. 8.8 to Fig. 8.9, 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

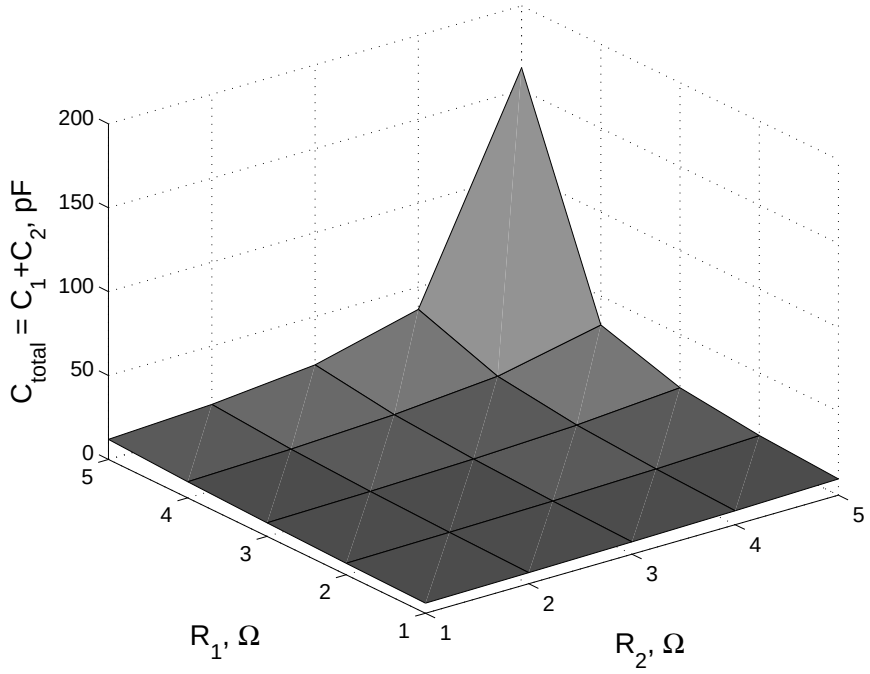


Figure 8.9: The total budgeted on-chip decoupling capacitance as a function of the parasitic resistance of the metal lines, R_1 and R_2: $I_{max} = 10\,\text{mA}$, $V_{dd} = 1\,\text{volt}$, $V_{load} = 0.9\,\text{volt}$, and $t_r = 100\,\text{ps}$. C_1 is fixed and set to 1 pF. The total budgeted on-chip 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.

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) [117].

8.5 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. 8.10. 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 (8.6) 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 (8.1) and (8.2), the voltage across C_1 and C_2 is estimated at the end of a switching cycle ($t = t_r$), as determined by (8.3) and (8.4). The parameters of the distributed on-chip decoupling capacitor network C_1, C_2, and R_2, are determined from (8.7)–(8.10). Note that different tradeoffs exist in a system of distributed on-chip decoupling capacitors, as discussed in Section 8.4. If the voltage across the terminals of a current load drops below the minimum tolerable level,

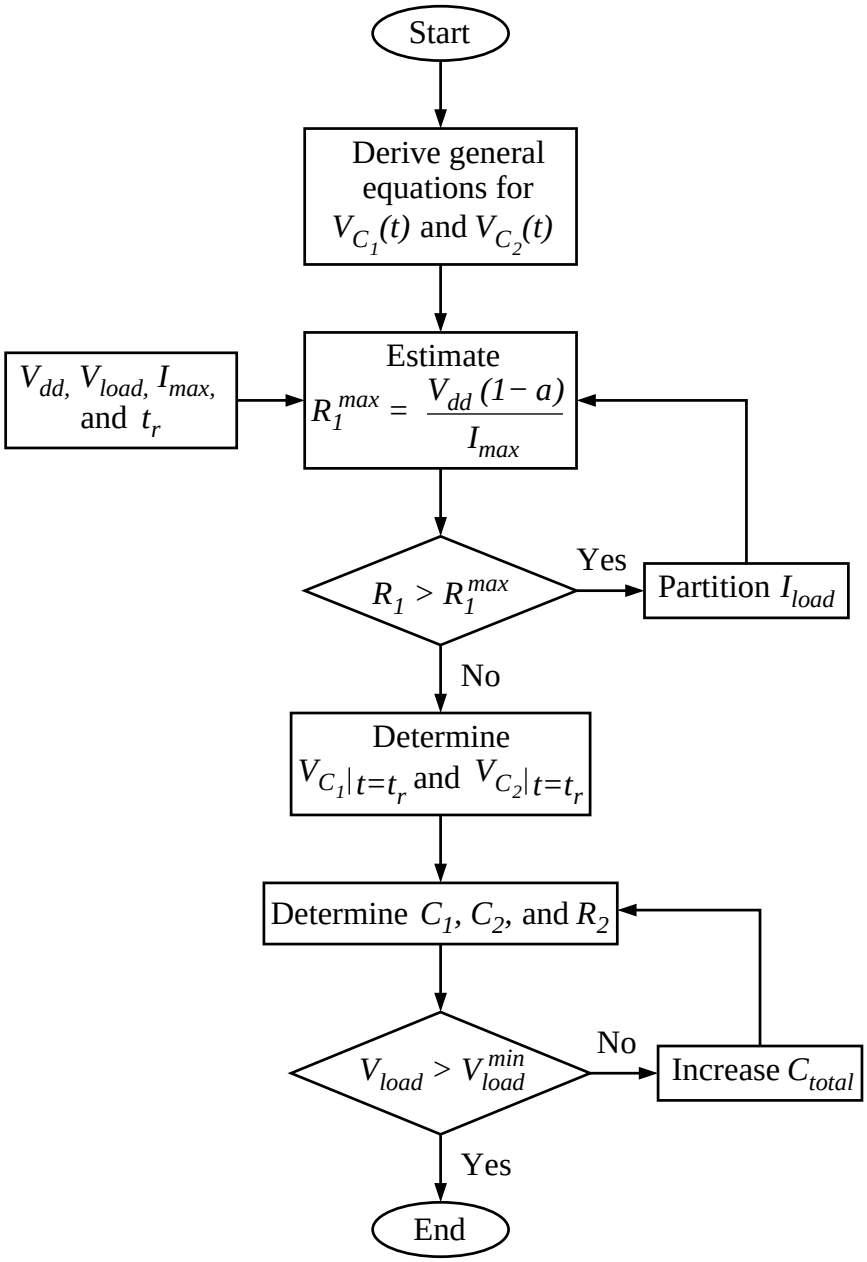


Figure 8.10: Design flow for determining the parameters of a system of distributed on-chip decoupling capacitors.

the total budgeted on-chip decoupling capacitance should be increased. The system of equations, (8.7) to (8.10), 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. 8.10.

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 proposed 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 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.

8.6 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. 8.6). 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 8.3. The parameters of the distributed on-chip decoupling capacitor network listed in Table 8.3 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 90 nm CMOS technology.

The parameters of the system of distributed on-chip decoupling capacitors are analytically determined from (8.7)–(8.10). 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 8.3, 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

Table 8.3: 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

$V_{dd} = 1\,\text{V}$ and $t_r = 100\,\text{ps}$

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 8.3 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 8.3 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 8.4. Note that in contrast to Table 8.3, 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 8.4 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.

Table 8.4: 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	(%)
$C_1 = 0.5\,\text{pF}$									
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
$C_1 = 1\,\text{pF}$									
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
$V_{dd} = 1\,\text{V}$ and $t_r = 100\,\text{ps}$									

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 V_{load}^{min} and $V_{C_2}^{min}$, respectively. 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%.

Comparing results from Table 8.4 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 Section 8.4.3.

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 [212]. 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 [121]. A tradeoff therefore exists between the magnitude and location of the on-chip decoupling capacitors comprising the distributed decoupling capacitor network.

8.7 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 proposed 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 value 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 9

Conclusions

The operation of an integrated circuit relies on the supply of power to the on-chip circuitry. Power distribution systems serve the purpose of supplying an integrated circuit with current while maintaining specific voltage levels. The power current is distributed across an integrated circuit through an on-chip power distribution network, an integral part of the overall power distribution system. The on-chip power distribution network delivers current to hundreds of millions of high speed transistors comprising a high complexity integrated circuit. Tens of amperes must be efficiently distributed to supply power to the on-chip circuits. Due to the high currents and high frequencies, the impedance of a power distribution system should be maintained sufficiently low over a wide range of frequencies in order to limit the voltage variations at the power load — the millions of on-chip transistors.

To maintain the impedance of a power distribution system below a specified level (the 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 result in antiresonance — an increase in the impedance of the power distribution system. 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 ratio of 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 cancel the antiresonance, shifting the antiresonant spikes to a frequency greater than the maximum operating frequency of the system.

Controlling the inductive characteristics of the interconnect comprising a power distribution network in a complex on-chip environment has become of significant importance in high speed circuits. The inductance is an essential parameter in determining the high frequency response of a power distribution grid. The significant inductive behavior of an on-chip power distribution network makes the power supply noise difficult to predict, exacerbating the analysis and verification process. On-chip

power distribution grids with multiple power supply voltages and multiple grounds are proposed in this dissertation. In this power delivery scheme, the loop inductance of a grid is effectively reduced, exploiting mutual coupling between power and ground paths.

The impedance characteristics of multi-layer grids, which consist of relatively thick and wide lines in the upper layers and fine lines in the lower layers, are particularly well suited for distributing power in high speed circuits. The upper layers provide a low impedance current path at low frequencies, while the lower layers serve as a low impedance path at high frequencies. This behavior facilitates maintaining low impedance characteristics over a wide frequency range.

Despite recent advancements in integrated circuit technologies and packaging solutions, on-chip decoupling capacitors remain an attractive cost effective solution for providing a low impedance power distribution network supplying current over a wide range of frequencies. A decoupling capacitor acts as a local reservoir of charge, which 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 sub-100 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 circuit area, on-chip MOS decoupling capacitors contribute

more than half of the total leakage power in modern high speed, high complexity ICs. New types of on-chip decoupling capacitors, such as MIM and lateral flux capacitors, have recently emerged as better candidates for on-chip decoupling capacitors.

On-chip decoupling capacitors have traditionally been allocated into 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 such that both the power supply and the current load are located inside 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 a significant portion of the on-chip area. A system of distributed on-chip decoupling capacitors should therefore be utilized in nanoscale ICs to satisfy technology and performance constraints. The methodologies for placing on-chip decoupling capacitors presented in this dissertation provide a computationally efficient way for determining the required on-chip decoupling capacitance to support expected current demands. In this way, the circuit area required for the on-chip decoupling capacitors is estimated at an early stage

of the design process, significantly reducing the number of design iterations and the eventual time to market.

The topics presented in this dissertation on the design of power distribution networks with decoupling capacitors are intended to provide insight into the electrical behavior and design principles of these high performance systems. A thorough understanding of the electrical phenomena in complex multi-layer power distribution networks with on-chip decoupling capacitors is therefore essential for developing effective methodologies and computer-aided tools for designing the next generation of nanoscale CMOS integrated systems.

Chapter 10

Future Research

The feature size of integrated circuits has been aggressively reduced in the pursuit of improved speed, power, and cost. Semiconductor technologies with feature sizes of several tens of nanometers are currently in development. The scaling of CMOS is expected to continue for at least another decade [213]. Modern nanometer circuits contain about a billion transistors and operate at clock speeds close to 10 GHz. Further improvements in circuit speed and integration density will exacerbate challenges in the design of power distribution systems, requiring lower impedance characteristics over a wider range of frequencies. On-chip decoupling capacitors are expected to occupy more than 50% of the total circuit area in future nanoscale ICs, dissipating a major portion of the total power. Existing techniques and tools for designing power distribution networks with on-chip decoupling capacitors need to be improved to meet these challenges.

Several future research directions are described in this chapter. An on-chip power distribution grid in high speed ICs is a complex structure, spanning several metal layers. A multi-layer model of an on-chip power distribution grid is therefore required to accurately analyze power noise and signal integrity in high complexity ICs. A model of a multi-layer on-chip power distribution grid for nanoscale ICs is suggested in Section 10.1. With recent developments in high performance packaging, on-chip power distribution grids are tightly coupled with the high performance packages. Chip-package co-design methodologies therefore need to be developed to accurately analyze power and signal integrity in nanoscale ICs, as described in Section 10.2. On-chip decoupling capacitors in mixed-signal and RF ICs can dramatically degrade substrate noise coupling. A design methodology for placing on-chip decoupling capacitors in mixed-signal ICs and systems-on-chip is discussed in Section 10.3. The required on-chip decoupling capacitance will be significantly reduced in 3-D ICs, as described in Section 10.4. Finally, a summary is provided in Section 10.5.

10.1 A Multi-Layer Model of On-Chip Power Distribution Grids

Determining the frequency dependent impedance characteristics of on-chip power distribution grids is essential to allocate a sufficient amount of metal resources in

high speed nanoscale integrated circuits. The power distribution network competes with signal routing for limited on-chip metal resources. A balanced decision should therefore be made to satisfy noise margins while consuming the minimum metal area dedicated to the power distribution network. As described in Chapter 7, allocating the on-chip decoupling capacitance also depends strongly upon the characteristics of the on-chip power distribution grid. On-chip power distribution systems have traditionally been modeled as a single layer mesh. In a distributed RLC mesh, the complexity of the model is significantly reduced, resulting in a highly inaccurate estimate of the grid impedance and power noise. Alternatively, complexity can be traded off with accuracy.

In complex power distribution grids, multiple parallel current return paths exist. Similar to multi-path current redistribution as described in [12], high frequency power current tends to flow in the bottom metal layers (the paths with the lowest inductance). At low frequencies, however, the power current flows in the top metal layers with minimum resistance, as shown in Fig. 10.1. As switching frequencies further increase, this phenomenon is expected to become more pronounced. A multi-layer frequency dependent model of the on-chip power distribution grids should therefore be developed, permitting the power noise to be accurately estimated. The impact of vias should also be included in the model.

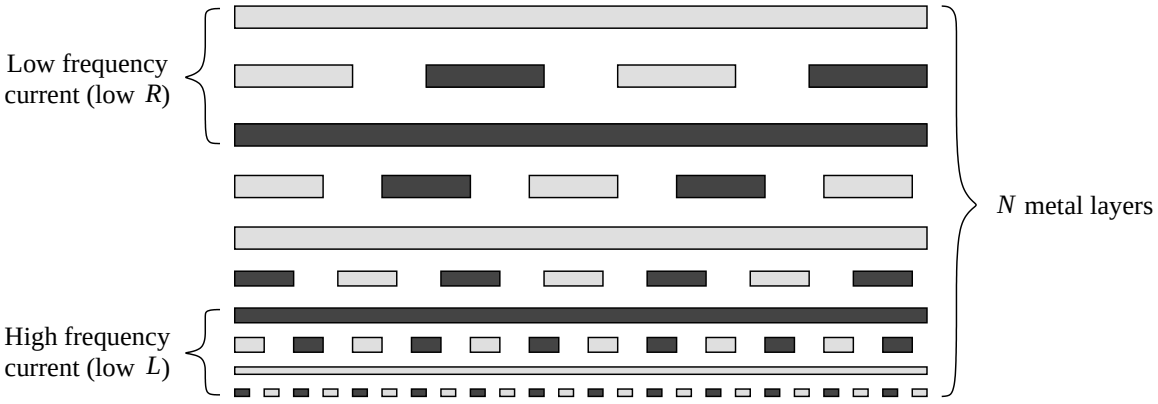


Figure 10.1: Current redistribution in a multi-layer power distribution grid. The power lines are dark grey and the ground lines are light grey (signal lines are not shown). In interdigitated power distribution grids, high frequency currents shift towards the low inductance bottom layers. At low frequencies, however, the currents flow in the top metal layers with the lowest resistance.

10.2 Chip-Package Co-Design Methodologies

The integration density and speed of modern high performance integrated circuits have substantially increased. The electrical performance of circuit packages has also improved, necessitating tight integration with the ICs. The power distribution network in an area array flip-chip package can be considered as an extension of the on-chip power distribution grid. Due to the relatively low impedance of the chip-to-package contacts, the package decoupling capacitors are more efficient, contributing to the total charge provided by the decoupling capacitance. The chip-package impedance characteristics also affect the frequency and magnitude of the (anti)resonance peaks. In this way, the local behavior of the on-chip power supply is greatly affected by the

package characteristics. If the electrical characteristics of the package are customized to match specific power requirements, the on-chip power supply integrity is enhanced while reducing the on-chip metal resources for distributing power and ground. As the circuit complexity further increases, co-designing the package with the circuit becomes more important. Packages and circuits, however, have traditionally been designed separately at different stages of the design process. As these trends are expected to continue, novel chip-package co-design methodologies will be required to efficiently distribute power and ground networks and allocate on-chip decoupling capacitors.

10.3 Substrate Noise-Aware Design Methodology for Placing On-Chip Decoupling Capacitors

As described in Chapters 7 and 8, in high complexity ICs, on-chip decoupling capacitors can occupy up to 40% of the total circuit area [206], [207], [208]. A greater portion of the total circuit area is expected to be occupied by the on-chip decoupling capacitors. As more functional blocks are integrated onto a single die, satisfying tighter noise margins will become a more challenging task. For example, on-chip decoupling capacitors placed in high speed circuits are connected to the

noisy power/ground paths, injecting noise into the substrate. The on-chip decoupling capacitors placed near noise sensitive (*e.g.*, analog) circuits will amplify the substrate noise, couple noise to the dedicated low noise power distribution network, as illustrated in Fig. 10.2. The signal integrity of an entire system will therefore be severely compromised. A design methodology for placing on-chip decoupling capacitors in mixed-signal circuits and systems-on-chip is therefore required to provide the required charge to the switching circuits while reducing substrate noise coupling.

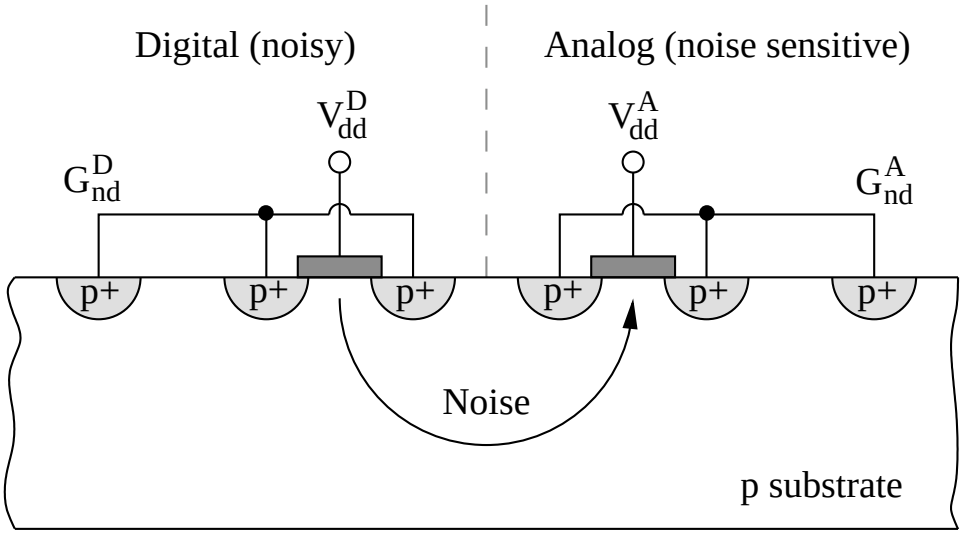


Figure 10.2: Substrate noise coupling in mixed-signal integrated circuits with on-chip decoupling capacitors. Noise from the power/ground paths in the high speed digital circuits is injected into the substrate. The on-chip decoupling capacitors placed near the noise sensitive analog circuits will amplify and couple the substrate noise into the low noise power distribution network.

10.4 Placement of On-Chip Decoupling Capacitors in 3-D ICs

Recently emerging 3-D IC technologies offer a promising solution for conventional 2-D ICs, reducing both the footprint and the interconnect length without requiring the transistors to be scaled. In 3-D ICs, several dies are stacked together, dramatically reducing the length of the interconnection among the circuits located on adjacent planes. In this way, on-chip decoupling capacitors placed on the adjacent planes are located physically closer to the current load(s) as compared to the on-chip decoupling capacitors in a 2-D circuit. Various types of heterogeneous circuits can be integrated into a 3-D IC, enhancing the potential functionality of a system. A larger intrinsic decoupling capacitance can be available in the vicinity of a switching circuit. The contribution of the on-chip decoupling capacitors placed on adjacent planes should therefore be considered in the design process. Techniques for placing and accurately estimating the on-chip decoupling capacitors in 3-D ICs need to be developed, reducing the total required on-chip decoupling capacitance.

10.5 Summary

Several research directions have been described in this chapter to improve existing methodologies for designing power distribution networks while enhancing signal

integrity by allocating on-chip decoupling capacitors. With the scaling of CMOS technologies, distributing power and ground in high speed, high complexity ICs will continue to be a challenging problem. As the power supply voltage is further reduced and the total current increases, on-chip decoupling capacitors will remain an indispensable solution to enhance signal integrity while controling the impedance characteristics of the on-chip power distribution networks. Significant research effort is therefore necessary to satisfy increasingly challenging requirements for designing high performance power distribution systems.

Bibliography

[1] C. Pirtle, *Engineering the World: Stories from the First 75 Years of Texas Instruments*. Dallas, TX: Southern Methodist University Press, 2005.

[2] T. R. Reid, *The Chip: How Two Americans Invented the Microchip and Launched a Revolution*. New York, NY: Random House, Inc., 2001.

[3] L. Berlin, *Man Behind the Microchip: Robert Noyce and the Invention of Silicon Valley*. New York, NY: Oxford University Press, 2005.

[4] J. S. Kilby, "Miniaturized Electronic Circuits," U.S. Patent # 3,138,743, June 23, 1964.

[5] J. A. Hoerni, "Planar Silicon Transistors and Diodes," *IRE Transactions on Electron Devices*, Vol. 8, No. 2, p. 178, March 1961.

[6] D. Kahng, "A Historical Perspective on the Development of MOS Transistors and Related Devices," *IEEE Transactions on Electron Devices*, Vol. 23, No. 7, pp. 655–657, July 1976.

[7] R. E. Kerwin, D. L. Klein, and J. C. Sarace, "Method for Making MIS Structures," U.S. Patent # 3,475,234, October 28, 1969.

[8] *International Technology Roadmap for Semiconductors, 2005 Edition.* Semiconductor Industry Association, 2005. [Online]. Available: http://public.itrs.net

[9] K. T. Tang and E. G. Friedman, "Estimation of Transient Voltage Fluctuations in the CMOS-Based Power Distribution Networks," *Proceedings of the IEEE International Symposium on Circuits and Systems*, pp. 463–466, May 2001.

[10] K. T. Tang and E. G. Friedman, "On-Chip ΔI Noise in the Power Distribution Networks of High Speed CMOS Integrated Circuits," *Proceedings of the IEEE International ASIC/SOC Conference*, pp. 53–57, September 2000.

[11] A. V. Mezhiba and E. G. Friedman, *Power Distribution Networks in High Speed Integrated Circuits*. Norwell, MA: Kluwer Academic Publishers, 2004.

[12] A. V. Mezhiba, "Design and Analysis of On-Chip High Performance Power Distribution Networks," Ph.D. Dissertation, University of Rochester, Rochester, NY, 2004.

[13] A. Mezhiba and E. G. Friedman, "Scaling Trends of On-Chip Power Distribution Noise," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, Vol. 12, No. 4, pp. 386–394, April 2004.

[14] M. Popovich and E. G. Friedman, "Decoupling Capacitors for Multi-Voltage Power Distribution Systems," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, Vol. 14, No. 3, pp. 217–228, March 2006.

[15] M. Popovich and E. G. Friedman, "Impedance Characteristics of Decoupling Capacitors in Multi-Power Distribution Systems," *Proceedings of the IEEE International Conference on Electronics, Circuits and Systems*, pp. 160–163, December 2004.

[16] K. T. Tang and E. G. Friedman, "Delay Uncertainty Due to On-Chip Simultaneous Switching Noise in High Performance CMOS Integrated Circuits," *Proceedings of the IEEE Workshop on Signal Processing Systems*, pp. 633–642, October 2000.

[17] K. T. Tang and E. G. Friedman, "Incorporating Voltage Fluctuations of the Power Distribution Network into the Transient Analysis of CMOS Logic Gates," *Analog Integrated Circuits and Signal Processing*, Vol. 31, No. 3, pp. 249–259, June 2002.

[18] M. Saint-Laurent and M. Swaminathan, "Impact of Power-Supply Noise on Timing in High-Frequency Microprocessors," *Proceedings of the IEEE Conference on Electrical Performance of Electronic Packaging*, pp. 261–264, October 2002.

[19] M. Saint-Laurent and M. Swaminathan, "Impact of Power-Supply Noise on Timing in High-Frequency Microprocessors," *IEEE Transactions on Advanced Packaging*, Vol. 27, No. 1, pp. 135–144, February 2004.

[20] A. Waizman and C.-Y. Chung, "Package Capacitors Impact on Microprocessor Maximum Operating Frequency," *Proceedings of the IEEE Electronic Components and Technology Conference*, pp. 118–122, June 2001.

[21] E. G. Friedman, *Clock Distribution Networks in VLSI Circuits and Systems*. Piscataway, NJ: IEEE Press, 1995.

[22] E. G. Friedman, *High Performance Clock Distribution Networks*. Norwell, MA: Kluwer Academic Publishers, 1997.

[23] I. S. Kourtev and E. G. Friedman, *Timing Optimization Through Clock Skew Scheduling*. Norwell, MA: Kluwer Academic Publishers, 2000.

[24] J. P. Eckhardt and K. A. Jenkins, "PLL Phase Error and Power Supply Noise," *Proceedings of the IEEE Conference on Electrical Performance of Electronic Packaging*, pp. 73–76, October 1998.

[25] A. W. Strong, *et al.*, *Reliability Wearout Mechanisms in Advanced CMOS Technologies*. New York, NY: John Wiley & Sons, Inc., 2006.

[26] L. Smith, "Reliability and Performance Tradeoffs in the Design of On-Chip Power Delivery and Interconnects," *Proceedings of the IEEE Conference on Electrical Performance of Electronic Packaging*, pp. 49–52, November 1999.

[27] K. T. Tang and E. G. Friedman, "Simultaneous Switching Noise in On-Chip CMOS Power Distribution Networks," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, Vol. 10, No. 4, pp. 487–493, August 2002.

[28] G. E. R. Lloyd, *Early Greek Science: Thales to Aristotle*. New York, NY: W W Norton & Co., Inc., 1974.

[29] A. D. Moore, *Electrostatics and Its Applications*. New York, NY: John Wiley & Sons, Inc., 1973.

[30] J. L. Heilborn, *Electricity in the 17th & 18th Centuaries: A Study in Early Modern Physics*. Mineola, NY: Dover Publications, 1999.

[31] A. Guillemin, *Electricity and Magnetism*. London: Macmillam, 1891.

[32] M. Faraday, *Experimental Researches in Electricity*. Mineola, NY: Dover Publications, 2004.

[33] J. D. Cutnell and K. W. Johnson, *Physics, 6th Edition*. New York, NY: John Wiley & Sons, Inc., 2003.

[34] T. H. Lee, *The Design of CMOS Radio-Frequency Integrated Circuits, 2nd Edition*. New York, NY: Cambridge University Press, 2004.

[35] C. P. Yuan and T. N. Trick, "A Simple Formula for the Estimation of the Capacitance of Two-Dimensional Interconnects in VLSI Circuits," *IEEE Electron Device Letters*, Vol. 3, No. 12, pp. 391–393, December 1982.

[36] T. Sakurai and K. Tamaru, "Simple Formulas for Two- and Three-Dimensional Capacitance," *IEEE Transactions on Electron Devices*, Vol. 30, No. 2, pp. 183–185, February 1983.

[37] J.-H. Chern *et al.*, "Multilevel Metal Capacitance Models for CAD Design Synthesis Systems," *IEEE Electron Device Letters*, Vol. 13, No. 1, pp. 32–34, January 1992.

[38] S.-C. Wong, G.-Y. Lee, and D.-J. Ma, "Modeling of Interconnect Capacitance, Delay, and Crosstalk in VLSI," *IEEE Transactions on Semiconductor Manufacturing*, Vol. 13, No. 1, pp. 108–111, January 2000.

[39] E. Barke, "Line-to-Ground Capacitance Calculation for VLSI: A Comparison," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 7, No. 2, pp. 295–298, February 1988.

[40] N. v.d. Meijs and J. T. Fokkema, "VLSI Circuit Reconstruction from Mask Topology," *Integration*, Vol. 2, No. 2, pp. 85–119, February 1984.

[41] C. R. Paul, *Introduction to Electromagnetic Compatibility*. New York, NY: John Wiley & Sons, Inc., 1992.

[42] R. R. Tummala, E. J. Rymaszewski, and A. G. Klopfenstein, *Microelectronics Packaging Handbook*. Boca Raton, FL: Chapman & Hall, 1997.

[43] T. Roy, L. Smith, and J. Prymak, "ESR and ESL of Ceramic Capacitor Applied to Decoupling Applications," *Proceedings of the IEEE Conference on Electrical Performance of Electronic Packaging*, pp. 213–216, October 1998.

[44] D. A. Neamen, *Semiconductor Physics and Devices: Basic Principles, 3rd Edition*. New York, NY: McGraw-Hill, 2002.

[45] Power Distribution System (PDS) Design: Using Bypass/Decoupling Capacitors. [Online]. Available: http://direct.xilinx.com/bvdocs/appnotes/xapp623.pdf

[46] L. D. Smith *et al.*, "Power Distribution System Design Methodology and Capacitor Selection for Modern CMOS Technology," *IEEE Transactions on Advanced Packaging*, Vol. 22, No. 3, pp. 284–291, August 1999.

[47] H. B. Bakoglu, *Circuits, Interconnections, and Packaging for VLSI*. Reading, MA: Addison-Wesley, 1990.

[48] W. Becker *et al.*, "Mid-Frequency Simultaneous Switching Noise in Computer Systems," *Proceedings of the IEEE Electronic Components and Technology Conference*, pp. 676–681, May 1997.

[49] W. D. Becker *et al.*, "Modeling, Simulation, and Measurement of Mid-Frequency Simultaneous Switching Noise in Computer Systems," *IEEE Transactions on Components, Packaging, and Manufacturing Technology, Part B: Advanced Packaging*, Vol. 21, No. 2, pp. 157–163, May 1998.

[50] T. Zhou, T. Strach, and W. D. Becker, "On Chip Circuit Model for Accurate Mid-Frequency Simultaneous Switching Noise Prediction," *Proceedings of the IEEE Conference on Electrical Performance of Electronic Packaging*, pp. 275–278, October 2005.

[51] M. Benoit, S. Taylor, D. Overhauser, and S. Rochel, "Power Distribution in High-Performance Design," *Proceedings of the IEEE International Symposium on Low Power Electronics and Design*, pp. 274–278, August 1998.

[52] S. Bobba, T. Thorp, K. Aingaran, and D. Liu, "IC Power Distribution Challenges," *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pp. 643–650, November 2001.

[53] S. Chun *et al.*, "Physics Based Modeling of Simultaneous Switching Noise in High Speed Systems," *Proceedings of the IEEE Electronic Components and Technology Conference*, pp. 760–768, May 2000.

[54] S. Chun *et al.*, "Modeling of Simultaneous Switching Noise in High Speed Systems," *IEEE Transactions on Advanced Packaging*, Vol. 24, No. 2, pp. 132–142, May 2001.

[55] L. Smith, "Simultaneous Switching Noise and Power Plane Bounce for CMOS Technology," *Proceedings of the IEEE Conference on Electrical Performance of Electronic Packaging*, pp. 163–166, October 1999.

[56] F. Y. Yuan, "Electromagnetic Modeling and Signal Integrity Simulations of Power/Ground Networks in High Speed Digital Packages and Printed Circuit Boards," *Proceedings of the IEEE/ACM Design Automation Conference*, pp. 421–426, June 1998.

[57] Z. Mu, "Simulation and Modeling of Power and Ground Planes in High Speed Printed Circuit Boards," *Proceedings of the IEEE International Symposium on Circuits and Systems*, pp. 459–462, May 2001.

[58] N. Na *et al.*, "Modeling and Transient Simulation of Planes in Electronic Packages," *IEEE Transactions on Advanced Packaging*, Vol. 23, No. 3, pp. 340–352, August 2000.

[59] D. J. Herrel and B. Beker, "Modeling of Power Distribution Systems in High-Performance Microprocessors," *IEEE Transactions on Advanced Packaging*, Vol. 22, No. 3, pp. 240–248, August 1999.

[60] T.-G. Yew, Y.-L. Li, C.-Y. Chung, and D. G. Figueroa, "Design and Performance Evaluation of Chip Capacitors on Microprocessor Packaging," *Proceedings of the IEEE Conference on Electrical Performance of Electronic Packaging*, pp. 175–178, October 1999.

[61] J. Kim *et al.*, "Separated Role of On-Chip and On-PCB Decoupling Capacitors for Reduction of Radiated Emission on Printed Circuit Board," *Proceedings of the IEEE International Symposium on Electromagnetic Compatibility*, pp. 531–536, August 2001.

[62] B. Garben, G. A. Katopis, and W. D. Becker, "Package and Chip Design Optimization for Mid-Frequency Power Distribution Decoupling," *Proceedings of the IEEE Conference on Electrical Performance of Electronic Packaging*, pp. 245–248, October 2002.

[63] M. Xu *et al.*, "Power-Bus Decoupling with Embedded Capacitance in Printed Circuit Board Design," *IEEE Transactions on Electromagnetic Compatibility*, Vol. 45, No. 1, pp. 22–30, February 2003.

[64] M. I. Montrose, "Analysis on Loop Area Trace Radiated Emissions from Decoupling Capacitor Placement on Printed Circuit Boards," *Proceedings of the IEEE International Symposium on Electromagnetic Compatibility*, pp. 423–428, August 1999.

[65] P. Muthana *et al.*, "Mid Frequency Decoupling Using Embedded Decoupling Capacitors," *Proceedings of the IEEE Conference on Electrical Performance of Electronic Packaging*, pp. 271–274, October 2005.

[66] O. P. Mandhana, "Design Oriented Analysis of Package Power Distribution System Considering Target Impedance for High Performance Microprocessors," *Proceedings of the IEEE Conference on Electrical Performance of Electronic Packaging*, pp. 273–276, October 2001.

[67] I. Novak *et al.*, "Distributed Matched Bypassing for Board-Level Power Distribution Networks," *IEEE Transactions on Advanced Packaging*, Vol. 25, No. 2, pp. 230–243, May 2002.

[68] L. D. Smith and D. Hockanson, "Distributed SPICE Circuit Model for Ceramic Capacitors," *Proceedings of the IEEE Electronic Components and Technology Conference*, pp. 523–528, May/June 2001.

[69] P. Larsson, "Resonance and Damping in CMOS Circuits with On-Chip Decoupling Capacitance," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, Vol. 45, No. 8, pp. 849–858, August 1998.

[70] L. D. Smith, R. E. Anderson, and T. Roy, "Chip-Package Resonance in Core Power Supply Structures for High Power Microprocessor," *Proceedings of the ASME International Electronic Packaging Conference and Exhibition*, July 2001.

[71] C. R. Paul, "Effectiveness of Multiple Decoupling Capacitors," *IEEE Transactions on Electromagnetic Compatibility*, Vol. 34, No. 2, pp. 130–133, May 1992.

[72] M. Popovich and E. G. Friedman, "Decoupling Capacitors for Power Distribution Systems with Multiple Power Supplies," *Proceedings of the IEEE EDS/CAS Activities in Western New York Conference*, p. 9, November 2004.

[73] A. Waizman and C.-Y. Chung, "Extended Adaptive Voltage Positioning (EAVP)," *Proceedings of the IEEE Conference on Electrical Performance of Electronic Packaging*, pp. 65–68, October 2000.

[74] A. Waizman and C.-Y. Chung, "Resonant Free Power Network Design Using Extended Adaptive Voltage Positioning (EAVP) Methodology," *IEEE Transactions on Advanced Packaging*, Vol. 24, No. 3, pp. 236–244, August 2001.

[75] M. Sotman, M. Popovich, A. Kolodny, and E. G. Friedman, "Leveraging Symbiotic On-Die Decoupling Capacitance," *Proceedings of the IEEE Conference on Electrical Performance of Electronic Packaging*, pp. 111–114, October 2005.

[76] H. H. Chen and J. S. Neely, "Interconnect and Circuit Modeling Techniques for Full-Chip Power Supply Noise Analysis," *IEEE Transactions on Components, Packaging, and Manufacturing Technology Part B: Advanced Packaging*, Vol. 21, No. 3, pp. 209–215, August 1998.

[77] S. Bobba and I. Hajj, "Input Vector Generation for Maximum Intrinsic Decoupling Capacitance of VLSI Circuits," *Proceedings of the IEEE International Symposium on Circuits and Systems*, pp. 195–198, May 2001.

[78] R. Panda, S. Sundareswaran, and D. Blaauw, "On the Interaction of Power Distribution Network with Substrate," *Proceedings of the IEEE International Symposium on Low Power Electronics and Design*, pp. 388–393, August 2001.

[79] R. Panda, S. Sundareswaran, and D. Blaauw, "Impact of Low-Impedance Substrate on Power Supply Integrity," *IEEE Transactions on Design & Test of Computers*, Vol. 20, No. 3, pp. 16–22, May/June 2003.

[80] R. Panda *et al.*, "Model and Analysis of Combined Package and On-Chip Power Grid Simulation," *Proceedings of the IEEE International Symposium on Low Power Electronics and Design*, pp. 179–184, August 2000.

[81] G. Steele, D. Overhauser, S. Rochel, and S. Z. Hussain, "Full-Chip Verification Methods for DSM Power Distribution Systems," *Proceedings of the IEEE/ACM Design Automation Conference*, pp. 744–749, June 1998.

[82] N. H. Pham, "On-Chip Capacitor Measurement for High Performance Microprocessor," *Proceedings of the IEEE Conference on Electrical Performance of Electronic Packaging*, pp. 65–68, October 1998.

[83] H. Seidl *et al.*, "A Fully Integrated Al_2O_3 Trench Capacitor DRAM for Sub-100 nm Technology," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 839–842, December 2002.

[84] K. V. Rao *et al.*, "Trench Capacitor Design Issues in VLSI DRAM Cells," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 140–143, December 1986.

[85] P. E. Gronowski *et al.*, "High-Performance Microprocessor Design," *IEEE Journal of Solid-State Circuits*, Vol. 33, No. 5, pp. 676–686, May 1998.

[86] W. J. Bowhill *et al.*, "Circuit Implementation of a 300 MHz 64-bit Second Generation CMOS Alpha CPU," *Digital Technical Journal*, Vol. 7, No. 1, pp. 100–118, 1995.

[87] P. Larsson, "Parasitic Resistance in an MOS Transistor Used as On-Chip Decoupling Capacitance," *IEEE Journal of Solid-State Circuits*, Vol. 32, No. 4, pp. 574–576, April 1997.

[88] A. Hastings, *The Art of Analog Layout*. Upper Saddle River, NJ: Prentice Hall, 2001.

[89] J. L. McCreary, "Matching Properties, and Voltage and Temperature Dependence of MOS Capacitors," *IEEE Journal of Solid-State Circuits*, Vol. 16, No. 6, pp. 608–616, December 1981.

[90] R. T. Howe and C. G. Sodini, *Microelectronics: An Integrated Approach*. Upper Saddle River, NJ: Prentice Hall, 1996.

[91] C. T. Black, *et al.*, "High-Capacity, Self-Assembled Metal-Oxide-Semiconductor Decoupling Capacitors," *IEEE Electron Device Letters*, Vol. 25, No. 9, pp. 622–624, September 2004.

[92] A. R. Alvarez, *BiCMOS Technology and Applications*. Norwell, MA: Kluwer Academic Publishers, 1993.

[93] A. Behr, M. Schneider, S. Filho, and C. Montoro, "Harmonic Distortion Caused by Capacitors Implemented with MOSFET Gates," *IEEE Journal of Solid-State Circuits*, Vol. 27, No. 10, pp. 1470–1475, October 1992.

[94] S. Rusu *et al.*, "A 1.5-GHz 130-nm Itanium 2 Processor with 6-MB On-Die L3 Cache," *IEEE Journal of Solid-State Circuits*, Vol. 38, No. 11, pp. 1887–1895, November 2003.

[95] Optimization of Metal-Metal Comb-Capacitors for RF Applications. [Online]. Available: http://www.oea.com/document/Optimiz_Metal.pdf

[96] M. J. Deen and T. A. Fjeldly, *CMOS RF Modeling, Characterization and Applications*. River Edge, NJ: World Scientific, 2004.

[97] B. Razavi, *RF Microelectronics*. Upper Saddle River, NJ: Prentice Hall, 1998.

[98] R. K. Ulrich and L. W. Schaper, *Integrated Passive Component Technology*. New York, NY: Wiley/IEEE Press, 2003.

[99] S. B. Chen *et al.*, "High-Density MIM Capacitors Using Al_2O_3 and $AlTiO_x$ Dielectrics," *IEEE Electron Device Letters*, Vol. 23, No. 4, pp. 185–187, April 2002.

[100] M. Y. Yang *et al.*, "High-Density MIM Capacitors Using $AlTaO_x$ Dielectrics," *IEEE Electron Device Letters*, Vol. 24, No. 5, pp. 306–308, May 2003.

[101] X. Yu *et al.*, "A High-Density MIM Capacitor (13 fF/μm^2) Using ALD HfO_2 Dielectrics," *IEEE Electron Device Letters*, Vol. 24, No. 2, pp. 63–65, February 2003.

[102] H. Hu *et al.*, "High Performance ALD HfO_2–Al_2O_3 Laminate MIM Capacitors for RF and Mixed Signal IC Applications," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 15.6.1–15.6.4, December 2003.

[103] S.-J. Ding *et al.*, "High-Density MIM Capacitor Using ALD High-k HfO_2 Laminate Dielectrics," *IEEE Electron Device Letters*, Vol. 24, No. 12, pp. 730–732, December 2003.

[104] S.-J. Kim *et al.*, "Metal-Insulator-Metal RF Bypass Capacitor Using Niobium Oxide (Nb_2O_5) With HfO_2–Al_2O_3 Barriers," *IEEE Electron Device Letters*, Vol. 26, No. 9, pp. 625–627, September 2005.

[105] Y. H. Wu *et al.*, "The Fabrication of Very High Resistivity Si with Low Loss and Cross Talk," *IEEE Electron Device Letters*, Vol. 21, No. 9, pp. 442–444, September 2000.

[106] A. Kar-Roy *et al.*, "High Density Metal Insulator Metal Capacitors Using PECVD Nitride for Mixed Signal and RF Circuits," *Proceedings of the IEEE International Conference on Interconnect Technology*, pp. 245–247, May 1999.

[107] J. A. Babcock *et al.*, "Analog Characteristics of Metal-Insulator-Metal Capacitors Using PECVD Nitride Dielectrics," *IEEE Electron Device Letters*, Vol. 22, No. 5, pp. 230–232, May 2001.

[108] P. Zurcher *et al.*, "Integration of Thin Film MIM Capacitors and Resistors into Copper Metallization Based RF-CMOS and Bi-CMOS Technologies," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 153–156, December 2000.

[109] M. Armacost *et al.*, "A High Reliability Metal Insulator Metal Capacitor for 0.18 μm Copper Technology," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 157–160, December 2000.

[110] C. H. Ng *et al.*, "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," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 241–244, December 2002.

[111] N. Inoue *et al.*, "High Performance High-k MIM Capacitor with Plug-in Plate (PiP) for Power Delivery Line of High-Speed MPUs," *Proceedings of the IEEE International Interconnect Technology Conference*, pp. 63–65, June 2006.

[112] T. Soorapanth, "CMOS RF Filtering at GHz Frequency," Ph.D. Dissertation, Stanford University, Stanford, California, 2002.

[113] Applications of Metal-Insulator-Metal (MIM) Capacitors, *International SEMATECH*, Technology Transfer # 00083985A-ENG, October 2000.

[114] O. E. Akcasu, "High Capacitance Structures in a Semiconductor Device," U.S. Patent # 5,208,725, May 4, 1993.

[115] B. B. Mandelbrot, *The Fractal Geometry of Nature*. New York, NY: Freeman, 1983.

[116] H. Samavati *et al.*, "Fractal Capacitors," *IEEE Journal of Solid-State Circuits*, Vol. 33, No. 12, pp. 2035–2041, December 1998.

[117] A. Mezhiba and E. G. Friedman, "Inductive Properties of High Performance Power Distribution Grids," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, Vol. 10, No. 6, pp. 762–776, December 2002.

[118] A. C. C. Ng and M. Saran, "Capacitor Structure for an Integrated Circuit," U.S. Patent # 5,583,359, December 10, 1996.

[119] R. Aparicio and A. Hajimiri, "Capacity Limits and Matching Properties of Integrated Capacitors," *IEEE Journal of Solid-State Circuits*, Vol. 37, No. 3, pp. 384–393, March 2002.

[120] M. Popovich and E. G. Friedman, "Decoupling Capacitors for Power Distribution Systems with Multiple Power Supply Voltages," *Proceedings of the IEEE International SOC Conference*, pp. 331–334, September 2004.

[121] M. Popovich, *et al.*, "Maximum Effective Distance of On-Chip Decoupling Capacitors in Power Distribution Grids," *Proceedings of the ACM/IEEE Great Lakes Symposium on VLSI*, pp. 173–179, March 2006.

[122] L. Smith, "Decoupling Capacitor Calculations for CMOS Circuits," *Proceedings of the IEEE Conference on Electrical Performance of Electronic Packaging*, pp. 101–105, November 1994.

[123] M. D. Pant, P. Pant, and D. S. Wills, "On-Chip Decoupling Capacitor Optimization Using Architectural Level Current Signature Prediction," *Proceedings of the IEEE International ASIC/SOC Conference*, pp. 288–292, September 2000.

[124] M. D. Pant, P. Pant, and D. S. Wills, "On-Chip Decoupling Capacitor Optimization Using Architectural Level Prediction," *Proceedings of the IEEE Midwest Symposium on Circuits and Systems*, pp. 772–775, August 2000.

[125] M. D. Pant, P. Pant, and D. S. Wills, "On-Chip Decoupling Capacitor Optimization Using Architectural Level Predictions," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, Vol. 10, No. 3, pp. 319–326, June 2002.

[126] S. Zhao, K. Roy, and C.-K. Koh, "Decoupling Capacitance Allocation for Power Supply Noise Suppression," *Proceedings of the ACM International Symposium on Physical Design*, pp. 66–71, April 2001.

[127] S. Zhao, K. Roy, and C.-K. Koh, "Power Supply Noise Aware Floorplanning and Decoupling Capacitance Placement," *Proceedings of the IEEE International Conference on VLSI Design*, pp. 489–495, January 2002.

[128] S. Zhao, K. Roy, and C.-K. Koh, "Decoupling Capacitance Allocation and Its Application to Power-Supply Noise-Aware Floorplanning," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 21, No. 1, pp. 81–92, January 2002.

[129] A. R. Conn, R. A. Haring, and C. Viswesvariah, "Noise Considerations in Circuit Optimization," *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pp. 220–227, November 1998.

[130] C. Viswesvariah, R. A. Haring, and A. R. Conn, "Noise Considerations in Circuits Optimization," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 19, No. 6, pp. 679–690, June 2000.

[131] H. Su, S. S. Sapatnekar, and S. R. Nassif, "An Algorithm for Optimal Decoupling Capacitor Sizing and Placement for Standard Cell Layouts," *Proceedings of the ACM International Symposium on Physical Design*, pp. 68–73, April 2002.

[132] S. Hu, S. S. Sapatnekar, and S. R. Nassif, "Optimal Decoupling Capacitor Sizing and Placement for Standard Cell Layout Designs," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 22, No. 4, pp. 428–436, April 2003.

[133] J. Rabaey and M. Pedram, *Low Power Design Methodologies*. Norwell, MA: Kluwer Academic Publishers, 1996.

[134] M. Pedram and J. Rabaey, *Power-Aware Design Methodologies*. Norwell, MA: Kluwer Academic Publishers, 2002.

[135] A. Chandrakasan, M. Potkonjak, J. Rabaey, and R. W. Brodersen, "HYPER-LP: A System for Power Minimization Using Architectural Transformations," *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pp. 300–303, November 1992.

[136] A. P. Chandrakasan, *et al.*, "Optimizing Power Using Transformations," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 14, No. 1, pp. 12–31, January 1995.

[137] A. P. Chandrakasan, S. Sheng, and R. W. Brodersen, "Low-Power CMOS Digital Design," *IEEE Journal of Solid-State Circuits*, Vol. 27, No. 4, pp. 473–484, April 1992.

[138] S. Mutoh, *et al.*, "1-V Power Supply High-Speed Digital Circuit Technology with Multithreshold-Voltage CMOS," *IEEE Journal of Solid-State Circuits*, Vol. 30, No. 8, pp. 847–854, August 1995.

[139] V. Kursun and E. G. Friedman, *Multi-Voltage CMOS Circuit Design*. New York, NY: John Wiley & Sons, Inc., 2006.

[140] T. Kuroda, *et al.*, "A High-Speed Low-Power 0.3 μm CMOS Gate Array with Variable Threshold Voltage (VT) Scheme," *Proceedings of the IEEE Custom Integrated Circuits Conference*, pp. 53–56, May 1996.

[141] V. Kursun and E. G. Friedman, "Domino Logic with Variable Threshold Keeper," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, Vol. 11, No. 6, pp. 1080–1093, December 2003.

[142] V. Kursun and E. G. Friedman, "Sleep Switch Dual Threshold Voltage Domino Logic with Reduced Standby Leakage Current," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, Vol. 12, No. 5, pp. 485–496, May 2004.

[143] K. Usami, T. Ishikawa, M. Kanazawa, and H. Kotani, "Low-Power Design Technique for ASIC's by Partially Reducing Supply Voltage," *Proceedings of the IEEE International ASIC Conference*, pp. 301–304, September 1996.

[144] D. Marculescu, "Power Efficient Processors Using Multiple Supply Voltages," *Proceedings of Workshop on Compilers and Operating Systems for Low Power*, October 2000.

[145] J.-M. Chang and M. Pedram, "Energy Minimization Using Multiple Supply Voltages," *Proceedings of the IEEE International Symposium on Low Power Electronics and Design*, pp. 157–162, August 1996.

[146] J.-M. Chang and M. Pedram, "Energy Minimization Using Multiple Supply Voltages," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, Vol. 5, No. 4, pp. 436–443, December 1997.

[147] R. I. Bahar, *et al.*, "An Application of ADD-Based Timing Analysis to Combinational Low Power Re-synthesis," *Proceedings of the ACM/IEEE International Workshop on Low Power Design*, pp. 39–44, April 1994.

[148] V. Kursun, R. M. Secareanu, and E. G. Friedman, "CMOS Voltage Interface Circuit for Low Power Systems," *Proceedings of the IEEE International Symposium on Circuits and Systems*, pp. 3667–3670, May 2002.

[149] K. Usami, *et al.*, "Automated Low-Power Technique Exploiting Multiple Supply Voltages Applied to a Media Processor," *IEEE Journal of Solid-State Circuits*, Vol. 33, No. 3, pp. 463–472, March 1998.

[150] K. Usami and M. Horowitz, "Clustered Voltage Scaling Technique for Low-Power Design," *Proceedings of the IEEE International Symposium on Low Power Electronics and Design*, pp. 3–8, April 1995.

[151] K. Usami, *et al.*, "Automated Low-Power Technique Exploiting Multiple Supply Voltages Applied to a Media Processor," *Proceedings of the IEEE Custom Integrated Circuit Conference*, pp. 131–134, May 1997.

[152] V. Kursun, S. G. Narendra, V. K. De, and E. G. Friedman, "Low-Voltage-Swing Monolithic DC – DC Conversion," *IEEE Transactions on Circuits and Systems II: Express Briefs*, Vol. 51, No. 5, pp. 241–248, May 2004.

[153] V. Kursun, V. K. De, E. G. Friedman, and S. G. Narendra, "Monolithic Voltage Conversion in Low-Voltage CMOS Technologies," *Microelectronics Journal*, Vol. 36, No. 9, pp. 863–867, September 2005.

[154] R. K. Krishnamurthy, A. Alvandpour, V. De, and S. Borkar, "High-Performance and Low-Power Challenges for Sub-70 nm Microprocessor Circuits," *Proceedings of the IEEE Custom Integrated Circuit Conference*, pp. 125–128, May 2002.

[155] S. H. Kulkarni and D. Sylvester, "High Performance Level Conversion for Dual V_{dd} Design," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, Vol. 12, No. 9, pp. 926–936, September 2004.

[156] V. Kursun, S. G. Narendra, V. K. De, and E. G. Friedman, "Analysis of Buck Converters for On-Chip Integration with a Dual Supply Voltage Microprocessor," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, Vol. 11, No. 3, pp. 514–522, June 2003.

[157] P. Hazucha, *et al.*, "A 233-MHz 80%–87% Efficient Four-Phase DC – DC Converter Utilizing Air-Core Inductors on Package," *IEEE Journal of Solid-State Circuits*, Vol. 40, No. 4, pp. 838–845, April 2005.

[158] M. Igarashi, *et al.*, "A Low-Power Design Method Using Multiple Supply Voltages," *Proceedings of the IEEE International Symposium on Low Power Electronics and Design*, pp. 36–41, August 1997.

[159] J.-S. Wang, S.-J. Shieh, J.-C. Wang, and C.-W. Yeh, "Design of Standard Cells Used in Low-Power ASIC's Exploiting the Multiple-Supply-Voltage Scheme," *Proceedings of the IEEE International ASIC Conference*, pp. 119–123, September 1998.

[160] M. Hamada, Y. Ootaguro, and T. Kuroda, "Utilizing Surplus Timing for Power Reduction," *Proceedings of the IEEE Conference on Custom Integrated Circuits*, pp. 89–92, May 2001.

[161] T. Sakurai and A. R. Newton, "Alpha-Power Law MOSFET Model and Its Application to CMOS Inverter Delay and Other Formulas," *IEEE Journal of Solid-State Circuits*, Vol. 25, No. 2, pp. 584–594, April 1990.

[162] W. Hung, *et al.*, "Total Power Optimization through Simultaneously Multiple-V_{dd} Multiple-V_{TH} Assignment and Device Sizing with Stack Forcing," *Proceedings of the IEEE International Symposium on Low Power Electronics and Design*, pp. 144–149, August 2004.

[163] S. K. Mathew, *et al.*, "A 4-GHz 300-mW 64-bit Integer Execution ALU with Dual Supply Voltages in 90-nm CMOS," *IEEE Journal of Solid-State Circuits*, Vol. 40, No. 1, pp. 44–51, January 2005.

[164] D. Nguyen, *et al.*, "Minimization of Dynamic and Static Power Through Joint Assignment of Threshold Voltages and Sizing Optimization," *Proceedings of the IEEE International Symposium on Low Power Electronics and Design*, pp. 158–163, August 2003.

[165] M. Takahashi, *et al.*, "A 60-mW MPEG4 Video Codec Using Clustered Voltage Scaling with Variable Supply-Voltage Scheme," *IEEE Journal of Solid-State Circuits*, Vol. 33, No. 11, pp. 1772–1780, November 1998.

[166] K. Zhang, *et al.*, "A 3-GHz 70-Mb SRAM in 65-nm CMOS Technology With Integrated Column-Based Dynamic Power Supply," *IEEE Journal of Solid-State Circuits*, Vol. 41, No. 1, pp. 146–151, January 2006.

[167] S. Raje and M. Sarrafzadeh, "Variable Voltage Scheduling," *Proceedings of the ACM International Symposium on Low Power Design*, pp. 9–14, April 1995.

[168] A. V. Mezhiba and E. G. Friedman, "Impedance Characteristics of Power Distribution Grids in Nanoscale Integrated Circuits," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, Vol. 12, No. 11, pp. 1148–1155, November 2004.

[169] W. S. Song and L. A. Glasser, "Power Distribution Techniques for VLSI Circuits," *IEEE Journal of Solid-State Circuits*, Vol. 21, No. 1, pp. 150–156, February 1986.

[170] D. A. Priore, "Inductance on Silicon for Sub-Micron CMOS VLSI," *Proceedings of the IEEE Symposium on VLSI Circuits*, pp. 17–18, May 1993.

[171] L.-R. Zheng and H. Tenhunen, "Effective Power and Ground Distribution Scheme for Deep Submicron High Speed VLSI Circuits," *Proceedings of the IEEE International Symposium on Circuits and Systems*, pp. 537–540, May 1999.

[172] M. Kamon, M. J. Tsuk, and J. White, "FastHenry: A Multipole-Accelerated 3-D Inductance Extraction Program," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 24, No. 9, pp. 1750–1758, September 1994.

[173] E. B. Rosa, "The Self and Mutual Inductance of Linear Conductors," *Bulletin of the National Bureau of Standards*, Vol. 4, No. 2, pp. 301–344, Government Printing Office, Washington, D.C., January 1908.

[174] A. V. Mezhiba and E. G. Friedman, "Properties of On-Chip Inductive Current Loops," *Proceedings of the ACM Great Lakes Symposium on VLSI*, pp. 12–17, April 2002.

[175] M. Popovich, E. G. Friedman, M. Sotman, and A. Kolodny, "On-Chip Power Distribution Grids with Multiple Supply Voltages for High Performance Integrated Circuits," *Proceedings of the ACM/IEEE Great Lakes Symposium on VLSI*, pp. 2–7, April 2005.

[176] G. Katopis, "ΔI Noise Specification for a High Performance Computer Machine," *Proceedings of the IEEE*, Vol. 73, No. 9, pp. 1405–1415, September 1985.

[177] R. Senthinathan and J. L. Prince, "Simultaneous Switching Ground Noise Calculation for Packaged CMOS Devices," *IEEE Journal of Solid-State Circuits*, Vol. 26, No. 11, pp. 1724–1728, November 1991.

[178] A. Vaidyanath, B. Thoroddsen, and J. L. Prince, "Effect of CMOS Driver Loading Conditions on Simultaneous Switching Noise," *IEEE Transactions on Components, Packaging, and Manufacturing Technology, Part B: Advanced Packaging*, Vol. 17, No. 4, pp. 480–485, November 1994.

[179] S. R. Vemuru, "Accurate Simultaneous Switching Noise Estimation Including Velocity-Saturation Effects," *IEEE Transactions on Components, Packaging, and Manufacturing Technology, Part B: Advanced Packaging*, Vol. 19, No. 2, pp. 344–349, May 1996.

[180] S.-J. Jou, W.-C. Cheng, and Y.-T. Lin, "Simultaneous Switching Noise Analysis and Low-Bounce Buffer Design," *IEE Proceedings on Circuits, Devices, and Systems*, Vol. 148, No. 6, pp. 303–311, December 2001.

[181] H.-R. Cha and O.-K. Kwon, "A New Analytic Model of Simultaneous Switching Noise in CMOS Systems," *Proceedings of the IEEE Electronic Components and Technology Conference*, pp. 615–621, May 1998.

[182] S. R. Vemuru, "Effects of Simultaneous Switching Noise on the Tapered Buffer Design," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, Vol. 5, No. 3, pp. 290–300, September 1997.

[183] P. Heydari and M. Pedram, "Ground Bounce in Digital VLSI Circuits," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, Vol. 11, No. 2, pp. 180–193, April 2003.

[184] T. J. Gabara, "Ground Bounce and Reduction Techniques," *Proceedings of the IEEE ASIC Conference*, pp. 13.-2.1–13.-2.2, September 1991.

[185] A. Vittal, H. Ha, F. Brewer, and M. Marek-Sadowska, "Clock Skew Optimization for Ground Bounce Control," *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pp. 395–399, November 1996.

[186] M. D. Pant, P. Pant, D. S. Wills, and V. Tiwari, "An Architectural Solution for the Inductive Noise Problem due to Clock-Gating," *Proceedings of the IEEE International Symposium on Low Power Electronics and Design*, pp. 255–257, August 1999.

[187] J. Oh and M. Pedram, "Multi-Pad Power/Ground Network Design for Uniform Distribution of Ground Bounce," *Proceedings of the ACM/IEEE Design Automation Conference*, pp. 287–290, June 1998.

[188] H. Chen, "Minimizing Chip-Level Simultaneous Switching Noise for High-Performance Microprocessor Design," *Proceedings of the IEEE International Symposium on Circuits and Systems*, pp. 544–547, May 1996.

[189] A. Zenteno, V. H. Champac, M. Renovell, and F. Azais, "Analysis and Attenuation Proposal in Ground Bounce," *Proceedings of the IEEE Asian Test Symposium*, pp. 460–463, November 2004.

[190] M. Badaroglu, *et al.*, "Digital Ground Bounce Reduction by Supply Current Shaping and Clock Frequency Modulation," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 24, No. 1, pp. 65–76, January 2005.

[191] J. Mermet and W. Nebel, *Low Power Design in Deep Submicron Electronics*. Norwell, MA: Kluwer Academic Publishers, 1997.

[192] A. Chandrakasan, W. J. Bowhill, and F. Fox, *Design of High-Performance Microprocessor Circuits*. New York, NY: Wiley-IEEE Press, 2000.

[193] S. R. Nassif and O. Fakhouri, "Technology Trends in Power-Grid-Induced Noise," *Proceedings of the ACM International Workshop on System Level Interconnect Prediction*, pp. 55–59, April 2002.

[194] L. D. Smith, "Packaging and Power Distribution Design Considerations for a Sun Microsystems Desktop Workstation," *Proceedings of the IEEE Conference on Electrical Performance of Electronic Packaging*, pp. 19–22, October 1997.

[195] M. Popovich and E. G. Friedman, "Noise Coupling in Multi-Voltage Power Distribution Systems with Decoupling Capacitors," *Proceedings of the IEEE International Symposium on Circuits and Systems*, pp. 620–623, May 2005.

[196] M. Popovich and E. G. Friedman, "Noise Aware Decoupling Capacitors for Multi-Voltage Power Distribution Systems," *Proceedings of the ACM/IEEE International Symposium on Quality Electronic Design*, pp. 334–339, March 2005.

[197] C. R. Paul, *Analysis of Linear Circuits*. New York, NY: McGraw-Hill, 1989.

[198] M. K. Gowan, L. L. Biro, and D. B. Jackson, "Power Considerations in the Design of the Alpha 21264 Microprocessor," *Proceedings of the IEEE/ACM Design Automation Conference*, pp. 726–731, June 1998.

[199] H. H. Chen and S. E. Schuster, "On-Chip Decoupling Capacitor Optimization for High-Performance VLSI Design," *Proceedings of the IEEE International Symposium on VLSI Technology, Systems, and Applications*, pp. 99–103, May 1995.

[200] F. Moll and M. Roca, *Interconnect Noise in VLSI Circuits*. Norwell, Massachusetts: Kluwer Academic Publishers, 2003.

[201] A. Dharchoudhury *et al.*, "Design and Analysis of Power Distribution Networks in PowerPC Microprocessors," *Proceedings of the IEEE/ACM Design Automation Conference*, pp. 738–743, June 1998.

[202] M. E. Van Valkenburg, *Network Analysis*. Upper Saddle River, NJ: Prentice Hall, 1974.

[203] M. Takamiya and 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 Journal of Solid-State Circuits*, Vol. 40, No. 3, pp. 719–725, March 2005.

[204] D. Lee, D. Blaauw, and D. Sylvester, "Gate Oxide Leakage Current Analysis and Reduction for VLSI Circuits," *IEEE Transactions on Very Large Scale Integration (VLSI) Circuits*, Vol. 12, No. 2, pp. 155–166, February 2004.

[205] M. Anis and Y. Massoud, "Power Design Challenges in Deep-Submicron Technology," *Proceedings of the IEEE International Midwest Symposium on Circuits and Systems*, pp. 1510–1513, December 2003.

[206] D. Deleganes, J. Douglas, B. Kommandur, and M. Patyra, "Designing a 3 GHz, 130 nm, Intel Pentium 4 Processor," *Proceedings of the IEEE Symposium on VLSI Circuits*, pp. 130–133, June 2002.

[207] R. McGowen, *et al.*, "Power and Temperature Control on a 90-nm Itanium Family Processor," *IEEE Journal of Solid-State Circuits*, Vol. 41, No. 1, pp. 229–237, January 2006.

[208] S. Naffziger, *et al.*, "The Implementation of a 2-Core, Multi-Threaded Itanium Family Processor," *IEEE Journal of Solid-State Circuits*, Vol. 41, No. 1, pp. 197–209, January 2006.

[209] T. Hubing, "Effective Strategies for Choosing and Locating Printed Circuit Board Decoupling Capacitors," *Proceedings of the IEEE International Symposium on Electromagnetic Compatibility*, pp. 632–637, August 2005.

[210] Mathematica 5.2, Wolfram Research, Inc.

[211] M. P. Goetz, "Time and Frequency Domain Analysis of Integral Decoupling Capacitors," *IEEE Transactions on Components, Packaging, and Manufacturing Technology, Part B: Advanced Packaging*, Vol. 19, No. 3, pp. 518–522, August 1996.

[212] T. Murayama, K. Ogawa, and H. Yamaguchi, "Estimation of Peak Current Trough CMOS VLSI Circuit Supply Lines," *Proceedings of the ACM Asia and South Pacific Design Automation Conference*, pp. 295–298, January 1999.

[213] *International Technology Roadmap for Semiconductors, 2006 Update*, Semiconductor Industry Association, 2006. [Online]. Available: http://public.itrs.net

Appendix A

Mutual Loop Inductance in Fully Interdigitated Power Distribution Grids with DSDG

Assuming $d_I^i = s_I^i = d$, from (4.3), the mutual inductances between the power and ground paths of the different voltage domains for a fully interdigitated power distribution grid with DSDG are

$$L_{Vdd1-Vdd2} = 0.2l\left(\ln\frac{2l}{2d} - 1 + \frac{2d}{l} - \ln\gamma + \ln k\right), \tag{A.1}$$

$$L_{Vdd1-Gnd2} = 0.2l\left(\ln\frac{2l}{3d} - 1 + \frac{3d}{l} - \ln\gamma + \ln k\right), \tag{A.2}$$

$$L_{Gnd1-Gnd2} = 0.2l\left(\ln\frac{2l}{2d} - 1 + \frac{2d}{l} - \ln\gamma + \ln k\right), \tag{A.3}$$

$$L_{Vdd2-Gnd1} = 0.2l\left(\ln\frac{2l}{d} - 1 + \frac{d}{l} - \ln\gamma + \ln k\right). \tag{A.4}$$

Substituting (A.1) – (A.4) into (4.8), the mutual inductive coupling M_{loop}^{intI} between the two current loops in a fully interdigitated power distribution grid with DSDG is

$$M_{loop}^{intI} = 0.2l\left(\ln\frac{2l}{2d} - 1 + \frac{2d}{l} - \ln\gamma + \ln k - \ln\frac{2l}{3d} + 1 - \frac{3d}{l} + \ln\gamma - \ln k + \ln\frac{2l}{2d} - 1 + \frac{2d}{l} - \ln\gamma + \ln k - \ln\frac{2l}{d} + 1 - \frac{d}{l} + \ln\gamma - \ln k\right). \tag{A.5}$$

Simplifying (A.5) and considering that $\ln\gamma$ and $\ln k$ are approximately the same for different distances between the lines, M_{loop}^{intI} is

$$\begin{aligned} M_{loop}^{intI} &= 0.2l\left(\ln\frac{2l}{2d} - \ln\frac{2l}{3d} + \ln\frac{2l}{2d} - \ln\frac{2l}{d}\right) \\ &= 0.2l\ln\frac{2l\times 3d\times 2l\times d}{2d\times 2l\times 2d\times 2l} \\ &= 0.2l\ln\frac{3}{4} < 0. \end{aligned} \tag{A.6}$$

Appendix B

Mutual Loop Inductance in Pseudo-Interdigitated Power Distribution Grids with DSDG

Assuming $d^i_{II} = 2d$ and $s^i_{II} = d$, from (4.3), the mutual inductances between the power and ground paths of the different voltage domains for a pseudo-interdigitated power distribution grid with DSDG are

$$L_{Vdd1-Vdd2} = 0.2l\left(\ln\frac{2l}{d} - 1 + \frac{d}{l} - \ln\gamma + \ln k\right), \tag{B.1}$$

$$L_{Vdd1-Gnd2} = 0.2l\left(\ln\frac{2l}{3d} - 1 + \frac{3d}{l} - \ln\gamma + \ln k\right), \tag{B.2}$$

$$L_{Gnd1-Gnd2} = 0.2l\left(\ln\frac{2l}{d} - 1 + \frac{d}{l} - \ln\gamma + \ln k\right), \tag{B.3}$$

$$L_{Vdd2-Gnd1} = 0.2l\left(\ln\frac{2l}{d} - 1 + \frac{d}{l} - \ln\gamma + \ln k\right). \tag{B.4}$$

Substituting (B.1) – (B.4) into (4.8), the mutual inductive coupling M_{loop}^{intII} between the two current loops in a pseudo-interdigitated power distribution grid with DSDG is

$$M_{loop}^{intII} = 0.2l\left(\ln\frac{2l}{d} - 1 + \frac{d}{l} - \ln\gamma + \ln k - \ln\frac{2l}{3d} + 1 - \frac{3d}{l} + \ln\gamma - \ln k + \ln\frac{2l}{d} - 1 + \frac{d}{l} - \ln\gamma + \ln k - \ln\frac{2l}{d} + 1 - \frac{d}{l} + \ln\gamma - \ln k\right). \tag{B.5}$$

Simplifying (B.5) and considering that $\ln\gamma$ and $\ln k$ are approximately the same for different distances between the lines, M_{loop}^{intII} is

$$\begin{aligned} M_{loop}^{intII} &= 0.2l\left(\ln\frac{2l}{d} - \ln\frac{2l}{3d} + \ln\frac{2l}{d} - \ln\frac{2l}{d} - \frac{2d}{l}\right) \\ &= 0.2l\left(\ln\frac{2l\times 3d\times 2l\times d}{d\times 2l\times d\times 2l} - \frac{2d}{l}\right) \\ &= 0.2l\left(\ln 3 - \frac{2d}{l}\right) > 0. \end{aligned} \tag{B.6}$$

Appendix C

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. 4.8), from (4.3), the mutual inductances between the power and ground paths of the different voltage domains for a fully paired power distribution grid with DSDG are

$$L_{Vdd1-Vdd2} = 0.2l\left[\ln\frac{2l}{(n+1)d} - 1 + \frac{(n+1)d}{l} - \ln\gamma + \ln k\right], \tag{C.1}$$

$$L_{Vdd1-Gnd2} = 0.2l\left[\ln\frac{2l}{(n+2)d} - 1 + \frac{(n+2)d}{l} - \ln\gamma + \ln k\right], \tag{C.2}$$

$$L_{Gnd1-Gnd2} = 0.2l\left[\ln\frac{2l}{(n+1)d} - 1 + \frac{(n+1)d}{l} - \ln\gamma + \ln k\right], \tag{C.3}$$

$$L_{Vdd2-Gnd1} = 0.2l\left(\ln\frac{2l}{nd} - 1 + \frac{nd}{l} - \ln\gamma + \ln k\right). \tag{C.4}$$

Substituting (C.1) – (C.4) into (4.8), the mutual inductive coupling M_{loop}^{prdI} between the two current loops in a fully paired power distribution grid with DSDG is

$$\begin{aligned} M_{loop}^{prdI} = 0.2l \Bigg[& \ln\frac{2l}{(n+1)d} - 1 + \frac{(n+1)d}{l} - \ln\gamma + \\ & \ln k - \ln\frac{2l}{(n+2)d} + 1 - \frac{(n+2)d}{l} + \ln\gamma - \\ & \ln k + \ln\frac{2l}{(n+1)d} - 1 + \frac{(n+1)d}{l} - \ln\gamma + \\ & \ln k - \ln\frac{2l}{nd} + 1 - \frac{nd}{l} + \ln\gamma - \ln k \Bigg]. \end{aligned} \tag{C.5}$$

Simplifying (C.5) and considering that $\ln\gamma$ and $\ln k$ are approximately the same for different distances between the lines, M_{loop}^{prdI} is

$$\begin{aligned} M_{loop}^{prdI} &= 0.2l \Bigg[\ln\frac{2l}{(n+1)d} + \frac{(n+1)d}{l} - \ln\frac{2l}{(n+2)d} - \\ &\quad \frac{(n+2)d}{l} + \ln\frac{2l}{(n+1)d} + \frac{(n+1)d}{l} - \ln\frac{2l}{nd} - \frac{nd}{l} \Bigg] \\ &= 0.2l \Bigg[\ln\frac{2l \times (n+2)d \times 2l \times nd}{(n+1)d \times 2l \times (n+1)d \times 2l} + \\ &\quad \frac{(n+1)d - (n+2)d + (n+1)d - nd}{l} \Bigg] \\ &= 0.2l \ln\left[\frac{(n+2)n}{(n+1)^2}\right] < 0 \ \text{ for } n \geq 1. \end{aligned} \tag{C.6}$$

Appendix D

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. 4.9), from (4.3), the mutual inductances between the power and ground paths of the different voltage domains for a pseudo-paired power distribution grid with DSDG are

$$L_{Vdd1-Vdd2} = 0.2l\left(\ln\frac{2l}{d} - 1 + \frac{d}{l} - \ln\gamma + \ln k\right), \tag{D.1}$$

$$L_{Vdd1-Gnd2} = 0.2l\left[\ln\frac{2l}{(n+2)d} - 1 + \frac{(n+2)d}{l} - \ln\gamma + \ln k\right], \tag{D.2}$$

$$L_{Gnd1-Gnd2} = 0.2l\left(\ln\frac{2l}{d} - 1 + \frac{d}{l} - \ln\gamma + \ln k\right), \tag{D.3}$$

$$L_{Vdd2-Gnd1} = 0.2l\left(\ln\frac{2l}{nd} - 1 + \frac{nd}{l} - \ln\gamma + \ln k\right). \tag{D.4}$$

Substituting (D.1) – (D.4) into (4.8), the mutual inductive coupling M_{loop}^{prdII} between the two current loops in a pseudo-paired power distribution grid with DSDG is

$$\begin{aligned} M_{loop}^{prdII} = 0.2l \Bigg[& \ln\frac{2l}{d} - 1 + \frac{d}{l} - \ln\gamma + \ln k - \\ & \ln\frac{2l}{(n+2)d} + 1 - \frac{(n+2)d}{l} + \ln\gamma - \ln k + \\ & \ln\frac{2l}{d} - 1 + \frac{d}{l} - \ln\gamma + \ln k - \ln\frac{2l}{nd} + 1 - \frac{nd}{l} + \ln\gamma - \ln k \Bigg]. \end{aligned} \tag{D.5}$$

Simplifying (D.5) and considering that $\ln\gamma$ and $\ln k$ are approximately the same for different distances between the lines, M_{loop}^{prdII} is

$$\begin{aligned} M_{loop}^{prdII} &= 0.2l \left[\ln\frac{2l}{d} + \frac{d}{l} - \ln\frac{2l}{(n+2)d} - \frac{(n+2)d}{l} + \ln\frac{2l}{d} + \frac{d}{l} - \ln\frac{2l}{nd} - \frac{nd}{l} \right] \\ &= 0.2l \left[\ln\frac{2l \times (n+2)d \times 2l \times nd}{d \times 2l \times d \times 2l} + \frac{2d - (n+2)d - nd}{l} \right] \\ &= 0.2l \left[\ln\left(n^2 + 2n\right) - \frac{2nd}{l} \right] > 0 \ \text{ for } n \geq 1. \end{aligned} \tag{D.6}$$