Enhancing Power and Signal Integrity in Three-Dimensional Integrated Circuits

A Dissertation Presented

by

Hailang Wang

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Electrical Engineering

Stony Brook University

May 2016

Stony Brook University

The Graduate School

Hailang Wang

We, the dissertation committee for the above candidate for the
Doctor of Philosophy degree, hereby recommend
acceptance of this dissertation.

Dr. Emre Salman - Dissertation Advisor
Assistant Professor, Department of Electrical and Computer Engineering

Dr. Sangjin Hong - Chairperson of Defense
Professor, Department of Electrical and Computer Engineering

Dr. Milutin Stanacevic - Defense Committee Member
Associate Professor, Department of Electrical and Computer Engineering

Dr. Thomas MacCarthy - Defense Committee Member
Assistant Professor, Department of Applied Mathematics and Statistics

This dissertation is accepted by the Graduate School

Charles Taber
Dean of the Graduate School

Abstract of the Dissertation

Enhancing Power and Signal Integrity in Three-Dimensional Integrated Circuits

by

Hailang Wang

Doctor of Philosophy

in

Electrical Engineering

Stony Brook University

2016

Three-dimensional (3D) integration has emerged as an enabling technology for integrated circuits (ICs) in the interconnect-centric design era, where the interconnects have become a bottleneck for the overall system performance. With 3D integration technology, multiple planar dies are stacked vertically while the communication among different dies is achieved by low impedance vertical connections such as through silicon vias (TSVs) or monolithic inter-tier vias (MIVs). Due to the unique characteristics of 3D ICs, various challenges exist in the fabrication, design and testing of 3D integrated systems. The research work proposed in this dissertation is focused on 3D design methodologies with emphasis on power and signal integrity.

Specifically, to reliably deliver the power supply voltage to each circuit module within a 3D system, *i.e.*, maintaining system-wide power integrity, the power distribution network in 3D ICs should be carefully designed with specific design considerations different from traditional 2D ICs. In this dissertation, two critical issues about the power distribution network are investigated to improve the power integrity of TSV-based 3D ICs. First, novel design topologies and analytic expressions are proposed for the physical implementation of *power gating* in 3D ICs. Power gating is an existing and effective low power design method to reduce leakage power consumption. It is demonstrated that the proposed methodology can effectively improve power integrity for 3D ICs with power gating. Alternatively, the efficacy of *decoupling capacitors*, which are intentional capacitors placed to reduce power supply noise, can be degraded due to power gating in 3D ICs. A reconfigurable decoupling capacitor topology that dynamically configures the connectivity of decoupling capacitors is investigated to achieve better utilization of the decoupling capacitors and further enhance power integrity of 3D ICs with power gating. In addition, to provide a fast yet accurate estimation of power supply noise, an analytic model and closed-form expressions are proposed, exhibiting significant improvement over existing analytical models for nanoscale ICs with fast transitions.

In addition to power integrity, TSV-based 3D ICs also introduce distinctive signal integrity issues. Electrical noise can couple from the TSVs into the silicon substrate of a die, which can disturb the operation of the active devices within the die. A methodology is developed to characterize the TSV induced noise coupling in 3D ICs. Design guidelines are also proposed based on the noise characterization results to improve signal integrity within 3D ICs.

Finally, the monolithic 3D integration technology based on MIVs (rather than TSVs) is investigated. To evaluate the benefits of MIV-based 3D ICs, *Mono3D*, a standard cell library for transistor-level monolithic 3D ICs, is developed in 45 nm CMOS technology. A complete back-end physical design flow utilizing the proposed *Mono3D* library is also demonstrated. As an example, a benchmark circuit is designed from gate-level netlist to physical layout using the *Mono3D* to evaluate the performance and power consumption of MIV-based 3D ICs.

Table of Contents

List of Figures

List of Tables

ACKNOWLEDGEMENTS

The past five years of my PhD experience is absolutely the most enjoyable five years in my life so far. There are these amazing people I would like to acknowledge here, who made pursuing PhD such a memorable journey.

First of all, I would like to pay my highest regards to my PhD advisor, Prof. Emre Salman. He is not only an academic advisor, but also a role model for me. Benefiting a lot from his academic perspective and professional knowledge, I also enjoyed the research environment he provided so I can focus on most interesting problems and make good progress in the area of 3D IC designs. From the numerous talks and discussions with Prof. Salman, I am so impressed that he can always quickly grasp the core points from a seemingly complicated problem and root-cause the issue. What I learned from Prof. Salman would for sure be a good fortune for me ever after.

I greatly appreciate the help from Prof. Sangjin Hong, Prof. Milutin Stanacevic and Prof. Thomas MacCarthy, especially for serving as my committee members and providing insightful advice to my dissertation.

I would also like to acknowledge all the members of NanoCAS Laboratory: Peirong, Zhihua, Suhas, Ziqi, Sateja, Mallika, Weicheng, Chen, Tutu, Yongwan... We shared so many good memories in our nice and comforting lab – our "home away from home". Thank you for making our academic adventure not only about experiments and circuit simulations, but also about friendship and trust. I wish you all have a fruitful research experience and a bright future.

Last, but not least, I feel so grateful for the love from my parents Bibo Wang, Lin Wang, and my fiancee Xi. I would not never accomplish this without the unconditional support from you.

Chapter 1

Introduction

In the past several decades, integrated circuit (IC) design process has shifted from a logic-centric methodology to an interconnect-centric approach. This shift is described in Section 1.1. Three-dimensional (3D) integration, an enabling technology in the interconnect-centric design era, is introduced in Section 1.2. Challenges related with 3D integration technology are discussed in Section 1.3. The outline of this dissertation is presented in Section 1.4.

1.1 Interconnect-Centric Design

The evolution of integrated circuit (IC) technology consists of interlaced limitations and breakthroughs to overcome these limitations. Since the development of the first IC in 1958, the primary objective has been cramming more transistors onto a single chip to achieve increasingly more complex functionality and better performance. The advances in semiconductor fabrication technology have enabled higher transistor density by scaling down the size of transistors. As Gordon Moore noted in 1965, the number of transistors on a single IC has doubled every 18 to 24 months [1]. This prediction has worked well for more than three decades. The number of transistors has increased from 2300 in the Intel 4004 processor released in 1971, by over 600 times, to 1.4 million in the Intel 486 processor released in 1992 [2].

However, limitations have emerged as fabrication technology has entered the sub-micron era, where over one million transistors can be integrated onto a single chip. The interconnects required to connect a large number of transistors have raised a challenge to the ongoing trend of performance improvement enabled by device scaling [3] [4]. Previously, in the micron technology era where the transistor minimum feature size is above 1 μm, the speed of transistor used to be the dominant factor for system performance. However, as the features size has decreased, the gate delay caused by transistors has been significantly reduced while the delay caused by interconnect has increased. As reported by the 2005 International Technology Roadmap for Semiconductors (ITRS) [5], at 0.25 μm technology node, the interconnect delay is approximately equal to the gate delay caused by transistors, as shown in Fig. 1.1. Beyond this technology node, interconnect delay (particularly the delay caused by the long global interconnects) becomes dominant over transistor delay. Therefore, long global interconnects have become a bottleneck for the overall system performance [4].

Various methods have been proposed to address the interconnect issue. Migrating from aluminum to copper as the interconnect material is an early example of this effort. By using copper interconnects, resistivity of the metal interconnects is reduced, partially compensating the delay caused by long global interconnects [6]. Another technology-level innovation is the application of ultra-low K dielectric ma-

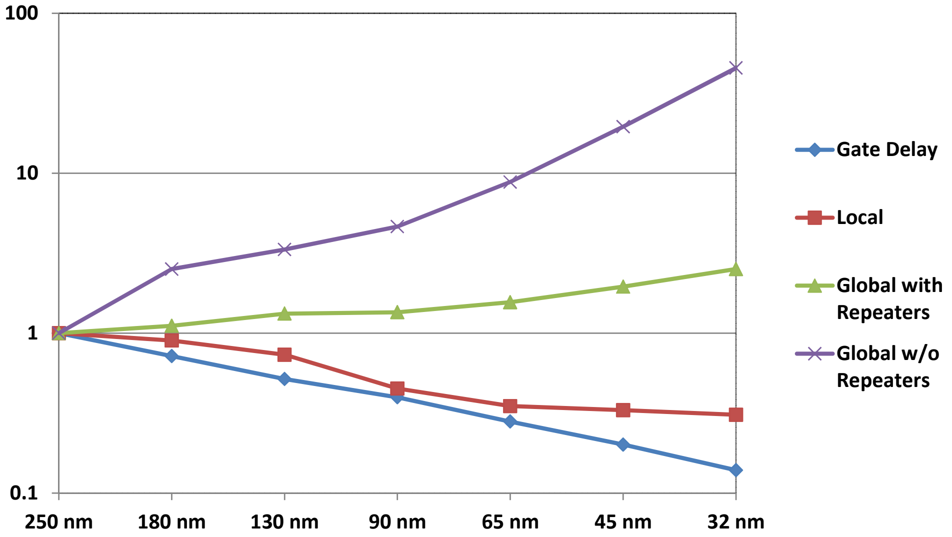


Figure 1.1: Delay comparison caused by transistor, local interconnect, and global interconnect in sub-micron technology nodes [5].

terials, which aims to address the parasitic capacitance of interconnects [7]. The high density of interconnect makes the lateral parasitic capacitance between neighboring metal lines an important issue, since the parasitic capacitance not only increases the signal delay but also causes additional power consumption [8]. The parasitic capacitance between adjacent metal lines is proportional to the dielectric constant *K* of the insulation material between the metal lines. Thus, low *K* dielectric materials such as carbon doped silicon and porous dielectric film are preferred over traditional silicon dioxide as the insulation material between interconnects [8].

The application of new materials brings a one-time reduction to the parasitic impedance of the interconnects. However, the limitations of long global interconnects subsequently emerge again as the scaling continues. Global interconnects with high resistance are required to connect different circuit modules within the die, which has become increasingly larger [9]. For example, the die size for Intel's

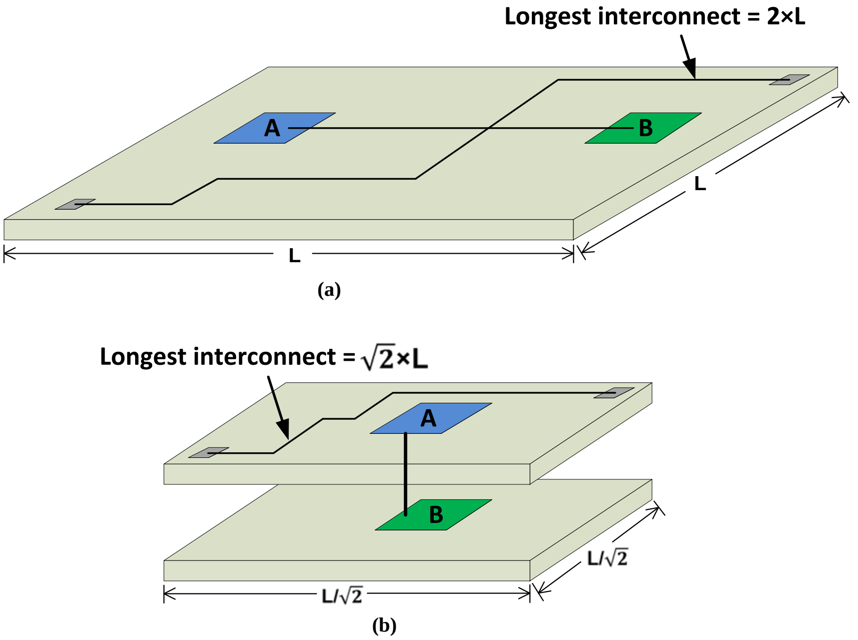


Figure 1.2: The communication between two blocks A and B is achieved by (a) long and high resistive global interconnect in 2D planar die, and (b) short and low resistive vertical connections in a 3D stack.

Haswell processors can be as high as 356 mm^2 [10]. Therefore, an intuitive solution is, instead of distributing the circuit modules in a 2D planar plane, the modules can be stacked vertically, *i.e.*, three-dimensional (3D) integration.

1.2 Introduction to 3D Integration

In 3D integration technology, multiple dies (also referred to as tiers) are bonded and vertically stacked on top of each other. The communication between multiple tiers in the stack is achieved by vertical connections such as wire bonds, soldering balls, through-silicon-vias (TSVs) or monolithic inter-tier vias (MIVs) depending on the specific 3D integration technology. These vertical connections have smaller resistance than the global planar interconnects. Therefore, as shown in Fig. 1.2, two circuit blocks A and B, which experience longer interconnect in 2D planar die, can be closer in the vertical dimension utilizing 3D integration technology. The communication between these two blocks can be achieved by short and relatively low resistance vertical connections in the 3D system, instead of the long and highly resistive interconnects in the 2D counterpart. According to simple estimations, the longest interconnect length exhibits a reduction approximately proportional to $\sqrt{N}$ when the 2D chip is converted into an N-plane 3D IC [9] [11]. The reduction in interconnect length not only reduces delay, but also power consumption due to less switching capacitance [12] [9].

In addition to achieving shorter interconnects, 3D integration enables heterogenous integration of multiple dies with different functionalities and possibly fabricated using different technologies, such as analog senors, digital logic, memory, radio frequency (RF) components or micro-electro-mechanical system (MEMS) modules [13]. Due to these advantages, 3D integration has been considered as an enabling technology to maintain performance improvement in the interconnect-centric design era of ICs.

1.3 Challenges in 3D Integrated Circuits

The advantages of 3D integration technology are highly desirable in future technology nodes to maintain higher integration densities while enhancing performance and reducing power. However, designing and fabricating 3D ICs have several important challenges, as outlined below.

In the past years, considerable research has been conducted on 3D integration at the fabrication level. A fundamental issue in 3D IC fabrication is to reliably manufacture the vertical connections among the multiple dies within a 3D stack. Various 3D integration technologies have been developed, with different structures and mechanisms for the vertical connections, such as through-silicon-vias (TSV) and monolithic inter-tier vias (MIVs). For instance, in TSV based 3D ICs, vertical TSVs travel through the substrate of each tier to achieve interconnections. The reliability of these vertical connections is a concern which affects the overall reliability of the 3D IC. Alternatively, to bond multiple dies, specialized bonding techniques and wafer thinning processes are required [14] [15] [16].

With the continuous progress in fabrication technologies for 3D integration, the primary obstacles in realizing complex 3D ICs now lie in multiple aspects of the design process [9]. For example, when designing a complex system such as a microprocessor, the architecture level considerations need to be tailored to effectively exploit the advantages of 3D technology. Different circuit modules in the microprocessor should be optimally partitioned and placed on different planes in the 3D stack to achieve optimal performance for the entire system. Another imminent issue is the insufficiency of existing electronic design automation (EDA) tools for 3D ICs. For modern VLSI circuits with several billion transistors, EDA tools are critical components in the design process. New tools need to be developed to accommodate the novel characteristics achieved by the vertical connections in 3D ICs.

During the testing phase, 3D ICs exhibit unique characteristics requiring novel testing methodologies. Traditionally, testing is performed at wafer level and die level. Therefore, only one plane of a 3D IC can be tested at a time, which means only part of the system functionality is tested. Thus, additional connection pads and pins are required to provide input pattern as well as power and clock signals to the tier under test [9] [17]. Alternatively, traditional testing methodologies of individual dies before bonding cannot guarantee the functionality of the overall 3D system. In 3D ICs, post-bonding testing methodologies which test the overall functionality of the system after bonding multiple dies are required [18] [19].

Another critical challenge in the physical design of 3D ICs is the power and signal integrity issues, which are the foci of this dissertation. To guarantee normal operation of a chip, power supply voltage should be reliably delivered to each transistor in the system without excessive power noise along the power delivery path, *i.e.*, maintaining system-wide power integrity [20]. Reliable delivery of the power supply voltage to every circuit module in a 3D IC is more challenging than a traditional 2D chip. The power distribution network on each die of a 3D stack is interconnected by the vertical connections (such as TSVs), generating a more complicated network. Higher parasitic impedance of this more complicated power distribution network exacerbates the power noise along the power delivery path. Furthermore, due to the heterogenous integration, certain low power design techniques such as power gating are often implemented, which further complicates the power distribution network due to the individual control of different power domains in the system. Therefore, novel design methodologies considering the unique characteristics of 3D ICs are developed in this dissertation to address power integrity issues.

Various forms of noise exist in an IC, which affect the operation of transistors and may cause functional failure. This issue is referred to as signal integrity. In addition to the traditional signal integrity issues encountered in 2D ICs (such as interconnect noise coupling [21] and substrate coupling [22]), 3D ICs experience new signal integrity challenges. For instance, in TSV-based 3D ICs, the TSVs inject noise into silicon substrate. This noise propagates and affects the active devices throughout the substrate. Therefore, accurate characterization of the noise coupling mechanisms and developing design guidelines for 3D ICs are essential to improve signal integrity, as targeted in this dissertation.

1.4 Outline

In this dissertation, research results about several critical issues related with power and signal integrity in 3D ICs are presented [23–28]. In Chapter 2, an overview of 3D ICs is provided. Power and signal integrity issues in 3D ICs are introduced to motivate the research described in this dissertation.

In Chapter 3, research results on implementing power gating in TSV-based 3D ICs are presented. Power gating is a well known low power design method to reduce leakage power consumption. Power gating requires additional transistors (referred to as sleep transistors) inserted along the power delivery path to realize individual control of the power delivery to each circuit module. In 3D ICs, the physical implementation of power gating is affected by the TSVs. In this chapter, several TSV-specific power gating schemes are proposed based on the characteristics of different TSV technologies to enhance power integrity.

In Chapter 4, a resource allocation methodology is proposed to simultaneously determine the size of sleep transistors and TSVs. The power supply noise is reduced with minimal area and power consumption overhead. The proposed methodology is important since both sleep transistors and TSVs occupy significant silicon area in 3D ICs.

In Chapter 5, the primary focus is on providing a fast yet accurate estimation of the power supply noise, a practical issue encountered during power integrity analysis. It is demonstrated that existing closed-form expressions have a common limitation in analyzing power supply noise with fast signal transitions. Since modern technologies exhibit signal transitions in the range of tens of pico seconds, it is highly critical to be able to consider this characteristic. Thus, an analytical model with closed-form expressions is developed to accurately estimate power supply noise. The proposed expressions exhibit significant improvement in accuracy over existing analytical models, particularly for nanoscale ICs with fast signal transitions.

In Chapter 6, design considerations about decoupling capacitors in 3D ICs are explored. Placing decoupling capacitors is a well known method to reduce power supply noise. However, the placement topology of decoupling capacitors should be carefully determined to guarantee the efficacy of decoupling capacitors. In 3D ICs, the vertical TSVs extend the effective range of decoupling capacitors from a single plane to multiple planes within a 3D stack. However, if conventional power gating is applied, the ability of decoupling capacitors to reduce noise in neighboring planes cannot be exploited. Therefore, two decoupling capacitor placement topologies are proposed to utilize the low resistance of TSVs. With the proposed topologies, the decoupling capacitor in a power gated plane is enabled to mitigate power noise of the neighboring active planes, improving the overall power integrity of a 3D IC.

In Chapter 7, TSV-induced noise coupling in 3D ICs is investigated as an important signal integrity issue. A methodology is proposed to effectively characterize the noise coupling for different TSV types and substrate schemes. The effect of different design parameters such as TSV type, placement of substrate contacts, signal slew rate, and voltage swing is investigated. Design guidelines are developed to improve system-wide signal integrity.

In Chapter 8, a relatively more recent type of 3D integration; monolithic 3D approach with inter-tier vias (MIVs), is explored. MIVs achieve significantly higher density as compared to TSVs. Thus, monolithic 3D technology is considered to be a highly promising option for vertical integration. The *Mono3D*, a standard cell library for transistor-level monolithic 3D ICs, is developed by using a 45 nm technology kit. It is demonstrated that the proposed Mono3D standard cell library can be integrated with existing standard physical design flows. An example 3D monolithic IC with two planes is designed and analyzed by utilizing the developed cell library and physical design flow. The advantages of the monolithic 3D approach in terms of physical area and power consumption are characterized.

Finally, the dissertation is concluded in Chapter 9. Possible future research directions are discussed.

Chapter 2

Power and Signal Integrity in 3D ICs

To investigate the power and signal integrity issues in 3D ICs, the unique characteristics of 3D ICs should be explored, which are the foci of this chapter. Several 3D integration technologies are introduced in Section 2.1 with emphasis on TSV-based 3D ICs. A brief overview of the power noise issue and power distribution network in 3D ICs is provided in Section 2.2. Several critical noise coupling mechanisms affecting the system signal integrity are discussed in Section 2.3.

2.1 3D Integration Technologies

Currently, there are several 3D integration technologies available including System-in-Package (SiP), TSV-based 3D integration, MIV-based 3D integration, and contactless coupled 3D ICs. A critical characteristic that distinguishes different 3D integration technologies is the vertical connectivity of different planes in a 3D stack.

In System-in-Package (SiP) technology, multiple pre-fabricated dies are encapsulated within the same package. Within the package, connections among multi-

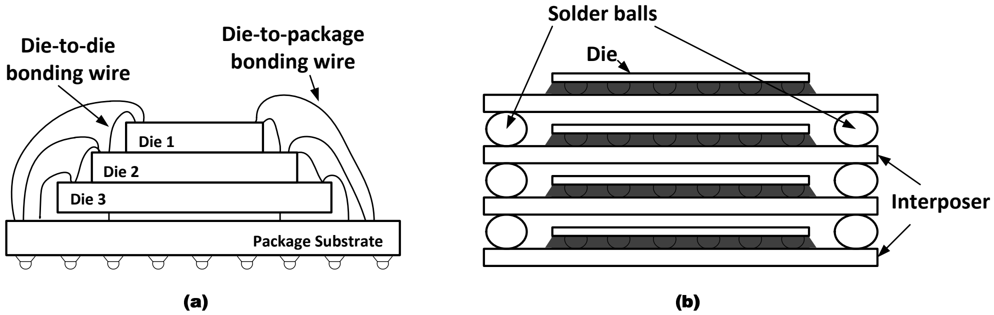


Figure 2.1: Example of the vertical connections among multiple planes in System-in-Package 3D Integration: (a) wire bonds, and (b) solder balls.

ple stacked dies are implemented using methods such as wire bonding or solder balls [29]. For example, as shown in Fig. 2.1(a), wire bonds are placed around the perimeter of each die to achieve die-to-die connections, as well as die-to-package connections. Similarly, in solder ball based method, an additional layer called interposer is inserted between two adjacent dies in the stack, as shown in Fig. 2.1(b). Solder balls are placed around the peripheral of the interposer layer to achieve connections between two interposers, while the interposer layer provides the connection between solder balls and the die. Compared with the wire bonding method, more dies can be stacked using solder ball based method due to the reduced parasitic impedance of solder balls [29].

However, although SiP reduces the off-chip interconnect length, the peripheral locations of bonding wires or solder balls undermine the advantage of reducing inter-plane interconnect length of 3D integration. As illustrated in Fig 2.2 (a) and (b), for two circuit blocks A and B which are located on different planes in a 3D stack, if wire bonds or solder balls are used to connect the two planes, the wiring between A and B detours to the peripheral of the die to reach the bonding wires or solder balls, even when A and B are physically close in the vertical dimension.

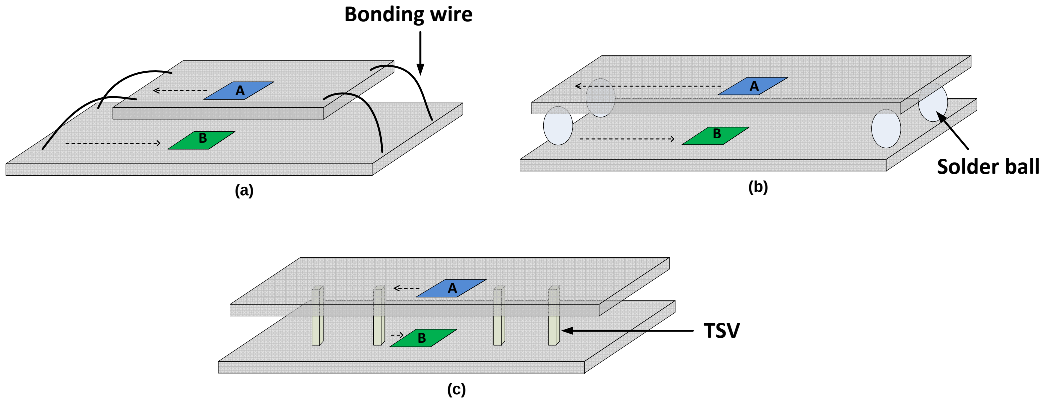


Figure 2.2: Inter-chip interconnects in different 3D integration technologies: (a) Wire bonding based SiP, (b) solder balls based SiP, and (c) TSV-based 3D IC.

Alternatively, in TSV-based 3D ICs, the communication between multiple dies is achieved by high density vertical TSVs. Due to the relatively small dimension of TSVs, a large number of TSVs can be placed throughout the entire die area. Therefore, as shown in Fig 2.2(c), to connect circuit blocks A and B, the wiring can go through a nearby TSV, instead of the longer distance toward the peripheral of the die. The vertical TSVs are fabricated with conductive materials with low resistivity, as introduced later. Therefore, the total impedance of inter-plane connections is significantly reduced in TSV-based 3D ICs. In MIV-based 3D ICs, connection topology of the MIVs is similar to TSVs, expect that the MIVs are smaller in size and therefore able to achieve fine-grained vertical connections. However, the MIV-based 3D ICs could not support a large number of tiers integrated vertically, due to the monolithic fabrication process, making it a less applicable solution than TSV-based 3D ICs for heterogenous systems integration.

In contactless coupled 3D ICs, the physical vertical vias are replaced by the coupling of electric or magnetic fields. Specifically, in capacitively coupled 3D integration, small on-chip parallel capacitors are placed on each plane, as shown in

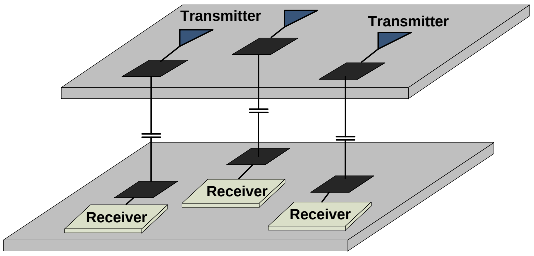


Figure 2.3: Capacitively coupled 3D integration.

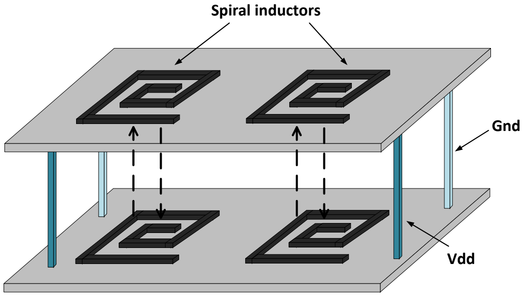


Figure 2.4: Inductively coupled 3D integration.

Fig. 2.3 [9]. Transmitter and receiver modules are required to drive the capacitors as well as modulate the received signals. Alternatively, in inductively coupled 3D ICs, spiral inductors are placed on each plane, as illustrated in Fig. 2.4 [9]. Specialized circuits are needed for transmitting and receiving signals. A common problem for the capacitively coupled and inductively coupled 3D ICs is the area overhead in fabricating the on-die capacitors and inductors, especially when CMOS technology is used where the fabrication of on-die capacitor and inductor is relatively area inefficient. Furthermore, the transmitter and receiver circuitry also consume power. Therefore, TSV-based 3D integration has been considered as the most promising approach in near future technology nodes.

2.1.1 TSV Structure and Parasitic Impedance

In TSV-based 3D ICs, multiple dies are stacked vertically while TSVs are the critical components that connect multiple dies in a 3D stack. As indicated by its name (through silicon via), TSVs are long vertical vias etched within the silicon substrate, as illustrated in Fig. 2.5(a). Conductive materials, such as polysilicon, tungsten or copper, are used to fill the etched vias. To prevent the conductive filling material from diffusing into the silicon substrate during the fabrication process, there is a dielectric insulation layer surrounding the TSVs [30]. The TSV physical characteristics such as depth, diameter, dielectric thickness are determined by the TSV fabrication technology, as introduced in the following section.

In 3D ICs with high integration level, thousands of TSVs are placed to provide vertical connections among the planes [31] [32]. These TSVs are utilized for different purposes, including distributing the power supply voltage (power TSVs), clock signal distribution (clock TSVs), as well as the communication between modules on different planes (signal TSVs). Therefore, it is important to reduce the parasitic impedance of the TSVs to achieve enhanced connectivity.

The parasitic impedance of a TSV consists of parasitic resistance R_{tsv}, parasitic inductance L_{tsv}, and parasitic capacitance C_{tsv}, as illustrated in Fig. 2.5(b) [33]. Various methods have been proposed to characterize the parasitic impedances of a TSV from both measurement and theoretical analysis [33–36]. The parasitic resistance and inductance are primarily determined by the TSV dimensions and filling material. Typically, a TSV is represented as a cylinder with a diameter W and depth D. The dielectric insulation, which is typically an oxide layer, has a thickness of t_{ox}. Therefore, the parasitic resistance R_{tsv} and inductance L_{tsv} can be estimated using the resistivity and permeability of the filling material and TSV dimensions [37]. The

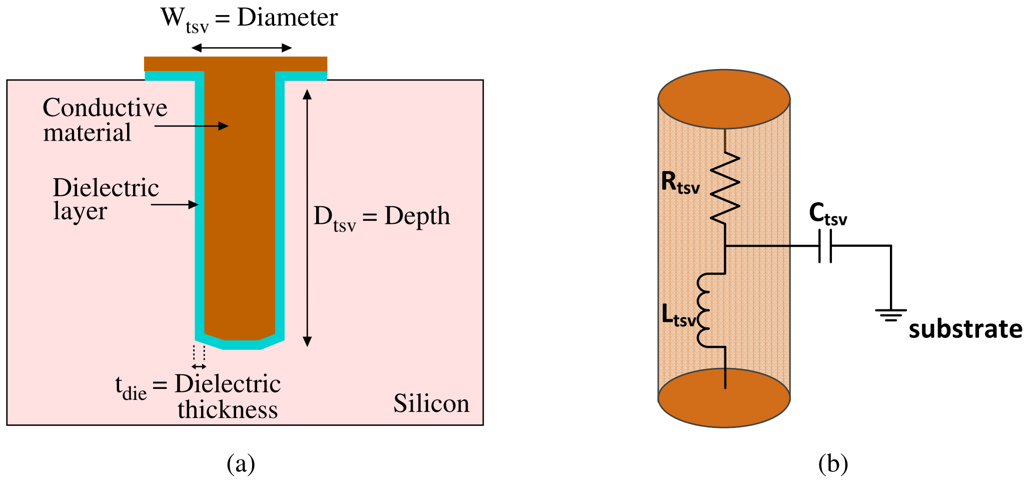


Figure 2.5: TSV characteristics: (a) cross-section of a TSV consisting of a conductive material and dielectric layer, (b) electrical model of a TSV illustrating the parasitic impedances.

parasitic capacitance C_{tsv} has two components: oxide capacitance and depletion capacitance. The oxide capacitance can be determined using the cylindrical capacitor formula [38]. When the TSV is charged at a bias voltage, there is also a depletion capacitance in series with the oxide capacitance [34]. A detailed method of estimating the parasitic impedances of a TSV with a distributed model is described in Chapter 6.

2.1.2 TSV Fabrication Technologies

In the process flow of TSV-based 3D ICs, the initial step is preparing the wafers to be bonded together. Each individual wafer is processed separately, which becomes a single plane of the 3D stack. During the wafer preparation, a certain amount of substrate area is reserved for TSVs. Deep reactive ion etching (DRIE) is used to etch the deep interplane vias within the silicon substrate [9]. These deep vias are later filled with conductive materials to form the TSVs. Depending on when the TSVs are formed in the process flow, TSV fabrication technologies can be categorized into via-first TSV, via-middle TSV, and via-last TSV technologies [39].

The preparation process of an individual wafer can be divided into two segments. The front-end-of-line (FEOL) includes the steps to pattern individual devices into silicon substrate. The back-end-of-line (BEOL) refers to all of the steps used to form the metal wires to interconnect the devices. The influence of FEOL and BEOL on the formation of TSV is primarily determined by temperature [9]. During FEOL, the temperatures involved in several process steps can be as high as 1000 °C, while the maximum temperature in BEOL is lowered to approximately 400 °C. A high temperature may impair the formed TSVs if the filling material of TSV cannot withstand the high temperatures. In via-first TSV fabrication approach, the TSVs are formed before FEOL. Therefore, via-first TSVs are typically filled with polysilicon, which can withstand the high temperature during the FEOL. Alternatively, in via-middle approach, TSVs are fabricated before BEOL but after FEOL. Without the necessity to withstand the high temperature during FEOL, the filling materials for via-middle TSVs are typically tungsten or copper, which have lower resistivity than polysilicon. A common characteristic of via-first and via-last TSVs is that, in both approaches, the TSVs are fabricated before the metal layers are formed. Therefore, the TSVs connect the bottom metal layer of the current plane and the top metal layer of the next plane, as shown in Fig. 2.6.

Alternatively, in via-last approach, the TSVs are fabricated after BEOL. Therefore, the high temperatures in FEOL and BEOL do not affect the filling material. Typically, copper is used as the filling material for via-last TSVs due to its low resistivity. Since the TSV formation occurs after the metalization is completed, via-last

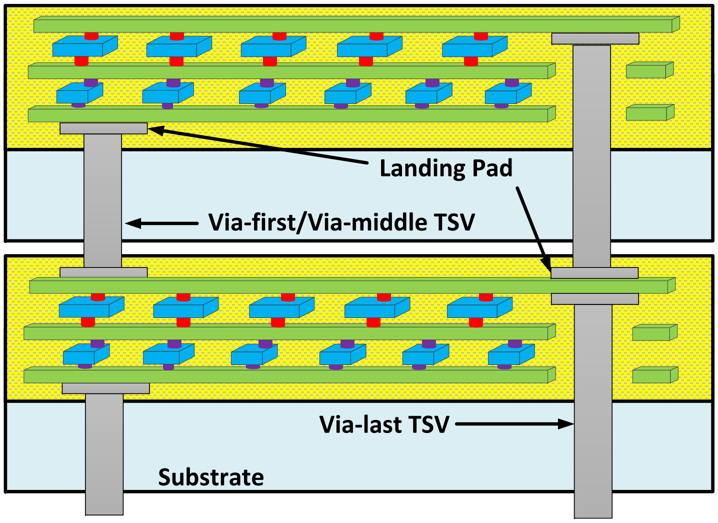


Figure 2.6: Comparison of the connectivity schemes of via-first/middle TSVs and via-last TSVs.

TSVs pass through not only the silicon substrate but also the metalization layers. In terms of connectivity, the via-last TSVs connect the top metal layer of the current plane and the top metal layer of the next plane, as shown in Fig. 2.6.

2.1.3 Progress in 3D Integrated Circuits

The development of 3D ICs has achieved significant progress in the past decade. In 2004, Intel explored the conversion of a traditional 2D implementation of a deeply pipelined Pentium 4 processor into a 3D chip [40]. The sub-modules in the 2D floorplan have been re-arranged and divided into two separate dies. These two dies have been bonded in a face-to-face fashion, *i.e.*, the metalization layers of the two dies are adjacent. Die-to-die vias connect the top metal layer of each die. Performance, power and thermal results of the 3D implementation have been compared with the original 2D chip. It has been demonstrated that 3D implementation provides 15% improvement in performance while reducing the power consumption by 15% [40].

In 2004, Tezzaron Semiconductor demonstrated a multiple-project wafer containing six designs of 3D ICs including memories, sensors and processors [41] [42] [43]. For instance, one of them is the world's first 3D IC RAM (reprogrammable memory) prototype chip [41]. Two dies of RAM circuitry are stacked in face-to-face fashion while tungsten filled via-first TSVs are used as the vertical interconnects. The 3D RAM exhibits 3 to 10 times better performance as compared to existing 2D RAMs at that time. The vertical via-first TSVs are only 10 micrometers in length, compared with the horizontal interconnects with a length of up to one centimeter in an equivalent 2D chip. Another notable prototype chip is a 8051-style processor/memory stack [42]. The vertical integration of logic processor and memory is shown to achieve higher speed and lower power consumption.

An example of 3D integration in a commercial application is the Apple A4 chip for first-generation iPad and iPhone 4 released in 2010 [44]. In the A4 chip, one processor die and two DRAM memory dies are vertically stacked and interconnected using wire bonding. Although no TSV is used in this design, placing RAM close to the processor reduces memory access latency.

In the 2012 International Solid-State Circuits Conference (ISSCC), additional 3D chips have been demonstrated [31] [32] [45]. One of these examples is the Centip3De chip from the University of Michigan [32], which explores the near-threshold computing (NTC) in 3D ICs. In the fabricated Centip3De chip, one of the two stacked dies contains 64 ARM M3 cores using near-threshold computing while the other die contains DRAM. Another 3D chip from Georgia Institute of Technology [31], the 3D-MAPS (3D Massively Parallel Processor with Stacked Memory), is also a two-tier 3D IC. The logic die on the top contains 64 general-purpose processor cores running at 277 MHz. The memory die on the bottom consists of 256KB SRAM. The 3D chip is fabricated using 130 nm technology while the die area is 5 mm × 5 mm. In this 25 mm^2 footprint, 50,000 TSVs and 50,000 face-to-face bondings are fabricated to achieve vertical interconnects.

2.2 Power Integrity

2.2.1 Power Supply Noise

With technology scaling, power supply voltage has also been reduced from 5 volts in the 1970s to sub one volt in modern technology nodes. Furthermore, due to increasing complexity and integration level, the overall current consumption of the chip has increased. This *low supply voltage* and *high current* characteristic causes a serious challenge in the design process of a power distribution network. If a large amount of current flows through a power distribution network, there is considerable voltage drop over the parasitic resistance and inductance that exist along the power delivery path. Therefore, the actual supply voltage delivered to local circuit blocks is less than the nominal power supply voltage, which is referred to as *power supply noise*. If the actual supply voltage delivered to the transistors is lower than a certain threshold, the circuit blocks may experience functional failure or performance degradation. Therefore, keeping the power supply noise within a tolerable limit, *i.e.*, maintaining system-wide power integrity, is critical to guarantee the normal operation of a chip. Depending on the cause of the voltage drop, the power supply noise can be categorized into IR drop and Ldi/dt noise.

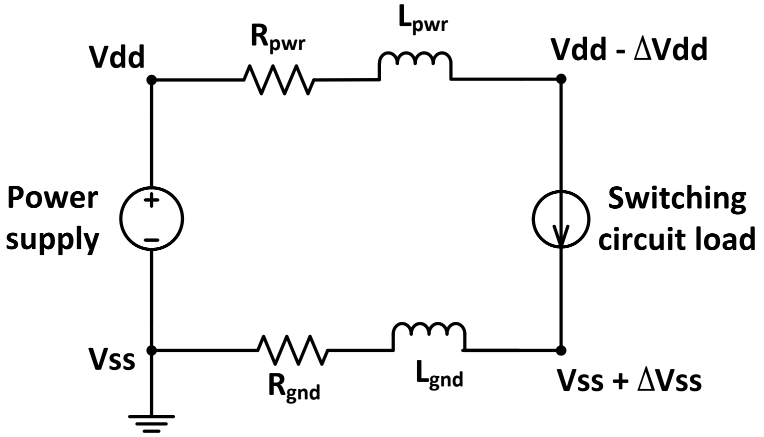


Figure 2.7: Power noise caused by the parasitic impedances of the power delivery path.

A simplified model of a power distribution network is shown in Fig. 2.7. It is assumed that the ideal power supply can provide a constant power supply voltage $V_{DD} - V_{SS}$. The overall resistance and inductance of the power (ground) delivery path is referred to as R_{pwr} and L_{pwr} (R_{gnd} and L_{gnd}). Due to the voltage drop on the parasitic impedances, the actual supply voltage delivered to the switching circuit load is $(V_{DD} - \Delta V_{DD}) - (V_{SS} + \Delta V_{SS})$. The power supply noise ΔV_{DD} and ΔV_{SS} can be formulated by

$$\Delta V_{DD} = V_{DD}^{\text{nominal}} - V_{DD}^{\text{delivered}} = I(t)R_{pwr} + L_{pwr}\frac{dI(t)}{dt}, \tag{2.1}$$

$$\Delta V_{SS} = V_{SS}^{\text{nominal}} - V_{SS}^{\text{delivered}} = I(t)R_{gnd} + L_{gnd}\frac{dI(t)}{dt}, \tag{2.2}$$

where $I(t)$ is the transient switching current drawn by the circuit load. The items $I(t)R_{pwr}$ and $I(t)R_{gnd}$ in (2.1) and (2.2) represent the IR drop, caused by the switching current flowing through the resistive impedance of the power delivery path.

Therefore, the *IR* drop is proportional to the magnitude of the switching current. Alternatively, the items $L_{pwr}\frac{dI(t)}{dt}$ and $L_{gnd}\frac{dI(t)}{dt}$ in (2.1) and (2.2) represent the di/dt noise. The voltage drop across an inductor is proportional to the *rate of change* of the current flowing through the inductor. It is therefore possible to have a large di/dt noise due to a high rate of change in current, even if the magnitude of the current is small.

Maintaining power integrity has been a challenging issue in traditional 2D ICs. However, in 3D ICs with additional stacked planes, maintaining system-wide power integrity for all of the planes becomes a more difficult task due to the complex characteristics of power distribution networks in 3D ICs, as introduced in the following section.

2.2.2 Power Distribution Networks in 3D ICs

A power distribution network can be divided into two levels: package level and on-die level. The package-to-die interface for power delivery is introduced in Section 2.2.2.1 whereas the on-die power grid structure is discussed in Section 2.2.2.2.

2.2.2.1 Package-to-die Interface

To achieve a reliable and low impedance package-to-die interface, various connection structures have been proposed. In early chip packages, wire bonding is the primary method for package-to-die connections, as illustrated in Fig. 2.8 [29]. The die is bonded to package substrate using glue with good thermal conductivity. Aluminum or gold wires are used to achieve connections between the on-die pads and lead frames on the package. However, the bonding wires typically have high

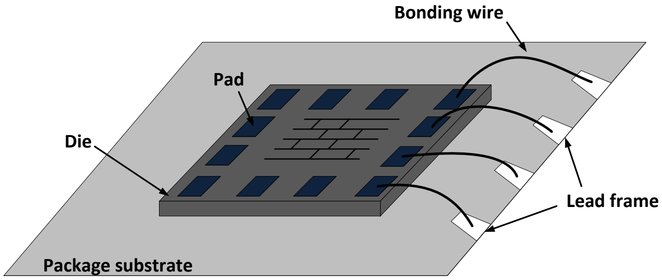


Figure 2.8: Wire bonding method for package-to-die interface [29].

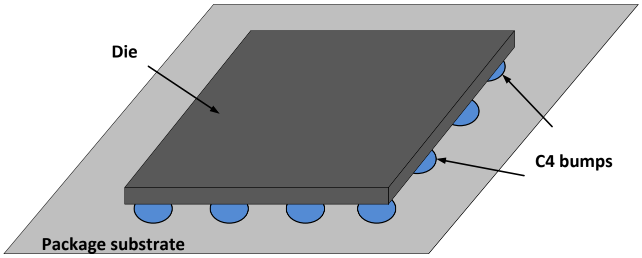


Figure 2.9: Flip-chip method for package-to-die interface [29].

self inductance as well as high mutual inductance between adjacent wires. In more advanced technologies, wire bonding method is gradually replaced by the flip-chip technology which has superior electrical properties. As illustrated in Fig. 2.9, the die is flipped (the silicon substrate is on the top while metallization layers are at the bottom) and attached to the package with an array of solder bumps, *i.e.*, controlled collapse chip connection (C4) bumps. The inductance of a C4 bump is typically in the range of 0.1 nH to 0.5 nH, which is significantly smaller than the inductance of a bonding wire (4 nH to 10 nH) [29].

2.2.2.2 On-Die Power Grid

After the power is delivered from package to die, the on-die power distribution is achieved by the on-chip power/ground metal interconnects. In modern CMOS fabrication technologies, the interconnects are implemented as multi-level metallization layers. The number of metal interconnect layers is typically in the range of 8 to 11 in recent nanoscale technology nodes [46] [47]. The multi-level interconnects are utilized to achieve different types of connections, including on-die power distribution to each circuit block, providing clock signal to all sequential circuits, and data communication among circuit blocks.

As power supply voltage decreases with technology scaling, distributing power over the entire die area with minimum voltage noise along the power delivery path has become more challenging. A common practice is to place abundant number of power lines over the entire die area to ensure that the noise at any spatial location is within the limit. The on-die power distribution network typically consists of hierarchical power grids, *i.e.*, the global power grid which distributes power at a coarse level over the entire die area, and the local power grids which receive power from the global grid and further distribute the power supply voltage to the local circuit blocks within a smaller area. In certain cases, a semi-global power grid is inserted between the global grid and local grid, producing a more hierarchical power distribution network.

In modern high performance ICs, several top metal layers in the multi-level interconnect metallization are typically reserved for power distribution. For instance, in a 65 nm technology node with eight metal layers, the top two metal layers (metal 8 and 7) are dedicated to the global power grid [20]. An illustrative schematic of the global power grid in a high performance microprocessor is shown in Fig. 2.10 [48].

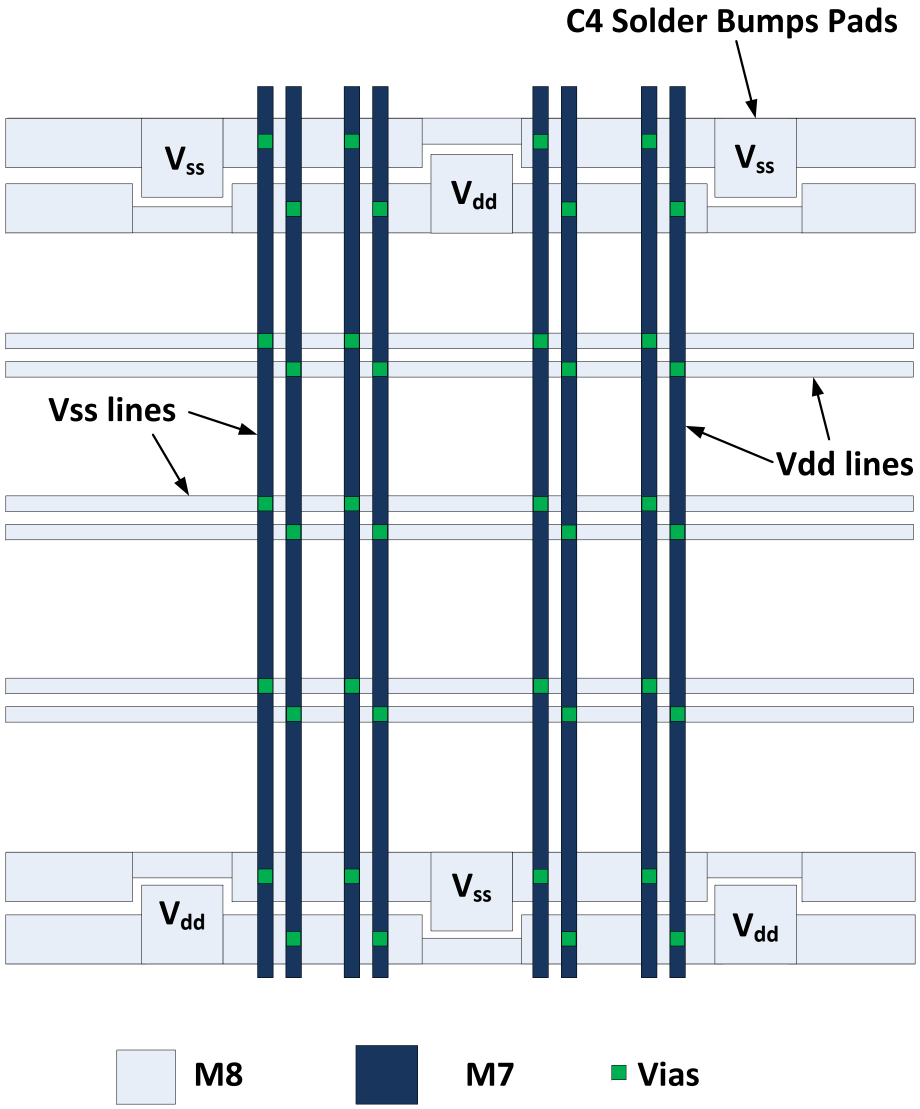


Figure 2.10: Structure of a global power grid in a high performance microprocessor [48].

In this structure, horizontal metal lines are placed on metal layer 8 (M8) while vertical lines are placed on metal layer 7 (M7). On each metal layer, VDD (power) lines and GND (ground) lines are interdigitated. Vias are placed at the intersection of horizontal and vertical metal lines to achieve connectivity.

2.2.2.3 3D Power Distribution Network with TSVs

In a 3D IC with multiple planes, the global and local power grids on each plane are similar to the power grids in 2D ICs introduced in previous section. TSVs connect the power grids on each plane. The TSVs are bonded with a metal layer through landing pads. Depending on whether via-first, via-middle, or via-last TSV technology is used, the landing pads are either at the bottom metal layer (for via-first/middle TSVs) or top metal layer (for via-last TSVs). The structure of a typical power distribution network in a 3D IC with three planes is shown in Fig. 2.11. In this structure, power supply is initially delivered from the package to the plane closest to the package (plane 1 in the figure) through C4 bumps. Depending upon the orientation of the dies and package substrate, it can be either the top plane or the bottom plane that receives the power supply voltage from the package. The global power grids on each plane are interconnected by the TSVs. Therefore, the power supply voltage is delivered sequentially from one plane to another through the TSVs. Within each plane, the global and local power grids distribute power to each circuit block within that plane.

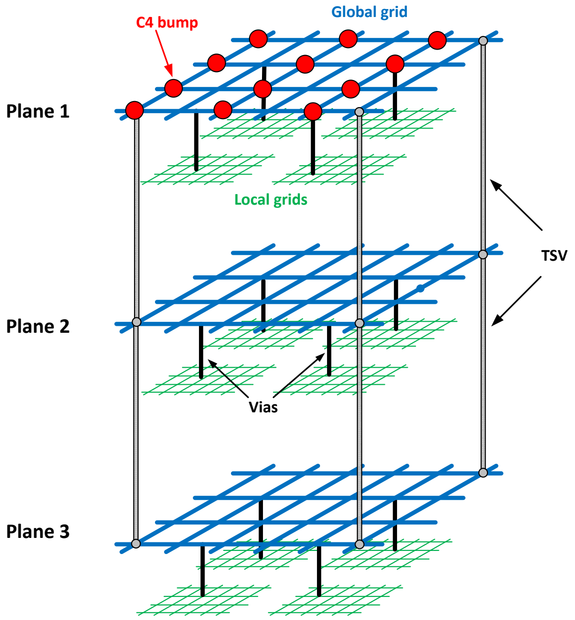


Figure 2.11: Illustrative structure of a power distribution network of a three-plane 3D IC consisting of C4 bumps, global power grids, local power grids, and TSVs.

2.2.3 Decoupling Capacitance

Due to the critical power integrity challenge, placing decoupling capacitors has become an important design technique to effectively reduce power supply noise in complex power distribution networks. Decoupling capacitance refers to the intentional or intrinsic capacitors placed across the power and ground rails, which provide instantaneous charge to the switching circuit loads within a short period of time. As shown in Fig. 2.12, the insertion of decoupling capacitor divides the power distribution network into two segments, *i.e.*, the upstream section from the power supply to decoupling capacitor (containing R_{p1}, L_{p1}, R_{g1}, L_{g1}), and the downstream section from decoupling capacitor to circuit load (containing R_{p2}, L_{p2}, R_{g2}, L_{g2}). At high frequencies, the decoupling capacitor path exhibits smaller impedance than the

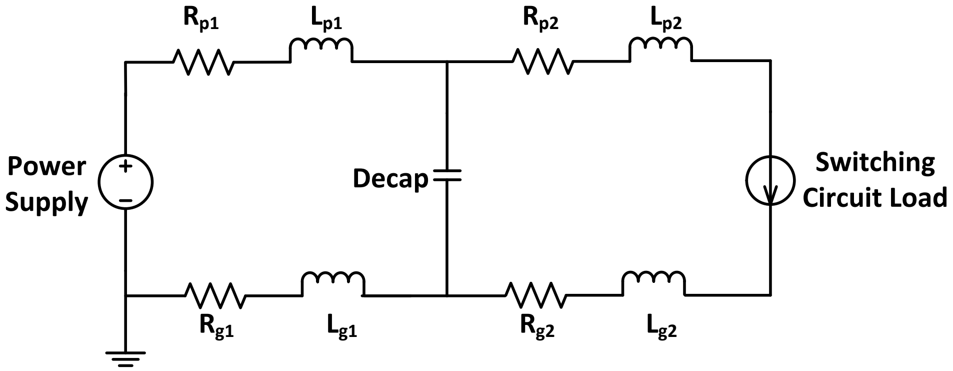


Figure 2.12: The circuit model of a power distribution network with decoupling capacitor.

upstream network, which effectively decouples the high impedance upstream parasitics from the load circuits. The decoupling capacitor serves as a temporary charge reservoir to provide transient charge to the switching load circuit. Alternatively, at low frequencies, the impedance of decoupling capacitor path is high. The switching load current is provided primarily by the power supply. Typically, decoupling capacitors are inserted in a hierarchical manner along a power distribution network, including board level, package level, and on-die level decoupling capacitors. In modern IC fabrication technologies, on-die decoupling capacitors can be implemented in different forms such as Metal-Oxide-Semiconductor (MOS) capacitor and Metal-Insulator-Metal (MIM) capacitor. It is reported that in high performance microprocessors, the on-die decoupling capacitors consume more than 20% of the overall die area [20].

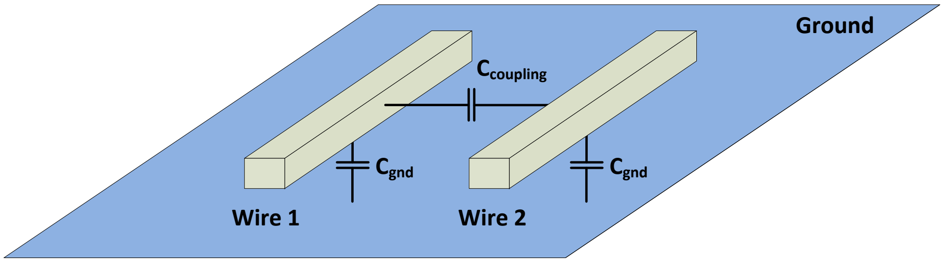


Figure 2.13: Crosstalk between two adjacent wires.

2.3 Signal Integrity

Another related critical issue in 3D ICs is the system-wide signal integrity. Multiple types of noise exist in 3D ICs which affect the integrity of signals transmitted within the system. These noise sources include crosstalk (Section 2.3.1), substrate coupling (Section 2.3.2), and the TSV-induced noise coupling (Section 2.3.3).

2.3.1 Crosstalk

Due to high density interconnects, the distance between adjacent wires is reduced. As shown in Fig. 2.13, there is capacitive coupling between two adjacent wires, referred to as crosstalk. The coupling capacitance $C_{coupling}$ is determined by the distance between the two adjacent wires and the insulation material. Suppose wire 1 is switching (the aggressor) whereas wire 2 is quiet (the victim), noise couples from the aggressor to the victim since a noise spike occurs in wire 2. The induced noise can be analyzed using the model in Fig. 2.14. If wire 2 is floating, the noise at victim can be formulated by

$$\Delta V_{victim} = \frac{C_{coupling}}{C_{gnd2} + C_{coupling}} \cdot \Delta V_{aggressor}, \tag{2.3}$$

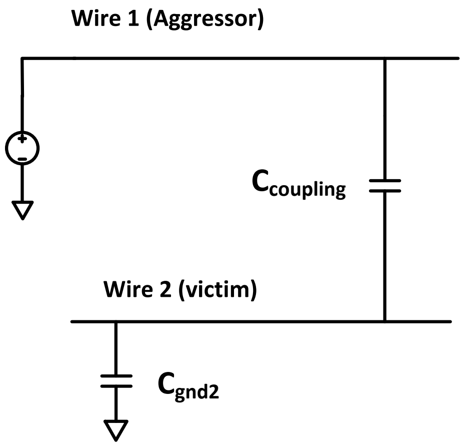


Figure 2.14: Electrical model for crosstalk analysis.

where C_{gnd2} is the capacitance of wire 2 to ground. Alternatively, when wire 2 is also driven by another driver, the noise is affected by the ratio of the time constants of the aggressor wire and victim wire [49].

2.3.2 Substrate Coupling

Substrate noise coupling is another noise propagation mechanism in traditional 2D ICs, particularly important in mixed-signal ICs where the noisy digital circuits and noise-sensitive analog/RF circuits coexist within the same die. The common substrate provides a medium for noise to propagate between the digital modules and analog/RF modules. As shown in Fig. 2.15, aggressor circuit injects noise into the substrate from the drain/source terminals of a transistor. The primary mechanisms of noise injection into substrate include impact ionization, coupling from source/drain junction capacitance, and coupling from the power/ground networks of the aggressor circuits [50].

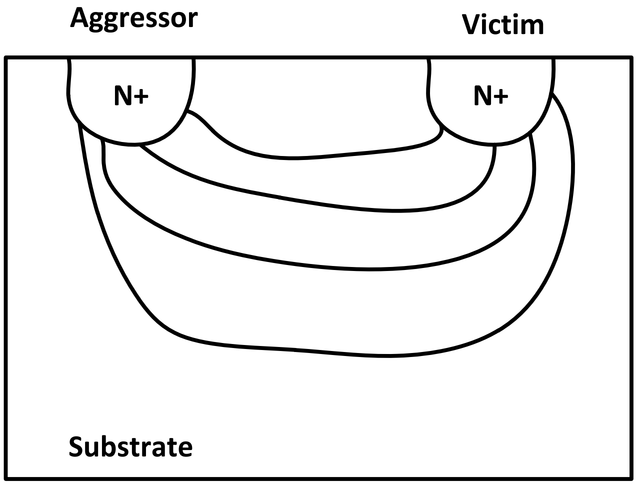


Figure 2.15: Current flows within the substrate causing noise coupling between aggressor and victim.

2.3.3 TSV-Induced Noise Coupling

In addition to the traditional crosstalk and substrate coupling mechanisms, 3D ICs also suffer from TSV-induced noise coupling [51] [52]. Specifically, signals with fast transitions are transmitted between different planes using TSVs. During a signal transition within a TSV, noise couples from the TSV into the substrate due to both dielectric and depletion capacitances. The coupling noise propagates throughout the substrate and disturbs the operation of nearby transistors.

This issue is important particularly for heterogeneous 3D ICs. For example, in a 3D chip shown in Fig. 2.16, multiple planes consisting of analog sensing front-end, digital logic circuit, memory, and communication modules are integrated within the same 3D stack. Note that in this 3D system, the front-end circuitry consisting of analog/RF blocks is typically located at the top plane (closer to the I/O pads) to reduce the overall impedance between the pads and analog inputs. However, analog/RF blocks and memory cells are among the most sensitive circuits to substrate

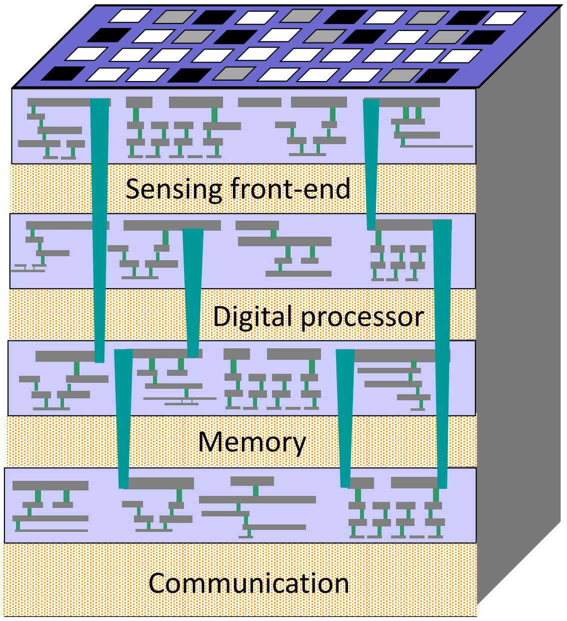


Figure 2.16: Three-dimensional integration of diverse planes using TSV technology [5].

noise coupling. The TSVs are required to transmit the digital signals (including the clock signal) to the digital plane. These TSVs travel through the substrate of the analog sensing front-end plane. Thus, TSV-induced noise becomes an important issue for the reliability of the analog/RF blocks. Digital transistors are also affected by TSV-induced noise if the physical distance between the TSV and active devices is sufficiently short [53]. The TSV-induced substrate noise coupling is investigated in detail in Chapter 7.

Chapter 3

Power Gating in 3D Integrated Circuits

In this chapter, the implementation of a well-known low power design technique, *i.e.*, power gating, in 3D ICs is investigated. A brief overview of low power design techniques is presented in Section 3.1. The power gating technique, which aims at reducing the leakage power consumption, is further discussed in Section 3.2. Two schemes for implementing power gating in 3D ICs are proposed in Section 3.3. Simulation results are provided in Section 3.4 to evaluate the proposed power gating schemes.

3.1 Low Power IC Design Methodologies

3D integration technology enables the continuation of Moore's Law in the nanoscale era. The number of transistors in a single chip has reached several billions. Even though the modern fabrication technologies permit integrating this number of transistors on the chip, all of the transistors cannot work simultaneously due to power consumption and thermal dissipation issues. It is reported that in a 3.3 GHz Intel Core i7 microprocessor, the thermal design power (TDP) is as high as 130 Watts [54]. Dissipating the heat produced from this amount of power from a chip with a 1.5 cm $\times$ 1.5 cm area reaches the practical limits of air cooling. Therefore, an imminent requirement is reducing the power consumption while maintaining the performance, *i.e.*, high energy efficiency.

The power consumption of a modern CMOS VLSI chip consists of two components: 1) dynamic power consumption, which is the power consumed during switching activity due to charging and discharging the capacitors, and 2) static power consumption, which is the power consumed when the chip is idle. Static power includes the subthreshold leakage of MOS transistors, gate leakage through the gate dielectrics, and junction leakage. Over the past several decades, significant effort for reducing dynamic and static power consumption has been made, spanning from device to architecture levels.

At the device level, novel device structures and technologies have been proposed. For example, in the HKMG (High-K and Metal Gate) technology, a high permittivity (high-K) dielectric material such as Hafnium is used as the insulator between the metal gate and the channel in MOS transistors, which reduces the gate leakage power [55] [56]. Alternatively, in multi-gate devices such as FinFET or tri-gate transistor, the gate surrounds the transistor channel to achieve enhanced control of the channel, thereby reducing the subthreshold leakage power [57].

At the architecture level, to achieve higher performance without excessive power consumption, parallelism has been extensively implemented in high performance microprocessors. For example, Intel abandoned single-core processors in 2004 [54]. As another example, in the big.LITTLE architecture from ARM, slower and lower power cores are coupled with more powerful and power-hungry cores within the same SoC [58]. Therefore, the processor can dynamically allocate the cores depending on real-time workload, preventing the excessive power consumption caused by the unnecessary usage of high performance cores for light workloads.

At the circuit level, multiple low power design methodologies have been proposed to reduce both dynamic and static power consumption, such as *power gating*, *near-threshold computing*, and *clock gating* [50, 59–61]. In power gating method, when a circuit block is in idle state, the power delivery to this block is prevented to save leakage power [59]. Alternatively, near-threshold computing is proposed for application scenarios where battery life has priority over performance. Transistors operate in the near-threshold region, where the power supply voltage is in the range of 300-500 mV in nanoscale technology nodes. With this low power supply voltage, CMOS logic circuits can operate at low MHz frequency range while the power consumption is several orders of magnitude lower than circuits operating at typical power supply voltages [60]. In clock gating, the clock signal to the flip-flops in a synchronous circuit block is disabled when the block is not active. Therefore, significant dynamic power from the switching of these flip-flops is saved [61].

In 3D ICs, due to the vertical connections achieved by TSVs, the power distribution network and clock distribution tree exhibit unique characteristics different from traditional 2D ICs. Therefore, the application of these circuit level low power design methodologies (such as power gating, clock gating and near-threshold computing) to 3D ICs requires additional considerations. In this chapter, the implementation of power gating in TSV-based 3D ICs is investigated.

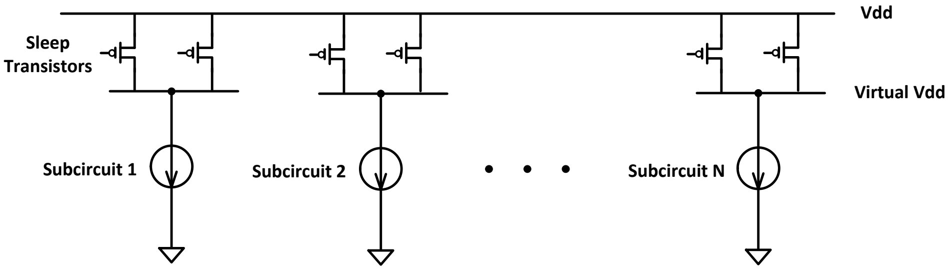


Figure 3.1: Sleep transistors to control the power delivery of individual circuit blocks.

3.2 Power Gating

In power gating method, in order to independently control the power delivery to each subcircuit, sleep transistors are inserted along the power delivery path, as illustrated in Fig. 3.1. Typically, sleep transistors are implemented as MOS transistors with high threshold voltage. Sleep transistors can be inserted either between the power rail and a virtual power rail (referred to as header), or between the ground rail and a virtual ground rail (referred to as footer). In this work, the header design style is considered, as shown in Fig. 3.1.

When a subcircuit is active, the sleep transistors controlling this subcircuit are switched on, connecting the virtual power rail with the actual power rail. Due to the nonnegligible resistance of the sleep transistors when turned on, there is voltage drop across the sleep transistors. Thus, the voltage on virtual power rail is lower than the supply voltage on the actual power rail, which exacerbates the issue of power supply noise. Alternatively, when the subcircuit is in sleep state, the sleep transistors controlling this subcircuit are turned off. The voltage on the virtual power rail (virtual V_{DD}) is pulled down to a sleep voltage that is close to zero volt, thereby significantly reducing the leakage current of the logic gates in this subcircuit. Note that, sleep transistor itself has nonnegligible subthreshold leakage current. Thus, the multi-threshold voltage CMOS (MTCMOS) technology is often utilized in circuits with power gating, where MOSFETs with low threshold voltage are used for logic gates to enhance performance, whereas the high threshold voltage MOSFETs are used for sleep transistors to reduce subthreshold leakage current [59].

A design consideration for power gating is the granularity at which power gating is achieved. In fine-grained power gating, each individual gate can be gated [62]. However, the area and power overhead of sleep transistors can be excessive [62]. Thus, power gating is typically achieved at the block level where the power supply voltage delivered to each macroblock can be be individually controlled. In 3D ICs, since the circuit blocks on a traditional 2D chip can be distributed to different planes within a 3D stack, an exclusive granularity level is the plane-level power gating, which controls the power delivery for each individual plane.

3.2.1 Sleep Transistors and Power Distribution Network

To physically implement power gating in ICs with a complex power distribution network, the interaction between the additional sleep transistors and existing power distribution network should be considered. A typical physical structure of a power distribution network in 2D ICs with block-level power gating is illustrated in Fig. 3.2. In this structure, the power supply voltage is initially delivered from package, through the C4 bumps, to the highest metal layer on die. It is assumed that there are 8 metalization layers in this illustrative example. Within the metalization layers, Metal 8 and 7 are dedicated to the global power grid while Metal 6 and 5 are used for the local power grid. The rest of the metal layers are used for local

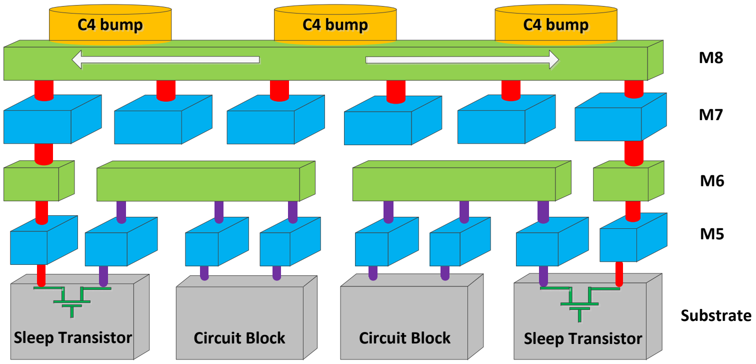


Figure 3.2: Physical structure of a power distribution network with power gating.

interconnects. Note that the sleep transistors, which are implemented as MOS transistors, are fabricated in the silicon substrate. Therefore, the power supply voltage has to be transmitted from the global power grid, through a stack of vias, to the sleep transistors, as shown in Fig. 3.2. If the sleep transistors are on, the power supply voltage propagates back to the local power grids, which further distribute the power within each circuit block. The sleep transistors and stacks of vias introduce additional impedance along the power delivery path.

3.3 Proposed Power Gating Topologies for 3D ICs

While fine-grained power gating in 3D ICs is similar to traditional 2D ICs, the coarse-grain (plane- and block-level) power gating in 3D ICs exhibits different characteristics due to the TSVs. To implement coarse-grained power gating in 3D ICs, two power gating topologies are proposed in Section 3.3.1, each tailored for specific TSV fabrication technology. In Section 3.3.2, the resistance of TSVs and stacks of metal vias are investigated to justify the proposed TSV-specific power gating topologies.

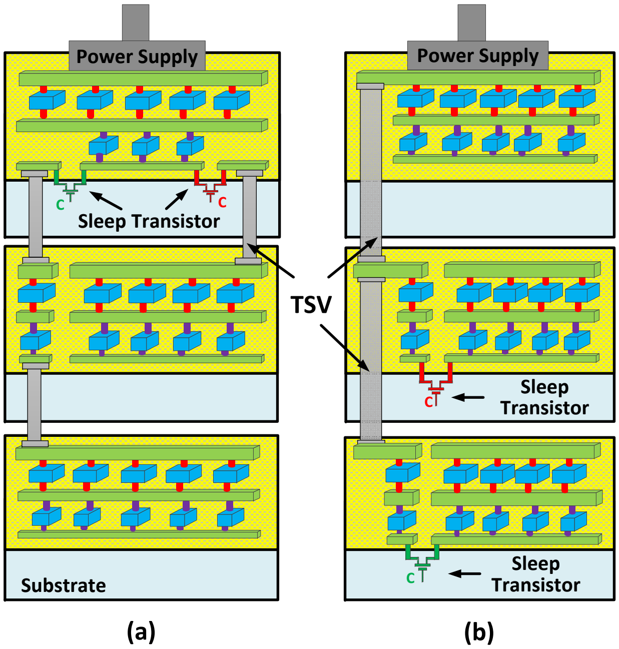


Figure 3.3: Physical structures of proposed power gating topologies: (a) lumped power gating topology, and (b) distributed power gating topology.

3.3.1 Lumped and Distributed Topologies

In the lumped power gating topology, as proposed for via-first and via-middle TSVs, all of the sleep transistors are fabricated on the top-most plane. As mentioned in Chapter 2, via-first/middle TSVs connect the lowest metal layer of a plane with the highest metal layer of an adjacent plane. Thus, sleep transistors are placed between the lowest metal layer of the top-most plane and TSVs, which also land on this layer, as depicted in Fig. 3.3(a). An advantage of the lumped topology is to maintain all of the control signals for sleep transistors within a single plane. With heterogeneous integration in 3D ICs, the devices on different planes can be fabricated with different technologies. Therefore, it is possible that, in the lumped topology, all the sleep transistors can be implemented as novel devices such as

MEMS-based (microelectromechanical systems) switch while the logic circuits are fabricated with traditional CMOS technology [63]. The implementation of sleep transistors as MEMS switches significantly reduces the leakage current of sleep transistors compared with traditional MOS transistor implementation, which enhances the power savings achieved by power gating.

Alternatively, in the distributed power gating topology, as proposed for via-last TSVs, sleep transistors are placed on each plane. Contrary to the lumped scheme, in the distributed topology, TSVs are continuously connected to the power supply. Note that distributed topology is advantageous for via-last TSVs since via-last TSVs connect to the power network at the highest metal layer in each plane. Thus, in the distributed scheme, additional stack of vias that transmits the virtual supply voltage (after the sleep transistor) back to the TSV is avoided, thereby enhancing the power supply noise during normal operation.

It is worth mentioning that in Fig. 3.3, the lumped topology is illustrated using via-first/via-middle TSVs, while the distributed topology is using via-last TSVs. However, the lumped topology can also use via-last TSVs. But additional stacks of vias are needed to connect the sleep transistors to the via-last TSVs, which connect to the top metal layer within a plane. The situation is similar to using via-first/via-middle TSVs with lumped power gating topology. Additional stacks of vias connect the top metal layer and bottom metal layer within a plane so that TSVs on different planes are connected. Because of the same connectivity of via-first and via-middle TSVs and the fact that via-first TSVs have larger impedance than via-last TSVs, via-middle TSVs are used in this work to represent via-first/via-middle TSV technology and compare with via-last TSVs.

	Minimum Width	**Minimum Spacing**	**Resistance**
Via 1	65 nm	75 nm	6 Ω
Via [2-3]	70 nm	85 nm	5 Ω
Via [4-6]	140 nm	160 nm	3 Ω
Via [7]	400 nm	440 nm	1 Ω

Table 3.1: Physical characteristics of vias in a 45 nm technology [66].

3.3.2 Comparison Between Metal Via Stack and TSV Resistance

As discussed in Section 3.3.1, the TSVs and metal via stacks are two critical components in the power delivery path for each plane. In previous works on power distribution network in 3D ICs, the focus is typically on TSVs while the resistance of the stacks of vias is neglected [64] [65]. In this section, it is demonstrated that the resistance of stacks of vias is nonnegligible as compared to TSVs.

Physical characteristics of the vias connecting different metal layers in a 45 nm technology are listed in Table 3.1 [66]. Via N represents a via that connects metal layers N and $N+1$. Assuming eight metal layers, the total resistance of a single stack of metal vias is approximately 26 ohms. A large number of vias is typically used to reduce the effective resistance.

Similarly, the physical characteristics of the TSVs are listed in Table 3.2. Note that only via-middle and via-last TSVs are considered since polysilicon-filled via-first TSVs have significantly high resistance [39]. As listed in this table, the resistance of a single copper-filled via-middle and via-last TSVs is, respectively, 66.8 mΩ and 12.8 mΩ. The differences in the physical dimensions of the TSVs and metal via stacks are also listed.

To achieve a fair comparison, the area is assumed to be constant for both TSVs and stack of metal vias. The number of vias that can be placed within an area consumed by a single TSV is determined. For a via-middle TSV, approximately

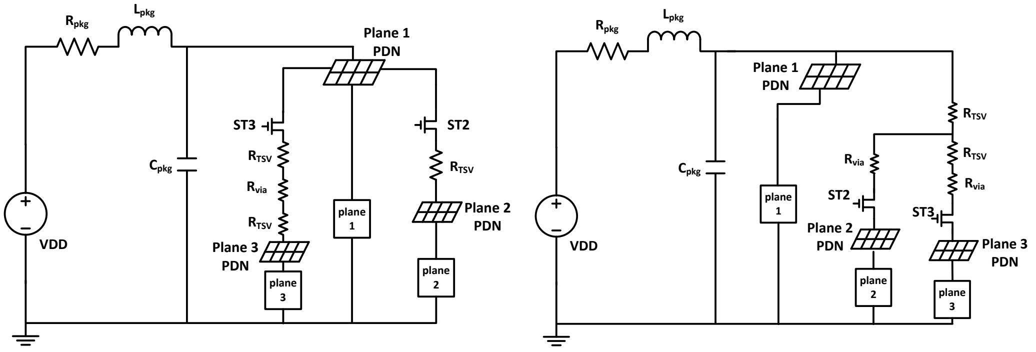
	Via-middle TSV	**Via-last TSV**
Filling material	Copper or tungsten	Copper
Material resistivity	16.8 nΩm or 56 nΩm	16.8 nΩm
Diameter	4 μm	10 μm
Pitch	8 μm	20 μm
Height	50 μm	60 μm
Single TSV resistance	66.8 mΩ or 222.8 mΩ	12.8 mΩ

Table 3.2: Physical characteristics of TSVs [39].

91 stacks of vias can be placed in parallel. Alternatively, approximately 567 via stacks can be placed in an area of a single via-last TSV. The effective resistance of the vias is 285.7 mΩ and 45.8 mΩ for, respectively, via-middle and via-last TSVs. Thus, under constant area, stack of metal via resistance is nonnegligible. This characteristic should be considered when developing power gating topologies for 3D ICs. A primary advantage of the proposed TSV-specific power gating topologies is avoiding the unnecessary metal via stacks along the power delivery path.

3.4 Simulation Results

To evaluate the proposed power gating topologies, analyses are performed in HSPICE using an industrial 65 nm CMOS technology with a nominal supply voltage of 1 V. In modern high performance ICs, the peak power can exceed 100 W while the die area can reach 250 mm^2 [67]. Thus, the average peak power density is approximated as 0.4 W/mm^2. In the analyses, a three-plane 3D IC with a die area of 1 mm^2 (for each plane) is considered to maintain reasonable computational complexity. The peak power for each 1 mm^2 plane is 400 mW. The leakage power consumption is assumed to be 20% of the overall power consumption. The clock



(a) (b)

Figure 3.4: Simplified equivalent model of a three-plane power distribution network consisting of: (a) via-middle TSVs with lumped power gating topology, and (b) via-last TSVs with distributed power gating topology.

Metal layer	Thickness	Width	Pitch
8	0.975 μm	9.0 μm	19.08 μm
7	0.650 μm	8.1 μm	16.92 μm

Table 3.3: Physical characteristics of the global power grid [69].

frequency is 2.5 GHz.

The analysis of the two power gating topologies is achieved based on the power distribution network models, as illustrated in Fig. 3.4. These models represent simplified equivalent circuits of the physical structures depicted in Fig. 3.3. In the models, the package level parasitic impedances of R_{pkg} = 1 mΩ, L_{pkg} = 120 pH, and C_{pkg} = 26 μF are used based on the measurement results from [68]. Metal layers 7 and 8 are dedicated to global power distribution grid, where an interdigitated metal lines structure is adopted. The primary physical characteristics of the global power grid are listed in Table 4.1 [69]. The power grid within each plane (1 mm^2 area) is simulated with the field solver FastHenry to extract the parasitic impedance [70]. The extracted impedance is used within the electrical models shown in Fig. 3.4.

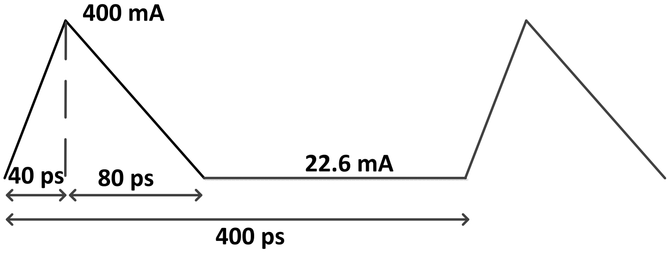


Figure 3.5: Current waveform used to mimic switching current within an area of 1 mm^2.

Parameter	Value
Gate voltage (V_{gs}^{st})	1 V
Threshold voltage (V_{th})	0.6 V
Channel length (L)	60 nm
C_{ox}	1.73 $\times 10^{-6}$ F/cm^2

Table 3.4: Physical characteristics of the sleep transistors [39, 71].

3.4.1 Power Supply Noise

When investigating power supply noise, the current waveform illustrated in Fig. 3.5 is used to mimic the switching current drawn by each plane. The period, idle time, and transition time of the waveform are determined from the clock frequency. Peak current and static leakage current are determined based on the peak and leakage power. These values are indicated in Fig. 3.5. Related physical characteristics of sleep transistors are listed in Table 3.4.

To evaluate the proposed power gating topologies, power supply noise is analyzed at the bottom-most plane as a function of sleep transistor width, as shown in Fig. 3.6. According to Fig. 3.6, lumped topology is favorable for via-middle TSVs, whereas distributed topology is preferred for via-last TSVs. Specifically, for via-middle TSVs, the preferred lumped topology achieves, on average, 14.8%

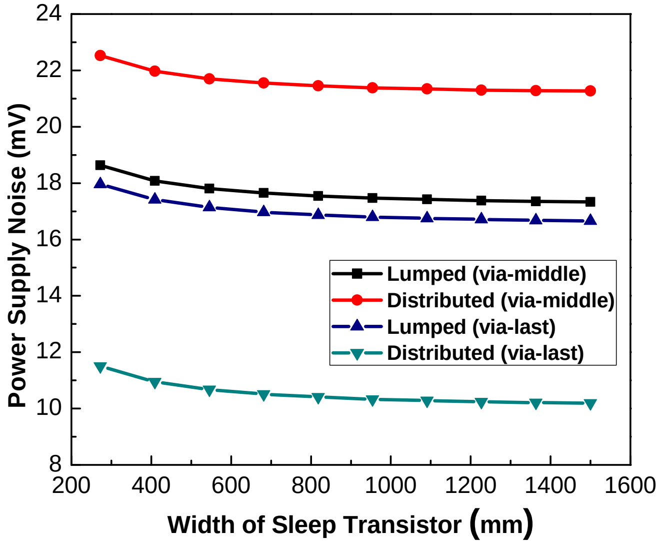


Figure 3.6: Power supply noise produced by the proposed power gating topologies: (a) effect of the size of sleep transistor, and (b) effect of number of TSVs.

reduction in power supply noise compared with the non-preferred topology. For via-last TSVs, reduction in power supply noise is approximately 22.9% with the preferred distributed topology compared with the non-preferred lumped topology. The reason is the additional resistance due to stack of metal vias if lumped topology is used for via-last TSVs and distributed topology is used for via-middle TSVs. Also note that via-last TSVs exhibit less power supply noise due to significantly lower TSV resistance.

3.4.2 Power Gating Noise

When a circuit block or an entire plane within a 3D IC transitions from sleep state to active state, a relatively large in-rush current is drawn to charge the capacitors in that block or plane, producing power noise at the semi-global or global power network. This noise affects the reliability of other active blocks and planes,

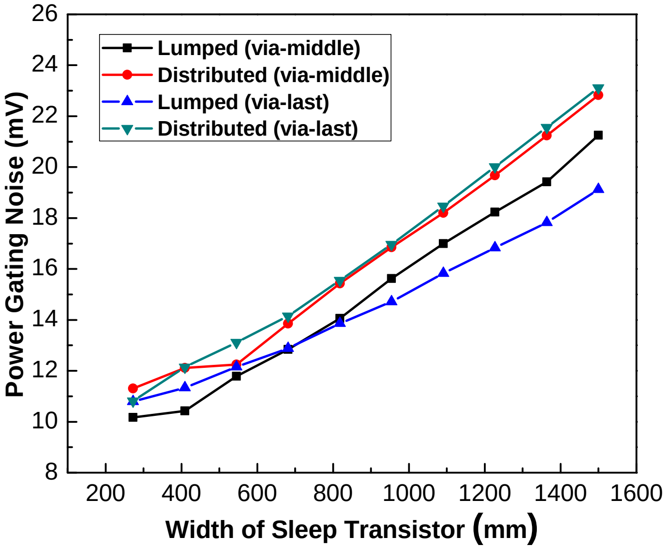


Figure 3.7: Power gating noise as a function of sleep transistor size for the proposed power gating topologies.

and is referred to as *power gating noise* or *in-rush current noise*. To accurately analyze power gating noise in 3D ICs, a ring oscillator consisting of five cascaded inverters is utilized to replace the current source used in previous analyses. Active devices are used rather than piecewise linear current source to accurately consider the dynamic turn-on behavior of the transistors. The sizing of the ring oscillator is decided to match the assumed power consumption. The same configuration is also used to analyze the turn-on time and leakage power reduction, as described in the following sections.

The power gating noise obtained by the proposed two topologies is shown in Fig. 3.7 as a function of sleep transistor width. One of the planes transitions from an off state to on state. The power gating noise at the neighboring planes is observed. Contrary to power supply noise, a larger sleep transistor produces greater power gating noise due to higher in-rush current. Furthermore, distributed power gating topology exhibits slightly more power gating noise than the lumped topology for both via-middle and via-last TSVs. When a plane is turned on, higher current is provided by the neighboring planes in a distributed topology since the impedance between the planes is relatively smaller (see Fig. 3.4). Thus, the neighboring planes suffer slightly more from power gating noise in the distributed topology due to the higher current.

3.4.3 Turn-on Time

Turn-on time is an important design objective while implementing a reliable power gating topology. The ratio of the turn-on time of the sleep transistors to idle period partially determines whether the block or plane should be power gated. Thus, a shorter turn-on time achieves higher reduction in the leakage power. The turn-on time for the proposed power gating topologies is shown in Fig. 3.8 for both via-middle and via-last TSVs. Note that turn-on time is analyzed from the start of the transition time until the virtual supply voltage reaches 90% of the nominal V_{DD}. As shown in Fig. 3.8, turn-on time is shorter for via-middle TSVs in lumped power gating topology and for via-last TSVs in distributed power gating topology. This trend is due to the reduced power supply noise in these cases, as illustrated in Fig. 3.6. A reduced power noise enables greater supply current for the gated plane to turn on, thereby reducing turn-on time.

3.4.4 Leakage Power Reduction

The leakage power reduction achieved by the two power gating topologies is shown in Fig. 3.9. The analysis is conducted for both lumped and distributed topolo-

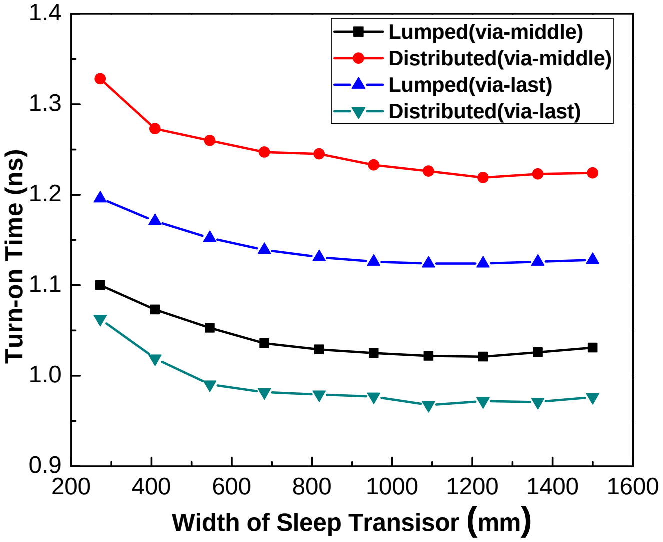


Figure 3.8: Turn-on time of different power gating topologies with varying sleep transistor size.

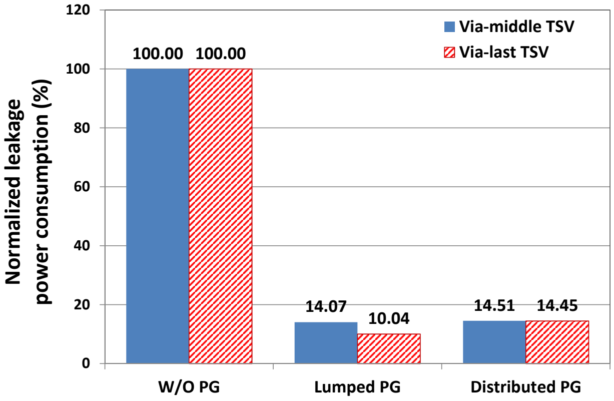


Figure 3.9: Normalized leakage power consumption achieved the proposed power gating topologies.

gies. The leakage power is normalized based on the leakage power consumption of the 3D IC without power gating. The sleep transistor width and number of TSVs are identical in both topologies while guaranteeing that the power supply noise is within 5% of the nominal V_{DD}. As illustrated in Fig. 3.9, for via-middle TSVs, the lumped and distributed power gating topologies can save, respectively, 85.93% and 85.49% leakage power. Alternatively, for via-last TSVs, the leakage power is reduced, respectively, by 89.96% and 85.55%, demonstrating the efficacy of the proposed power gating topologies.

3.5 Summary

In this chapter, two topologies are proposed at the physical level to achieve reliable power gating in 3D ICs. It is demonstrated that a lumped topology is more advantageous for via-middle TSVs, whereas a distributed topology produces enhanced results for via-last TSVs. The proposed topologies are validated with HSPICE simulations. Up to 22.9% reduction in power supply noise is demonstrated by using the proposed method, while reducing the leakage power, on average, by 86%.

Chapter 4

Resource Allocation Methodology for TSVs and Sleep Transistors in 3D ICs with Power Gating

In 3D ICs with power gating, the sleep transistors (essential component of power gating) have significant impact on power supply noise, and also introduce area contention with the through silicon vias (TSVs), both of which consume considerable substrate area. Therefore, in this chapter, a co-optimization methodology is proposed that simultaneously considers the sizing of sleep transistors and TSVs to achieve optimal power integrity with minimal area overhead. The design tradeoffs between sleep transistors and TSVs are discussed in Section 4.1. The proposed co-optimization methodology is presented in Section 4.2. A case study of the power distribution network of a three-plane 3D IC is developed in Section 4.3 to validate the efficacy of the proposed methodology.

4.1 Design Tradeoffs between TSVs and Sleep Transistors

In TSV-based 3D ICs, the data communication, power and clock distribution among the dies is achieved by high density TSVs [39]. The number of TSVs in high performance ICs exceeds 50,000 while the diameter of a single TSV is in the micrometer range [31]. Furthermore, a keep-out zone exists surrounding each TSV where no active devices are placed to prevent the TSV-induced mechanical stress from affecting the reliable operation of these devices [72]. Thus, TSVs consume considerable silicon substrate area. Alternatively, the parasitic impedance of power TSVs (TSVs delivering power supply voltage) causes power supply noise as the switching load current flows through the TSVs. To reduce the effective impedance of TSVs, multiple TSVs are placed in parallel, which further increases the area overhead.

When power gating is implemented in 3D ICs, sleep transistors are inserted to individually control the power delivery to each subcircuit. In practice, in coarse-grain power gating, a large number of sleep transistors is placed around the intended area to be power gated, forming a ring structure, as illustrated in Fig. 4.1 [46]. The effective resistance of sleep transistors determines the voltage drop between the global power rail and virtual power rail when the sleep transistors are turned on, as depicted in Fig. 3.1. The total effective width of sleep transistors in the ring can exceed one meter to ensure a sufficiently low impedance and capability to carry the required supply current [46]. Therefore, the sleep transistors and TSVs both consume significant physical area and both have significant effect on the power integrity. Thus, co-optimization of sleep transistors and TSVs plays an important

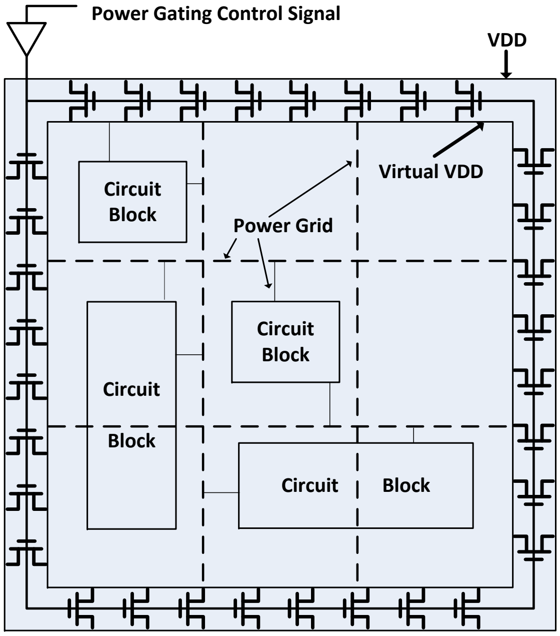


Figure 4.1: Ring style sleep transistor placement for coarse-grain power gating [46].

role in the design process of power distribution networks to improve power integrity while maintaining minimum physical area overhead.

4.2 Proposed Methodology

4.2.1 Summary of the Proposed Flow

The proposed resource allocation methodology is summarized in Fig. 4.2. The first step is to determine the available physical area A for power/ground TSVs and sleep transistors. Assume that the ratio of this area allocated to sleep transistors is k [*i.e.* sleep transistors occupy an area of $k \times A$ whereas TSVs occupy an area of $(k-1) \times A$]. Next, the minimum and maximum sleep transistor sizes are deter-

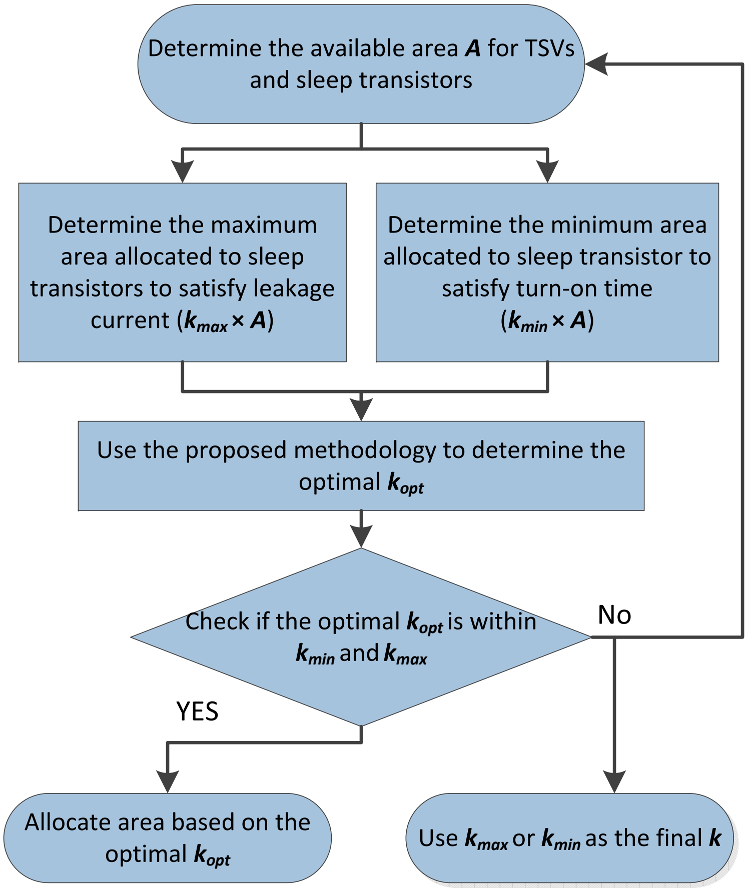

Determine the available area $\boldsymbol{A}$ for TSVs and sleep transistors
Determine the maximum area allocated to sleep transistors to satisfy leakage current ($\boldsymbol{k_{max}} \times \boldsymbol{A}$)
Determine the minimum area allocated to sleep transistor to satisfy turn-on time ($\boldsymbol{k_{min}} \times \boldsymbol{A}$)
Use the proposed methodology to determine the optimal $\boldsymbol{k_{opt}}$
Check if the optimal $\boldsymbol{k_{opt}}$ is within $\boldsymbol{k_{min}}$ and $\boldsymbol{k_{max}}$
No
YES
Allocate area based on the optimal $\boldsymbol{k_{opt}}$
Use $\boldsymbol{k_{max}}$ or $\boldsymbol{k_{min}}$ as the final $\boldsymbol{k}$

Figure 4.2: Summary of the proposed resource allocation methodology.

mined to satisfy the constraints on, respectively, turn-on time and leakage current. If less area is allocated to sleep transistors than the minimum required ($k_{min} \times A$), turn-on time may not be satisfied due to insufficient drive current. Alternatively, if more area is allocated to sleep transistors than the maximum permitted ($k_{max} \times A$), the constraint on leakage current may not be satisfied since subthreshold leakage current is proportional to device size. This tradeoff has been well characterized in the literature [59] [73].

In the next step, the proposed analytic expressions are used to determine the optimum resource allocation to minimize power supply noise. If this optimum allocation is within the permitted range determined in the previous steps, (*i.e.*, $k_{min} \leq k_{opt} \leq k_{max}$), then the number of TSVs and sleep transistor size are finalized, enhancing the system-wide power integrity of the 3D IC, while satisfying turn-on time and leakage current. If the optimum value is outside the range, the minimum power supply noise cannot be obtained under the area constraint A determined in step 1. In this case, the available area may be adjusted to achieve minimum power supply noise, or either k_{min} or k_{max} can be used to allocate the area.

4.2.2 Proposed Analytic Expressions to Determine k_{opt}

Assuming a TSV diameter of d, the area of a single TSV is $\pi d^2/4$. The substrate area occupied by a TSV is increased by α due to keep-out zone. Thus, the number of TSVs that can be reliably fabricated within an area of $(1-k) \times A$ is

$$N_{\text{TSV}} = \frac{(1-k)A}{\alpha(\pi d^2/4)}. \tag{4.1}$$

Similarly, the physical area $k \times A$ consumed by the sleep transistors (each having a channel width W and length L) is estimated as

$$kA = \beta \sum_{1}^{N_{st}} (W \times L), \tag{4.2}$$

where N_{st} is the overall number of sleep transistors and the parameter β considers the area overhead of the transistors due to drain/source contacts and required spacing between adjacent transistors.

The effective resistance of N TSVs in a bundle is approximated as

$$R_{\mathrm{TSV}}^{\mathrm{eff}} = \frac{4\rho h_{\mathrm{TSV}}}{\pi d^2 N_{\mathrm{TSV}}} = \frac{\rho h_{\mathrm{TSV}} \alpha}{(1-k)A}, \tag{4.3}$$

where h_{TSV} is the height of a single TSV and ρ is the resistivity of the TSV filling material.

When turned on, the sleep transistors operate in the linear region due to small voltage drop across the source and drain terminals. The channel resistance in this region can be estimated as

$$R_{st} \approx \frac{L}{\mu C_{ox} W (V_{gs}^{st} - V_{th})}, \tag{4.4}$$

where μ is the electron (or hole) mobility, V_{gs}^{st} is the control signal applied to the gate of the sleep transistor, and V_{th} is the threshold voltage. Combining (4.2) and (4.4), the effective resistance of the sleep transistor is

$$R_{\mathrm{st}}^{\mathrm{eff}} \approx \frac{L^2 \beta}{\mu C_{ox} (V_{gs}^{st} - V_{th})} \times \frac{1}{kA}. \tag{4.5}$$

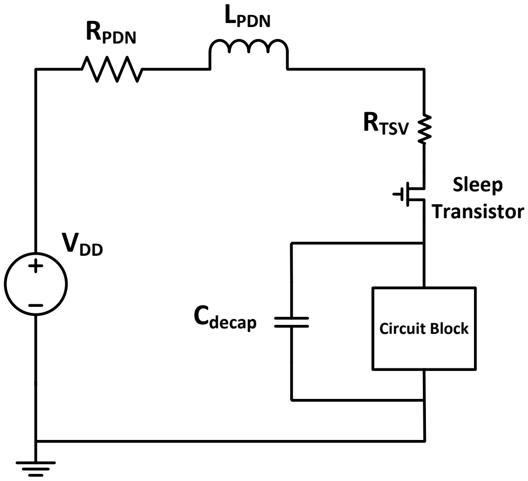


Figure 4.3: Simplified model of a power delivery path illustrating sleep transistors and TSVs.

The simplified model shown in Fig. 4.3 is used to gain an intuitive understanding of the proposed methodology. This model describes the power delivery path for a single plane in a 3D stack with distributed power gating topology. The resistance of the TSVs delivering power to this plane is modeled by R_{TSV}. The sleep transistor is placed between the TSV and circuit blocks in the power delivery path. All of the other impedances including the package level parasitics and impedances of the neighboring planes are represented by R_{PDN} and L_{PDN}. The decoupling capacitance of the plane is modeled by C_{decap}.

When evaluating a power distribution network, considering only the static IR drop is not sufficient due to resonance. The method described in [73] is adopted in this work, which simultaneously considers the static IR drop and resonant supply noise. According to the model in Fig. 4.3, the impedance of the power supply network is

$$Z(\omega) \quad = \quad (R_{\mathrm{PDN}} + j\omega L_{\mathrm{PDN}} + R_{TSV} + R_{ST}) \| \frac{1}{j\omega C_{decap}}. \tag{4.6}$$

Thus, the impedance at DC can be expressed as

$$Z_{DC} = R_{\mathrm{PDN}} + R_{TSV} + R_{ST}, \tag{4.7}$$

whereas the magnitude of the impedance at the resonant frequency can be approximated as

$$Z_{res} \approx \frac{L_{\mathrm{PDN}}}{(R_{\mathrm{PDN}} + R_{TSV} + R_{ST})C_{decap}}. \tag{4.8}$$

According to [73], the worst case power supply noise is determined by the sum of DC noise and resonant noise,

$$\begin{aligned} V_{noise}^{worstcase} &= V_{noise(DC)} + V_{noise(res)} \\ &= Z_{DC} \cdot I_{dc} + Z_{res} \cdot I_{res}, \end{aligned} \tag{4.9}$$

where I_{dc} and I_{res} are, respectively, the current flow at DC and resonant frequency. Assume that the ratio between I_{dc} and I_{res} is μ/ν, where $\mu + \nu = 1$. Therefore, the worst case impedance of the power network is

$$Z_{worst} = \frac{V_{noise}}{I_{dc} + I_{res}} = \mu Z_{DC} + \nu Z_{res}. \tag{4.10}$$

Z_{worst} is used as a metric to evaluate the power distribution network where static

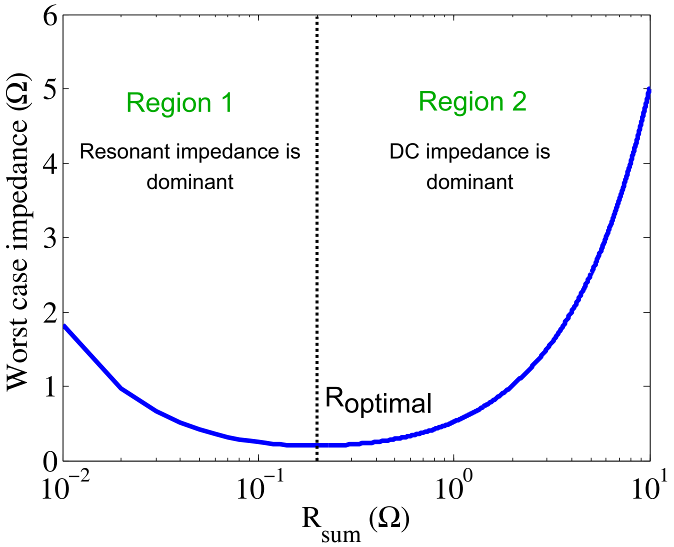


Figure 4.4: The magnitude of worst case impedance Z_{worst} varying with R_{sum} values.

IR drop and resonant noise are both considered. Using (4.7) and (4.8) in (4.10),

$$Z_{worst} \approx \mu(R_{\mathrm{PDN}} + R_{TSV} + R_{ST}) + \frac{\nu L_{\mathrm{PDN}}}{(R_{\mathrm{PDN}} + R_{TSV} + R_{ST})C_{decap}}. \quad (4.11)$$

$R_{\mathrm{ST}} + R_{TSV}$ is defined as R_{sum} and Z_{worst} is plotted as a function of R_{sum} in Fig. 4.4. As demonstrated in this figure, if R_{sum} is relatively small (region 1), Z_{worst} decreases as R_{sum} increases since the resonant impedance is more dominant than the DC impedance in this region. A larger R_{sum} reduces the resonant impedance (due to greater damping), which decreases Z_{worst}. As R_{sum} further increases, Z_{worst} reaches a minimum and starts increasing (in region 2). The increase in region 2 is due to the DC impedance, which is the dominant term in this region. Z_{worst} is minimized when the two terms on the right-hand-side of (4.11) are equal. This condition is satisfied when

$$R_{sum} = R_{optimal} = \sqrt{\frac{\nu}{\mu} \cdot \frac{L_{\mathrm{PDN}}}{C_{decap}}} - R_{ST}. \tag{4.12}$$

Thus, k should be chosen such that R_{sum} is equal to the $R_{optimal}$. Using (4.3) and (4.5), R_{sum} is be expressed in terms of k as

$$R_{sum} = \frac{T_1}{kA} + \frac{T_2}{(1-k)A}, \tag{4.13}$$

where

$$T_1 = \frac{L^2\beta}{\mu C_{ox}(V_{gs}^{st} - V_{th})}, \quad T_2 = \rho h_{\mathrm{TSV}}\alpha.$$

Replacing (4.12) in (4.13),

$$\frac{T_1}{kA} + \frac{T_2}{(1-k)A} = \sqrt{\frac{\nu}{\mu} \cdot \frac{L_{\mathrm{PDN}}}{C_{decap}}} - R_{ST}. \tag{4.14}$$

The two real roots of (4.14), given by (4.15), achieve $R_{optimal}$, thereby minimizing the worst case impedance of the power network, Z_{worst}.

$$k_{optimal1,2} = \frac{(R_{optimal}A + T_1 - T_2) \pm \sqrt{(R_{optimal}A + T_1 - T_2)^2 - 4R_{optimal} \cdot AT_1}}{2R_{optimal} \cdot A} \tag{4.15}$$

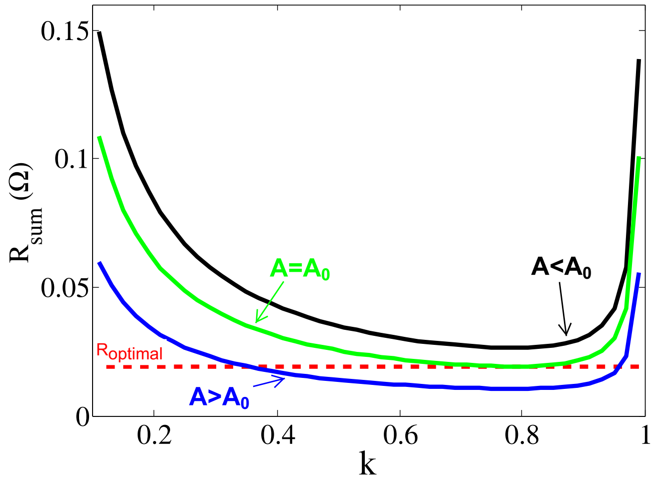


Figure 4.5: The R_{sum} under different k values with three area constraint A = 11000 μm^2, A = 8000 μm^2 and A = 20000 μm^2.

4.2.3 Area Dependence

It is important to note that $R_{optimal}$, determined by (4.12), is independent of the physical area available to sleep transistors and TSVs. Thus, if the overall area allocated to sleep transistors and TSVs is smaller than a certain value, $R_{optimal}$ cannot be achieved by any k value. To illustrate this phenomenon, R_{sum} is plotted in Fig. 4.5 as a function of k under three different area constraints. As shown in this figure, if the available area is smaller than the critical area, indicated as A_0, the R_{sum} curve does not reach $R_{optimal}$ for any value of k. Alternatively, if the area is larger or equal to A_0, then $R_{optimal}$ is achieved at a particular k, as analytically determined by (4.15). Note that for the R_{sum} curve when $A > A_0$, there are two k values where $R_{optimum}$ is achieved. According to the proposed flow, as shown in Fig. 4.2, the optimum k should be between k_{min} and k_{max} to satisfy the constraints on turn-on time and leakage current. Thus, the selected k should be within this range. Note that a larger k favors turn-on time due to wider sleep transistors whereas a smaller

k favors leakage current due to smaller sleep transistors.

The critical area A_0 can also be analytically expressed. According to Fig. 4.5, the minimum value of R_{sum} is equal to $R_{optimal}$ if $A = A_0$. The k value that achieves minimum R_{sum} is

$$\frac{dR_{sum}}{dk} = 0 \rightarrow k_{min} = \frac{1}{\sqrt{\frac{T_2}{T_1}} + 1}. \tag{4.16}$$

If this k_{min} is used, the minimum R_{sum} is obtained as,

$$R_{sum}^{min} = \frac{(\sqrt{T_1} + \sqrt{T_2})^2}{A}. \tag{4.17}$$

Since $R_{sum}^{min} = R_{optimal}$ at $A = A_0$, A_0 can be determined by equating (4.12) with (4.17), and solving for A,

$$A_0 = \frac{(\sqrt{T_1} + \sqrt{T_2})^2}{(\sqrt{\frac{\nu}{\mu} \cdot \frac{L_{\mathrm{PDN}}}{C_{decap}}} - R_{ST})}. \tag{4.18}$$

Thus, if the overall area allocated to TSVs and sleep transistors is equal or greater than (4.18), the worst case impedance, as determined by (4.11), can be minimized.

4.3 Simulation Results

To evaluate the proposed methodology and analytic expressions, a comprehensive case study is developed. The analysis setup is described in Section 4.3.1. Simulation results are presented in Section 4.3.2 to verify the proposed methodology.

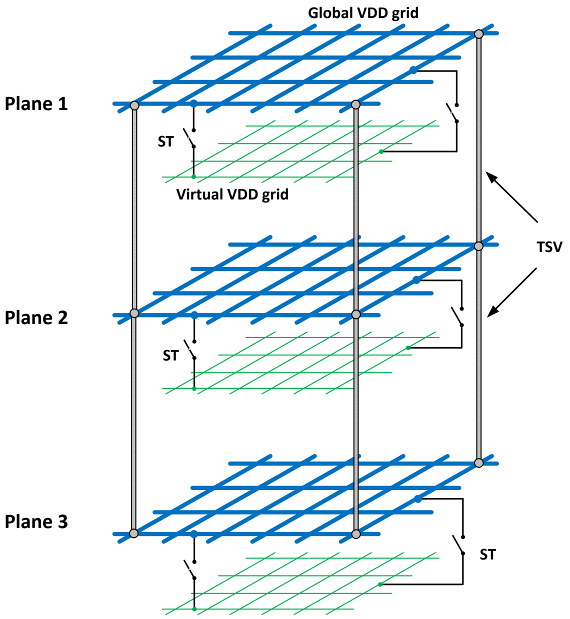


Figure 4.6: Power distribution network of a three-plane 3D IC with via-last TSVs and power gating illustrating the global and virtual power grids, sleep transistors (ST), and TSVs.

4.3.1 Simulation Setup

A power distribution network for a three-plane 3D IC with via-last TSVs is developed, as conceptually illustrated in Fig. 4.6. A 45 nm CMOS technology with 10 available metal layers in each plane is adopted [66]. A portion of the power network with an area of 1 mm by 1 mm is analyzed. Each plane consists of a global power network, virtual power network, distributed PMOS sleep transistors, distributed decoupling capacitance (implemented as MOS capacitors), and distributed switching load circuit consisting of inverter gates, as depicted in Fig. 4.7. Note that the top plane also consists of C4 bumps to connect the on-chip grid with the flip-chip substrate.

The top two metal layers (9 and 10) on each plane are dedicated to global power

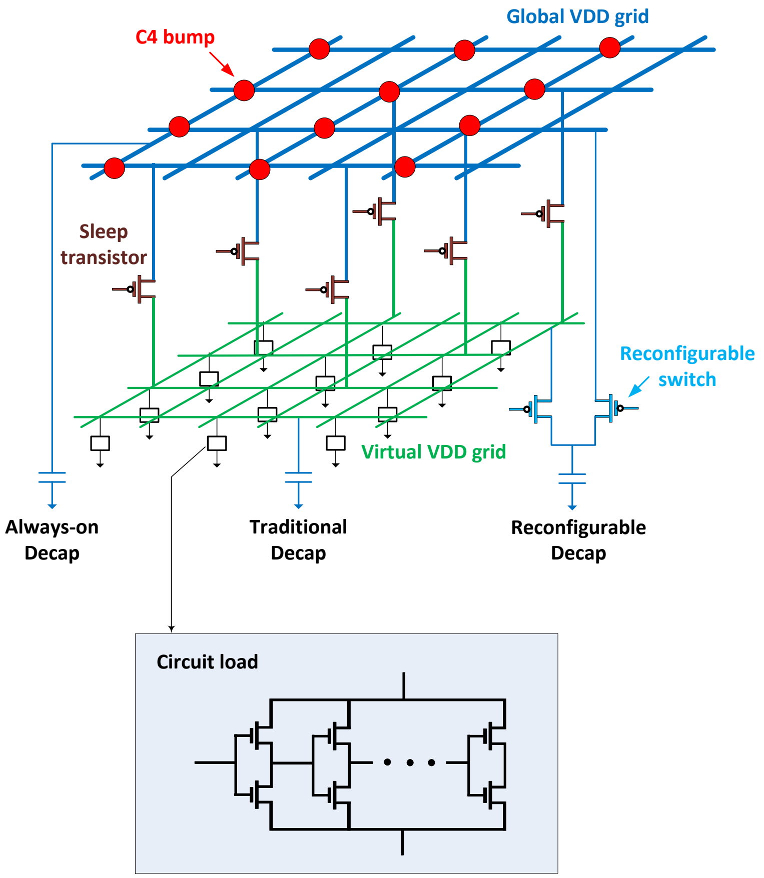


Figure 4.7: Plane-level power network illustrating distributed sleep transistors, decoupling capacitors (traditional and proposed topologies), switching load circuits (gates with active devices), and the C4 bumps (for the top plane only).

Parameters		Values
Metal 10 & 9	Pitch	45 μm
	Width	40 μm
	Resistivity (ohm/sq)	0.03
Metal 8 & 7	Pitch	23.5 μm
	Width	20 μm
	Resistivity (ohm/sq)	0.075

Table 4.1: Primary physical characteristics of the global and virtual power grids [66].

	$R_{\textbf{unit}}$ (mΩ)	$C_{\textbf{unit}}$ (fF)	$L_{\textbf{unit}}$ (pH)
Global grid	67.50	27.00	37.70
Virtual grid	176.25	7.05	20.13

Table 4.2: Parasitic impedances of the unit interconnect segment within the global and virtual power grids.

distribution network with an interdigitated grid of 11×11 metal lines [69]. Metal layers 8 and 7 are used as the virtual power grid that is connected to the global grid through sleep transistors. Virtual V_{DD} network is also represented by an interdigitated grid of 21×21. Power gating is achieved using a distributed method where the sleep transistors that control a plane are placed within that plane [23].

Primary physical characteristics of the 3D power grid are listed in Table 4.1. The pitch and width of the metal lines are determined based upon the technology design rules [66] while also considering routing constraints. For each interconnect segment, an *RLC* π circuit is used to model the parasitic impedances of the power grid. The unit parasitic capacitance (extracted from FastCap [74]), inductance (extracted from FastHenry [70]) and resistance (based on sheet resistance [66]) values are listed in Table 4.2.

Physical characteristics of the via-last TSVs (with copper as the filling material),

Parameters	**Values**
Nominal power supply voltage V_{DD}	1.0 V
Lumped package resistance R_{package}	1 mΩ
Lumped package inductance L_{package}	120 pH
Single C4 bump resistance R_{C4}	5 mΩ
Single C4 bump inductance L_{C4}	200 pH
Via-last TSV diameter W	10 μm
Via-last TSV height H	60 μm
Via-last TSV dielectric thickness t_{ox}	0.2 μm
Resistivity of TSV filling material (copper) ρ_f	16.8 nΩ·m
Single via-last TSV resistance R_{tsv}	20mΩ
Single via-last TSV capacitance C_{tsv}	283 fF
Single via-last TSV inductance L_{tsv}	35 pH

Table 4.3: Package, TSV, and C4 bump parasitic impedances and physical characteristics [68, 75]

C4 bumps, and the package impedances are listed in Table 4.3. A flip-chip package is assumed and modeled with a lumped resistance of 1 mΩ and inductance of 120 pH [68]. C4 bumps are regularly placed with a pitch of 200 μm over the 1 mm × 1 mm area [75]. Each C4 bump has a resistance of 5 mΩ and inductance of 200 pH [75]. Clustered via-last TSVs are distributed throughout the area as a 5×5 array and connect the global power grid on each plane. Each TSV cluster consists of four TSVs. Thus, 100 power TSVs are used to connect two planes. The overall number of power TSVs in the three-plane stack is 200.

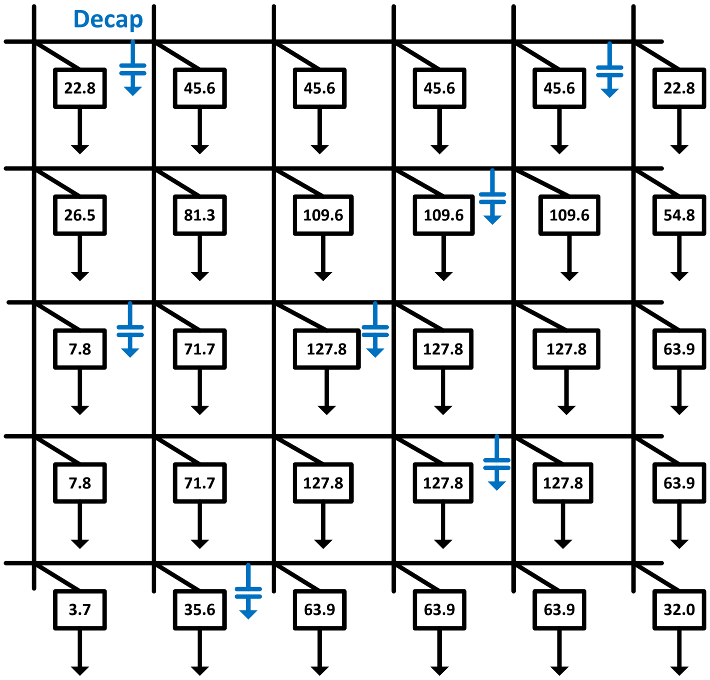
4.3.1.1 Switching Load Circuit

As opposed to using piecewise linear (PWL) current sources to model the switching load circuit (typical practice in existing work [76,77]), gates with active devices are used since power gating is considered. Note that power gating noise cannot be accurately analyzed when PWL sources are used since the transient turn-on characteristics of the active devices play an important role in power gating noise. Furthermore, savings in the subthreshold leakage current cannot be estimated with PWL sources. Finally, an active load enables to consider the negative feedback between supply noise and load current (larger supply noise reduces load current, which in turn reduces supply noise), enhancing the accuracy of the analysis.

Similar to [78], inverter pairs with varying number and size are used to model the switching load circuit. The overall area is divided into 30 sub-areas and a switching circuit is connected to each sub-area to consider the spatial heterogeneity of the current loads. The spatial load current distribution and power densities are based on [48]. As an example, the current distribution of the top plane is illustrated in Fig. 4.8 where the peak current for each block is indicated. For the middle and bottom planes, the same switching circuits are used, but these circuits are placed at different locations throughout the power network. Note that according to the current profiles of the inverter pairs, the peak power density reaches 40 W/cm^2, which is comparable to the power density in modern processors [79]. Decoupling capacitors are also conceptually shown in Fig. 4.8. Decoupling capacitors are inserted at multiple locations depending upon the power noise distribution.

4.3.2 Verification of the Proposed Methodology

According to Fig. 4.2, the first step is to determine the minimum and maximum values of k to ensure that k_{opt} is within this range. The simulation setup described in the previous section is analyzed using HSPICE. k_{min} is determined as 0.06 to guarantee that leakage current is reduced by three orders of magnitude, similar to [59]. Note that higher reduction can be achieved by increasing k_{min}. Similarly, k_{max} is determined as 0.97 to guarantee that the worst case turn-on time for a plane



(Unit of peak current: mA)

Figure 4.8: Current distribution within the top plane [48]. The numbers refer to the peak current drawn by the digital gates at each node. The peak power density reaches 40 W/cm^2, which is comparable to the power density in modern processors [79].

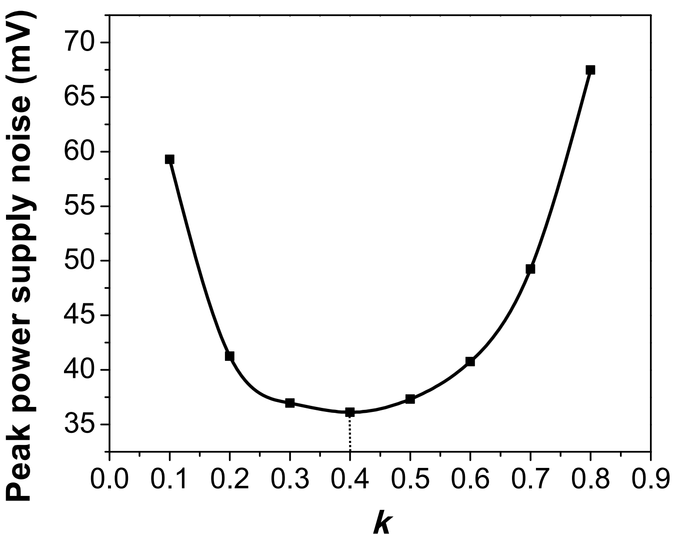


Figure 4.9: Peak transient power supply noise of a circuit load in the bottommost plane under different values of k.

does not exceed 400 ps. A shorter turn-on time can be achieved by decreasing k_{max}.

In the next step, the simulation setup is analyzed to experimentally determine k_{opt}. The worst case power supply noise on the bottom plane is shown in Fig. 4.9 as a function of k. According to this figure, peak power supply noise is minimized (36 mV) if k is approximately 0.6, *i.e.*, 60% of the overall area is allocated to sleep transistors and the remaining area is used for TSVs. Note that a nonoptimal k can significantly increase the power supply noise. For example, at $k = 0.2$, power supply noise is 67 mV, beyond the 5% constraint.

In the last step, the proposed analytic expressions are utilized to determine k_{opt}. AC simulations are performed to extract the effective impedances required by the proposed model. The calculated worst case impedance is shown in Fig. 4.10 as a function of k. As shown in this figure, according to the proposed expressions, k_{opt} is approximately equal to 0.64. The estimated value is sufficiently close to the

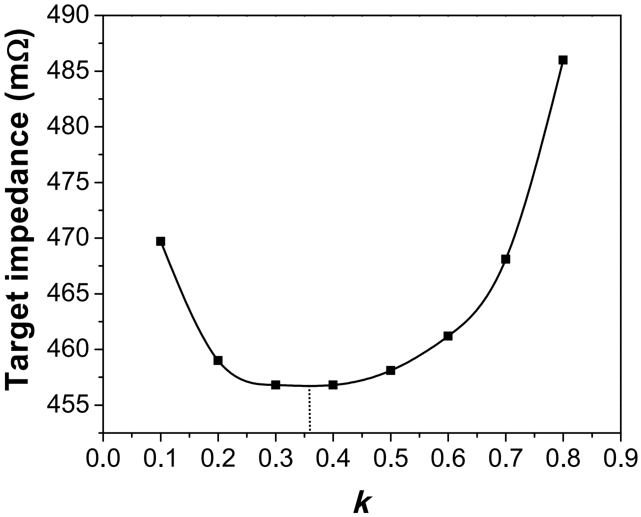


Figure 4.10: Worst case impedance for a circuit load in the bottommost plane predicted from the proposed analytical model.

simulations where the error is 4%. The error is due to the approximations made to extract the effective impedances from the highly distributed simulation setup.

4.4 Summary

In this chapter, a methodology and analytic expressions have been proposed to appropriately allocate available physical area between sleep transistors and TSVs. The methodology minimizes power supply noise while simultaneously satisfying leakage current and turn-on time. The proposed expressions have been verified through a comprehensive simulation setup where the error is less than 4%. The proposed resource allocation achieves more than 46% reduction in supply noise.

Chapter 5

Closed-Form Expressions to Estimate Power Supply Noise with Fast Signal Transitions

Since power supply noise has been a primary concern in the design of high performance integrated circuits with low supply voltages and high current demands, accurate estimation and characterization of power supply noise has become essential. In Chapters 3 and 4, simulation based methods are utilized to characterize the power supply noise. Alternatively, closed-form expressions have also attracted considerable attention to analytically determine power supply noise as a function of multiple circuit parameters [80–88]. Unlike simulation based methods, analytical models can provide intuition on the effect of various circuit characteristics such as parasitic impedances of the power delivery path and switching load circuit. However, as demonstrated later in this chapter, all of the previous closed-form expressions [80–88] share a common limitation/assumption which makes these expressions significantly inaccurate when the signal transitions are fast.

A new modeling methodology is developed in this chapter to fix this limitation and enhance the accuracy of previous methods for nanoscale technology nodes with signal transition times in the range of several tens of picoseconds. A brief summary of the previous work focusing on closed-form expressions to estimate power supply noise is provided in Section 5.1. The relationship between damping characteristics and peak power supply noise is investigated in Section 5.3 with emphasis on input transition time. The proposed modeling methodology and closed-form expressions are provided in Section 5.4. Simulation results are presented in Section 5.5 to evaluate the proposed model and compare the model with existing closed-form expressions.

5.1 Previous Work in Power Supply Noise Modeling

In existing closed-form expressions, the power supply noise is typically modeled using the schematic shown in Fig. 5.1 [80–86]. In this schematic, inductance L, resistance R and capacitance C represent the parasitic impedances of the power/ground distribution network. The switching circuit is represented by CMOS gates. The input is a saturated ramp signal that remains at logic low until $t = t_0$, and linearly increases after $t > t_0$, reaching logic high at $t = t_r$. The time interval between t_0 and t_r is referred to as the *transition period*, whereas the time interval after t_r is referred to as the *post-transition period*, as depicted in Fig. 5.1. Note that a saturated ramp signal is sufficiently accurate in representing practical input signals in CMOS gates provided that the slew rate of the saturated ramp is the same

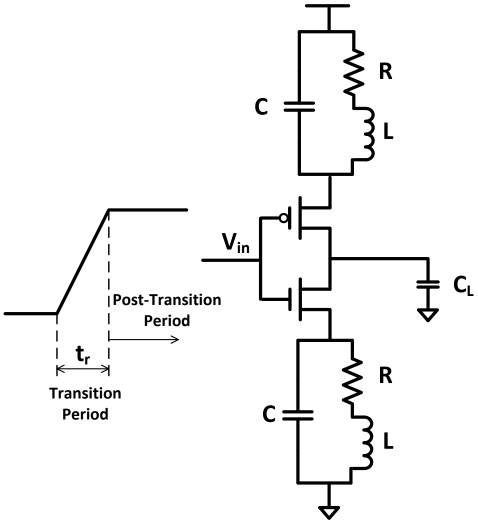


Figure 5.1: Equivalent circuit typically used in existing work to develop closed-form expressions for power supply noise.

as the slew rate of the practical signal. Also note that slew rate should be calculated while considering the threshold voltage. Three considerations play an important role in the accuracy of closed-form expressions: 1) whether all of the *R*, *L*, and *C* components of the power/ground network are considered in the model, 2) $I-V$ relationship of the switching transistors, and 3) the assumptions used to obtain a closed-form expression.

In [80] [81], Senthinathan *et al.* have investigated the negative feedback effect between ground noise and switching current (*i.e.* higher ground noise reduces switching current due to a smaller gate-to-source voltage), and provided closed-form expressions for power supply noise. In this model, it has been assumed that the switching current during the transition period can be represented by a triangular wave, which is not sufficiently accurate in estimating power supply noise. Furthermore, capacitance *C* within the power network has been ignored assuming that the current through the capacitor is sufficiently small. In addition, the parasitic resistance R has also been neglected. Thus, only the parasitic inductance L has been considered. For the switching gate, Shockley's $I-V$ equation has been utilized [89]. Even though it has been indicated that the ground noise can be an underdamped signal, this characteristic has not been modeled.

In [82], Vaidyanath *et al.* have used a similar modeling approach as [80, 81] where only the parasitic inductance L has been considered. It has been assumed that the power supply noise linearly increases during the transition period. This assumption can cause considerable inaccuracy, as shown in this paper.

In [83], Vemuru *et al.* have adopted Sakurai's α-power law model [90] and assumed that the first-order derivative of the transistor current is constant in submicron technologies. In [84], Tang and Friedman have used a polynomial expansion approximation to obtain a closed-form expression for the simultaneous switching noise, also utilizing the α-power law model. All of the parasitic impedances of the power network have been considered.

A primary limitation in all of these works, *i.e.*, [80] to [84], is the assumption that *the peak noise occurs at the end of input signal transition*, *i.e.*, at $t = t_r$ in Fig. 5.1. As demonstrated in this paper, this assumption does not hold if the input signal transition time is sufficiently fast, as typically encountered in modern technologies.

In relatively more recent work, rather than using α-power law model (which can predict transistor current in all of the operation regions), an application specific device modeling (ASDM) methodology has been used to obtain a simple, but efficient model for the saturation region, which is the region of interest for power supply noise analysis [85] [86]. Linear mathematical formulation provided by the ASDM method facilitates the analysis of damping characteristics with sufficient accuracy.

In [85], Ding *et al.* have discussed the effect of parasitic capacitance on the damping characteristics of an inductive power network. It has been indicated that under a certain range of parasitic capacitance, the system is underdamped during the transition period. The peak noise, therefore, *does not* occur at the end of transition, but *at the first local maximum within the transition period.* In [86], Hekmat *et al.* have improved the model in [85] by also considering the parasitic resistance within the power network.

More recently, closed-form expressions have been developed to estimate power supply noise at the package and board levels [87, 88]. These works, however, have not considered the nonlinear on-chip switching circuit characteristics. Instead, an ideal linear current source has been used.

5.2 Contributions of This Work

None of the previous works mentioned above have investigated the damping characteristics after the transition is over, *i.e.*, within the post-transition period in Fig. 5.1. When the transition time of input signal is short, the power supply noise can continue increasing after the transition is complete. Thus, analysis of the power supply noise only within the transition period can significantly underestimate the peak noise under specific damping conditions, as described later in Section 5.3.

Therefore, this work is intended to provide closed-form expressions for power supply noise when the input transition times are sufficiently short, as typically encountered in nanoscale technologies. The effect of signal transition time on damping behavior is investigated while considering resistive, capacitive, and inductive characteristics of a power network. The validity of a widely acknowledged assumption in previous work (*i.e.*, power noise reaches the peak value at the end of the transition period) is evaluated and it is demonstrated that this assumption is valid only under specific damping conditions when the transition time is sufficiently long. Closed-form expressions are developed to more accurately model the damping behavior with fast input signals. The proposed expressions can accurately predict not only the peak noise, but also the entire voltage noise waveform, as validated by SPICE simulations. The accuracy of existing closed-form expressions is significantly enhanced.

Note that the proposed closed-form expressions are not intended for chip-level analysis of power supply noise. Instead, these expressions provide intuition on the effect of fast signal transitions on peak noise, enabling more accurate design methodologies and optimization frameworks [88, 91]. Furthermore, it is important to note that the proposed expressions can be accurately utilized to estimate simultaneous switching noise caused by I/O drivers where the lumped model is relatively more applicable.

5.3 Motivational Example on Peak Supply Noise and Damping Behavior

An example is provided in this section to illustrate the effect of input transition time on peak noise characteristics. The power supply noise generated by an I/O driver consisting of a CMOS inverter in 45 nm technology and driving a capacitive load of 10 pF is analyzed, as shown in Fig. 5.1. In case 1, the schematic is simu-

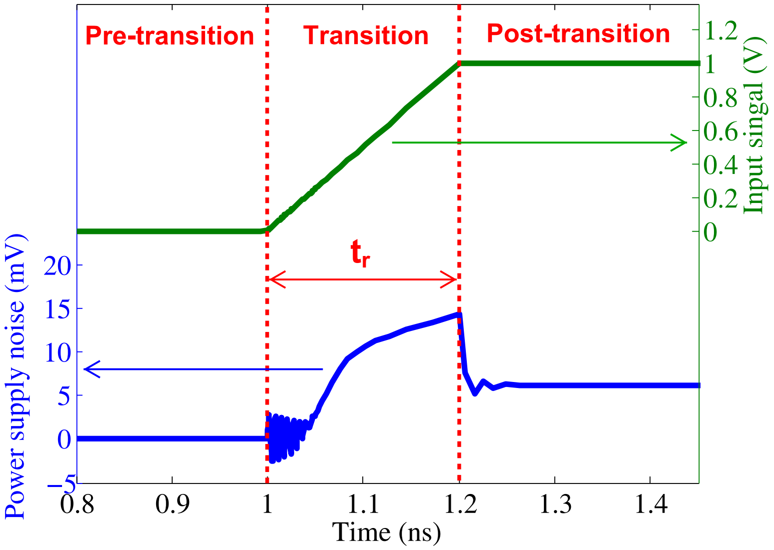


Figure 5.2: Transient power supply noise waveform when parasitic capacitance is ignored: R=5 Ω, L=1 nH, W_n=1 μm and W_p=2 μm, C_L = 10 pF, transition time t_r is 200 ps.

lated when input transition time is 200 ps while ignoring the parasitic capacitance C. The transient noise waveform is depicted in Fig. 5.2. Since C is not considered, the circuit represents a first-order system (ignoring the transistor parasitic capacitances) and the power noise decreases after t_r. Peak supply noise occurs at $t = t_r$, as assumed in existing work.

In case 2, the same circuit is simulated using a parasitic capacitance C of 1 pF, also with 200 ps input transition time. The transient power noise waveform is shown in Fig. 5.3. Since the circuit represents a second-order system, the damping behavior depends upon the R, L, C impedances and transistor characteristics. As shown in Fig. 5.3, in case 2, the system is underdamped. Peak noise does not occur at the end of transition. Instead, the peak noise occurs at the first local maximum of the decaying sine wave, as also observed in [85]. In this case, modeling the circuit damping behavior within the transition period is sufficient because the power noise

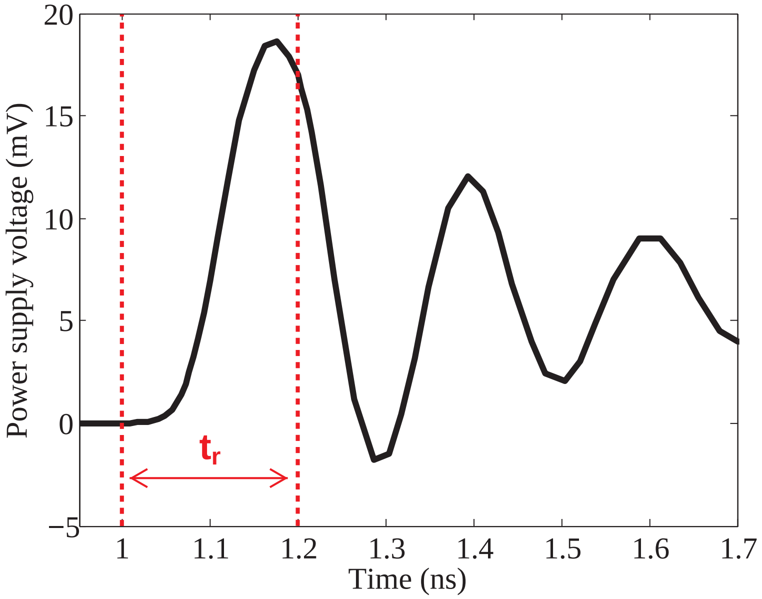


Figure 5.3: Transient power supply noise waveform when parasitic capacitance is considered: R=5 Ω, L=1 nH, C=1 pF, W$_n$=1 μm and W$_p$=2 μm, C_L = 10 pF, transition time t_r is 200 ps.

after $t = t_r$ is always smaller than the first local maximum before $t = t_r$, as depicted in Fig. 5.3.

In case 3, the same circuit is simulated, but the transition time t_r is reduced from 200 ps to 20 ps. The transient noise waveform is illustrated in Fig. 5.4. As shown in this figure, when the transition time is shorter, the power supply noise does not reach the first local maximum at or before $t = t_r$. Alternatively, the noise continues to increase and reaches the peak value *within the post-transition period*. Therefore, modeling the system only until $t = t_r$ significantly underestimates the peak noise (by 73.8% in this example). It is therefore highly important to consider the circuit damping and noise characteristics after $t = t_r$.

This characteristic can be intuitively described by considering the extreme case when the gate input is a step signal with zero transition time. In this extreme case,

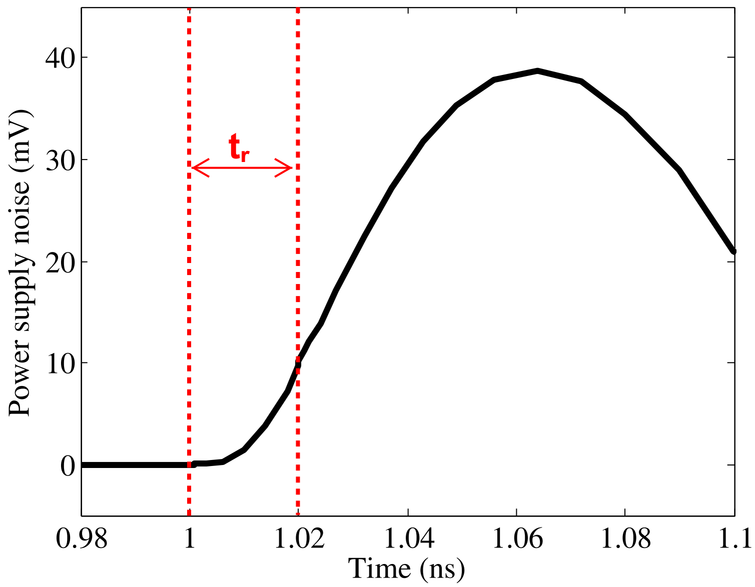


Figure 5.4: Transient power supply noise waveform when the transition time t_r of the input signal is reduced to 20 ps: R=5 Ω, L=1 nH, C=1 pF, W_n=1 μm and W_p=2 μm, C_L = 10 pF.

the peak noise would always occur after the (zero) transition period. Since the system reacts to the step function with natural response, when the input transition time is sufficiently short, it is again expected that the peak noise would occur after the transition period. An analytical analysis of this phenomenon is provided in the following section.

Note that since the input signal function is different within the transition and post-transition periods, the system damping behavior needs to be modeled separately, as described in the following section.

5.4 Proposed Model for On-Chip Power Supply Noise

The closed-form expressions for the post-transition period are described in Section 5.4.1. Using the proposed analytical model, the effect of signal transition time t_r on power supply noise characteristics is discussed in Section 5.4.2.

5.4.1 Proposed Closed-Form Expressions for Power Supply Noise

As mentioned in Section 5.3, if the transition time of the input signal is short, the post-transition period ($t > t_r$) should be investigated to obtain an accurate peak noise value. To model the transition period ($t < t_r$), an existing method as in [86] can be used, while considering all of the R, L, C parasitic impedances of the power network. Two important transient circuit characteristics need to be determined from the transition period that serve as the initial conditions for the analysis of the post-transition period: 1) the power supply noise at the end of transition $V_{n(t=t_r)}$, and 2) the current flowing through the capacitance $I_{c(t=t_r)} = CdV_n/dt$.

Referring to Fig. 5.1, since PMOS current is negligible when the input signal is stable at high voltage during the post-transition period, current through the PMOS transistor is ignored. Similar to [85, 86], an application specific device modeling (ASDM) methodology for MOS transistor is used. When the load capacitance at the output is sufficiently large, the drain-to-source voltage V_{ds} of the NMOS transistor is larger than $V_{gs} - V_{th}$ when the peak noise occurs during the post-transition period (*i.e.* NMOS transistor is in saturation region). Note that this assumption holds particularly for I/O drivers where the load capacitances are typically high. Thus, the drain current I_{ds} increases approximately linearly with V_{gs}. A linear equation is therefore used to predict the NMOS current in saturation region,

$$I = K_1(V_{in} - V_n - V_0), \tag{5.1}$$

where K_1 and V_0 are constant parameters determined from experimental or simulation data. V_n is the voltage of the source node that exhibits power supply noise. Note that the ASDM method is different from Sakurai's α-power law model with α=1 since V_0 does not correspond to the threshold voltage of the transistor, as further discussed in [85].

When a large number of I/O drivers switch, a parameter N is used as the effective number of I/O gates that switches simultaneously. Therefore, the overall drain current of those N gates during the post-transition period is

$$I_N = NK_1(K_2 - V_n), \tag{5.2}$$

where K_2 is $V_{\mathrm{DD}} - V_0$ since the gate input voltage is stable at V_{DD} during the post-transition period. The damping characteristics of the system can be investigated by

$$I_L = I_N - C\frac{dV_n}{dt}, \tag{5.3}$$

$$V_n = R \cdot I_L + L\frac{dI_L}{dt}, \tag{5.4}$$

where I_L and I_C are, respectively, the transient current flowing through inductance L and capacitance C. Solving (5.2), (5.3), and (5.4) together, the following expression is obtained,

$$LC\frac{dV_n^2}{dt^2}+(RC+LNK_1)\frac{dV_n}{dt}+(RNK_1+1)V_n=RNK_1K_2. \tag{5.5}$$

Since the right hand side of the equation consists of a constant term, the particular solution of (5.5) is in the form of $V_p = B_0$, where B_0 is

$$B_0=\frac{RNK_1K_2}{RNK_1+1}. \tag{5.6}$$

The characteristic equation of (5.5) is

$$LCr^2+(RC+LNK_1)r+(RNK_1+1)=0. \tag{5.7}$$

The discriminant of the characteristic equation is

$$\Delta=(RC-LNK_1)^2-4LC. \tag{5.8}$$

Depending on the magnitude of (5.8), the system is either underdamped, overdamped, or critically-damped, as discussed below.

5.4.1.1 Underdamped Case

If $\Delta < 0$, the system is underdamped, and the power supply noise during the post-transition period can be expressed as

$$V_n=e^{-\alpha t}(B_1\cdot cos\omega_d t+B_2\cdot sin\omega_d t)+B_0, \tag{5.9}$$

where

$$\alpha = \frac{RC + LNK_1}{2LC}, \tag{5.10}$$

$$\omega_d = \sqrt{\frac{RNK_1 + 1}{LC} - \alpha^2}. \tag{5.11}$$

B_1 and B_2 are two coefficients determined from the initial conditions,

$$B_1 = V_{n(t=t_r)} - B0, \tag{5.12}$$

$$B_2 = \frac{I_{c(t=t_r)}/C + \alpha B_1}{\omega_d}. \tag{5.13}$$

5.4.1.2 Overdamped Case

If $\Delta > 0$, the system is overdamped. The waveform of the power supply noise during the post-transition period can be expressed as the sum of general and particular solutions of (5.5),

$$V_n = A_1 e^{\lambda_1 t} + A_2 e^{\lambda_2 t} + B_0, \tag{5.14}$$

where

$$\lambda_{1,2} = -\frac{RC + LNK_1}{2LC} \pm \sqrt{(\frac{RC + LNK_1}{2LC})^2 - \frac{RNK_1 + 1}{LC}}. \tag{5.15}$$

B_0 is the particular solution as provided by (5.6). A_1 and A_2 are two coefficients determined from the initial conditions,

$$A_1 = \frac{I_{c(t=t_r)}/C - (V_{n(t=t_r)} - B_0) \cdot \lambda_2}{\lambda_1 - \lambda_2}, \tag{5.16}$$

$$A_2 = V_{n(t=t_r)} - B_0 - A_1. \tag{5.17}$$

<table>
<tr><th></th><th>Overdamped</th><th>Underdamped</th></tr>
<tr><td><b>R</b></td><td>$R > \frac{LNK_1 + \sqrt{4LC}}{C}$</td><td>$R < \frac{LNK_1 + \sqrt{4LC}}{C}$</td></tr>
<tr><td><b>C</b></td><td>$C < C_1$ or $C > C_2$
or $\Delta_C < 0$</td><td>$C_1 < C < C_2$</td></tr>
<tr><td><b>L</b></td><td>$L < L_1$ or $L > L_2$
or $\Delta_L < 0$</td><td>$L_1 < L < L_2$</td></tr>
<tr><td colspan="3">$C_1 = \frac{(2RLNK_1 + 4L) - \sqrt{\Delta_C}}{2R^2}, C_2 = \frac{(2RLNK_1 + 4L) + \sqrt{\Delta_C}}{2R^2}$
(when $\Delta_C = (2RLNK_1 + 4L)^2 - 4(RLNK_1)^2 > 0$)

$L_1 = \frac{(2RCNK_1 + 4C) - \sqrt{\Delta_L}}{2(NK_1)^2}, L_2 = \frac{(2RCNK_1 + 4C) + \sqrt{\Delta_L}}{2(NK_1)^2}$
(when $\Delta_L = (2RCNK_1 + 4C)^2 - 4(NK_1RC)^2 > 0$)</td></tr>
</table>

Table 5.1: Boundary conditions on damping characteristics as obtained from (5.8).

Using (5.9) and (5.14), the voltage noise waveform and the peak power supply noise for the post-transition period can be determined. Note that when $\Delta = 0$, the system is critically-damped. In this case, power supply noise can be estimated by either (5.9) or (5.14) since the two expressions are equivalent at $\Delta = 0$.

According to (5.8), damping behavior is affected by the parasitic impedances and the transistor characteristics (indicated by the K_1 and N terms). Thus, the boundary conditions on damping characteristics can be obtained from (5.8). These boundary conditions for parasitic impedances are listed in Table 5.1.

5.4.2 Effect of Input Transition Time on Peak Power Supply Noise

The effect of signal transition time on peak noise characteristics can be analyzed using the proposed model described in the previous section. Underdamped and overdamped cases are separately investigated, as discussed below.

5.4.2.1 Underdamped Case

According to [86], if the system is underdamped, the power supply noise waveform during the transition period ($t_0 < t < t_r$) is predicted by

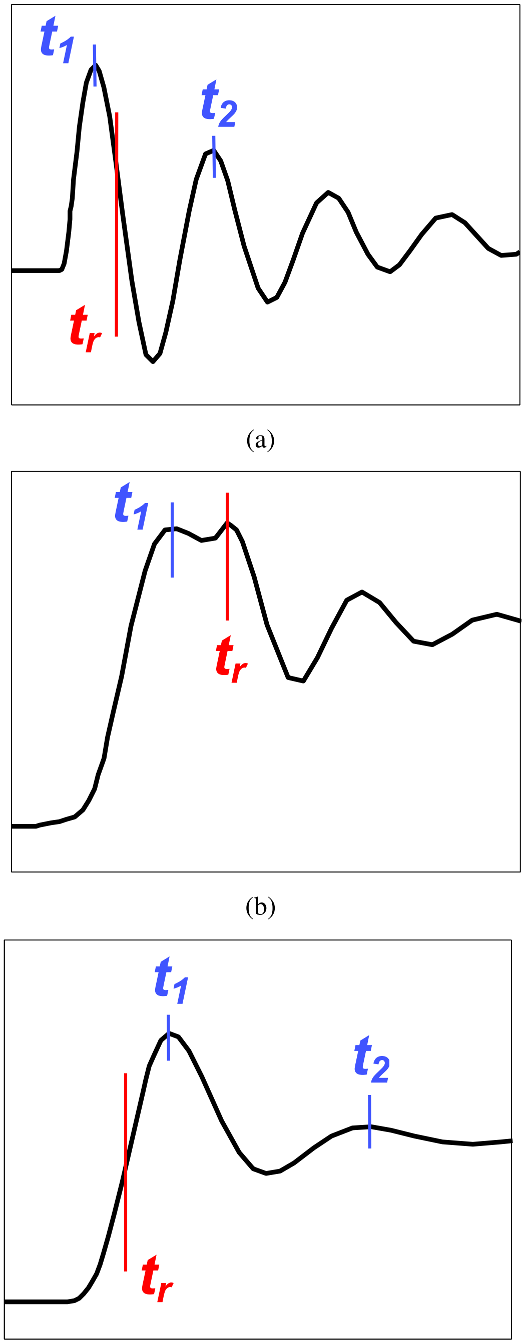
$$\begin{aligned} V_n &= e^{-\alpha t'}(B'_1 \cdot cos\omega_d t' + B'_2 \cdot sin\omega_d t') + at' + b \\ &= e^{-\alpha t'}\sqrt{B'^2_1 + B'^2_2} sin(\omega_d t' + \arctan\frac{B'_2}{B'_1}) + at' + b \end{aligned} \tag{5.18}$$

where $t' = t - t_0$ and $t_0 = tV_0/V_{DD}$. B'_1, B'_2, a and b are determined by initial conditions. Therefore, the noise waveform during the transition period is a decaying sine wave superposed with a linearly increasing term $at' + b$. The local maxima of the waveform occur at $t = t_n$ $(n = 1, 2, ...)$ that satisfy

$$\omega_d(t - t_0) + \arctan\frac{B'_2}{B'_1} = 2n\pi - 3/2\pi \quad (n = 1, 2, ...). \tag{5.19}$$

According to [85], if the parasitic resistance R is not considered, the linearly increasing term $at' + b$ in (5.18) is reduced to a constant term. Therefore, the power supply noise at a local maximum is always smaller than its preceding local maximum $[V_n(t_1) > V_n(t_2) > V_n(t_3) > \cdots > V_n(t_n)]$. However, if the parasitic resistance R is included in the model, it is possible that a local maximum is larger than or equal to its preceding local maximum [$e.g.$ $V_n(t_3) > V_n(t_2)$], due to the linearly increasing term in (5.18).

Alternatively, during the post-transition period, as investigated in this paper, the particular solution of (5.9) is constant since the gate input signal is in steady state.



(a)

(b)

(c)

Figure 5.5: Illustrative examples on the occurrence time of peak power supply noise in underdamped case: (a) peak noise occurs within the transition period, (b) peak noise occurs at end of the transition period, and (c) peak noise occurs within the post-transition period.

The power supply noise waveform during the post-transition period is represented by a decaying sine wave superposed with a constant term. Therefore, if $t_r > t_1$, the peak noise can occur either at one of the local maxima t_n [see Fig. 5.5(a)], or at the end of transition period t_r [see Fig. 5.5(b)]. Alternatively, if $t_r < t_1$, the noise continues to increase after $t = t_r$. The peak power supply noise occurs at the first local maximum *within the post-transition period*, as predicted by the proposed closed-form expression (5.9) [see Fig. 5.5(c)].

5.4.2.2 Overdamped Case

If the system is overdamped, the power noise during the transition period is predicted by

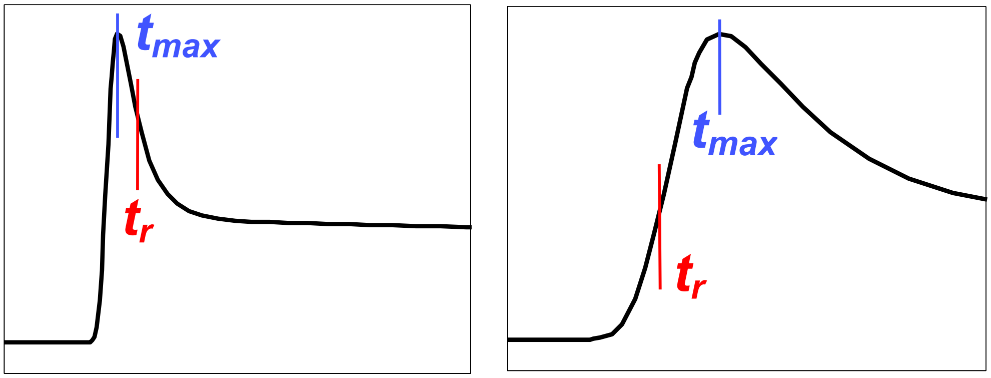
$$V_n = A'_1 e^{\lambda_1 t'} + A'_2 e^{\lambda_2 t'} + at' + b, \tag{5.20}$$

where t' = $t - t_0$. A'_1 and A'_2 are determined by initial conditions of the transition period. a and b are the same as in (5.18). Thus, a single t = t_{max} exists that satisfies

$$\frac{dV_n}{dt'} = 0 \Longrightarrow A'_1 \lambda_1 e^{\lambda_1 t'} + A'_2 \lambda_2 e^{\lambda_2 t'} + a = 0. \tag{5.21}$$

The noise first increases monotonically until $t = t_{max}$ and then decreases monotonically. Therefore, if t_r is larger than t_{max}, the peak noise occurs before $t = t_r$, as shown in Fig. 5.6(a). Alternatively, if t_r is smaller than t_{max}, the noise monotonously increases during the transition period, and continues to increase within the post-transition period, as shown in Fig. 5.6(b).

Thus, in either overdamped or underdamped cases, it is possible that the peak noise can occur after $t = t_r$. It is therefore necessary to model both the transition and the post-transition periods, as proposed in this paper. The validation of the pro-



(a) (b)

Figure 5.6: Illustrative examples on the occurrence time of peak power supply noise in overdamped case: (a) peak noise occurs within the transition period, and (b) peak noise occurs within the post-transition period.

posed closed-form expressions with SPICE and comparison with existing models are provided in the following section.

5.5 Simulation Results

To evaluate the proposed model, the schematic shown in Fig. 5.1 is simulated with a 45 nm CMOS technology using SPICE [66]. For a single inverter, W_n=100 nm, W_p=200 nm, L=50 nm. To investigate the effect of number of simultaneously switching gates on power noise, N inverters are considered where N varies from 100 to 5000. Note that the number of parallel output drivers ranges from several hundreds to several thousands, making 5000 a reasonable number for estimating noise in I/O drivers [92,93]. Multiple cases are investigated with various RLC parasitic impedances to consider each damping characteristic of the system. Estimation of the entire power supply noise waveform and comparison with SPICE are presented in Section 5.5.1. Estimation of the peak power supply noise, comparison

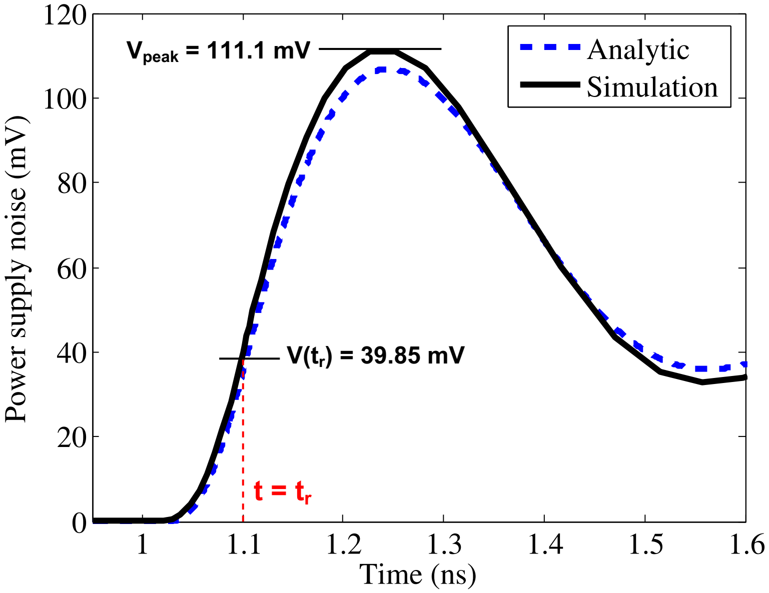


Figure 5.7: Comparison of the power supply noise waveform in underdamped case as predicted by the proposed expressions and SPICE simulations. R=5 Ω, L=1 nH, C=10 pF, t_r=100 ps, W$_n$=100 nm, W$_p$=200 nm, C_L = 10 pF, and N=100.

with SPICE and existing work are provided in Section 5.5.2.

5.5.1 Estimation of the Power Supply Noise Waveform

5.5.1.1 Underdamped Case

The simulation results are compared with (5.9) for an underdamped case where R=5 Ω, L=1 nH, C=10 pF, N=100, and t_r=100 ps. The comparison is shown in Fig. 5.7, demonstrating an average error of less than 5%. Note that since $t_r < t_1$ (see Section 5.4.2.1), the noise does not reach first local maximum at $t = t_r$. Instead, the noise continues to increase during the post-transition period, reaching up to 111.1 mV. Thus, if the noise at $t = t_r$ (39.85 mV) is used as the peak noise, as in previous works, the peak noise is underestimated by 64.1%.

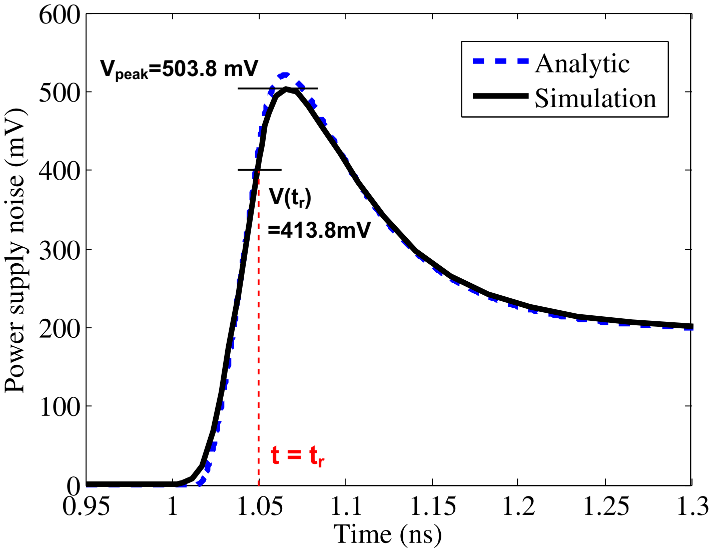


Figure 5.8: Comparison of the power supply noise waveform in overdamped case as predicted by the proposed expressions and SPICE simulations. R=5 Ω, L=1 nH, C=1 pF, t_r=50 ps, W$_n$=100 nm, W$_p$=200 nm, C_L = 10 pF, and N=500.

5.5.1.2 Overdamped Case

The simulation results are compared with (5.14) for an overdamped case where R=5 Ω, L =1 nH, C=1 pF, N=500, and t_r=50 ps. According to Fig. 5.8, the proposed closed-form expression accurately predicts power supply noise with an error less than 4%. Similar to the underdamped case, estimating the peak noise at $t = t_r$ underestimates the actual peak noise by 17.8%.

5.5.2 Estimation of the Peak Noise

The peak power supply noise predicted by the proposed closed-form expressions (5.9) and (5.14) is compared with the simulation results (SPICE), and the models proposed in [85] and [86] for a variety of RLC impedances. The results

Parasitic impedances			**SPICE**	**Ding [85]**		**Hekmat [86]**		**This paper**	
R (Ω)	L (pH)	C (pF)		Peak (mV)	Error (%)	Peak (mV)	Error (%)	Peak (mV)	Error (%)
0.5	100	5	32.89	29.40	-10.61	31.78	-3.38	31.96	-2.82
		10	29.84	24.00	-19.56	24.79	-16.93	28.65	-4.00
	500	5	94.42	60.07	-36.38	60.22	-36.22	89.87	-4.82
		10	72.81	33.76	-53.63	33.80	-53.57	68.71	-5.63
	1000	5	131.90	65.23	-50.54	65.27	-50.52	125.29	-5.01
		10	100.90	35.18	-65.14	35.19	-65.13	94.92	-5.92
1	100	5	35.92	29.40	-18.15	33.92	-5.57	34.60	-3.67
		10	32.19	24.00	-25.43	25.49	-20.82	30.75	-4.48
	500	5	95.18	60.07	-36.89	60.36	-36.58	91.20	-4.18
		10	74.20	33.76	-54.50	33.84	-54.39	70.07	-5.57
	1000	5	133.10	65.23	-50.99	65.31	-50.93	126.38	-5.05
		10	102.20	35.18	-65.58	35.20	-65.56	96.12	-5.95
5	100	5	60.68	29.40	-51.55	45.36	-25.24	59.27	-2.33
		10	55.11	24.00	-56.44	29.18	-47.05	54.87	-0.43
	500	5	106.90	60.07	-43.81	61.39	-42.57	102.99	-3.66
		10	85.75	33.76	-60.63	34.12	-60.21	82.90	-3.32
	1000	5	141.90	65.23	-54.03	65.60	-53.77	135.75	-4.33
		10	111.10	35.18	-68.34	35.27	-68.28	106.83	-3.84
Average error (%)				-45.68		-42.04		-4.17	
Maximum error (%)				-68.34		-68.28		-5.95	

Table 5.2: Comparison of the simulated and estimated peak power supply noise for a variety of *RLC* parasitic impedances when N=100 and t_r=100 ps.

Parasitic impedances			SPICE	Ding [85]		Hekmat [86]		This paper	
R (Ω)	L (pH)	C (pF)		Peak (mV)	Error (%)	Peak (mV)	Error (%)	Peak (mV)	Error (%)
0.5	100	5	45.13	29.79	-33.99	30.02	-33.49	42.97	-4.79
		10	34.82	16.83	-51.67	16.89	-51.50	32.83	-5.73
	500	5	100.60	35.18	-65.03	35.19	-65.02	94.92	-5.64
		10	75.37	18.27	-75.76	18.28	-75.75	70.56	-6.38
	1000	5	136.60	35.90	-73.72	35.90	-73.72	128.31	-6.07
		10	102.80	18.46	-82.04	18.16	-82.04	96.03	-6.59
1	100	5	46.70	29.79	-36.21	30.23	-35.26	44.63	-4.43
		10	36.25	16.83	-53.58	16.95	-53.25	34.48	-4.87
	500	5	101.50	35.18	-65.34	35.20	-65.32	96.12	-5.30
		10	76.52	18.27	-76.12	18.28	-76.11	71.88	-6.06
	1000	5	137.70	35.90	-73.93	35.91	-73.92	129.38	-6.05
		10	103.70	18.46	-82.20	18.46	-82.20	97.24	-6.23
5	100	5	63.49	29.79	-53.08	31.58	-50.25	62.21	-2.01
		10	55.39	16.83	-69.62	17.31	-68.75	55.15	-0.44
	500	5	111.50	35.18	-68.45	35.27	-68.36	106.83	-4.19
		10	86.92	18.27	-78.98	18.30	-78.95	84.25	-3.07
	1000	5	145.40	35.90	-75.31	35.93	-75.29	138.46	-4.77
		10	112.60	18.46	-83.61	18.47	-83.60	107.89	-4.18
Average error (%)				-66.59		-66.26		-4.82	
Maximum error (%)				-83.61		-83.60		-6.59	

Table 5.3: Comparison of the simulated and estimated peak power supply noise for a variety of *RLC* parasitic impedances when N=100 and t_r=50 ps.

are listed in Table 5.2 (for t_r=100 ps) and Table 5.3 (for t_r=50 ps). Note that a wide range of *RLC* parastic impedances is covered in these tables to demonstrate the accuracy of the proposed expressions in both realistic and extreme cases.

As listed in these tables, the proposed expressions significantly enhance the accuracy in estimating the peak noise as compared to existing models. Specifically, in Table 5.2, the maximum error of the proposed model is 5.95% whereas the maximum error for [85] and [86] are, respectively, 68.34% and 68.28%. Similarly, the average error of the proposed model over all cases is 4.17% whereas the average error for [85] and [86] are, respectively, 45.68% and 42.04%.

If the signal transition time is reduced to 50 ps, as in Table 5.3, the accuracy of existing models is further degraded whereas the proposed model can still accurately estimate the peak power supply noise. Specifically, the maximum error of the proposed model is 6.59% whereas the maximum error for [85] and [86] are, respectively, 83.61% and 83.60%. Similarly, the average error of the proposed model over all cases is 4.82% whereas the average error for [85] and [86] are, respectively, 66.59% and 66.26%.

Note that in some cases (*e.g.* R=0.5 Ω, L=100 pH, C=1 pF in Table 5.2), the peak noise obtained from the proposed model is the same as the peak noise obtained from [86]. In these cases, the peak noise occurs within the transition period ($t \leq t_r$) where the accuracy of the existing models is sufficient. In most of the cases (and particularly when transition time is short), however, this condition does not hold and considering only the transition period can produce an error up to 83% whereas the maximum error of the proposed expressions is 6.59%.

Finally, the accuracy of the model is evaluated when the number of simultaneously switching gates is increased. As shown in Fig. 5.9, peak noise is accurately

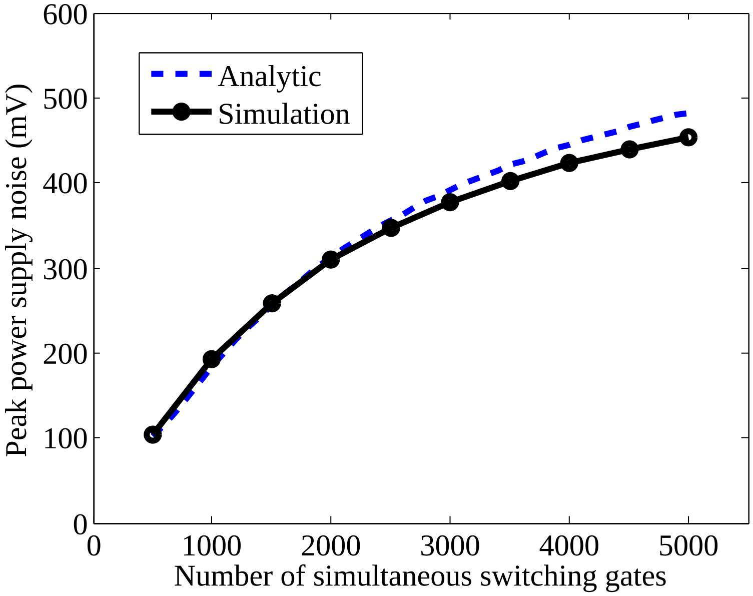


Figure 5.9: Comparison of the peak power supply noise predicted by the proposed expressions and obtained by SPICE simulations as the number of simultaneously switching gates increases. R=0.5 Ω, L=100 pH, C=1 pF, t_r=100 ps, W$_n$=100 nm, W$_p$=200 nm, C_L = 10 pF.

estimated as the number of simultaneously switching gates increases (up to 5000). The error between the proposed expressions and SPICE is less than 6%.

5.6 Summary

In this chapter, accurate closed-form expressions have been developed to estimate on-chip power supply noise with fast signal transitions. It has been demonstrated that the peak noise can occur after the transition is complete, invalidating existing assumptions and models. The damping characteristics of the system have been investigated while considering the *post-transition* period. The accuracy of the proposed closed-form expressions has been evaluated by comparing the results with SPICE. Maximum error of 6.59% has been demonstrated. The proposed expressions have also been compared with existing models and the accuracy of the existing models has been enhanced by up to 79.4%.

Chapter 6

Decoupling Capacitor Topologies for TSV Based 3D ICs with Power Gating

Utilizing decoupling capacitors is an essential technique in modern high performance ICs to satisfy the stringent power integrity requirement. In 3D ICs, the vertical interconnects with low impedance increase the utility of decoupling capacitors. In 3D ICs, however, power gating can significantly degrade the system-wide power integrity since the decoupling capacitance associated with the power gated block/plane becomes ineffective for the neighboring, active planes. In this chapter, novel decoupling capacitor topologies are investigated to alleviate this issue by exploiting 1) relatively low resistance TSVs and 2) ability of TSVs to bypass plane-level power networks when delivering the power supply voltage. The related background about decoupling capacitors in 3D ICs with power gating is discussed in Section 6.1. The proposed decoupling capacitor topologies are presented in Section 6.2. These topologies are evaluated in Section 6.3 through a comprehensive case study.

6.1 Decoupling Capacitors in 3D ICs with Power Gating

As introduced in Chapter 2, decoupling capacitors serve as charge reservoir to provide instantaneous charge for neighboring switching circuit loads. The efficacy of a decoupling capacitor is determined by the impedance of the metal lines connecting the decoupling capacitor and switching load circuits. Placing the decoupling capacitors sufficiently close to the switching load is helpful in reducing the power supply noise [94].

In 3D ICs, the efficacy of a decoupling capacitor on the neighboring planes is critical since it becomes increasingly challenging to satisfy power supply noise as the number of planes increases. In existing work [95], it has been observed that the decoupling capacitance placed within a plane is highly effective in reducing the power supply noise of the neighboring planes, as also observed in this work. This behavior, however, holds unless the related blocks (or entire plane) are power gated, as described below.

In traditional topologies, the decoupling capacitors within a power gated block (or plane) cannot provide charge to neighboring planes since these capacitors are typically connected to a virtual power grid that is closer to the switching circuit. However, if the block or plane is power gated, those capacitors are disconnected from the global power network, making these capacitors ineffective for the neighboring planes. Since system-wide power integrity is a critical challenge in 3D ICs, effective use of intentional decoupling capacitance is crucial, even when power gating is adopted.

Note that a similar issue exists in 2D ICs with multiple power domains. However, due to longer global interconnects, it is relatively impractical to utilize the capacitance of a power gated domain for the remaining, active domains. As demonstrated by [96], the required decoupling capacitance increases exponentially if the resistance between the capacitor and switching load exceeds a certain threshold. In TSV-based 3D ICs, the low impedance vertical TSVs connections facilitate the utilization of decoupling capacitors on a plane to suppress power supply noise for the neighboring planes. To exploit this advantage, novel decoupling capacitors topologies are proposed in next section for 3D ICs with power gating.

6.2 Decoupling Capacitor Topologies for Power-Gated 3D ICs

In this section, two decoupling capacitor placement topologies are discussed: 1) reconfigurable topology, as described in Section 6.2.1 and 2) always-on topology, as described in Section 6.2.2. In both topologies, decoupling capacitors on a power gated plane can be utilized to suppress power supply and power gating noise within the neighboring active planes.

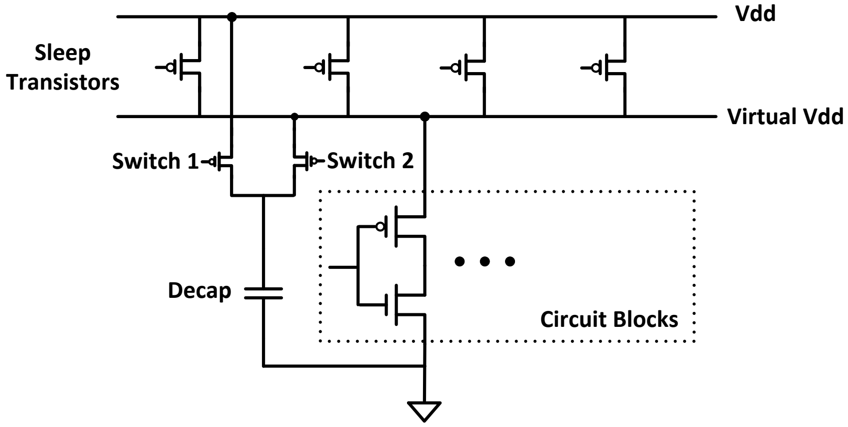


Figure 6.1: Conceptual representation of the reconfigurable decoupling capacitor topology with power gating.

6.2.1 Reconfigurable Topology

In the reconfigurable topology, two switches are introduced (similar to [97]) to form a configurable decoupling capacitor, as conceptually illustrated in Fig. 6.1. If a certain plane is active, the decoupling capacitors on that plane are connected to the virtual V_{DD} grid through switch 2, thereby reducing the power supply noise on that plane. Alternatively, if the plane is power gated (sleep transistors are turned off), the decoupling capacitors are connected to the global V_{DD} grid, bypassing the sleep transistors. Thus, even if the plane is power gated, the decoupling capacitors are effective for the remaining planes. The overhead of this topology includes the reconfigurable switches, metal resources required to route the related control signals, and a possible increase in overall power consumption depending upon how the capacitors are implemented, as further discussed and quantified in Section 6.3.

The design process (sizing and choosing an appropriate threshold voltage) of these switches exhibits similar and well known tradeoffs as the design process of the sleep transistors [98, 99]. Similar to sleep transistors, high-V_{th} switches are

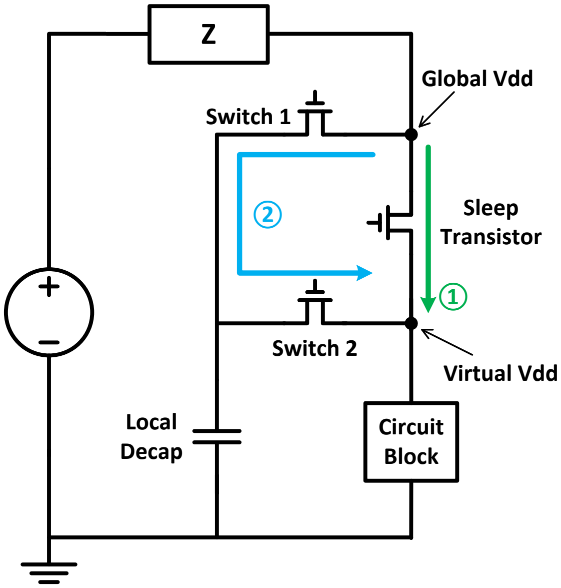


Figure 6.2: Illustration of the additional resistive path between the global and virtual power networks formed by the reconfigurable switches.

used to minimize the voltage at the virtual V_{DD} grid when the plane is power gated (switch 1 is on and switch 2 is off). Note that the two reconfigurable switches form an additional path from global V_{DD} grid to virtual V_{DD} grid, as depicted in Fig. 6.2. Thus, the effective resistance of the sleep transistors and the effective resistance of the reconfigurable switches are in parallel, partially reducing the off-resistance between global and virtual V_{DD} grids. High-V$_{th}$ switches are therefore required to maintain significant savings in the leakage current when the plane is power gated.

6.2.2 Always-on Topology

The reconfigurable placement methodology described above provides flexibility to exploit a decoupling capacitor located in a power gated plane for the remaining planes. This flexibility is achieved at the expense of additional reconfigurable

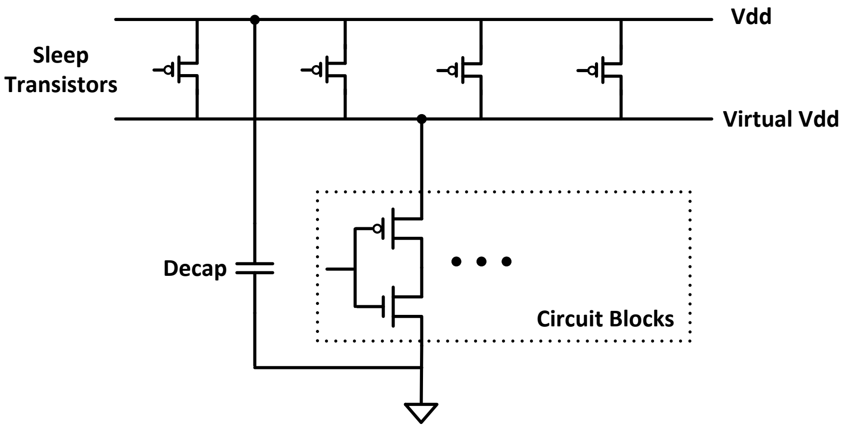


Figure 6.3: Conceptual representation of the always-on decoupling capacitor topology with power gating.

switches. To mitigate the overhead of reconfigurable topology, an always-on topology is considered for planes with low switching activity (such as a sensing plane of a heterogeneous 3D IC that is periodically activated). In this topology, the decoupling capacitors are always connected to the global V_{DD} grid, thereby bypassing the sleep transistors, as conceptually depicted in Fig. 6.3. Thus, the decoupling capacitance within a power gated plane is available to suppress power supply and power gating noise of the neighboring active planes. The limitation of this topology is a possible increase in the power supply noise of the plane where decoupling capacitors are located due to a greater impedance between the capacitor and the switching circuit [94]. This tradeoff and the additional decoupling capacitance required to mitigate this limitation are characterized in the following section.

6.3 Case Study

To investigate the benefits and tradeoffs of the decoupling capacitor topologies discussed in this chapter, the same simulation setup developed in previous Chapter 4 is used. The design issue of simultaneously sizing decoupling capacitors and switches are investigated in Section 6.3.1. Simulation results are presented in Section 6.3.2 where the proposed topologies are compared with the traditional topology in terms of power supply noise, power gating noise, physical area, turn-on time, and overall power consumption. The effect of number of planes is also investigated.

6.3.1 Reconfigurable Switch and Decoupling Capacitor Sizing

The size of the reconfigurable switches should be sufficiently large to minimize the shield effect of these switches on the decoupling capacitors [50, 94]. Analyses demonstrate that multiple pairs of decoupling capacitor and switch size satisfy the power supply noise constraint. Since both decoupling capacitors and switches consume area, it is important to choose a pair that minimizes the physical area overhead. This characteristic is illustrated in Fig. 6.4. Each pair of switch size and decoupling capacitance (on the curve with square markers) satisfies the 5% power supply noise constraint (50 mV). The area overhead (determined as a percentage of the overall area) is shown by the curve with triangle markers. The decoupling capacitors are implemented as MOS capacitors in the 45 nm technology with an oxide thickness of 1 nm [66]. As illustrated in this figure, a small decoupling capacitor requires a very large switch size to satisfy the noise constraint. A large switch size not only increases the area overhead, but also increases the voltage at the virtual V_{DD} grid when the block/plane is power gated, thereby increasing the leakage current.

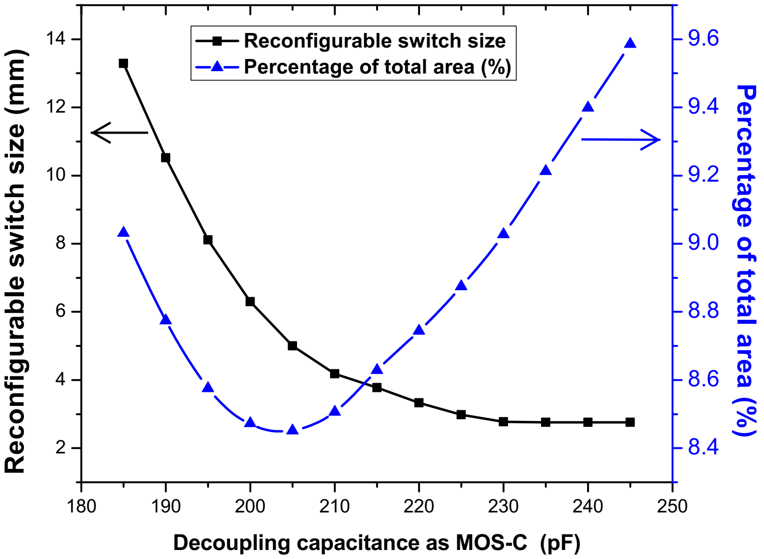


Figure 6.4: Area consumption (as percent of the overall area) of different pairs of decoupling capacitance and reconfigurable switch size. Note that each pair satisfies the power supply noise constraint of 50 mV (5% of the supply voltage). Area overhead is minimized at a specific pair. Decoupling capacitors are implemented as MOS-C.

Alternatively, if decoupling capacitance exceeds a certain threshold, the switch size cannot be reduced further, thereby increasing the overall area overhead. For this case study, the minimum area overhead (8.45%) occurs when the equivalent decoupling capacitor (from MOS-C) is approximately 205 pF and switch size is 5 mm. Note that 205 pF and 5 mm represent, respectively, a single decoupling capacitor and a single switch in the power network. The overall decoupling capacitance (per plane) is equal to approximately 3.3 nF whereas the overall switch size (per plane) is equal to 160 mm since there are two switches per capacitor and the total number of capacitors per plane is 16. This amount of decoupling capacitance and switch size is used for the reconfigurable topology in the remaining portions of this chapter.

6.3.2 Simulation Results

The efficacy of the decoupling capacitor topologies discussed in this chapter is demonstrated by comparing these methods with the traditional topology. Design criteria such as area overhead, power supply noise, power gating noise, and turn-on time are analyzed. The effect of number of planes on the capacitor topologies is investigated. Power overhead of each topology is also quantified to demonstrate that the proposed topologies do not undermine the leakage savings achieved by power gating. All of the simulations have been performed using SPICE accurate SPECTRE simulator [100].

Several different scenarios are considered:

- **Scenario 1**: All of the three planes are active, representing the greatest workload.
- **Scenario 2**: The top and bottom planes are active, while the middle plane is power gated.
- **Scenario 3**: Only the bottom plane is active, while the middle and top planes are power gated.
- **Scenario 4**: The middle and bottom planes are active, while the top plane is power gated.
- **Scenario 5**: Only the middle plane is active, while the top and bottom planes are power gated.

	ST	Decap	Switch
Traditional	36.3 mm	2.90 nF	N/A
Reconfigurable	36.3 mm	3.29 nF	160 mm
Always-on	36.3 mm	5.21 nF	N/A

Table 6.1: Size of the decoupling capacitors and reconfigurable switches.

Area (μm^2)		**ST**	**Decap**	**Switch**	**Area per plane**
Traditional	MOS	1815	65413	N/A	67228 (6.72%)
	MIM		233870		235685 (23.56%)
Reconfigurable	MOS	1815	76733	8000	86548 (8.65%)
	MIM		265322		275137 (27.51%)
Always-on	MOS	1815	118158	N/A	119973 (11.99%)
	MIM		420161		421976 (42.19%)

Table 6.2: Comparison of the physical area overhead of the traditional, reconfigurable, and always-on topologies.

6.3.2.1 Area Overhead

The size of the distributed decoupling capacitors (implemented as MOS-C) is determined based on Scenario 1 where all of the planes are active. For each topology, the decoupling capacitors are sized to ensure that the worst case power supply noise is within 5% of the V_{DD} (50 mV) throughout the entire power network. This constraint ensures that no additional performance penalty is introduced with the reconfigurable and always-on topologies. For the reconfigurable topology, the size of the switches and decoupling capacitors is determined to minimize the physical area overhead while satisfying the power supply noise constraint, as described in Section 6.3.1. These sizes are listed in Table 6.1. Note that the size of the decoupling capacitors, sleep transistors, and switches (for the reconfigurable topology) listed in this table refers to per plane.

MOS and metal-insulator-metal (MIM) capacitors are considered to estimate the area overhead of the decoupling capacitors to demonstrate the tradeoff between area and leakage overhead. Note that in all of the simulation results, capacitors are implemented as MOS-C using NMOS transistors in 45 nm technology with an oxide thickness of 1 nm [66]. For MIM capacitors, the area overhead is analytically estimated by assuming a capacitance density of 12.4 fF/μm^2 based on [101]. The itemized area overhead of the traditional, reconfigurable, and always-on topologies are listed in Table 6.2 for both MOS and MIM capacitors. If MOS capacitor is used (as in the simulations), the area overhead (due to capacitors and sleep transistors) is 6.70% of the overall area for the traditional topology where the decoupling capacitors are connected to the virtual V_{DD} grid. The area overhead increases to 8.65% in the reconfigurable topology since the size of the decoupling capacitors should be moderately increased to compensate for the shield effect of the reconfigurable switches (which also contributes to the area overhead). Finally, in the always-on topology, the area overhead increases to 11.99% due to an increase in the decoupling capacitance, as listed in Table 6.1. For MIM capacitor, the area overhead (analytically determined) is significantly greater than MOS capacitor (23.56%, 27.51%, 42.19%, respectively, for traditional, reconfigurable and always-on topologies). The leakage current of MIM capacitor, however, is significantly less than MOS capacitor, as discussed in Section 6.3.2.5. Also, note that the MIM capacitors consume area within the metal layers rather than consuming transistor area.

6.3.2.2 Power Integrity

Power supply noise and power gating noise are analyzed for each scenario. The power gating status of each individual plane is summarized in Table 6.3. The re-

	Power gating status		
	Top	**Middle**	**Bottom**
Scenario 1	on	on	on
Scenario 2	on	off (→on)	on
Scenario 3	off	off (→on)	on
Scenario 4	off (→on)	on	on
Scenario 5	off	on	off (→on)

Table 6.3: Power gating status of each plane in different scenarios. The parentheses indicate one plane is in transition from off state to on state for the measurement of power gating noise.

	Power supply noise (mV)				
	Traditional	Reconfigurable		Always-on	
	Peak	Peak	Redtn.	Peak	Redtn.
Scenario 1	50	50	N/A	50	N/A
Scenario 2	48.16	43.48	9.7%	43.40	9.9%
Scenario 3	52.22	39.64	24.1%	38.07	27.1%
Scenario 4	48.55	44.50	8.3%	42.89	11.7%
Scenario 5	52.51	38.78	26.1%	37.55	28.5%

Table 6.4: Peak power supply noise obtained from each scenario and noise reduction achieved by the proposed topologies (All of the decoupling capacitors are implemented as MOS capacitors).

configurable and always-on decoupling capacitor topologies achieve significant reduction in both peak and RMS power supply noise, as listed in Table 6.4 and 6.5. Note that, in the simulations, all of the decoupling capacitors are implemented as MOS-C.

In scenario 1 where all of the planes are switching, the peak power supply noise is equal to 50 mV for each topology since the decoupling capacitor, sleep transistor, and switch sizes are determined based on this scenario. Note that power supply noise is observed in the bottom plane except scenario 5 where bottom plane is power

	Power supply noise (mV)				
	Traditional	Reconfigurable		Always-on	
	RMS	RMS	Redtn.	RMS	Redtn.
Scenario 1	30.34	26.26	13.5%	26.45	12.8%
Scenario 2	23.40	17.2	26.5%	15.83	32.4%
Scenario 3	16.93	9.19	45.7%	8.57	49.4%
Scenario 4	23.03	17.15	25.5%	16.65	27.7%
Scenario 5	17.03	9.21	45.9%	8.46	50.3%

Table 6.5: RMS power supply noise obtained from each scenario and noise reduction achieved by the proposed topologies (All of the decoupling capacitors are implemented as MOS capacitors).

	Power gating noise (mV)				
	Traditional	Reconfigurable		Always-on	
	Peak	Peak	Redtn.	Peak	Redtn.
Scenario 1	N/A	N/A		N/A	
Scenario 2	93.46	19.45	79.2%	17.98	80.8%
Scenario 3	112.28	19.61	82.5%	17.31	84.6%
Scenario 4	93.65	19.40	79.3%	17.24	81.6%
Scenario 5	112.0	19.60	82.5%	17.39	84.5%

Table 6.6: Peak power gating noise obtained from each scenario and noise reduction achieved by the proposed topologies (All of the decoupling capacitors are implemented as MOS capacitors).

	Power gating noise (mV)				
	Traditional	Reconfigurable		Always-on	
	RMS	RMS	Redtn.	RMS	Redtn.
Scenario 1	N/A	N/A		N/A	
Scenario 2	74.55	12.94	82.6%	11.47	84.6%
Scenario 3	83.60	12.80	84.7%	11.49	86.3%
Scenario 4	74.34	12.93	82.6%	11.41	84.7%
Scenario 5	83.65	12.82	84.7%	11.49	86.3%

Table 6.7: RMS power gating noise obtained from each scenario and noise reduction achieved by the proposed topologies (All of the decoupling capacitors are implemented as MOS capacitors).

gated. In this case, noise is observed in the middle plane. For scenarios 3 and 5 (where two planes are power gated), both the reconfigurable and always-on topologies reduce the peak power supply noise by more than 20%. In these scenarios, the reconfigurable topology reduces the RMS noise by 46% whereas the always-on topology achieves 50% reduction in RMS noise. For scenarios 2 and 4 (where only one plane is power gated), the reduction in peak noise is approximately 9% and 10% for, respectively, reconfigurable and always-on topologies. For the same scenarios, proposed topologies achieve, respectively, at least 25% and 27% reduction in RMS noise.

It is important to note that in the traditional topology, the peak noise in scenarios 3 and 5 exceeds 50 mV despite a reduction in the overall switching current due to power gating. This characteristic is due to less decoupling in the power network since the decoupling capacitors in the power gated planes cannot behave as charge reservoirs for the remaining, active planes.

Transient behavior of voltage noise at a specific node within the bottom plane is depicted in Fig. 6.5 for each topology for scenario 3, demonstrating the reduction in peak and RMS noise. Similarly, the spatial distribution of peak power supply noise is shown in Fig. 6.6 for scenario 3 where the first two planes are power gated and the bottom plane is active. Reduction in peak noise throughout the power network is illustrated for both reconfigurable and always-on topologies.

To investigate power gating noise, one of the power gated planes transitions from sleep to active state in each scenario (except scenario 1), as also indicated in Table 6.3. The voltage fluctuation due to in-rush current during the wake-up process is analyzed. Note that a gradual wake-up strategy is adopted where switching circuits on each plane are divided into 5 segments and each segment sequentially

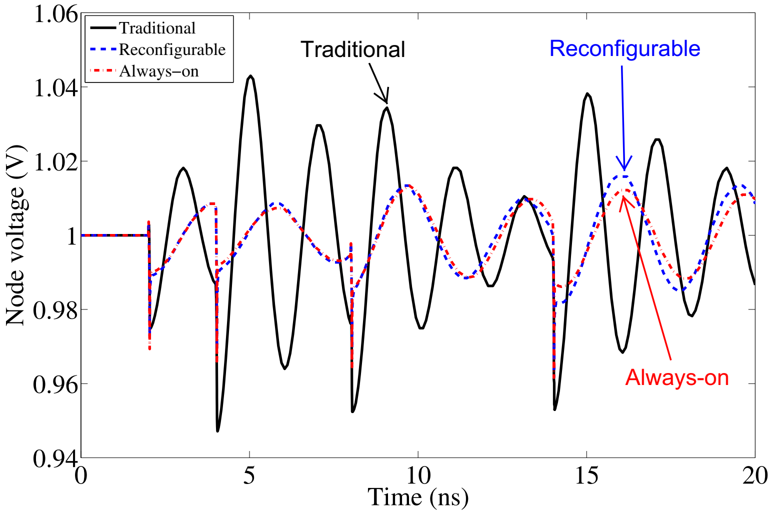
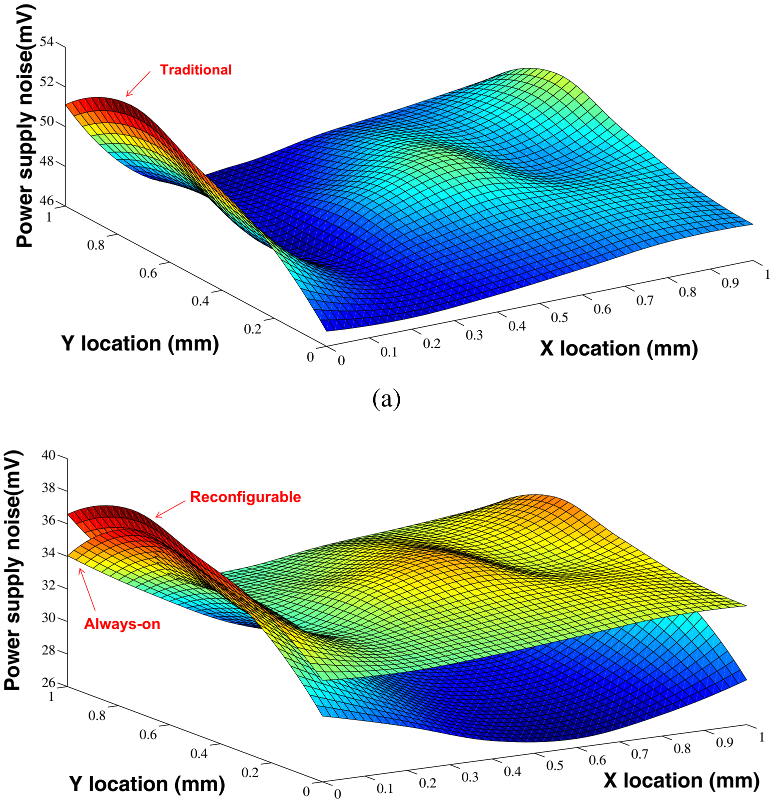


Figure 6.5: Transient behavior of the on-voltage at a specific node within virtual grid of the bottom plane for each topology for scenario 3 (first two planes are power gated and the bottom plane is active).

wakes up with a time interval of 100 ps based on [102]. Peak power gating noise is observed in the bottom plane except scenario 5 where bottom plane has a transition. In this case, noise is observed in the middle plane. Results are listed in Table 6.6 and 6.7. Both reconfigurable and always-on topologies achieve approximately 80% reduction in peak and RMS power gating noise. This considerable reduction in power gating noise is due to a significant amount of in-rush current in traditional topology that flows not only for the activated circuit, but also to charge the associated decoupling capacitors, as also observed in [97]. Thus, a greater in-rush current produces significantly high power gating noise (particularly due to parasitic inductance). Alternatively, in both reconfigurable and always-on topologies, the decoupling capacitors are connected to the global V_{DD} grid when the plane is power gated. Thus, even if the plane is power gated, these capacitors remain charged (significantly reducing in-rush current) and can behave as charge reservoir once the



(a)

(b)

Figure 6.6: Spatial distribution of the peak power supply noise on the bottom plane for scenario 3 (first two planes are power gated and the bottom plane is active): (a) traditional topology, (b) reconfigurable and always-on topologies.

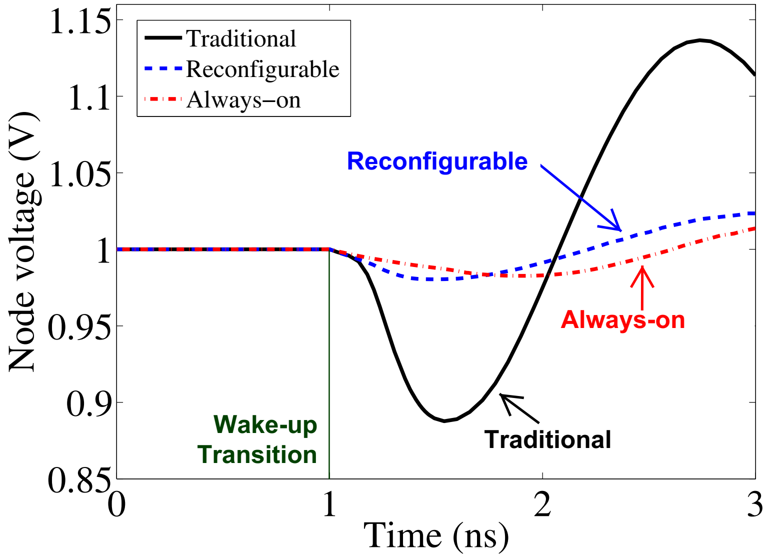


Figure 6.7: Transient behavior of power gating noise at a specific node within the bottom plane for each topology for scenario 3 (top plane is power gated and middle plane transitions from off to on state at 1 ns).

plane transitions to active state. The transient behavior of the power gating noise is illustrated in Fig. 6.7 for scenario 3. The middle plane transitions from sleep to active state at 1 ns and the in-rush current noise is observed on the bottom plane.

6.3.2.3 Turn-on Time

Turn-on time for each topology is investigated. A gradual wake-up strategy described in the previous subsection is adopted. Scenario 3 is considered where the middle plane is turned on while the top plane is power gated and bottom plane is active. Power supply voltage variation on the virtual grid of one of the circuit blocks is illustrated in Fig. 6.8 during the wake-up process. The wake-up time is determined when the voltage reaches 90% of the nominal V_{DD}. For the traditional topology, the wake-up time of the circuit block is 0.80 ns. Alternatively, with the

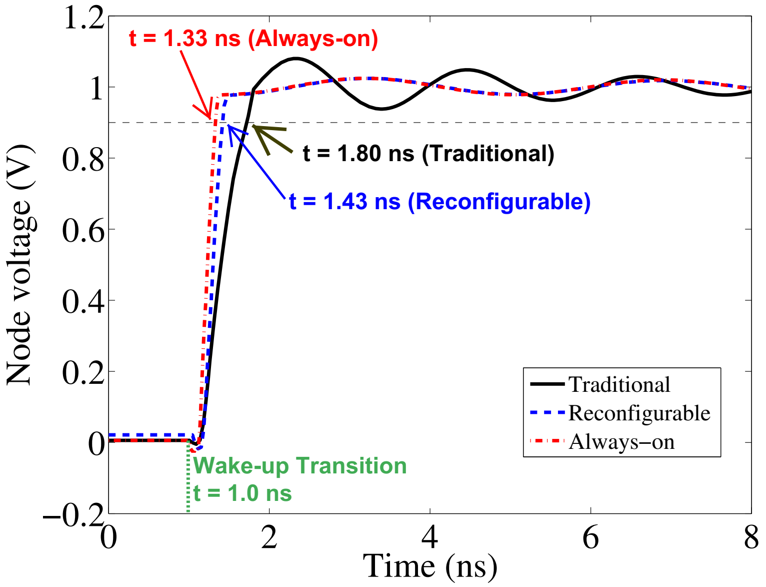


Figure 6.8: Power supply voltage variation on the virtual grid of one of the circuit blocks (located within the middle plane) during the wake-up process.

reconfigurable and always-on topologies, the wake-up time is reduced, respectively, to 0.43 ns and 0.33 ns. As mentioned before, in these topologies, the decoupling capacitors within a power gated plane remain charged, thereby reducing the turn-on time. As mentioned in [97], a shorter wake-up time enables larger leakage power savings. The overall time required to turn on the entire plane is 1.17 ns, 0.84 ns, and 0.74 ns for, respectively, traditional, reconfigurable and always-on topologies.

6.3.2.4 Effect of Number of Planes

The effect of the number of planes on the efficacy of the decoupling capacitor topologies is discussed. The simulation setup of three-plane 3D IC is extended to increase the number of planes with the same physical characteristics. Switching circuit loads within the top plane (closest to the package) are maintained active whereas the additional planes beneath the first plane are power gated. Peak power supply noise on the top plane is illustrated in Fig. 6.9 as the number of power gated planes increases. As shown in this figure, for the reconfigurable and always-on topologies, the decoupling capacitors within the second and third planes are highly effective in reducing the supply noise of the top plane. If, however, the number of planes further increases, the supply noise starts to slightly increase. The decoupling capacitors within the fourth and farther planes are not effective for the top plane due to greater impedance. Since these capacitors are implemented as MOS-C and are connected to the global grid, the overall current drawn from the power supply slightly increases with increasing number of planes (due to nonnegligible MOS-C leakage current). Thus, power noise of the top plane slightly increases if the number of planes increases beyond three. For the traditional topology, power noise slightly decreases with increasing number of power gated planes due to the parasitic capacitance of the power grid and TSVs within the power gated planes. Note that the negative impact of high MOS-C leakage current can be alleviated if capacitors are implemented with the MIM technique. This reduction in leakage current is achieved at the expense of a significant increase in physical area, as analytically determined in Section 6.3.2.1.

6.3.2.5 Power Overhead

It is important to quantify the power overhead of the decoupling capacitor topologies to ensure that the proposed topologies do not undermine the reduction in leakage current. Each topology is simulated for each scenario and the overall average power consumption is determined. All of the decoupling capacitors are implemented as MOS-C. Results are listed in Table 6.8. The smallest overhead occurs in

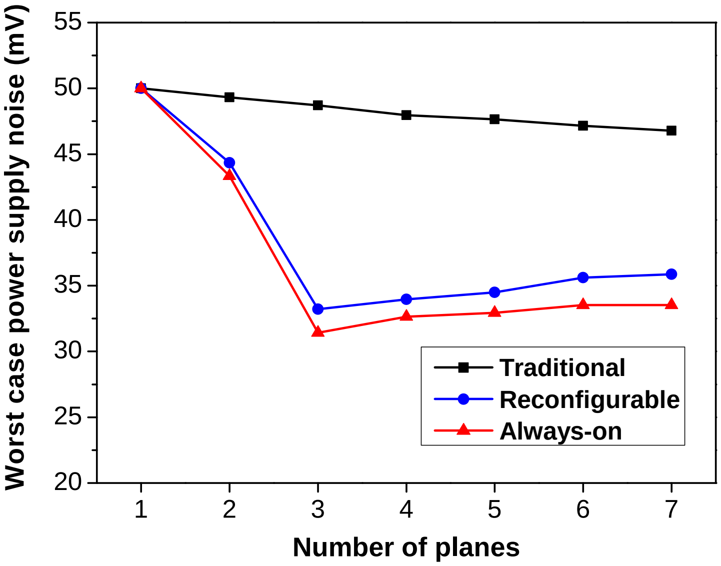


Figure 6.9: Peak power supply noise on the top plane as the number of power gated planes increases.

scenario 1 where all of the planes are active. This overhead is due to increased capacitance and switches (for the reconfigurable topology only). Power overhead increases in scenarios 3 and 5 where two planes are power gated. Specifically, for the reconfigurable topology, power overhead is approximately 5% due to the leakage current of MOS capacitors that are connected to the global grid. For the always-on topology, the power overhead increases to approximately 10% since more capacitance is required in this topology. Thus, reconfigurable topology exhibits less power overhead than the always-on topology. Note that if MIM capacitors are utilized, the power overhead can be significantly reduced. For example, assuming a leakage current of 1 nA/cm^2 for an MIM capacitor based on [101], the power consumption increases by only 1.25% and 1.38%, respectively, for the reconfigurable and always-on topologies in Scenario 3. This small power overhead is achieved at the expense of a significant increase in area, as listed in Table 6.2.

	Traditional	**Reconfigurable**		**Always-on**	
	Power (mW)	Power (mW)	Overhead (%)	Power (mW)	Overhead (%)
Scenario 1	26.56	27.01	1.69%	27.14	2.18%
Scenario 2	17.62	18.15	2.99%	18.52	5.11%
Scenario 3	8.86	9.43	6.46%	9.80	10.64%
Scenario 4	17.53	18.12	3.37%	18.35	4.68%
Scenario 5	8.98	9.60	6.90%	9.86	9.80%

Table 6.8: Overall Average Power Consumption (when all decoupling capacitors are implemented as MOS capacitors).

6.4 Summary

3D ICs are expected to be heavily power gated due to higher integration and substantial subthreshold leakage current in modern CMOS processes. In 3D ICs with power gating, system-wide power integrity can be compromised if traditional decoupling capacitor placement topology is utilized since these capacitors are typically placed sufficiently close to the switching circuit, *i.e.*, connected to the virtual power network. When a block within a plane or the entire plane is power gated, related decoupling capacitors cannot provide charge to the neighboring, active planes, degrading both power supply noise and power gating noise. Two characteristics of TSVs are exploited to alleviate this issue: (1) low resistivity and (2) ability to bypass plane-level power network when delivering the power supply voltage to farther planes. Utilizing these two characteristics, two decoupling capacitor topologies are investigated with significant reductions in power supply and gating noise at the expense of a moderate increase in area and power consumption. The turn-on time of the proposed topologies and the effect of number of planes on the efficacy of these topologies are also investigated.

Chapter 7

Compact Model to Efficiently Characterize TSV-to-Transistor Noise Coupling in 3D ICs

As discussed in Chapter 2, in addition to the issue of power integrity, another critical challenge in 3D ICs is to ensure system-wide signal integrity, which is exacerbated due to the multiple tiers interconnected with TSVs. The TSV-to-substrate coupling is a unique noise generation mechanism in 3D ICs. Specifically, when the signal transmitted by a TSV transitions from high to low or low to high, noise couples from TSV into the substrate due to both dielectric and depletion capacitances. The coupling noise propagates throughout the substrate and affects the reliability of nearby transistors. This issue is exacerbated for TSVs that carry signals with high switching activity factors and fast transitions such as clock signals. In this chapter, TSV-induced substrate noise is investigated. To characterize the TSV-induced

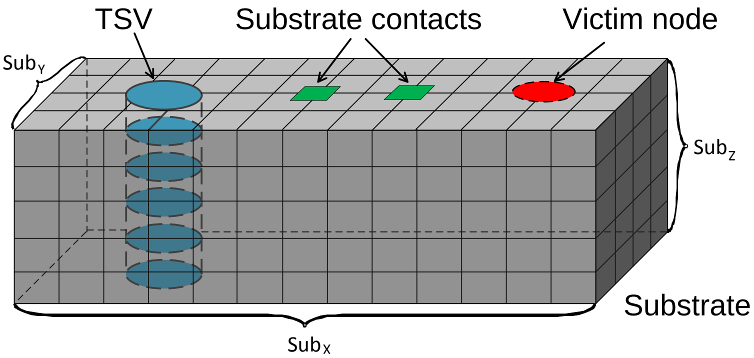


Figure 7.1: Physical structure used to analyze TSV induced noise coupling.

noise coupling, a highly distributed electrical model is described in Section 7.1. A compact π model is proposed in Section 7.2 for efficient estimation of TSV induced noise at a victim transistor. Each admittance within the compact model is expressed in Section 7.3 as a function of multiple physical parameters. This approach permits to consider different substrate biasing schemes and TSV types. Design guidelines are provided in Section 7.4 based on the analysis results obtained from the compact model.

7.1 Distributed Model for Characterizing TSV Induced Noise Coupling

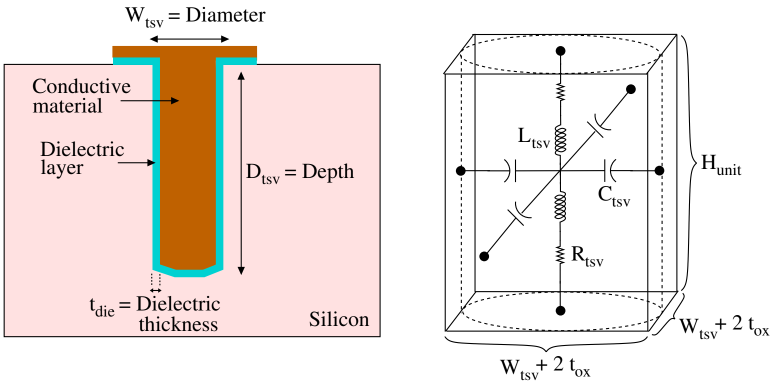
To characterize TSV induced noise coupling as a function of multiple design parameters, the physical structure depicted in Fig. 7.1 is used. This structure consists of a noise injector (TSV), noise transmitter (substrate), and a noise receptor (victim transistor). Substrate contacts are also included to bias the substrate. Note that the number and placement of substrate contacts between the TSV and victim transistor play an important role in the noise coupling analysis and safe zone characterization, as demonstrated in this chapter.

To analyze this physical structure, several approaches have been adopted such as using an electromagnetic field solver, device simulator, and a highly distributed model using 3D TLM method [103–105]. In the distributed model, the physical structure is discretized into unit cells (for both TSV and substrate) and each unit cell is modeled with lumped parasitic impedances. A distributed model based on 3D-TLM is described in this section. This model is used as a reference to validate the proposed compact model (see Section 7.2) and closed-form expressions (see Section 7.3). The TSV and substrate models are described, respectively, in Sections 7.1.1 and 7.1.2. The advantages and limitations of a 3D-TLM based distributed model are discussed in Section 7.1.3.

7.1.1 TSV Model

A typical TSV is represented as a cylinder with a diameter and depth, as illustrated in Fig. 7.2(a). Two primary components of a TSV are (1) conductive filling material such as polysilicon, tungsten or copper (varies depending upon the specific TSV fabrication technology), (2) a dielectric layer that surrounds the conductive part to prevent the filling material from diffusing into the silicon [30].

A TSV unit cell consisting of parasitic resistance R_{tsv}^{unit}, parasitic inductance L_{tsv}^{unit}, and capacitance to substrate C_{tsv}^{unit} is illustrated in Fig. 7.2(b) [33]. The x and y dimensions of the unit cell are both equal to $W + 2t_{ox}$, as determined by the TSV diameter W and thickness of the oxide layer t_{ox}. The z dimension is equal to H_{unit}, as determined by the TSV height and the required resolution in the transmission line matrix method.



(a)

(b)

Figure 7.2: TSV representations: (a) cross-section of a TSV consisting of a conductive material and dielectric layer, (b) electrical model of a unit TSV cell used for discretization.

Considering the skin effect, the unit TSV resistance R_{tsv}^{unit} is determined by [37]

$$R_{tsv}^{unit} = \frac{1}{2}\sqrt{(R_{AC}^{tsv,unit})^2 + (R_{DC}^{tsv,unit})^2}, \tag{7.1}$$

where the DC resistance $R_{DC}^{tsv,unit}$ and AC resistance $R_{AC}^{tsv,unit}$ are, respectively,

$$R_{DC}^{tsv,unit} = \frac{1}{2}\frac{\rho_f H_{unit}}{\pi(W/2)^2}, \tag{7.2}$$

$$R_{AC}^{tsv} = \frac{\rho_f H_{unit}}{4\pi(W/2)\delta_{tsv}}. \tag{7.3}$$

ρ_f is the resistivity of the filling material and the skin depth δ_{tsv} is

$$\delta_{tsv} = \sqrt{\frac{\rho_f}{\pi f \mu_f}}, \tag{7.4}$$

where f is the frequency and μ_f is the permeability of the filling material. The unit

TSV inductance L_{tsv}^{unit} is

$$\begin{aligned} L_{tsv}^{unit} &= \frac{1}{2}\frac{\mu_o}{4\pi}[2H_{unit}\ln(\frac{2H_{unit}+\sqrt{(W/2)^2+(2H_{unit})^2}}{W/2})+ \\ & (W/2-\sqrt{(W/2)^2+(2H_{unit})^2})], \end{aligned} \tag{7.5}$$

where μ_o is vacuum permeability. The unit TSV capacitance C_{tsv}^{unit} has two series components: oxide and depletion capacitance. The oxide capacitance is determined from the cylindrical capacitor formula as [38]

$$C_{ox}^{unit} = \frac{1}{4}\frac{2\pi\varepsilon_{ox}H_{unit}}{\ln(\frac{W/2+t_{ox}}{W/2})}, \tag{7.6}$$

where ε_{ox} is the oxide permittivity. Assuming that the TSV voltage is at V_{DD}, TSV depletion capacitance is [34]

$$C_{dep}^{unit} = \frac{1}{4}\frac{2\pi\varepsilon_s H_{unit}}{\ln(\frac{W/2+t_{ox}+t_{dep}}{W/2+t_{ox}})}, \tag{7.7}$$

where ε_s is the dielectric permittivity of silicon and t_{dep} is the depletion width within the substrate when the TSV voltage is at V_{DD}. The overall TSV capacitance is

$$C_{tsv}^{unit} = \frac{1}{4}\frac{C_{ox}^{unit}C_{dep}^{unit}}{C_{ox}^{unit}+C_{dep}^{unit}}. \tag{7.8}$$

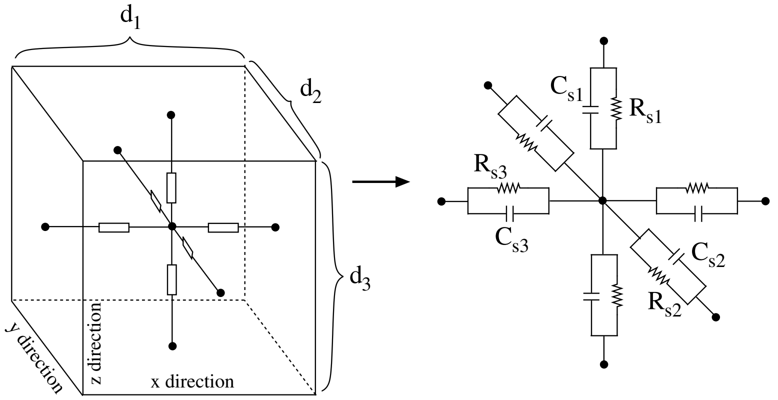


Figure 7.3: Distributed model of a substrate network where each unit cell is represented by six resistances and capacitances.

7.1.2 Substrate Model

A lightly doped bulk type substrate is assumed. Note that an epi type substrate with a heavily doped bulk beneath the lightly doped silicon layer typically produces greater noise coupling. Thus, epi type substrate is less applicable to 3D heterogeneous integration where circuits with distinct electrical characteristics coexist. Also note that a lightly doped silicon substrate produces a lower TSV capacitance due to a larger depletion width [34].

A similar discretization technique is applied to model the lightly doped substrate. A unit substrate cell consisting of six parallel *RC* admittances is illustrated in Fig. 7.3. Referring to this figure, the three substrate resistances R_{s1}, R_{s2}, and R_{s3} are, respectively,

$$R_{s1} = \frac{1}{2}\frac{\rho_s d_3}{d_1 d_2}, \tag{7.9}$$

$$R_{s2} = \frac{1}{2}\frac{\rho_s d_2}{d_1 d_3}, \tag{7.10}$$

$$R_{s3} = \frac{1}{2}\frac{\rho_s d_1}{d_2 d_3}, \tag{7.11}$$

where ρ_s is the substrate resistivity. Similarly, the three substrate capacitances C_{s1}, C_{s2}, and C_{s3} are, respectively,

$$C_{s1} = 2\frac{\varepsilon_s d_1 d_2}{d_3}, \tag{7.12}$$

$$C_{s2} = 2\frac{\varepsilon_s d_3 d_1}{d_2}, \tag{7.13}$$

$$C_{s3} = 2\frac{\varepsilon_s d_2 d_3}{d_1}. \tag{7.14}$$

7.1.3 Accuracy and Limitations of the Distributed Model

The TSV and the substrate unit cells are combined to produce a highly distributed mesh based on 3D-TLM. Substrate contacts are also considered in the model to properly bias the substrate.

Previous studies have demonstrated the accuracy of the distributed model using 3D-TLM [103, 106]. In [106], a fabricated test vehicle using an industrial via-last TSV technology is used to measure TSV induced noise coupling. The transfer function from TSV to victim transistor is measured and compared with the transfer function obtained from the distributed model. The comparison results demonstrate that below 100 MHz, the 3D-TLM matches reasonably well with the experimental results where the error is less than 1 dB. As the frequency increases (up to 10 GHz), the discrepancy increases, but remains within 3 dB. Similarly, in [103], a 3D field solver is used to analyze TSV-to-TSV noise coupling. The result is compared with the distributed 3D-TLM model. The error in the noise transfer function is within 2 dB until approximately 10 GHz. Both the measurement and 3D field solver results demonstrate that the 3D-TLM model can accurately model the 3D physical structure including a TSV, substrate contact, and victim transistor.

Despite the reasonable accuracy achieved by the distributed model, the computational complexity is significantly high, particularly when the dimensions of the unit cells are small. This issue is exacerbated as the distance between the TSV and victim transistor increases. Furthermore, the number and location of the substrate contacts play an important role in characterizing the TSV safe zone. Re-analysis of the distributed structure when these characteristics change is computationally prohibitive. Therefore, a compact model is proposed to alleviate these limitations, as described in the following section. A highly distributed model based on a 3D-TLM method is used as a reference to evaluate the accuracy of the proposed compact model.

7.2 Compact Π Model For Efficient TSV Noise Coupling Analysis

A two-port, linear time-invariant network can be generally characterized with four admittances: $Y_{11}(j\omega)$, $Y_{12}(j\omega)$, $Y_{21}(j\omega)$, and $Y_{22}(j\omega)$. Utilizing this characteristic, the proposed compact model consists of a single TSV cell and an equivalent two-port π network to model noise propagation, as depicted in Fig. 7.4. Each electrical element within the π network consists of a parallel RC circuit, producing an admittance $(1/R) + j\omega C$. These admittances can be obtained from the distributed mesh (based on 3D-TLM method), as described in the previous section. Specifically, the four $Y(j\omega)$ parameters of the distributed mesh are obtained through an

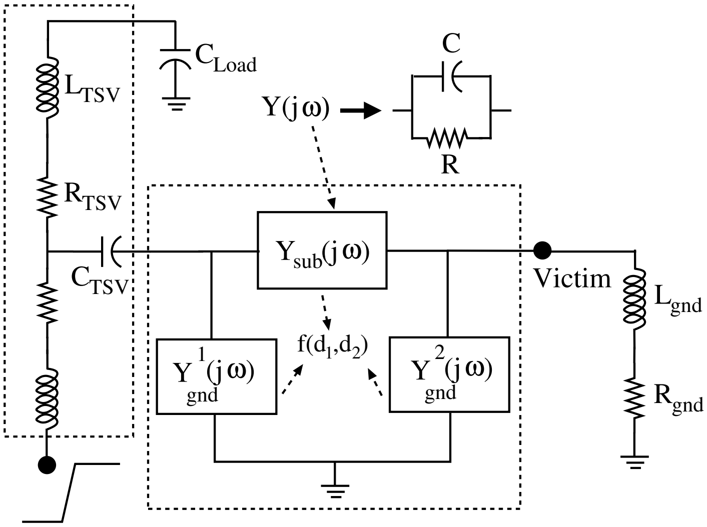


Figure 7.4: Compact π model to efficiently estimate the noise at the victim node in the presence of a TSV and substrate contacts.

AC analysis. The resistances and capacitances within the π network are determined such that the four $Y(j\omega)$ parameters of the compact π network are equal to the respective $Y(j\omega)$ parameters of the distributed mesh. According to this procedure, the admittances within the π network $Y_{sub}(j\omega)$, $Y^1_{gnd}(j\omega)$, and $Y^2_{gnd}(j\omega)$ are determined as follows:

- $Y_{sub}(j\omega) = (1/R_{sub}) + j\omega C_{sub} = Y_{21}(j\omega)$: represents the equivalent substrate admittance between the TSV and victim transistor.
- $Y^1_{gnd}(j\omega) = (1/R^1_{gnd}) + j\omega C^1_{gnd} = Y_{11}(j\omega) - Y_{21}(j\omega)$: represents the equivalent substrate admittance between the TSV and ground node.
- $Y^2_{gnd}(j\omega) = (1/R^2_{gnd}) + j\omega C^2_{gnd} = Y_{22}(j\omega) - Y_{21}(j\omega)$: represents the equivalent substrate admittance between the victim node and ground node.

Note that in this study, Y_{11}, Y_{12}, Y_{21}, and Y_{22} are obtained by simulating the distributed mesh. Another approach is to obtain these Y parameters directly from a

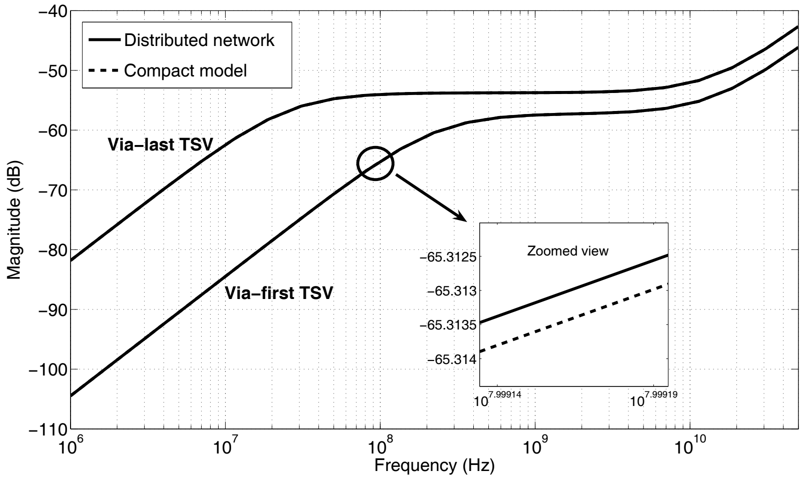


Figure 7.5: Comparison of the proposed compact π model with high complexity distributed mesh for both via-first and via-last TSVs. The solid line represents noise at the victim node obtained from distributed mesh whereas the dashed line represents noise at the victim node obtained from the compact π model.

3D field solver or measurement results. Also note that a single *RC* value is chosen for each admittance since the variation of the resistance and capacitance with frequency is negligible in the frequency range of interest. Specifically, the maximum change is less than 0.1% up to 100 GHz.

7.2.1 Accuracy Analysis

The accuracy of the compact model is demonstrated by comparing the transfer function of the compact model with the transfer function of the distributed mesh with significantly higher complexity. Assuming a lightly doped substrate (with 10 Ω·cm resistivity and 103.4×10^{-12} F/m absolute permittivity), the two transfer functions are compared in Fig. 7.5 for both via-first and via-last TSVs. In the distributed model, the dimensions L_{sub}, W_{sub}, and H_{sub} of the unit substrate cell are each 1 μm. The distance between the TSV and victim node is 10 μm and a single substrate contact is placed in the middle of the two ports. The overall length of the substrate is 100 μm. The height of the substrate (determined by the TSV height) is 10 μm for a via-first TSV and 50 μm for a via-last TSV. Alternatively, the width of the substrate (partly determined by the TSV diameter) is 5 μm for a via-first TSV and 12 μm for a via-last TSV. Note that via-last TSVs have greater dimensions as compared to via-first TSVs, significantly affecting the noise at the victim node, as further discussed in Section 7.4.

As illustrated in Fig. 7.5, noise coupling due to TSVs is accurately estimated by the compact model with negligible error within the frequency range of interest. Note that the noise magnitude at the victim node is higher for via-last TSVs due to higher TSV capacitance and greater substrate dimensions. This difference is more than 20 dB at low frequencies and decreases to approximately 4 dB in the gigahertz range.

7.2.2 Complexity Analysis

For a via-first TSV unit cell, the x and y dimensions are both equal to 4.4 μm ($W + 2t_{ox}$ where $W = 4$ and $t_{ox} = 0.2$), whereas the z dimension is 1 μm. Alternatively, for a via-last TSV unit cell, the x and y dimensions are equal to 10.4 μm ($W + 2t_{ox}$ where $W = 10$ and $t_{ox} = 0.2$) and the z dimension is 1 μm. In the distributed model with a via-first TSV, these dimensions produce 60,080 number of circuit elements (resistance, capacitance, and inductance). For a via-last TSV, this number increases to 720,400 due to greater y and z dimensions of the substrate. Alternatively, the compact π model contains only 11 number of elements for both via-first and via-last TSVs.

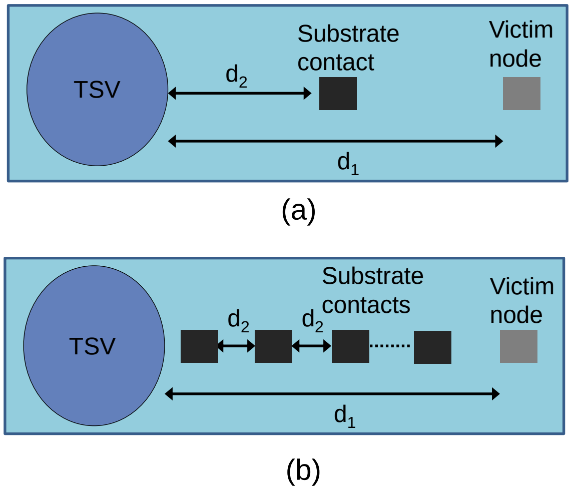
Sufficient accuracy and significantly lower complexity of the proposed compact model support the analysis of TSV induced noise. To consider the effect of various design parameters on coupling noise, each *RC* element within the compact model is expressed as a function of two physical design parameters, as described in the following section.

7.3 TSV Safe Zone Characterization

To determine TSV safe zone, the dependence of TSV induced noise on design parameters such as distance between TSV and victim node, and the number and location of substrate contacts should be characterized. Two substrate biasing schemes are considered.

In the first scheme, as depicted in Fig. 7.6(a), a single substrate contact is placed between the TSV and victim node. The physical distance between the TSV and victim node is d_1 and the distance between the TSV and substrate contact is d_2. In the second scheme, as depicted in Fig. 7.6(b), substrate contacts are regularly placed between the TSV and victim node. In this case, d_2 refers to the distance between each substrate contact. The second scenario is considered since substrate contacts can be regularly placed in an automated manner based on latch-up constraints of the technology [107]. These two scenarios are separately investigated for both via-first and via-last TSVs, producing four different cases, as summarized below:

- Case 1: via-first TSV with a single substrate contact between TSV and victim node
- Case 2: via-first TSV with regularly placed substrate contacts between TSV and victim node



(a)

(b)

Figure 7.6: Two substrate biasing schemes used to characterize noise coupling: (a) single substrate contact between TSV and victim node, (b) regular placement of the substrate contacts between TSV and victim node.

- Case 3: via-last TSV with a single substrate contact between TSV and victim node
- Case 4: via-last TSV with regularly placed substrate contacts between TSV and victim node

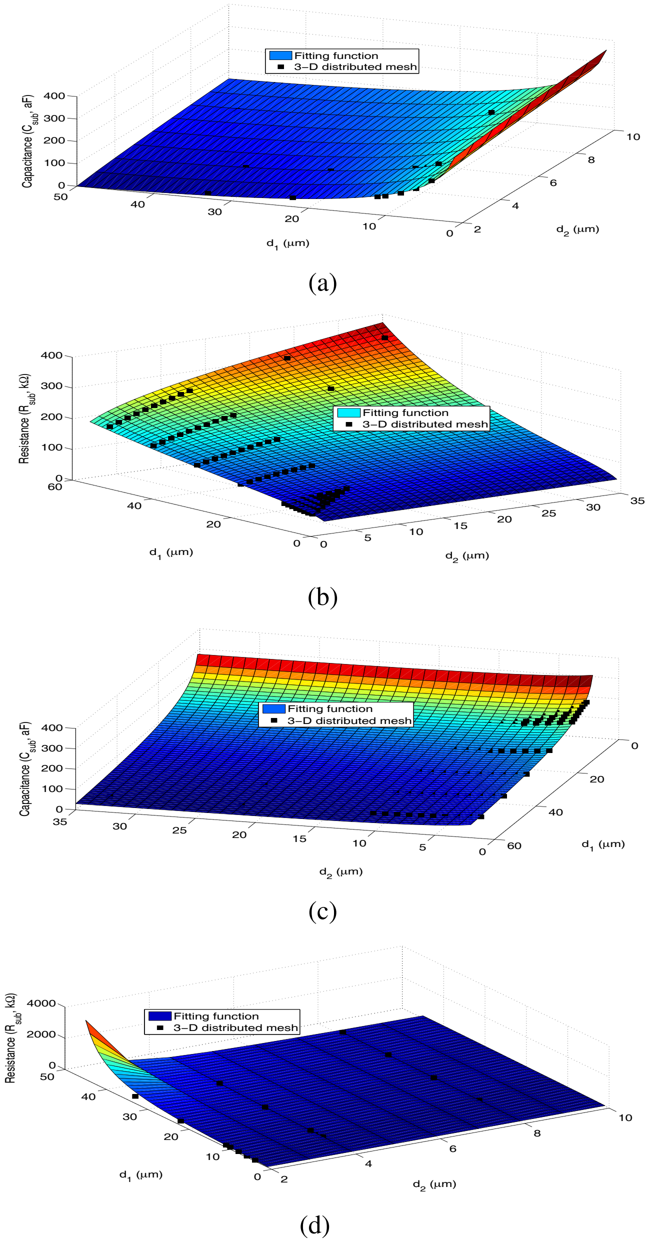
For each case, the $Y(j\omega)$ parameters of the π network are characterized as a function of d_1 and d_2. To evaluate these dependencies, AC analyses of the distributed mesh (based on 3D-TLM) described in Section 7.1 are performed with different values of d_1 and d_2. Note that a 3D field solver can also be used to perform these analyses. The data obtained in this step are used to generate a 3D surface for each resistance and capacitance within $Y_{sub}(j\omega)$, $Y^1_{gnd}(j\omega)$, and $Y^2_{gnd}(j\omega)$. This surface is approximated with a logarithmic function using a 3D least square regression analysis. The logarithmic function $F(d_1, d_2)$ used to approximate the admittances of the π network as a function of the physical distances d_1 and d_2 is

$$F(d_1,d_2) = A + Bd_1 + Cd_2 + D\ln d_2 + E\ln d_1, \tag{7.15}$$

where A, B, C, D, and E are fitting coefficients. These fitting coefficients are determined such that the resistor/capacitor value obtained from this expression reasonably approximates the actual resistor/capacitor value (in the compact π model) that is obtained from the highly distributed 3-D TLM model (or a field solver). Note that both the resistance (in kilo Ωs) and capacitance (in atto Farads) of each $Y(j\omega)$ within the π network are represented by the function F. Also note that the distances d_1 and d_2 are in μm. Since the π network has three admittances each consisting of a parallel RC circuit, six logarithmic functions are developed for each case, producing a total of 24 functions. The fitting coefficients for each function are listed in Table 7.1.

As an example, the resistance R_{sub} and capacitance C_{sub} of the $Y_{sub}(j\omega)$ are plotted, respectively, in Figs. 7.7(a) and 7.7(b) for a via-first TSV with a single substrate contact (case 1). The same parameters are plotted for a via-last TSV with regularly placed substrate contacts (case 4) in Figs. 7.7(c) and 7.7(d). The dotted points represent the data obtained from the analysis of the distributed mesh and the surface represents the function F that approximates these data. The procedure is similar for other cases and the RC elements of the remaining admittances [$Y^1_{gnd}(j\omega)$ and $Y^2_{gnd}(j\omega)$] within the compact model. Note that in all cases, d_1 is greater than d_2 since substrate contacts are placed between the TSV and victim node.

The sufficient accuracy of the fitting method is demonstrated by quantifying the average percent error (as compared to the distributed mesh based on 3D-TLM) for



(a)

(b)

(c)

(d)

Figure 7.7: Comparison of the data obtained from the analysis of the distributed mesh with the function F that approximates these data: (a) resistance R_{sub} of the $Y_{sub}(j\omega)$ with a via-first TSV and a single substrate contact (case 1), (b) capacitance C_{sub} of the $Y_{sub}(j\omega)$ with a via-first TSV and a single substrate contact (case 1), (c) resistance R_{sub} of the $Y_{sub}(j\omega)$ with a via-last TSV and regularly placed substrate contacts (case 4), and (d) capacitance C_{sub} of the $Y_{sub}(j\omega)$ with a via-last TSV and regularly placed substrate contacts (case 4).

Cases	Admittances	Fitting coefficients					Average error (%)
		A	B	C	D	E	
Case 1	$R_{sub} = 1000/F$ (kΩ)	27.09	0	0	-0.98	-5.11	6.6
	$C_{sub} = F$ (aF)	326.7	0.41	0.48	-17.26	-67.55	2.8
	$R^1_{gnd} = 1000/F$ (kΩ)	28.31	0	0	-8.14	1.98	1.9
	$C^1_{gnd} = F$ (aF)	301.3	-0.28	0.66	-96.94	30.09	1.6
	$R^2_{gnd} = F$ (kΩ)	45.91	2.30	3.08	-218.3	154.9	9.6
	$C^2_{gnd} = 1000/F$ (aF)	8.05	0.28	0.29	-20.39	12.49	10.4
Case 2	$R_{sub} = 1000/F$ (kΩ)	29.18	0.28	0.38	1.34	-11.78	8.3
	$C_{sub} = F$ (aF)	317.1	2.99	4.24	14.2	-127.6	10.8
	$R^1_{gnd} = 1000/F$ (kΩ)	69.24	-0.046	1.97	-35.05	4.73	0.7
	$C^1_{gnd} = F$ (aF)	758.5	-0.49	21.13	-380.8	51.33	0.7
	$R^2_{gnd} = 1000/F$ (kΩ)	9.50	-0.19	-0.99	-1.95	9.31	1.6
	$C^2_{gnd} = F$ (aF)	106.4	-2.05	-10.99	-20.83	100.7	1.6
Case 3	$R_{sub} = 1000/F$ (kΩ)	27.35	-0.082	0.036	-2.11	-1.12	2.4
	$C_{sub} = F$ (aF)	296.6	-0.89	0.83	-20.78	-13.12	1.9
	$R^1_{gnd} = 1000/F$ (kΩ)	36.16	0.028	0.39	-10.16	1.18	2.4
	$C^1_{gnd} = F$ (aF)	235.6	-1.18	-2.16	-61.86	51.91	6.9
	$R^2_{gnd} = 1000/F$ (kΩ)	11.57	0.11	-0.19	4.43	-5.86	11.6
	$C^2_{gnd} = F$ (aF)	129.1	1.18	-2.22	49.1	-65.23	10.9
Case 4	$R_{sub} = 1000/F$ (kΩ)	24.41	0.13	0.42	1.69	-8.36	8.5
	$C_{sub} = F$ (aF)	265.3	1.40	4.67	17.88	-91.2	7.2
	$R^1_{gnd} = 1000/F$ (kΩ)	117.6	-0.03	0.40	-35.66	3.64	0.3
	$C^1_{gnd} = F$ (aF)	998	0.03	-68.78	37.31	32.11	2.9
	$R^2_{gnd} = 1000/F$ (kΩ)	14.56	-0.01	-0.73	-3.49	5.93	2.4
	$C^2_{gnd} = F$ (aF)	162	-0.13	-8.17	-37.74	64.26	2.5

Table 7.1: Fitting coefficients for the function F that approximates the admittances within the compact model (see Fig. 7.4) for each case. The function F is given by (7.15).

each resistance and capacitance within the π network. Specifically, for each case, d_1 and d_2 are varied and the highly distributed model is simulated to obtain the RC values within the π model. The difference between these RC values and those obtained by (7.15) determines the error. The average error is listed in the last column of Table 7.1 for each case. Note that the fitting coefficients listed in Table 7.1 are obtained for a certain range of d_1 and d_2. Specifically, for case 1 and case 3 (where a single substrate contact exists between the TSV and victim node), d_1 (distance between TSV and victim node) varies from 4 μm to 55 μm and d_2 (distance between TSV and substrate contact) varies from 2 μm to 33 μm. Alternatively, for case 2 and case 4 (where multiple substrate contacts are regularly placed between the TSV and victim node), d_1 varies from 4 μm to 44 μm and d_2 (distance between two substrate contacts) varies from 2 μm to 8 μm. Note that the maximum average error is slightly over 10% for certain resistances and capacitances. This error, however, does not significantly affect the electrical characteristics (and noise estimation at the victim node) since the maximum error occurs at the extreme cases when the resistance is sufficiently large and capacitance is sufficiently small. Also note that the average error over four cases is 4.8%. The proposed model and the function F can be used to efficiently characterize TSV-to-transistor noise coupling, as discussed in the following section.

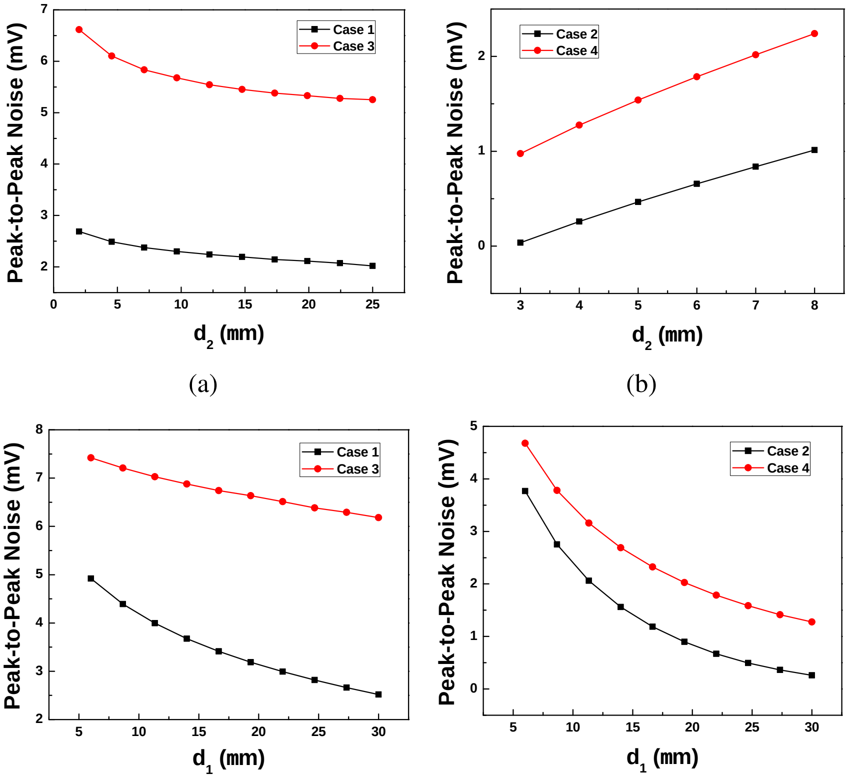
7.4 Design Guidelines

The compact model illustrated in Fig. 7.4 and fitting parameters obtained in Section 7.3 are used to investigate the effect of various design and fabrication parameters such as placement of substrate contacts and TSV type (see Section 7.4.1), slew rate (see Section 7.4.2) and voltage swing (see Section 7.4.3) of the TSV signals, and differential signaling (see Section 7.4.4). Design guidelines are developed based on the analysis results to improve signal integrity in TSV based 3D ICs.

7.4.1 Placement of Substrate Contacts

Peak-to-peak noise at the victim transistor due to TSV activity is analyzed using (7.15) and the compact model. This noise is depicted in Figs. 7.8(a) and 7.8(b) as a function of d_2 when d_1 is constant at 30 μm.

According to Fig. 7.8(a), where a single substrate contact exists between the



(a)

(b)

(c)

(d)

Figure 7.8: TSV induced switching noise at the victim node: (a) as a function of d_2 at constant d_1 for case 1 and case 3, (b) as a function of d_2 at constant d_1 for case 2 and case 4, (c) as a function of d_1 at constant d_2 for case 1 and case 3, and (d) as a function of d_1 at constant d_2 for case 2 and case 4.

TSV and victim node, switching noise is reduced as the substrate contact is placed closer to the victim node as opposed to the TSV. This characteristic is due to TSV height and distributed TSV capacitance to substrate. Thus, a single substrate contact closer to the TSV is not sufficiently effective since noise is injected into the substrate along the entire TSV depth. Note that based on Fig. 7.8(a), this characteristic is stronger in via-last TSVs since the height of a via-last TSV is five times greater than a via-first TSV. In traditional 2D circuits, it is typically a physical design decision to place the substrate contacts (or guard rings) around an aggressor noise source or around a sensitive victim block. In 3D circuits where TSVs are primary source of switching noise, placing the substrate contacts closer to the victim block is more advantageous, as demonstrated in Fig. 7.8(a).

According to Fig. 7.8(b), where multiple, regularly placed substrate contacts exist between the TSV and victim node, switching noise is significantly less as compared to Fig. 7.8(a) and is further reduced as d_2 decreases, *i.e.*, number of substrate contacts increases. Also note that in both figures, switching noise due to via-last TSVs is significantly greater than via-first TSVs since the diameter is larger and height is longer.

Peak-to-peak switching noise at the victim transistor is shown in Figs. 7.8(c) and 7.8(d) as a function of d_1 when d_2 is constant at 4 μm. As illustrated in Fig. 7.8(c), when only a single substrate contact exists, placing the victim transistor farther from the switching TSV is an effective method for via-first TSVs. Alternatively, for via-last TSVs, the noise exhibits low sensitivity to the distance between TSV and victim transistor. This phenomenon is due to longer height (therefore smaller substrate resistances) and larger diameter (therefore larger capacitances) of via-last TSVs.

According to Fig. 7.8(d), when multiple substrate contacts are regularly placed, increasing the physical distance between the switching TSV and victim transistor is helpful for both via-first and via-last TSVs. In this case, the effective impedance between the TSV and ground node becomes significantly lower since the number of substrate contacts increases as d_1 is increased.

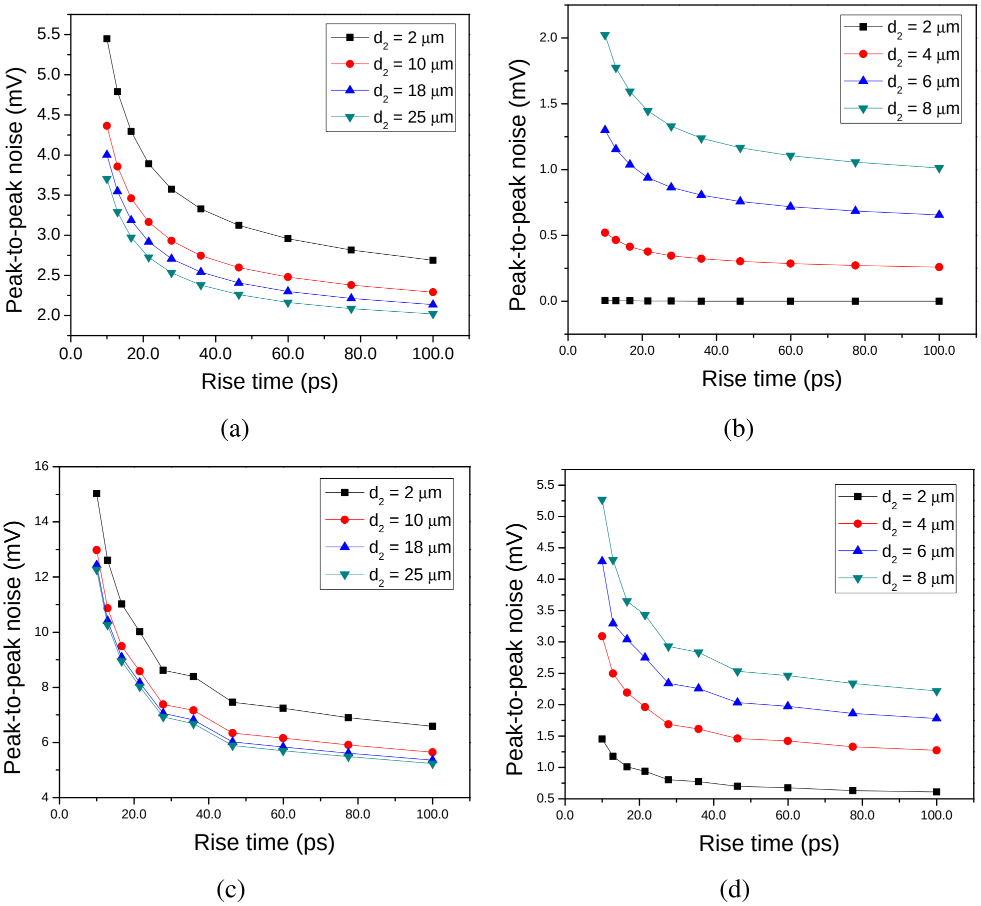
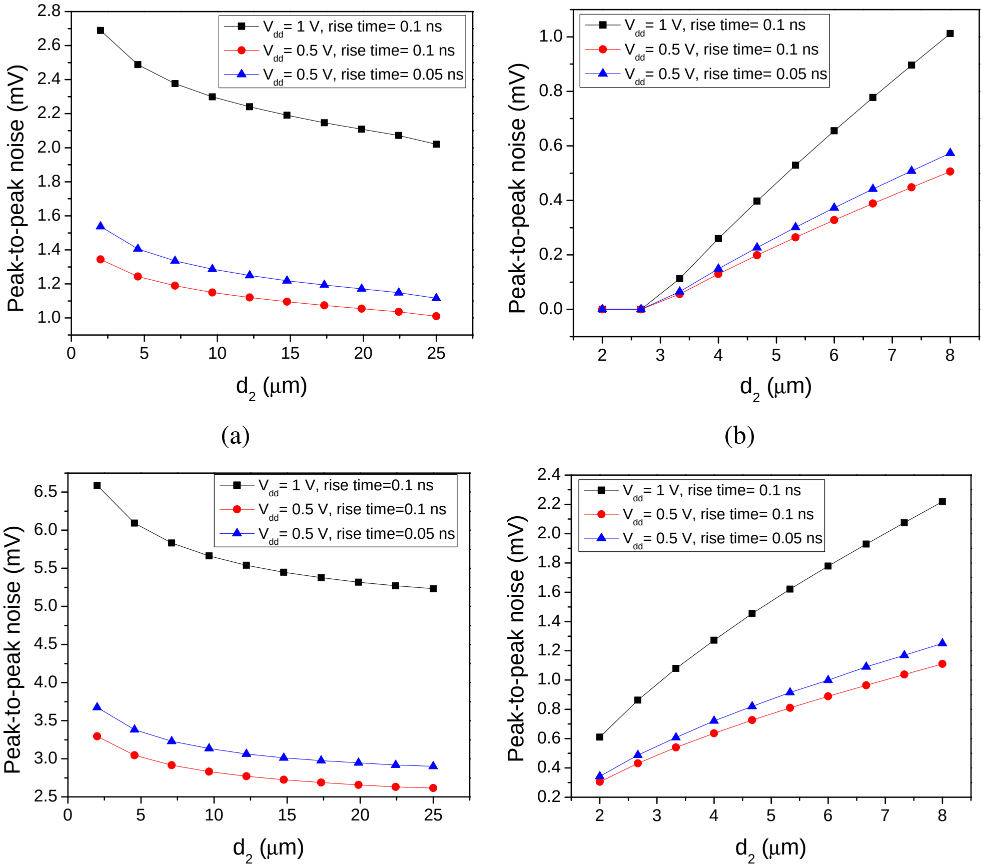


Figure 7.9: TSV induced switching noise at the victim node as a function of rise time: (a) case 1: via-first TSV with a single substrate contact, (b) case 2: via-first TSV with regularly placed substrate contacts, (c) case 3: via-last TSV with a single substrate contact, and (d) case 4: via-last TSV with regularly placed substrate contacts.

7.4.2 Slew Rate of the TSV Signal

Slew rate of a signal within a TSV not only affects the circuit speed and power consumption, but also TSV induced noise coupling into the substrate. Specifically, the transient characteristics of a TSV signal determine the frequency range of interest and therefore the coupling strength between TSV and a victim transistor. In this analysis, slew rate is varied by changing the rise/fall time of a transient signal applied to a TSV, as shown in Fig. 7.4. Each of the four cases described in the previous section are considered. In each case, four typical values of d_2 (distance between TSV and substrate contact in cases 1 and 3, distance between two substrate contacts in cases 2 and 4) are chosen based on the results obtained in the previous section. The voltage swing of the transient signal is constant at 1 V while d_1 (distance between TSV and victim node) is constant at 30 μm. The rise time varies from 10 ps to 100 ps.

As demonstrated in Fig. 7.9, a significant initial reduction in noise is achieved when the rise time increases from 10 ps to approximately 30 ps. For example, in case 1 where a via-first TSV and a single substrate contact exist, if the rise time of the transient signal increases from 10 ps to 30 ps, the peak noise is reduced by 34.4%, 32.7%, 32.2%, and 31.6% when d_2 (distance between TSV and substrate contact) is, respectively, 2 μm, 10 μm, 18 μm, and 25 μm. Similar reduction in noise is also observed for a via-last TSV. In regularly placed substrate contacts (cases 2 and 4), the effect of rise time is weaker as the number of substrate contacts increases.



(a)

(b)

(c)

(d)

Figure 7.10: TSV induced switching noise at the victim node at different voltage swings and rise times : (a) case 1: via-first TSV with a single substrate contact, (b) case 2: via-first TSV with regularly placed substrate contacts, (c) case 3: via-last TSV with a single substrate contact, and (d) case 4: via-last TSV with regularly placed substrate contacts.

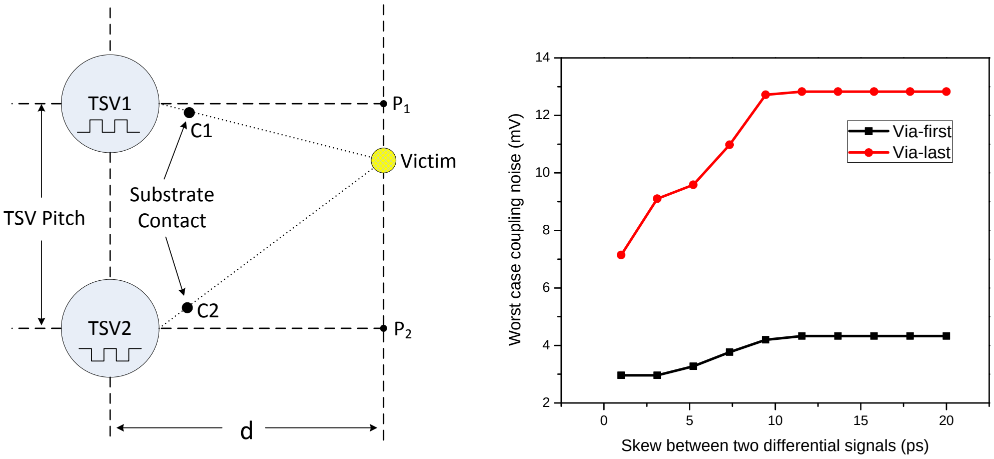
7.4.3 Voltage Swing of the TSV Signal

Another parameter that affects the TSV power, delay, and noise characteristics is the voltage swing. Low swing TSVs reduce both power consumption and delay since the oxide and depletion capacitances require less charge. Noise coupling into the substrate is also reduced due to weaker dv/dt. Since the voltage swing and slew rate are interdependent, three configurations are investigated: (1) V_{dd} = 1 V, rise time = 100 ps, (2) V_{dd} = 0.5 V, rise time = 100 ps; (3) V_{dd} = 0.5 V, rise time = 50 ps. d_1 (distance between TSV and victim node) remains constant at 30 μm. The simulation results for each case are shown in Fig. 7.10.

If the voltage swing is reduced from 1 V to 0.5 V, while the slew rate is the same (rise time is proportionally reduced), peak noise at the victim node is reduced by over 40% for case 1. If the voltage swing remains at 0.5 V and the rise time is increased from 50 ps to 100 ps, peak noise is reduced by an additional 10%. The other three cases exhibit a similar pattern. Thus, if the rise time of a TSV signal is above a certain threshold ($\approx$ 30 ps), reducing the voltage swing is a more effective method to reduce TSV induced noise than increasing the transition time.

7.4.4 Differential TSV Signaling

An active substrate noise reduction method has been proposed in [108] where the phase of the noise is reversed and reinjected into the substrate, producing up to 83% reduction in noise. A similar result can be achieved in TSV related noise coupling through differential signaling, as investigated in this section. For example, differential clocking can significantly reduce the effective noise injected into



(a) (b)

Figure 7.11: Effect of differential signaling on TSV related noise coupling, (a) analysis setup and (b) noise at the victim node as a function skew between the two signals.

the substrate by the TSVs. The proposed compact model is used to evaluate this behavior. As shown in Fig. 7.11(a), two TSVs carrying out-of-phase signals are placed on the substrate where the distance between the TSVs is equal to the minimum pitch (8 μm for via-first and 20 μm for via-last). Distance between each TSV and points P_1 and P_2 (d in the figure) is 30 μm. Worst case noise between points P_1 and P_2 is observed as a function of skew between the two out-of-phase TSV signals. This noise is plotted in Fig. 7.11(b). Note that the effect of substrate contact C1 on TSV2 and the effect of C2 on TSV1 are neglected since the model cannot consider contacts placed outside the trajectory between TSV and victim. This assumption is valid if pitch is sufficiently greater than the distance between TSV and substrate contact. In practice, noise at the victim node is expected to be slightly lower than the estimated value since C1 can filter a small amount of noise that originates from TSV2. The analysis therefore provides a pessimistic estimation.

According to Fig. 7.11(b), the efficiency of differential signaling in reducing TSV related noise is strongly dependent upon the skew between the two out-of-phase signals. If the two signals are almost exactly out-of-phase (1 ps skew), differential signaling achieves 32.1% and 44.9% reduction in peak noise, respectively, for via-first and via-last TSVs. However, when the skew reaches approximately 10 ps, the advantage of differential signaling diminishes. Thus, emphasis should be placed on ensuring small skew if differential signaling is utilized to cancel TSV related noise coupling in 3D ICs.

7.5 Summary

TSV-to-transistor noise coupling has been evaluated and quantified in 3D ICs. A compact π model has been proposed to estimate noise at the victim transistor as a function of different substrate biasing schemes (single substrate contact and multiple regularly placed substrate contacts) and TSV fabrication methods (via-first and via-last). A closed-form expression has been developed to approximate each admittance within the π model with a logarithmic function. Both the compact model and the closed-form expression have been validated using a 3D transmission line matrix method with an average error of 4.8%. These expressions and the model have been utilized to better understand the effect of different design parameters on noise for both via-first and via-last TSVs, such as substrate contact placement, slew rate and voltage swing of the TSV signals, and differential TSV signaling.

Chapter 8

Design and Characterization of a Standard Cell Library for Monolithic 3D ICs

In this chapter, a different vertical integration technology, *i.e.*, monolithic interlayer via (MIV) based 3D ICs, is investigated. Contrary to TSVs, MIVs have comparable size to on-chip metal vias and therefore offer a highly promising vertical integration technology. An overview of the MIV based monolithic 3D ICs and multiple design styles are explored in Section 8.1. The procedure of developing a transistor-level monolithic 3D IC standard cell library is explained in Section 8.2. Utilizing this standard cell library, a benchmark IC is implemented from gate level netlist to physical layout in monolithic 3D style, as described in Section 8.3.

8.1 Monolithic Inter-layer Vias (MIV)

In TSV-based 3D integration, as discussed in previous chapters, the multiple planes are processed separately, and then bonded together using TSVs as the vertical connections. An important advantage of this technology is the ability to fabricate each plane in different process technologies, enabling heterogenous integration. The primary limitation, however, is the significantly larger size of the vertical TSVs as compared to MIVs discussed in this chapter [109].

In MIV-based 3D integration, multiple planes are fabricated sequentially. The MIVs are fabricated using a similar process as the regular local metal vias. Thus, monolithic 3D ICs enable ultra fine-grained vertical integration.

8.1.1 Fabrication of MIV-based 3D ICs

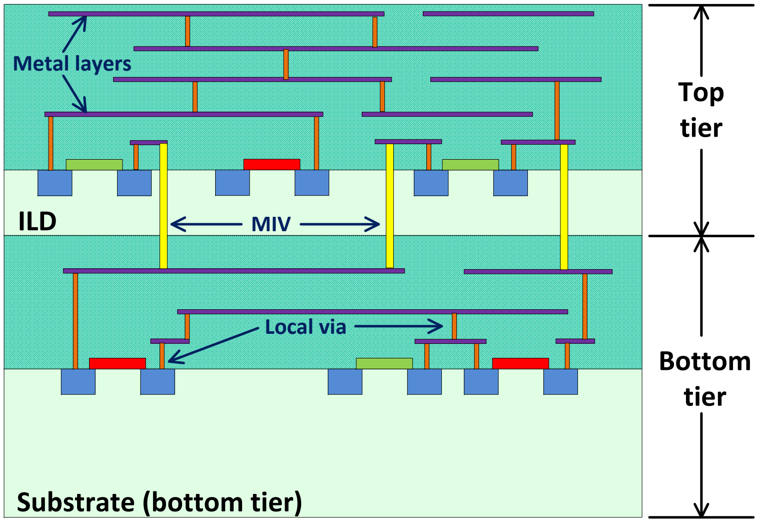


Figure 8.1: Monolithic 3D integration with MIVs as the inter-tier vertical connections.

An illustrative cross-section diagram of a two-tier monolithic 3D IC is shown in Fig 8.1 [110–112]. Initially, the devices on the bottom plane are fabricated within the silicon substrate of the bottom plane. The multiple metal layers for the bottom plane are then deposited on top of this substrate. After that, an inter-layer dielectric (ILD) is deposited. This layer serves as the isolation between the top and bottom planes. The MIVs are then fabricated, which pass through the ILD layer. On top of the ILD layer, a second silicon layer is formed where the transistors for the top plane are fabricated. Metal layers for top plane are fabricated after the devices of the top plane are formed. The inter-tier MIVs connect the topmost metal layer in the bottom tier and bottommost metal layer in the top tier, in a similar fashion as via-first and via-middle TSVs.

The most critical challenge in the fabrication of monolithic 3D ICs is to minimize the detrimental effect of the top plane on bottom plane. In the past few years, significant research has focused on developing the low-thermal budget process [111] [110]. For example, in [111], for the bottom plane, PMOS and NMOS transistors are fabricated using FDSOI (Fully Depleted Silicon On Insulator) with a silicon thickness of approximately 30 nm. A thin ILD is then deposited. The thickness of the ILD can be in the range of 60 nm to 110 nm, depending on the pre-bonding planarization. The critical point is using a low temperature molecular bonding process to bond the SOI substrate which serves as the active layer for top plane devices. For all the fabrication steps related with top plane devices, the thermal budget is limited to 600 °C.

8.1.2 Different Design Styles in MIV-based 3D ICs

Due to the small diameter and higher density of MIVs, fine-grain vertical integration can be achieved, enabling distinct 3D integration design styles, some of which are not possible with TSV-based 3D ICs [112, 113]. One of the design styles is the transistor-level monolithic 3D IC (TL-Mono3D), which is the most fine-grained integration [113]. As illustrated in Fig. 8.2, in TL-Mono3D, the monolithic 3D IC typically consists of two tiers, each of which only contains either N-channel MOSFET or P-channel MOSFET. The NMOS pull-down network and PMOS pull-up network in a CMOS standard cell are separated and located in the top and bottom tiers, respectively. The ultra fine-grain MIVs provide the connections between NMOS and PMOS transistors within a cell, as well as other inter-cell connections. Note that this TL-Mono3D design style is not feasible in the TSV-based 3D ICs, since the density of TSVs cannot provide sufficient vertical connections for this type of integration.

Other design styles include the gate-level monolithic integration (GL-Mono3D) and block-level monolithic integration (BL-Mono3D), as shown, respectively, in Fig. 8.3 and Fig. 8.4 [112, 113]. In GL-Mono3D, each standard cell can be the same as in 2D ICs. Multiple standard cells within a functional block are distributed on multiple tiers, while MIVs are utilized for the inter-cell connections. The block level floorplan for 2D ICs can therefore be reused. BL-Mono3D represents a more coarse-grain level integration. In this case, the system partitioning is achieved based on individual functional blocks. The gates/cells within a functional block are placed on the same tier wheras different functional blocks are distributed on other tiers.

In the last two design styles (GL-Mono3D and BL-Mono3D), the advantage of ultra small size and high density MIVs is not fully utilized since these design

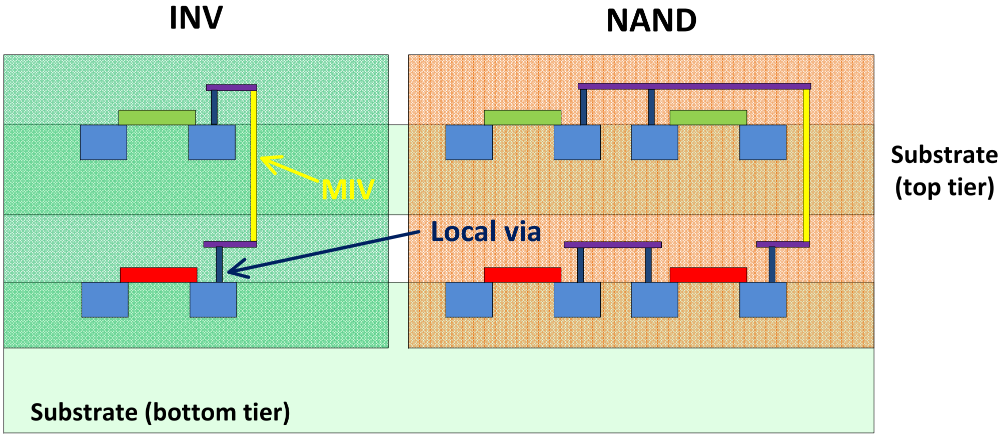


Figure 8.2: Transistor-level monolithic 3D integration design style, where all of the PMOS transistors are fabricated on one tier and all of the NMOS transistors are fabricated on the other tier.

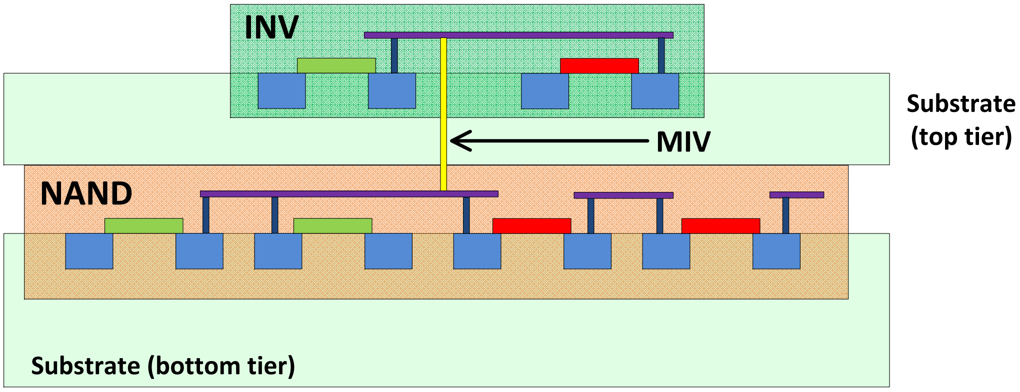


Figure 8.3: Gate-level monolithic 3D integration design style.

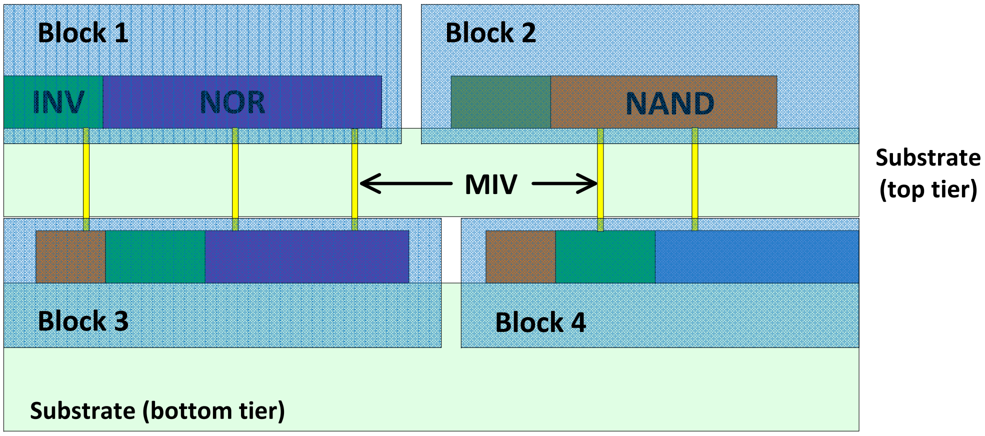


Figure 8.4: Block-level monolithic 3D integration design style.

styles are also available in TSV-based 3D ICs. The transistor level (TL-Mono3D) is therefore the focus of this work to explore the design considerations and performance characteristics of monolithic 3D ICs.

8.2 Monolithic 3D Standard Cell Library

Standard cell library is one of the primary components of application specific integrated circuit (ASIC) design flow, which enables multiple design procedures during the physical design process such as logic synthesis, place and route, and power/clock synthesis. The essential information in a standard cell library includes the physical layout for each cell and their electrical characteristics such as power consumption and timing information. To explore the various physical design challenges in monolithic 3D ICs such as power distribution network design and clock tree synthesis, larger scale benchmark circuits should be implemented with the monolithic 3D integration technology. This approach requires a standard cell li-

AND2X1	AOI21X1
BUFX2	CLKBUF1
DFFPOSX1	FILL
INVX1	INVX2
INVX4	MUX2X1
NAND2X1	NOR2X1
OAI21X1	OR2X1
XNOR2X1	XOR2X1

Table 8.1: Standard cells available in the Mono3D library.

brary specific for monolithic 3D ICs.

Thus, the *Mono3D*, a standard cell library for transistor-level monolithic 3D ICs is developed. The Mono3D standard cell library is based on the baseline 2D standard cell library “FreePDK45” in 45 nm technology [66]. The process and physical characteristics for each 2D plane in the Mono3D are retrieved from FreePDK45, including the transistor models and physical characteristics such as metal layer parameters and parasitic information. The standard cells provided in FreePDK45 serve as the baseline to evaluate the proposed Mono3D standard cell library.

The development of the Mono3D is a continuous effort that requires iterative improvements and updates as more manufacturing data becomes available. Preliminary results are described in this chapter by relying on the first version of the proposed cell library. Currently, 16 standard cells exist in the Mono3D library, as listed in Table 8.1. The design flow describing the development of the standard cell library is illustrated in Fig. 8.5.

Initially, for each cell listed in the table, the layout is drawn in full custom methodology with the monolithic 3D approach. Cadence Virtuoso [114] is used as the primary tool for drawing the layouts. In a traditional 2D cell, the PMOS pull-up network and NMOS pull-down network are placed on the same substrate, as shown

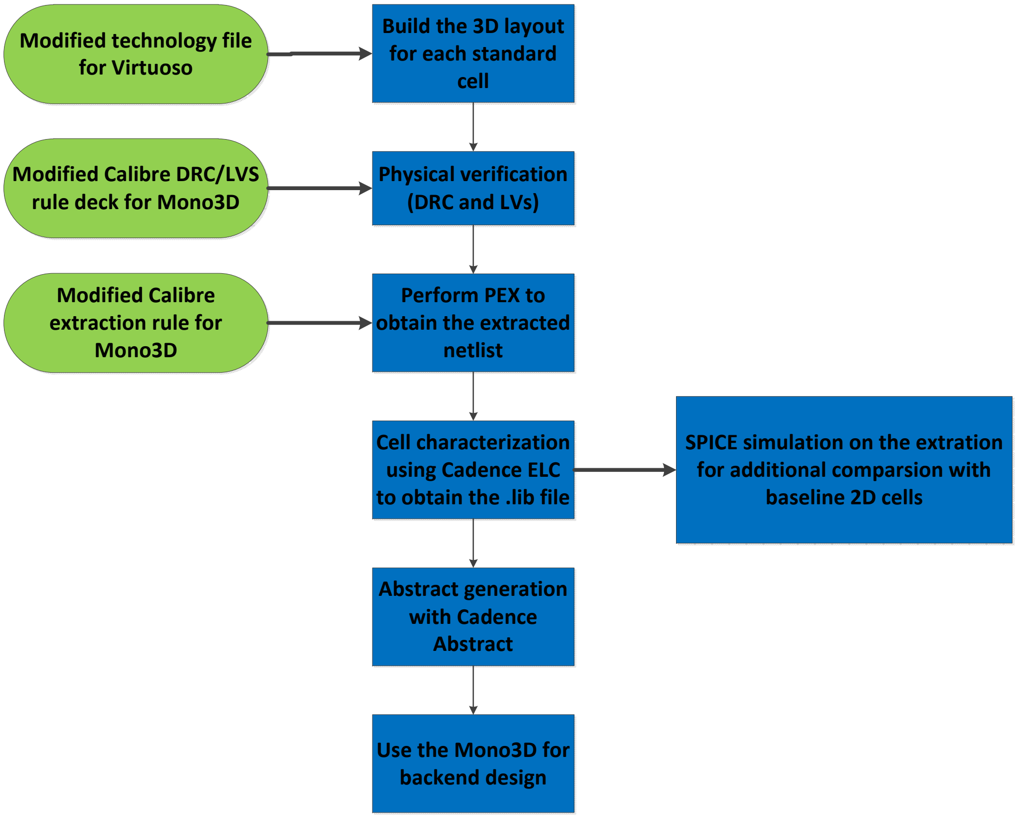

Modified technology file for Virtuoso
Build the 3D layout for each standard cell
Modified Calibre DRC/LVS rule deck for Mono3D
Physical verification (DRC and LVs)
Modified Calibre extraction rule for Mono3D
Perform PEX to obtain the extracted netlist
Cell characterization using Cadence ELC to obtain the .lib file
SPICE simulation on the extration for additional comparsion with baseline 2D cells
Abstract generation with Cadence Abstract
Use the Mono3D for backend design

Figure 8.5: Process of building the standard cell library.

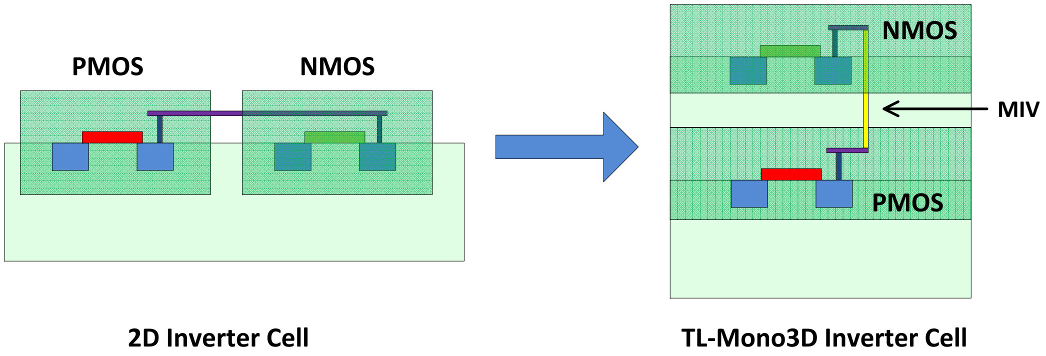


Figure 8.6: Comparison of the physical structure of an inverter standard cell in traditional 2D and TL-Mono 3D.

in Fig. 8.6. In a Mono3D cell, the PMOS transistors are placed on the bottom tier and NMOS transistors are placed on the top tier. The reason for this placement is related with the high temperature processing steps. Note that in monolithic 3D integration, multiple planes are fabricated sequentially. Thus, during the fabrication of the upper tier, the temperature of the processing steps should be carefully controlled to protect the devices on the bottom tier. Due to this thermal limitation, the transistors on the upper tier typically have degraded electrical characteristics as compared to transistors on the bottom layer. Since PMOS transistors already suffer from low mobility of the holes, they are placed on the bottom tier. This placement partially balances these two effects and achieves comparable pull up and pull down network sizes within a cell.

After the 3D layout of each standard cell is completed, physical verification including design rule check (DRC) and layout versus schematic (LVS) are performed using Calibre from Mentor Graphics [115]. The Calibre DRC/LVS rule files are modified based on the rule files provided with the FreePDK45.

There are several critical differences between the 3D and 2D rule files that need to be addressed. These differences are 1) distinguishing the metal layers above the bottom tier from the metal layers above the top tier, and 2) new information related with MIVs that does not exist in 2D rule files. In the current version of Mono3D, there are 10 metal layers above the top tier (full metal stack available in this technology) and 5 metal layers above the bottom tier. There are several reasons why the bottom tier has less number of metal layers: 1) due to the sequential fabrication of monolithic 3D ICs, high number of metal layers on the bottom tier makes the fabrication of the top tier more challenging and possibly reduces the reliability of the transistors located on the top tier; 2) high number of metal layers above the bottom tier increases the height of the MIVs, which in turn increases the MIV parasitic impedances [109]; 3) in the TL-Mono3D integration style, the metal layers above the bottom tier are primarily used for the inter-cell routing and power grid for bottom tier. Global signal routing is achieved above the top tier. Thus, a subset of the full metal stack above the bottom tier is sufficient for monolithic 3D ICs, particularly for TL-Mono3D.

After each standard cell successfully passes the DRC and LVS steps, an *RC* parasitic extraction of each 3D layout is performed using Calibre PEX [115]. The PEX rule file is also modified based on the 2D rule file in FreePDK45. The purpose of the *RC* extraction is to represent the 3D physical layout of each standard cell by an extracted netlist that contains all of the parasitic impedances including MIV parasitics. The extracted netlist is used for cell characterization and is compatible with existing EDA tools.

The files required to use a standard cell library in conventional design flows include the .LIB file which contains the timing and power information for each cell, .DB file which is a compiled binary version of the .LIB file, and the .LEF file which contains the layout physical abstraction of each cell needed by the place and route tool. In this work, Cadence Encounter Library Characterizer (ELC) is used for the cell characterization [116] to analyze the timing and power characteristics of each cell, and generate the .LIB and .DB files. HSPICE simulations are also performed to compare the 3D cells with the 2D baseline cells, as presented in Section 8.2.1.

To enable automatic place and route, physical dimension data of each cell should be available. The abstract view of a cell contains the blockage shape and the pin connectivity information, which are sufficient for a place and route tool. In this work, Cadence Abstract is used to obtain the abstract view of each standard cell and generate the .LEF file [117].

8.2.1 Comparison of Mono3D cells and 2D Cells

With the extracted netlist of each cell, the performance metrics (area, delay, and power consumption) between Mono3D cells and baseline 2D cells can be compared using HSPICE simulations, as listed in Table 8.2.

According to this table, transforming the 2D cells into monolithic 3D cells, the layout area can be reduced, on average, by 32.7%. Area reduction does not reach 50% since the MIVs within a standard cell consume nonnegligible area and cause routing congestion. Thus, although the height of a standard cell row in Mono3D is reduced to half of the height of the 2D cells, the width of a 3D cell may expand to accommodate the additional MIVs and ensure the inner-cell routing within the reduced cell footprint.

As listed in Table 8.2, the delay and power characteristics of the Mono3D cells are comparable to 2D cells. Specifically, Mono3D cells have, on average, 2.8% higher propagation delay and 2.4% higher power consumption than the 2D cells.

	Area (μm^2)		**Delay (ps)**		**Power (μW)**	
	Mono3D	2D	Mono3D	2D	Mono3D	2D
AND2X1	1.8590	3.5778	19.012	18.537	2.066	2.097
AOI21X1	2.2495	3.0705	15.858	14.557	2.739	2.531
BUFX2	1.7422	2.5632	19.496	18.786	10.673	10.513
CLKBUF1	4.5323	6.1143	30.608	27.750	50.078	48.210
DFFPOSX1	4.7559	7.6362	37.239	43.936	22.352	21.261
INVX1	1.2816	2.0559	6.498	6.837	3.212	3.324
INVX2	1.5620	2.0559	6.548	6.504	6.388	6.540
INVX4	1.4819	2.5632	6.757	6.307	12.587	12.74
MUX2X1	3.2541	4.5924	17.777	17.214	4.599	4.413
NAND2X1	1.3751	2.5632	10.312	10.967	1.142	1.193
NOR2X1	1.7489	2.5632	12.212	11.913	1.262	1.217
OAI21X1	2.1961	3.0705	13.562	13.608	2.424	2.494
OR2X1	2.1961	3.5778	21.621	19.484	2.115	1.933
XNOR2X1	4.1151	5.0997	42.163	38.446	10.450	9.610
XOR2X1	3.9116	5.0997	41.315	37.979	10.213	9.578

Table 8.2: Performance comparison of each individual standard cell in Mono3D and baseline 2D PDK.

This increase can be described by the denser cell layout, producing additional coupling capacitances and the MIV impedances. Also note that the primary advantage of 3D integration is the reduction of the global interconnect length due to reduced form factor. Shorter global interconnects consume less power and have lower propagation delay. This characteristic cannot be observed in a standard cell since the cells are sufficiently small and the transistor power/delay dominates the interconnect power/delay.

The Mono3D is used to develop a larger IC by utilizing a standard design flow, as described in the following section.

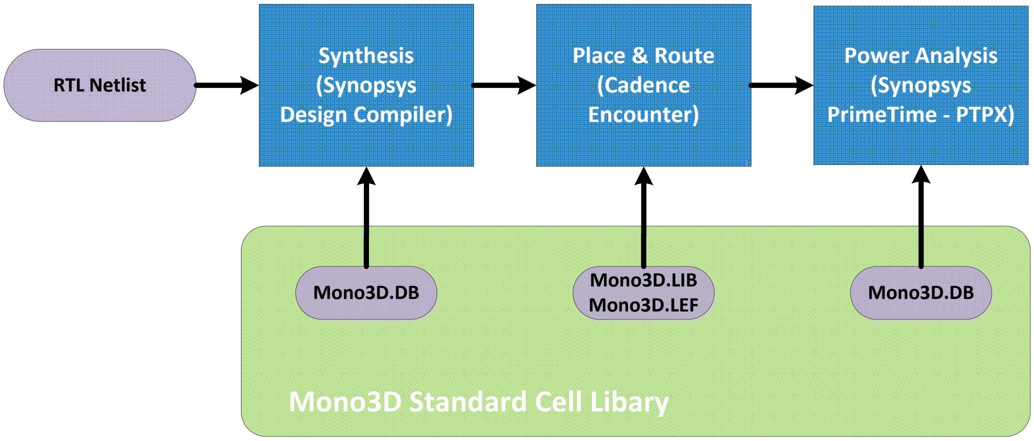

RTL Netlist
Synthesis (Synopsys Design Compiler)
Place & Route (Cadence Encounter)
Power Analysis (Synopsys PrimeTime - PTPX)
Mono3D.DB
Mono3D.LIB
Mono3D.LEF
Mono3D.DB
Mono3D Standard Cell Libary

Figure 8.7: Physical design flow utilizing the Mono3D standard cell library.

8.3 Developing a Monolithic 3D IC with the Mono3D Standard Cell Library

As mentioned in Section 8.2, the purpose the Mono3D standard cell library is to facilitate the automated physical design of 3D monolithic ICs. Conventional physical design flow utilizing the proposed Mono3D standard cell library is described in Section 8.3.1. The performance characteristics of the IC designed with the Mono3D are provided in Section 8.3.2.

8.3.1 Physical Design Flow with the Mono3D Library

Due to the unique design characteristic of transistor-level monolithic 3D ICs, the proposed Mono3D standard cell library is compatible with existing EDA tools that are originally developed for 2D chips. The integration of Mono3D into existing design flows is illustrated in Fig. 8.7.

For the synthesis stage, the extracted netlist of each cell contains the parasitic information of both the top and bottom tiers, and the vertical MIVs. The .LIB file generated after the cell characterization step has the same format as its 2D counterpart, while including the 3D-specific parasitic impedances that may affect the timing/power results. The compiled .DB file is fed to the synthesis tool (Synopsys Design Compiler is used in this work) to translate a hardware description language to a gate-level netlist consisting of the 3D cells in Mono3D.

For place and route, since the pull-up network within the bottom tier has the same footprint as the pull-down network within the top tier, a place and route tool needs to only consider the top tier in terms of cell boundary and blockage. All of the signal pins within Mono3D cells are placed on the top tier. The ground rail (VSS) is routed only within the top tier whereas the power rail (VDD) is routed only within the bottom tier. Note however that both the VDD/VSS pin information and the signal pin information are preserved in the .LEF file.

After the physical implementation of the IC is completed, power analysis can be performed. In this work, Synopsys Primetime PX is used for the power analysis. Since the power estimation is based on the gate-level netlist, with parasitic information properly annotated by the SPEF file, Mono3D can support the power analysis for monolithic 3D ICs.

8.3.2 Monolithic 3D IC Design with the Mono3D Library

Using the developed Mono3D standard cell library and the aforementioned physical design flow, an IC is designed from register transfer language (RTL) netlist to physical layout. The functionality of Mono3D is validated.

A benchmark circuit s38584 from ISCAS89 is used to evaluate the developed

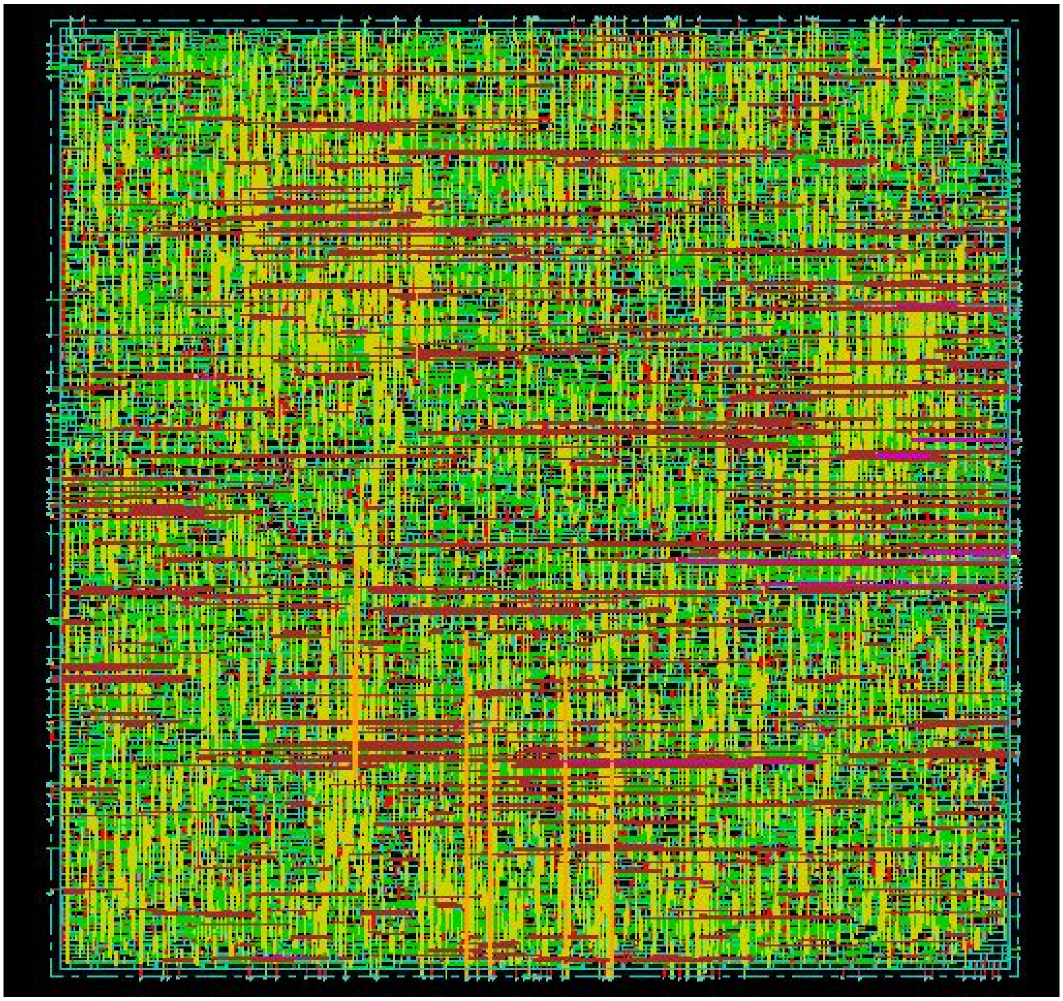


Figure 8.8: Layout of a benchmark circuit designed with the Mono3D standard cell library.

Mono3D standard cell library. The circuit contains 1426 D-flip-flops, 19253 gates, 38 input pins and 304 output pins. A transistor-level monolithic 3D implementation of s38584 is developed using Mono3D, as shown in Fig. 8.8. Alternatively, the 2D implementation of the same circuit is also developed using FreePDK45 [66]. The performance comparison of the Mono3D approach and the 2D counterpart is listed in Table 8.3.

	Mono3D	**Baseline 2D**	**Improvement**
Area	0.0264 mm^2	0.0372 mm^2	29.0%
Standard cell area	0.0179 mm^2	0.0227 mm^2	21.4%
Power consumption	6.78 mW	7.43 mW	8.7%

Table 8.3: Performance comparison of the Mono3D based monolithic 3D circuit and the 2D counterpart.

As listed in Table 8.3, the Mono3D approach achieves 29.0% reduction in die area, which correlates well with the average 32.7% reduction in individual cell area. Due to the smaller circuit footprint, there is more routing congestion within the Mono3D circuit. In terms of power consumption, due to the reduced interconnect length (and therefore reduced parasitic impedance) the power consumption of Mono3D circuit is 8.7% less than the 2D circuit. Note that the power savings are expected to grow as the size of the circuit increases. The benchmark circuit evaluated here is significantly smaller in size as compared to industrial ICs.

8.4 Summary

In this chapter, Mono3D, a standard cell library for designing transistor-level monolithic 3D ICs is developed. For each standard cell, full *RC* extraction is performed on the 3D-specific layout using modified deck files. Complete timing, power, and physical characteristics are obtained through cell characterization and abstract generation. It has been demonstrated that the proposed Mono3D library can fully support the existing physical design flows. A sequential benchmark circuit is designed with the proposed Mono3D library, demonstrating 29.0% reduction in circuit area and 8.7% reduction in power consumption.

Chapter 9

Conclusion and Future Work

3D integration is a fundamental enabling technology in the interconnect-centric design era. Furthermore, 3D technology is considered to be the most feasible candidate to ensure higher integration densities despite the fundamental limits of classical technology scaling.

Significant research efforts have been made in the past years to overcome the challenges in the fabrication, design and testing of 3D ICs. In this dissertation, several novel design methodologies have been proposed to enhance the power and signal integrity in 3D ICs. These contributions are summarized in Section 9.1. Several possible future directions are discussed in Section 9.2.

9.1 Dissertation Summary

With higher and heterogeneous integration achieved by 3D ICs, the design process of a power distribution network has become highly challenging. A 3D power network delivers the power supply voltage to each circuit module within each plane of a 3D stack. The power supply noise degrades system performance and may cause functional failure. Furthermore, low power design techniques such as power gating are typically implemented in 3D ICs, increasing the complexity of a power distribution network.

In this dissertation, the design implications of power gating in TSV-based 3D ICs have been investigated. The physical structure of a power distribution network including global and local power grids, metal vias, and TSVs should be considered when implementing power gating in 3D ICs. For TSVs fabricated with different via-first/via-middle/via-last technologies, the connectivity scheme of TSVs and metal layers varies. Lumped and distributed sleep transistor placement topologies have been proposed in Chapter 3, each specifically tailored to certain TSV types to exploit the difference in TSV connectivity schemes. It is demonstrated that the TSV-specific topologies can improve power integrity with reasonable power and area overhead.

In addition to the placement topologies, the size of sleep transistors is another design consideration that exhibits a tradeoff between multiple design constraints such power integrity, physical area, and power consumption. In TSV-based 3D ICs, it is observed that the sleep transistors have significant effect on the power supply noise, and also introduce area contention with TSVs. Note that both sleep transistors and TSVs consume considerable substrate area. A resource allocation methodology has been proposed in Chapter 4 to simultaneously determine the size of sleep transistors and number of TSVs. The proposed method enhances power integrity while minimizing the area overhead.

In analyzing power supply noise, it is observed that the existing closed-form expressions to estimate power supply noise do not produce accurate results for nanoscale circuits with fast transitions. An enhanced analytic model with closed-form expressions has been proposed to address this issue. The proposed expressions have been validated with comprehensive SPICE simulations. Significant improvement in accuracy has been demonstrated, particularly for circuits with fast transitions.

Power gating introduces certain issues to decoupling capacitors in 3D ICs. In coarse-grain power gating, power delivery to a subcircuit is turned off to save leakage power when the subcircuit is in idle mode. However, a nondesirable impact of this method is that the decoupling capacitors located within this subcircuit are disconnected from the global power network. In 3D ICs, the low resistive vertical TSVs extend the effective range of decoupling capacitors from a single plane to multiple planes. It is therefore desirable to effectively utilize the decoupling capacitors even when the subcircuit is power gated. A reconfigurable decoupling capacitor topology which can dynamically configure the decoupling capacitors depending upon the status of power gating status has been investigated in Chapter 5. Two reconfigurable switches connect the decoupling capacitors to either the global power network (when the plane is power gated) or the virtual power network (when the plane is active). It has been demonstrated that the reconfiguration mechanism significantly improves the overall power integrity in a 3D IC with power gating.

TSV-based 3D ICs also exhibit signal integrity issues. As the critical component in 3D integration, vertical TSVs pass through the silicon substrate. In 3D systems with heterogeneous integration, a TSV carrying high frequency switching signal from digital logic plane injects noise into the substrate of a noise-sensitive analog plane. A methodology has been proposed to characterize the TSV-induced noise coupling in 3D ICs. A compact model has been developed to efficiently estimate the coupling noise in substrate with sufficient accuracy. The proposed model with closed-form expressions has been utilized to better understand TSV-induced noise as a function of multiple parameters such as TSV type, placement of substrate contacts, signal slew rate, and voltage swing. Design guidelines have been developed based on these results to improve system-wide signal integrity in 3D ICs.

The other 3D integration option – monolithic 3D ICs is also explored. Multiple design styles within the monolithic 3D ICs are discussed. Mono3D, a 45 nm standard cell library is developed to enable the back-end physical design of transistor-level monolithic 3D ICs. A benchmark IC is designed and characterized to validate the functionality of the developed Mono3D library and also demonstrate the advantages of monolithic 3D ICs.

9.2 Future Work

The future work of this research can focus on the 3D monolithic integration technology. Additional standard cells can be added into the Mono3D standard cell library to enhance the performance of the circuits built with this library. For each standard cell, the layout structure can be modified to understand the impact of different design options. For example, design considerations such as the relative location of the VDD/VSS power rails, placement strategy of the MIVs require additional analysis. Related tradeoffs can also be identified.

Furthermore, larger benchmark circuits with more complex functions can be designed to evaluate the performance characteristics of monolithic 3D ICs. These larger benchmark circuits can also be used to investigate primary global networks (such as power and clock networks) in monolithic 3D ICs. Related design guidelines and methodologies can be developed.

Bibliography

[1] G. Moore, "Cramming More Components Onto Integrated Circuits," *Proceedings of the IEEE*, vol. 86, no. 1, pp. 82–85, January 1998.

[2] "Intel Microprocessor Historical Reference." [Online]. Available: http://www.intel.com/pressroom/kits/quickreffam.htm

[3] H. Bakoglu and J. Meindl, "Optimal Interconnection Circuits for VLSI," *IEEE Transactions on Electron Devices*, vol. 32, no. 5, pp. 903–909, May 1985.

[4] K. Saraswat and F. Mohammadi, "Effect of Scaling of Interconnections on the Time Delay of VLSI Circuits," *IEEE Transactions on Electron Devices*, vol. 29, no. 4, pp. 645–650, April 1982.

[5] "International Technology Roadmap for Semiconductors (ITRS)," 2011.

[6] S. Merchant, S. Kang, M. Sanganeria, B. van Schravendijk, and T. Mountsier, "Copper Interconnects for Semiconductor Devices," *JOM*, vol. 53, no. 6, pp. 43–48, June 2001.

[7] A. Brown, "Fast Films [IC Interconnect Insulation]," *IEEE Spectrum*, vol. 40, no. 2, pp. 36–40, February 2003.

[8] D.-Y. Kim, “Study on Reliability of VLSI Interconnection Structures,” Ph.D. dissertation, Stanford University, 2003.

[9] V. F. Pavlidis and E. G. Friedman, *Three-Dimensional Integrated Circuit Design*. Morgan Kaufmann, 2009.

[10] “Intel High End Desktop Processors.” [Online]. Available: http://ark.intel.com/products/family/79318/Intel-High-End-Desktop-Processors

[11] J. Joyner, P. Zarkesh-Ha, J. Davis, and J. Meindl, “A Three-dimensional Stochastic Wire-length Distribution for Variable Separation of Strata,” in *Proceedings of the IEEE International Interconnect Technology Conference*, June 2000, pp. 126–128.

[12] J. Knickerbocker, P. Andry, E. Colgan, B. Dang, T. Dickson, X. Gu, C. Haymes, C. Jahnes, Y. Liu, J. Maria, R. Polastre, C. Tsang, L. Turlapati, B. Webb, L. Wiggins, and S. Wright, “2.5D and 3D Technology Challenges and Test Vehicle Demonstrations,” in *Proceedings of the Electronic Components and Technology Conference*, May 2012, pp. 1068–1076.

[13] M. Koyanagi, H. Kurino, K.-W. Lee, K. Sakuma, N. Miyakawa, and H. Itani, “Future System-on-silicon LSI Chips,” *IEEE Micro*, vol. 18, no. 4, pp. 17–22, July 1998.

[14] C. Tan, R. Gutmann, and L. Reif, *Wafer Level 3-D ICs Process Technology*, ser. Integrated Circuits and Systems. Springer, 2009.

[15] A. Sigl, S. Pargfrieder, C. Pichler, C. Scheiring, and P. Kettner, “Advanced Chip To Wafer Bonding: A Flip Chip to Wafer Bonding Technology for

High Volume 3DIC Production Providing Lowest Cost of Ownership," in *Proceedings of the IEEE International Conference on Electronic Packaging Technology & High Density Packaging*, August 2009, pp. 932–936.

[16] V. Rao, S. C. Chong, C. Zhaohui, J. L. Aw, E. Ching, H. Gilho, and D. Fernandez, "Development of Bonding Process for High Density Fine Pitch Micro Bump Interconnections with Wafer Level Underfill for 3D Applications," in *Proceedings of the IEEE Electronics Packaging Technology Conference*, December 2013, pp. 543–548.

[17] H. Lee and K. Chakrabarty, "Test Challenges for 3D Integrated Circuits," *IEEE Design Test*, vol. PP, no. 99, pp. 1–1, September 2013.

[18] E. Marinissen, J. Verbree, and M. Konijnenburg, "A Structured and Scalable Test Access Architecture for TSV-based 3D Stacked ICs," in *Proceedings of the IEEE VLSI Test Symposium*, April 2010, pp. 269–274.

[19] C.-W. Chou, J.-F. Li, J.-J. Chen, D.-M. Kwai, Y.-F. Chou, and C.-W. Wu, "A Test Integration Methodology for 3D Integrated Circuits," in *IEEE Asian Test Symposium*, December 2010, pp. 377–382.

[20] M. Popovich, A. Mezhiba, and E. G. Friedman, *Power Distribution Networks with On-Chip Decoupling Capacitors*. Springer, 2008.

[21] R. Ho, K. Mai, and M. Horowitz, "The Future of Wires," *Proceedings of the IEEE*, vol. 89, no. 4, pp. 490–504, April 2001.

[22] D. Su, M. Loinaz, S. Masui, and B. Wooley, "Experimental Results and Modeling Techniques for Substrate Noise in Mixed-signal Integrated Circuits," *IEEE Journal of Solid-State Circuits*, vol. 28, no. 4, pp. 420–430, April 1993.

[23] H. Wang and E. Salman, “Power Gating Methodologies in TSV Based 3D Integrated Circuits,” in *Proceedings of the ACM International Conference on Great Lakes Symposium on VLSI*, May 2013, pp. 327–328.

[24] H. Wang, M. H. Asgari, and E. Salman, “Efficient Characterization of TSV-to-transistor Noise Coupling in 3D ICs,” in *Proceedings of the ACM International Conference on Great Lakes Symposium on VLSI*, 2013, pp. 71–76.

[25] H. Wang and E. Salman, “Decoupling Capacitor Topologies for TSV-Based 3-D ICs With Power Gating,” *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 23, no. 12, pp. 2983–2991, December 2015.

[26] H. Wang and E. Salman, “Compact Model to Efficiently Characterize TSV-to-transistor Noise Coupling in 3D ICs,” *Integration, the VLSI Journal*, vol. 47, no. 3, pp. 296 – 306, June 2014.

[27] H. Wang and E. Salman, “Resource Allocation Methodology for Through Silicon Vias and Sleep Transistors In 3D ICs,” in *Proceedings of the International Symposium on Quality Electronic Design*, March 2015, pp. 528–532.

[28] H. Wang and E. Salman, “Enhancing System-wide Power Integrity In 3D ICs With Power Gating,” in *Proceedings of the International Symposium on Quality Electronic Design*, March 2015, pp. 322–326.

[29] B. N. Jan M. Rabaey, Anantha P. Chandrakasan, *Digital Integrated Circuits: A Design Perspective*. Pearson Education, 2003.

[30] S. Ramaswami *et al.*, “Process Integration Considerations for 300 mm TSV Manufacturing,” *IEEE Transactions on Device and Materials Reliability*, vol. 9, no. 4, pp. 524–528, December 2009.

[31] D. H. Kim *et al.*, “3D-MAPS: 3D Massively Parallel Processor with Stacked Memory,” in *IEEE International Solid-State Circuits Conference, Digest of Technical Papers*, February 2012, pp. 188–190.

[32] D. Fick, R. Dreslinski, B. Giridhar, G. Kim, S. Seo, M. Fojtik, S. Satpathy, Y. Lee, D. Kim, N. Liu, M. Wieckowski, G. Chen, T. Mudge, D. Sylvester, and D. Blaauw, “Centip3De: A 3930DMIPS/W Configurable Near-threshold 3D Stacked System with 64 ARM Cortex-M3 Cores,” in *IEEE International Solid-State Circuits Conference, Digest of Technical Papers*, February 2012, pp. 190–192.

[33] I. Savidis and E. G. Friedman, “Closed-Form Expressions of 3-D Via Resistance, Inductance, and Capacitance,” *IEEE Transactions on Electron Devices*, vol. 56, no. 9, pp. 1873–1881, September 2009.

[34] G. Katti, M. Stucchi, K. De Meyer, W. Dehaene, “Electrical Modeling and Characterization of Through Silicon via for Three-Dimensional ICs,” *IEEE Transactions on Electron Devices*, vol. 57, no. 1, pp. 256–262, January 2010.

[35] R. Weerasekera, M. Grange, D. Pamunuwa, H. Tenhunen, and L.-R. Zheng, “Compact Modelling of Through-Silicon Vias (TSVs) in Three-dimensional (3-D) Integrated Circuits,” in *Proceedings of the IEEE International Conference on 3D System Integration*, September 2009, pp. 1–8.

[36] J. Kim and J. Kim, “Signal Integrity Modeling and Measurement of TSV in 3D IC,” in *Proceedings of Asia and South Pacific Design Automation Conference*, January 2013, pp. 13–16.

[37] J. Kim, J. Cho, and J. Kim, "TSV Modeling and Noise Coupling in 3D IC," in *Proceedings of the IEEE Electronics System-Integration Technology Conference*, September 2006, pp. 1–6.

[38] J. S. Pak *et al.*, "PDN Impedance Modeling and Analysis of 3D TSV IC by Using Proposed P/G TSV Array Model Based on Separated P/G TSV and Chip-PDN Models," *IEEE Transactions on Components, Packaging, and Manufacturing Technology*, vol. 1, no. 2, pp. 208–219, February 2011.

[39] S. M. Satheesh and E. Salman, "Power Distribution in TSV-Based 3-D Processor-Memory Stacks," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 2, no. 4, pp. 692–703, December 2012.

[40] B. Black, D. Nelson, C. Webb, and N. Samra, "3D Processing Technology and Its Impact on IA32 Microprocessors," in *Proceedings of the IEEE International Conference on Computer Design: VLSI in Computers and Processors*, October 2004, pp. 316–318.

[41] "Tezzaron 3D SRAM Prototype." [Online]. Available: http://www.tezzaron.com/3d-sram-prototype/

[42] "Tezzaron 3D Microcontroller Prototype." [Online]. Available: http://www.tezzaron.com/3d-ic-microcontroller-prototype/

[43] "Tezzaron 3D Sensor Demo/Test." [Online]. Available: http://www.tezzaron.com/3d-sensor-demotest/

[44] "Apple A4 Chip." [Online]. Available: https://www.ifixit.com/Teardown/Apple+A4+Teardown/2204

[45] M. Wordeman, J. Silberman, G. Maier, and M. Scheuermann, “A 3D System Prototype of An Edram Cache Stacked over Processor-like Logic Using Through-Silicon Vias,” in *IEEE International Solid-State Circuits Conference, Digest of Technical Papers*, February 2012, pp. 186–187.

[46] R. Jotwani, S. Sundaram, S. Kosonocky, A. Schaefer, V. Andrade, A. Novak, and S. Naffziger, “An x86-64 Core in 32 nm SOI CMOS,” *IEEE Journal of Solid-State Circuits*, vol. 46, no. 1, pp. 162 –172, January 2011.

[47] P. Bai, C. Auth, S. Balakrishnan, M. Bost, and *et al.*, “A 65nm Logic Technology Featuring 35nm Gate Lengths, Enhanced Channel Strain, 8 Cu Interconnect Layers, Low-k ILD and 0.57 um^2 SRAM Cell,” in *Technical Digest of International Electron Devices Meeting*, December 2004, pp. 657 – 660.

[48] Q. Zhu, *Power Distribution Network Design for VLSI*. Wiley, 2004.

[49] D. M. H. Neil H. E. Weste, *CMOS VLSI Design: A Circuits and Systems Perspective*. Addison Wesley, 2011.

[50] E. Salman and E. Friedman, *High Performance Integrated Circuit Design*. McGraw-Hill, 2012.

[51] E. Salman, “Noise Coupling Due to Through Silicon Vias (TSVs) in 3-D Integrated Circuits,” in *Proceedings of the IEEE International Symposium on Circuits and Systems*, May 2011, pp. 1411–1414.

[52] E. Salman, M. Asgari, and M. Stanacevic, “Signal Integrity Analysis of a 2-D and 3-D Integrated Potentiostat for Neurotransmitter Sensing,” in *Biomedical Circuits and Systems Conference (BioCAS), 2011 IEEE*, November 2011, pp. 17–20.

[53] P. LeDuc *et al.*, “Enabling Technologies for 3D Chip Stacking,” in *Proceedings of the IEEE International Symposium on VLSI Technology, Systems and Applications*, April 2008, pp. 76–78.

[54] J. L. Hennessy and D. A. Patterson, *Computer Architecture: A Quantitative Approach*. Elsevier, 2011.

[55] G. D. Wilk, R. M. Wallace, and J. M. Anthony, “High-K Gate Dielectrics: Current Status and Materials Properties Considerations,” *Journal of Applied Physics*, vol. 89, no. 10, pp. 5243–5275, 2001.

[56] R. Chau, S. Datta, M. Doczy, B. Doyle, J. Kavalieros, and M. Metz, “High-k Metal-gate Stack and its MOSFET Characteristics,” *IEEE Electron Device Letters*, vol. 25, no. 6, pp. 408–410, June 2004.

[57] D. Hisamoto *et al.*, “FinFET-a Self-aligned Double-gate MOSFET Scalable to 20 nm,” *IEEE Transactions on Electron Devices*, vol. 47, no. 12, pp. 2320–2325, Dec 2000.

[58] “ARM big.LITTLE Technology.” [Online]. Available: http://www.arm.com/products/processors/technologies/biglittleprocessing.php

[59] S. Mutoh, T. Douseki, Y. Matsuya, T. Aoki, S. Shigematsu, and J. Yamada, “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.

[60] R. Dreslinski, M. Wieckowski, D. Blaauw, D. Sylvester, and T. Mudge, “Near-Threshold Computing: Reclaiming Moore’s Law Through Energy Ef-

ficient Integrated Circuits," *Proceedings of the IEEE*, vol. 98, no. 2, pp. 253–266, February 2010.

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

[62] D. Ikebuchi, N. Seki, Y. Kojima, M. Kamata, L. Zhao, H. Amano, T. Shirai, S. Koyama, T. Hashida, Y. Umahashi, H. Masuda, K. Usami, S. Takeda, H. Nakamura, M. Namiki, and M. Kondo, "Geyser-1: A MIPS R3000 CPU Core with Fine Grain Runtime Power Gating," in *Proceedings of the IEEE Asian Solid-State Circuits Conference*, November 2009, pp. 281–284.

[63] M. Henry and L. Nazhandali, "NEMS-Based Functional Unit Power-Gating: Design, Analysis, and Optimization," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 60, no. 2, pp. 290–302, February 2013.

[64] K. Kim, W. Lee, J. Kim, T. Song, J. Kim, J. S. Pak, J. Kim, H. Lee, Y. Kwon, and K. Park, "Analysis of Power Distribution Network in TSV-based 3D-IC," in *Proceedings of IEEE Conference on Electrical Performance of Electronic Packaging and Systems*, October 2010, pp. 177–180.

[65] A. Shayan, X. Hu, H. Peng, M. Popovich, W. Zhang, C.-K. Cheng, L. Chua-Eoan, and X. Chen, "3D Power Distribution Network Co-design for Nanoscale Stacked Silicon ICs," in *Proceedings of the IEEE Conference on Electrical Performance of Electronic Packaging*, October 2008, pp. 11–14.

[66] "FreePDK45." [Online]. Available: http://www.eda.ncsu.edu/wiki/FreePDK45:Contents

[67] “White Paper of Intel Xeon Processor: Measuring Processor Power,” Intel Corporation, Tech. Rep., April 2011.

[68] M. Gupta, J. Oatley, R. Joseph, G.-Y. Wei, and D. Brooks, “Understanding Voltage Variations in Chip Multiprocessors using a Distributed Power-Delivery Network,” in *Proceedings of The Design, Automation Test in Europe Conference Exhibition*, April 2007, pp. 1–6.

[69] R. Jakushokas and E. Friedman, “Multi-Layer Interdigitated Power Distribution Networks,” *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 19, no. 5, pp. 774 –786, May 2011.

[70] “FastHenry,” http://www.fastfieldsolvers.com/.

[71] S. K. Lim, “3D Circuit Design with Through-Silicon-Via: Challenges and Opportunities,” Georgia Institute of Technology, Tech. Rep., April 2010.

[72] M.-Y. Tsai, P.-S. Huang, C.-Y. Huang, H. Jao, B. Huang, B. Wu, Y.-Y. Lin, W. Liao, J. Huang, L. Huang, S. Shih, and J. P. Lin, “Investigation on Cu TSV-Induced KOZ in Silicon Chips: Simulations and Experiments,” *IEEE Transactions on Electron Devices*, vol. 60, no. 7, pp. 2331–2337, July 2013.

[73] J. Gu, H. Eom, J. Keane, and C. Kim, “Sleep Transistor Sizing and Adaptive Control for Supply Noise Minimization Considering Resonance,” *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 17, no. 9, pp. 1203–1211, September 2009.

[74] K. Nabors and J. White, “FastCap: A Multipole Accelerated 3-D Capacitance Extraction Program,” *IEEE Transactions on Computer-Aided Design*

of Integrated Circuits and Systems, vol. 10, no. 11, pp. 1447–1459, November 1991.

[75] M. Healy and S.-K. Lim, “Distributed TSV Topology for 3-D Power-Supply Networks,” *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 20, no. 11, pp. 2066–2079, November 2012.

[76] H. Jiang, M. Marek-Sadowska, and S. Nassif, “Benefits and Costs of Power-gating Technique,” in *Proceedings of IEEE International Conference on Computer Design*, October 2005, pp. 559 – 566.

[77] H. H. Chen and D. D. Ling, “Power Supply Noise Analysis Methodology for Deep-submicron VLSI Chip Design,” in *Proceedings of the ACM/IEEE Design Automation Conference*, June 1997, pp. 638–643.

[78] X. Zhang *et al.*, “Characterizing and Evaluating Voltage Noise in Multi-Core Near-Threshold Processors,” in *Proceedings of International Symposium on Low Power Electronics and Design*, September 2013, pp. 82–87.

[79] H. Wei *et al.*, “Cooling Three-dimensional Integrated Circuits Using Power Delivery Networks,” in *Proceedings of the IEEE International Electron Devices Meeting*, December 2012, pp. 14.2.1–14.2.4.

[80] R. Senthinathan, G. Tubbs, and M. Schuelein, “Negative Feedback Influence on Simultaneous Switching CMOS Outputs,” in *Proceedings of the IEEE Custom Integrated Circuits Conference*, May 1988, pp. 5.4/1–5.4/5.

[81] R. Senthinathan and J. 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.

[82] A. Vaidyanath, B. Thoroddsen, and J. 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.

[83] 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.

[84] 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.

[85] L. Ding and P. Mazumder, "Accurate Estimating Simultaneous Switching Noises by using Application Specific Device Modeling," in *Proceedings of the Design, Automation and Test in Europe Conference & Exhibition*, March 2002, pp. 1038–1043.

[86] M. Hekmat, S. Mirabbasi, and M. Hashemi, "Ground Bounce Calculation due to Simultaneous Switching in Deep Sub-micron Integrated Circuits," in *Proceedings of the IEEE International Symposium on Circuits and Systems*, May 2005, pp. 5617–5620.

[87] J. Kim *et al.*, "Closed-Form Expressions for the Maximum Transient Noise Voltage Caused by an IC Switching Current on a Power Distribution Network," *IEEE Transactions on Electromagnetic Compatibility*, vol. 54, no. 5, pp. 1112–1124, October 2012.

[88] J. Kim, J. Lee, S. Ahn, and J. Fan, "Closed-Form Expressions for the Noise Voltage Caused by a Burst Train of IC Switching Currents on a Power Distribution Network," *IEEE Transactions on Electromagnetic Compatibility*, vol. 56, no. 6, pp. 1585–1597, December 2014.

[89] W. Shockley, "The Theory of P-N Junctions in Semiconductors and P-N Junction Transistors," *Bell System Technical Journal*, vol. 28, no. 3, pp. 435–489, July 1949.

[90] T. Sakurai and A. Newton, "Alpha-Power Law MOSFET Model and its Applications to CMOS Inverter Delay and other Formulas," *IEEE Journal of Solid-State Circuits*, vol. 25, no. 2, pp. 584–594, April 1990.

[91] L. Chen, M. Marek-Sadowska, and F. Brewer, "Buffer Delay Change in the Presence of Power and Ground Noise," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 11, no. 3, pp. 461–473, June 2003.

[92] S. Takaya, M. Nagata, A. Sakai, T. Kariya, S. Uchiyama, H. Kobayashi, and H. Ikeda, "A 100GB/s Wide I/O with 4096b TSVs Through an Active Silicon Interposer with In-place Waveform Capturing," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2013 IEEE International*, Feb 2013, pp. 434–435.

[93] D. U. Lee, K. W. Kim, K. W. Kim, K. S. Lee, S. J. Byeon, J. H. Kim, J. H. Cho, J. Lee, and J. H. Chun, "A 1.2 V 8 Gb 8-Channel 128 GB/s High-Bandwidth Memory (HBM) Stacked DRAM With Effective I/O Test Circuits," *IEEE Journal of Solid-State Circuits*, vol. 50, no. 1, pp. 191–203, Janurary 2015.

[94] M. Popovich, M. Sotman, A. Kolodny, and E. Friedman, “Effective Radii of On-Chip Decoupling Capacitors,” *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 16, no. 7, pp. 894–907, July 2008.

[95] A. Todri *et al.*, “A Study of Tapered 3-D TSVs for Power and Thermal Integrity,” *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 21, no. 2, pp. 306–319, February 2013.

[96] E. Wong, J. Minz, and S. K. Lim, “Decoupling Capacitor Planning and Sizing For Noise and Leakage Reduction,” in *Proceedings of the IEEE International Conference on Computer-Aided Design*, November 2006, pp. 395–400.

[97] T. Xu, P. Li, and B. Yan, “Decoupling for Power Gating: Sources of Power Noise and Design Strategies,” in *Proceedings of the ACM/IEEE Design Automation Conference*, June 2011, pp. 1002 –1007.

[98] W. Wang, M. Anis, and S. Areibi, “Fast Techniques for Standby Leakage Reduction in MTCMOS Circuits,” in *Proceedings of the IEEE International System-on-Chip Conference*, September 2004, pp. 21–24.

[99] D.-S. Chiou, S.-H. Chen, and S.-C. Chang, “Sleep Transistor Sizing for Leakage Power Minimization Considering Charge Balancing,” *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 17, no. 9, pp. 1330–1334, September 2009.

[100] “Cadence Spectre.” [Online]. Available: {http://www.cadence.com/products/cic/spectrecircuit}

[101] S.-U. Park *et al.*, "Analysis of Reliability Characteristics of High Capacitance Density MIM Capacitors with SiO_2-HfO_2-SiO_2 Dielectrics," *Microelectronic Engineering*, vol. 88, no. 12, pp. 3389 – 3392, December 2011.

[102] K. Kawasaki *et al.*, "A Sub-μs Wake-Up Time Power Gating Technique With Bypass Power Line for Rush Current Support," *IEEE Journal of Solid-State Circuits*, vol. 44, no. 4, pp. 1178–1183, April 2009.

[103] J. Cho *et al.*, "Guard Ring Effect for Through Silicon Via (TSV) Noise Coupling Reduction," in *Proceedings of the IEEE CPMT (Components, Packaging, and Manufacturing Technology) Symposium Japan*, August 2010, pp. 1–4.

[104] M. Brocard, P. Le Maitre, C. Bermond, P. Bar, R. Anciant, A. Farcy, T. Lacrevaz, P. Leduc, P. Coudrain, N. Hotellier, H. Ben Jamaa, S. Cheramy, N. Sillon, J. Marin, and B. Flechet, "Characterization and Modelling of Si-substrate Noise Induced by RF Signal Propagating in TSV of 3D-IC Stack," in *Proceedings of the Electronic Components and Technology Conference*, May 2012, pp. 665–672.

[105] P. Saguet, "The 3D Transmission-Line Matrix Method: Theory and Comparison of the Processes," *International Journal of Numerical Modelling: Electronic Networks, Devices and Fields*, vol. 2, no. 4, pp. 191–201, December 1989.

[106] J. Cho *et al.*, "Modeling and Analysis of Through-Silicon Via (TSV) Noise Coupling and Suppression Using a Guard Ring," *IEEE Transactions on Com-*

ponents, Packaging, and Manufacturing Technology, vol. 1, no. 2, pp. 220–233, February 2011.

[107] J. M. Hutson, R. D. Schrimpf, and L. M. Massengill, "The Effects of Scaling and Well and Substrate Contact Placement on Single Event Latchup in Bulk CMOS Technology," in *Proceedings of the IEEE European Conference on Radiation and Its Effects on Components and Systems*, September 2005, pp. 19–23.

[108] T. Liu, J. Carothers, and W. Holman, "Active Substrate Coupling Noise Reduction Method for ICs," *Electronics Letters*, vol. 35, no. 19, pp. 1633–1634, September 1999.

[109] Y.-J. Lee, D. Limbrick, and S. K. Lim, "Power Benefit Study for Ultra-high Density Transistor-level Monolithic 3D ICs," in *Proceedings of the ACM/EDAC/IEEE Design Automation Conference*, May 2013, pp. 1–10.

[110] S.-M. Jung, H. Lim, K. Kwak, and K. Kim, "A 500-MHz DDR High-Performance 72-Mb 3-D SRAM Fabricated With Laser-Induced Epitaxial c-Si Growth Technology for a Stand-Alone and Embedded Memory Application," *IEEE Transactions on Electron Devices*, vol. 57, no. 2, pp. 474–481, February 2010.

[111] P. Batude *et al.*, "Advances in 3D CMOS Sequential Integration," in *Proceedings of the IEEE International Electron Devices Meeting*, December 2009, pp. 1–4.

[112] C. Liu and S. K. Lim, "A Design Tradeoff Study with Monolithic 3D Integration," in *Proceedings of the International Symposium on Quality Electronic Design*, March 2012, pp. 529–536.

[113] S. Panth, K. Samadi, Y. Du, and S. K. Lim, "Design and CAD Methodologies for Low Power Gate-level Monolithic 3D ICs," in *Proceedings of the IEEE/ACM International Symposium on Low Power Electronics and Design*, August 2014, pp. 171–176.

[114] "Cadence Virtuoso." [Online]. Available: http://www.cadence.com/products/cic/Pages/default.aspx

[115] "Mentor Calibre." [Online]. Available: https://www.mentor.com/products/ic_nanometer_design/verification-signoff/

[116] "Cadence Encounter Library Characterizer (ELC)." [Online]. Available: http://www.cadence.com/products/di/library_characterizer/pages/default.aspx

[117] "Cadence Abstract Generation." [Online]. Available: http://www.cadence.com/rl/resources/conference_papers/ctp_cndlivesv2006_sharma_abstracts6.1.0.pdf