Modeling and Design Techniques for 3-D ICs under Process, Voltage, and Temperature Variations

THÈSE N^{O} 5543 (2012)

PRÉSENTÉE LE 21 NOVEMBRE V2012

À LA FACULTÉ INFORMATIQUE ET COMMUNICATIONS

LABORATOIRE DES SYSTÈMES INTÉGRÉS (IC/STI)

PROGRAMME DOCTORAL EN INFORMATIQUE, COMMUNICATIONS ET INFORMATION

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Hu XU

acceptée sur proposition du jury:

Prof. B. Rimoldi, président du jury
Prof. G. De Micheli, directeur de thèse
Prof. L. Benini, rapporteur
Prof. Y. Leblebici, rapporteur
Prof. D. Soudris, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2012

Try not to become a man of success,
but rather try to become a man of value.
— Albert Einstein

To my parents...

Acknowledgements

Pursuing and completing my PhD degree is definitely one of the most exciting and challenging activities in the first 30 years of my life. I have shared both the sweet and difficult parts of this journey with many people. It has been a pleasant challenge and enjoyment to spend four years in EPFL, and its members will always remain dear to me.

First of all, I want to express my deep thanks to my thesis supervisor, Professor Giovanni De Micheli, for his continuous encouragement and thoughtful guidance along my pursuit of PhD. Professor De Micheli has set an excellent example for my research and future career. I want to show my sincere thanks to Professor Vasilis F. Pavlidis, for his kind patience and indispensable help to my work in EPFL. Professor Pavlidis' inspiring and scrupulous guidance has led me into the world of 3-D ICs. I also thank the other members of my PhD thesis committee: Professors Bixio Rimoldi, Yusurf Leblebici, Luca Benini, and Dimitris Soudris, for their precious time and helpful suggestions to improve my research and this thesis. I am very grateful to Professor Wayne P. Burleson, whose beneficial advices and kind concerns have helped me to carry forward my work and improve my research capability.

I want to express my gratitude to all my colleagues in Integrated Systems Laboratory, EPFL. Thanks to their warm help and kindness, I have spent a pleasant time in Lausanne, in both my PhD research and daily life. I am especially happy to thank Mme. Christina Govoni for her careful and tireless administrative service. I also want to thank the students I have worked with, Mr. Xifan Tang, Mr. Cheng Zhang, Mr. Muhammad Waqas Chaudhary, and Mr. Dimitrios Anagnostos, for their help to improve my research works.

Last but not the least, I am honoured to acknowledge my gratitude to Swiss National Science Foundation and Intel Braunschweig Lab (Germany), for their generous support.

Lausanne, 26 October 2012 Hu XU

Abstract

Three-dimensional (3-D) integration is a promising solution to further enhance the density and performance of modern *integrated circuits* (ICs). In 3-D ICs, multiple dies (tiers or planes) are vertically stacked. These dies can be designed and fabricated separately. In addition, these dies can be fabricated in different technologies. The effect of different sources of variations on 3-D circuits, consequently, differ from 2-D ICs. As technology scales, these variations significantly affect the performance of circuits. Therefore, it is increasingly important to accurately and efficiently model different sources of variations in 3-D ICs.

The process, voltage, and temperature variations in 3-D ICs are investigated in this dissertation. Related modeling and design techniques are proposed to design a robust 3-D IC. Process variations in 3-D ICs are first analyzed. The effect of process variations on synchronization and 3-D clock distribution networks, is carefully studied. A novel statistical model is proposed to describe the timing variation in 3-D clock distribution networks caused by process variations. Based on this model, different topologies of 3-D clock distribution networks are compared in terms of skew variation. A set of guidelines is proposed to design 3-D clock distribution networks with low clock uncertainty.

Voltage variations are described by power supply noise. Power supply noise in 3-D ICs is investigated considering different characteristics of potential 3-D power grids in this thesis. A new algorithm is developed to fast analyze the steady-state *IR*-drop in 3-D power grids. The first droop of power supply noise, also called resonant supply noise, is usually the deepest voltage drop in power distribution networks. The effect of resonant supply noise on 3-D clock distribution networks is investigated. The combined effect of process variations and power supply noise is modeled by skitter consisting of both skew and jitter. A novel statistical model of skitter is proposed. Based on this proposed model and simulation results, a set of guidelines has been proposed to mitigate the negative effect of process and voltage variations on 3-D clock distribution networks.

Thermal issues in 3-D ICs are considered by carefully modeling *thermal through silicon vias* (TTSVs) in this dissertation. TTSVs are vertical vias which do not carry signals, dedicated to facilitate the propagation of heat to reduce the temperature of 3-D ICs. Two analytic models are proposed to describe the heat transfer in 3-D circuits related to TTSVs herein, providing proper closed-form expressions for the thermal resistance of the TTSVs. The effect of different physical and geometric parameters of TTSVs on the temperature of 3-D ICs is analyzed. The proposed models can be used to fast and accurately estimate the temperature to avoid the overuse of TTSVs occupying a large portion of area.

Abstract

A set of models and design techniques is proposed in this dissertation to describe and mitigate the deleterious effects of process, voltage, and temperature variations in 3-D ICs. Due to the continuous shrink in the feature size of transistors, the large number of devices within one circuit, and the high operating frequency, the effect of these variations on the performance of 3-D ICs becomes increasingly significant. Accurately and efficiently estimating and controlling these variations are, consequently, critical tasks for the design of 3-D ICs.

Keywords: 3-D ICs, process variations, power supply noise, temperature variation, clock distribution networks, power distribution networks, clock skew, clock jitter, thermal model

Résumé

L'intégration dans la *Troisième-Dimension* (3-D) est une solution prometteuse pour continuer d'accroitre la densité et la performance des Circuits Intégrés (Integrated Circuits - ICs) modernes. Dans les circuits 3-D, de multiples puces sont empilées verticalement. Chacune de ces puces peut être conçue et fabriquée séparément. De plus, ces puces peuvent être fabriquées dans des technologies variées. Les effets des différentes sources de variations sur les circuits 3-D, diffèrent ainsi des circuits 2-D. A mesure que les technologies évoluent, les variations affectent significativement les performances des circuits. Ainsi, il est d'une important capitale de modéliser de façon précise et efficace les différentes sources de variations des circuits 3-D. Les variations dues aux procédés de fabrication, aux alimentations et là température sont étudiées dans ce manuscrit. La modélisation associée, ainsi que des techniques de conception adaptées sont proposées afin de concevoir des circuits 3-D robustes.

Les variations dues aux procédés de fabrication des circuits 3-D sont dans un premier temps analysées. Les effets des procédés de fabrication sur la synchronisation et la distribution de l'horloge sont étudiés minutieusement. Un nouveau modèle statistique est alors proposé pour décrire les variations de timing dans les réseaux 3-D de distribution d'horloge, sous l'influence des variations dans les procédés de fabrication. S'appuyant sur ce modèle, différentes topologies de réseaux d'horloge 3-D sont comparées en termes de skew. Une méthodologie de conception est alors proposée pour des distributions d'horloge 3-D avec une faible incertitude sur l'horloge.

Les variations de la tension sont décrites au moyen de bruit sur l'alimentation. Le bruit sur l'alimentation dans les circuits 3-D est étudié en considérant différents maillages pour la distribution d'alimentation. Un nouvel algorithme est alors développé pour permettre une analyse rapide de la chute statique du produit courant-résistance dans les mailles d'alimentation des circuits 3-D. Le premier pic sur le bruit d'alimentation, aussi appelé bruit résonnant d'alimentation, correspond habituellement à la plus importanter chute de tension du réseau. L'effet de la résonnance d'alimentation sur le réseau de distribution d'horloge est étudié. L'effet combiné de la variation des procédés de fabrication et du bruit sur l'alimentation est modélisé par du skew et du jitter. Un nouveau modèle statistique de skitter est alors proposé. Basé sur le modèle proposé et sur des résultats de simulation, un jeu de règles a été décrit afin de réduire l'effet négatif des variations sur le réseau de distribution d'horloge 3-D.

Les contraintes thermiques dans les circuits 3-D sont considérées par une modélisation soignée des contacts thermiques 3-D (Thermal Through Silicon Vias - TTSVs). Les TTSVs sont des interconnexions verticales ne transmettant pas de signaux mais dédiées à faciliter

l'évacuation de chaleur et donc de réduire la température des circuits 3-D. Deux modèles analytiques sont proposés pour décrire les transferts thermiques dans les circuits 3-D en présence de TTSVs, en proposant une forme close pour la résistance thermique des TTSVs. L'effet sur la température de différentes géométries et dimensions de TTSVs a été analysé. Les modèles proposés sont utilisables pour une estimation rapide et précise de la température afin d'éviter une utilisation abusive de TTSVs, qui occupent une large surface.
Dans cette thèse, un jeu de modèles et de techniques de conception est proposé afin de décrire et de réduire les effets nuisibles de la variation des procédés de fabrication, de l'alimentation et de la température dans les circuits 3-D. En raison de la réduction des dimensions des transistors, du grand nombre de composants au sein d'un même circuit et de l'importante fréquence de fonctionnement, les effets de ces variations sur les performances des circuits 3-D deviennent de plus en plus importants. L'estimation fine et efficace, ainsi que le contrôle de ces variations sont ainsi une tâche critique dans la conception de circuits 3-D.

Mots-clés : Circuits 3-D, variations de procédés, buit d'alimentation, variation de température, réseau de distribution d'horloge, réseau de distribution d'alimentation, skew, jitter, modèles thermiques

Contents

List of Figures

List of Tables

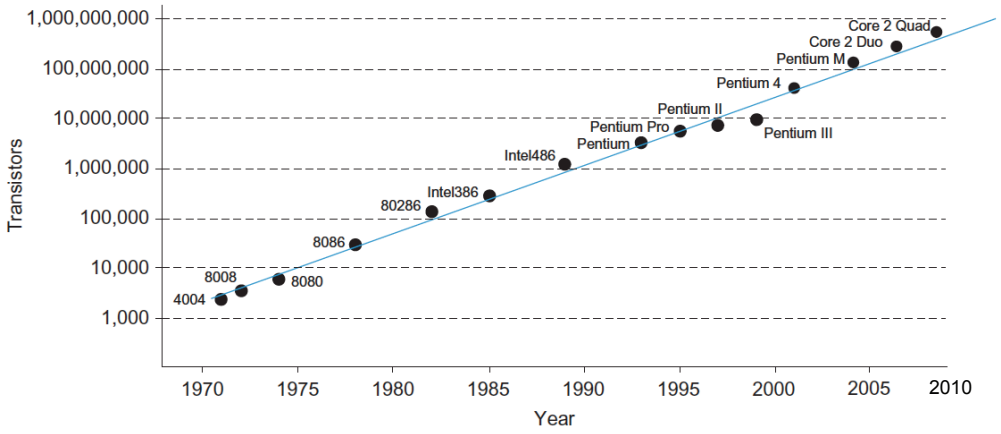
1 Introduction

Three-dimensional (3-D) integration emerges as a promising system integration approach to increase the density of devices, to shorten the interconnects of circuits, and, thus, to enhance the performance of circuits. Since multiple circuits are vertically stacked to form a 3-D system, the combined variations of these circuits significantly differ from traditional 2-D circuits. In addition, synchronizing devices across tiers becomes challenging. The modeling methods for different sources of variations and the design techniques to increase robustness in 3-D ICs, in particular 3-D clock distribution networks, are the focus of this dissertation.

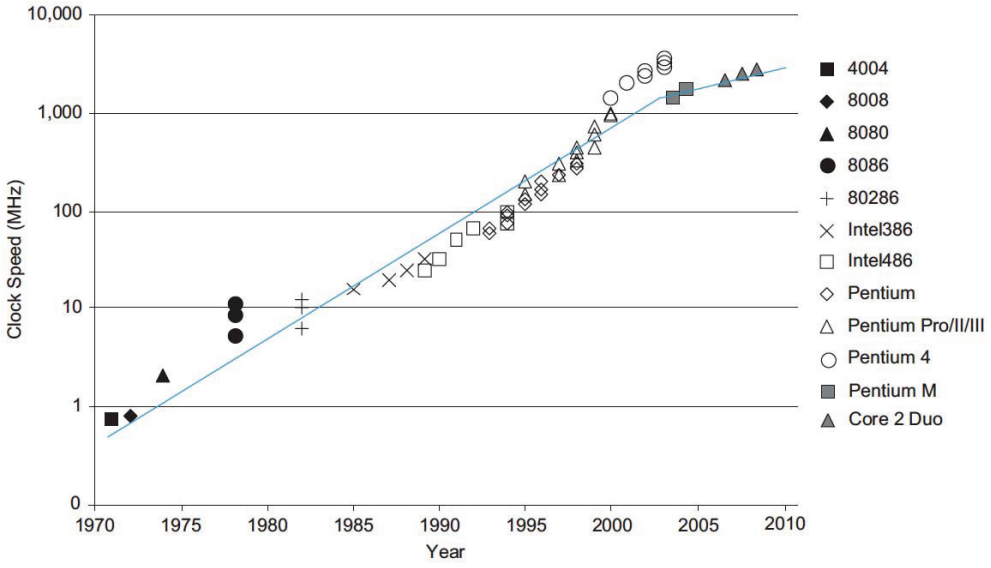
The fundamentals of 3-D IC design and fabrication are introduced in the following section. The classification and manufacturing technologies of 3-D ICs are introduced in Sections 1.2 and 1.3, respectively. Clock distribution topologies for 3-D circuits are introduced in Section 1.4, where the synchronization approaches of digital circuits are discussed. The contributions of this dissertation are presented in Section 1.5. The assumptions and limitations of this work are summarized in Section 1.6. The organization of this thesis is listed in the last section.

1.1 Background of 3-D ICs

During the last decades, integrated circuits have experienced a tremendous increase in density, functionality, and speed. One of the most important driving forces for this continuous growth is the persistent downscaling of the size of transistors. The rapid development of electronic devices originated from 1947, where the first transistor was fabricated by J. Bardeen, W. Brattain, and W. Shockley [42]. After this first transistor with a size of over 10 cm, the size of transistors has fast scaled down to 22 nm for massive production in 2012 [43]. When the first integrated circuit was invented in 1958 by Jack Kilby [2], only two transistors were integrated within this circuit. In 2012, however, Intel Core i7-990X processor contains more than one billion transistors integrated within one chip [1]. The number of transistors within one circuit, the speed of circuit, and the functionality of integrated circuits all increase dramatically as technology scales , as illustrated in Fig. 1.1 [1, 2].



(a)



(b)

Figure 1.1: The development of integrated circuits over time, where (a) and (b) are the increase in the number of transistors and the clock frequency of Intel processors, respectively [1, 2].

Nevertheless, as the feature size of transistors becomes smaller than 20 nm, traditional planar ICs have encountered new challenges. First, due to the limitation of manufacturing technologies and materials, it is extremely difficult and expensive to continue the scaling of feature size [44]. Second, the interconnect delay becomes dominant over the gate delay due to the increase in the interconnect length and *RC* delay, as depicted in Fig. 1.2 [3]. This increase is mainly due to the increase in the footprint size of circuits and the decrease in the width, space, and height of wires. Therefore, it is challenging for 2-D ICs to further enhance the density and speed of circuits simply by scaling. Consequently, new technologies are required to provide higher density, higher performance, and more functionalities for integrated systems.

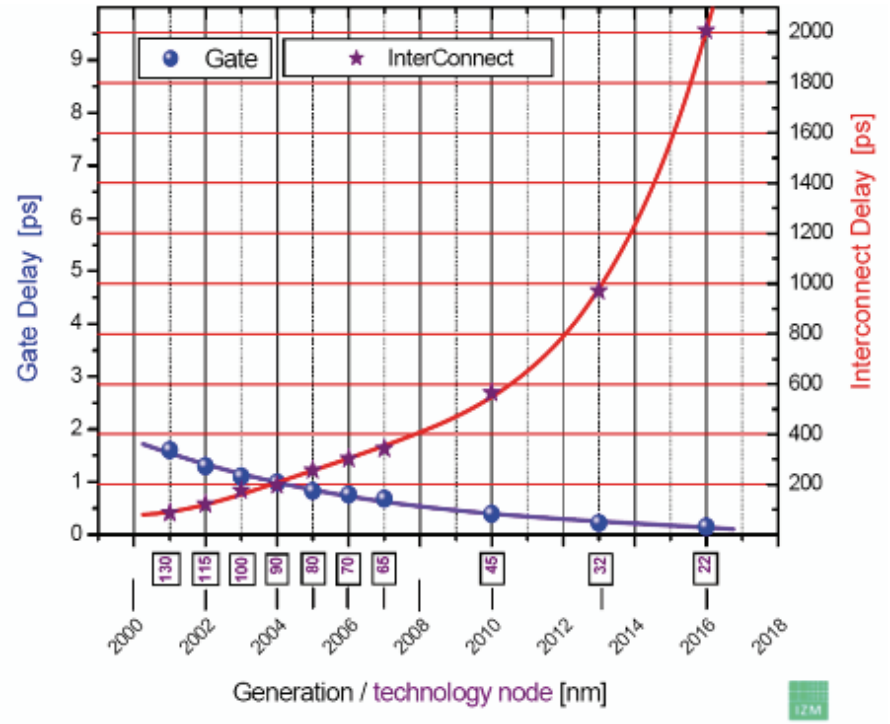


Figure 1.2: The increase in interconnect delay with technology generations [3].

Three-dimensional integration provides a promising solution to continue improving the performance of integrated circuits. In 3-D ICs, vertical integration can be implemented at different levels. For instance, multiple chips/dies can be fabricated separately and then vertically stacked together. Each chip forms a plane or tier of the final circuit. An example of this type of 3-D circuits is illustrated in Fig. 1.3(a) [4], where a processor layer is stacked on top of a memory layer. At lower levels, devices (transistors) can be stacked together to provide higher density. For instance, the first vertically-stacked device was fabricated in early 1980s as illustrated in Fig. 1.3(b) [5]. By utilizing vertical integration, 3-D ICs exhibit three major advantages over traditional planar circuits: higher density, shorter interconnects, and easier heterogeneous integration.

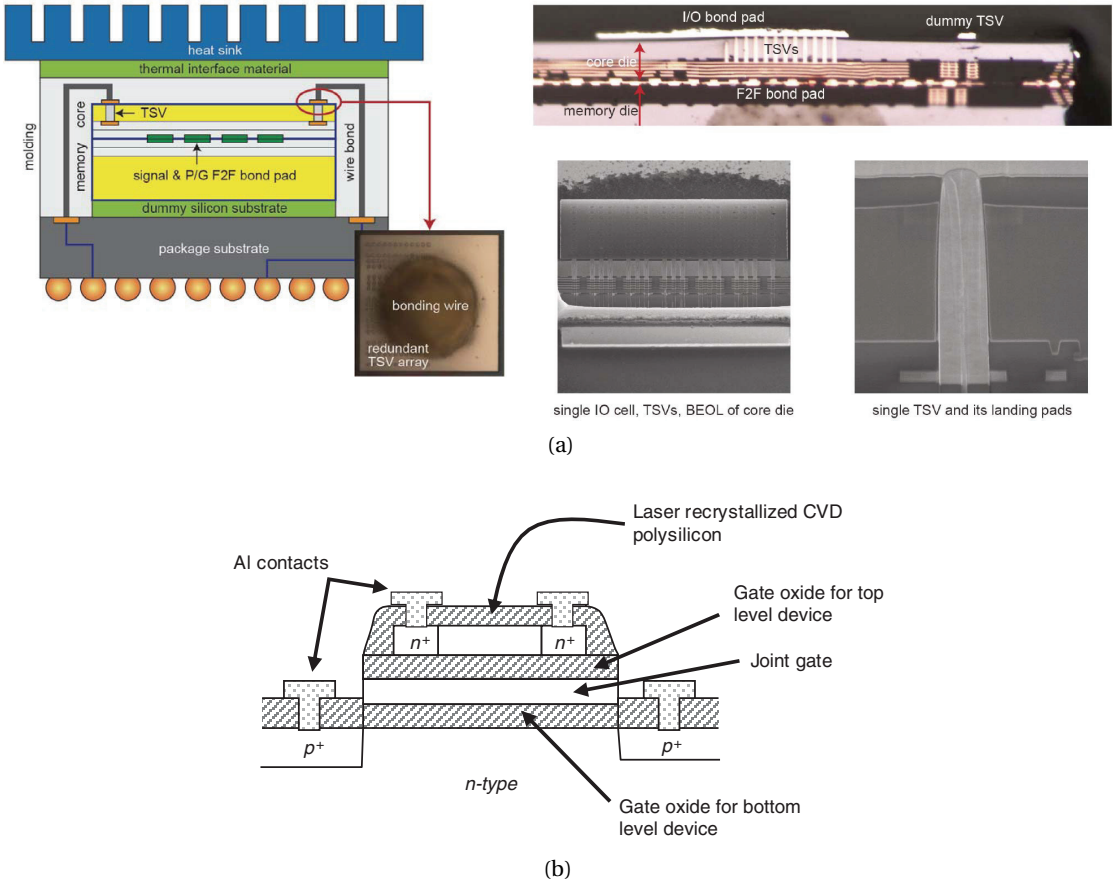


Figure 1.3: Different levels of 3-D integration, where (a) is the cross-section of a 3-D circuit consisting of two dies [4] and (b) is a vertically stacked inverter [5].

Higher density

By vertically stacking devices or entire circuits, a larger number of transistors can be integrated with the same area as compared to a planar IC. Assuming that n dies with a similar number of transistors are vertically stacked in a 3-D circuit, the resulting density is n times that of the corresponding 2-D circuit. Consequently, the density of circuits can significantly be increased by 3-D integration. For instance, a 3-D DRAM fabricated by Samsung has achieved a capacity of 8 GB with a 50% increase in density [45, 46].

Shorter interconnects

For the same number of transistors, 3-D ICs exhibit a smaller footprint than 2-D circuits due to the higher density. In addition, vertical interconnects are used in 3-D circuits. Consequently, the on-chip interconnect length can be greatly reduced. The decrease in the length of interconnects with the number of tiers is plotted in Fig. 1.4. As shown in this figure, the

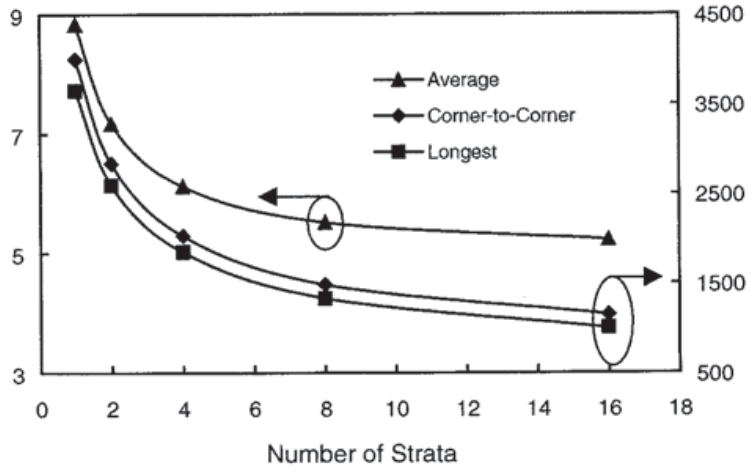


Figure 1.4: Lengths of the longest and average interconnects *vs.* the number of tiers [6].

interconnect length decreases significantly as multiple circuits are vertically stacked. This reduction in interconnect length, consequently, decreases the interconnect delay. As shown in Fig. 1.2, since the interconnect delay dominates the overall delay of a circuit, decreasing the interconnect delay helps to enhance the speed of circuits.

Heterogeneous integration

There are currently three main application domains of 3-D ICs: memory stack, memory-on-microprocessor, and analog/digital mixed-signal circuits. A memory stack is a homogeneous 3-D circuit consisting of multiple memory dies [45]. Nevertheless, memory-on-microprocessor [4] and analog/digital mixed-signal 3-D ICs are heterogeneous 3-D circuits. Heterogeneous integration, where different types of circuit blocks are integrated within one circuit, provides multiple functionalities within one system [47, 48]. In 2-D ICs, these different circuits have to be fabricated with the same technology. In 3-D ICs, however, these circuits can be located in different tiers and fabricated with different technologies. This feature greatly eases heterogeneous integration. Consequently, DRAM and processors can be fabricated separately with different technologies and stacked together [4]. Analog sensors and digital process units can also be designed and fabricated separately and then integrated into one 3-D circuit [48].

Several challenging issues of heterogeneous integration can be mitigated with 3-D ICs. For instance, analog and digital circuits can coexist in heterogeneous circuits. In 2-D ICs, since the same silicon substrate is shared by the entire circuit, substrate crosstalk can introduce large noise during the operation of these circuits. 3-D ICs, consequently, provides a novel way to mitigate this noise. As illustrated in Fig. 1.5 [7], the digital and analog parts of a 3-D circuit can be located in different tiers manufactured on different substrates. The substrate crosstalk,

therefore, can significantly be mitigated. In addition, due to the higher density, heterogeneous 3-D ICs provide several functions with a smaller footprint as compared to 2-D systems.

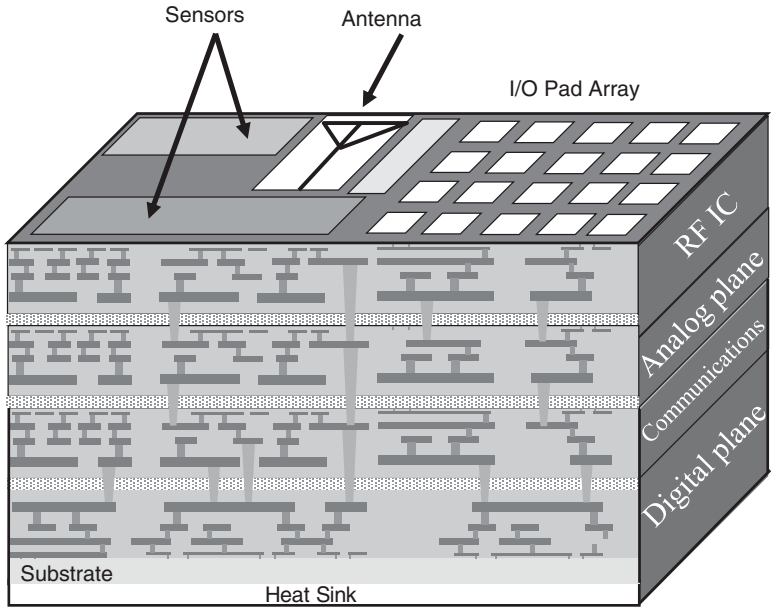


Figure 1.5: An example of a heterogeneous 3-D circuit containing both digital and analog circuitries [7].

Due to the significant advantages of 3-D ICs over conventional 2-D ICs, a strong research effort has been put in both academia and industry [7, 49]. Currently, different types of 3-D circuits have been proposed and fabricated. The classification of 3-D systems is introduced in the following section.

1.2 Classification of 3-D Circuits

Different types of 3-D integrated systems are introduced in this section. As previously mentioned, vertical integration can be implemented at different levels. Based on the integration level and the interconnection technologies among tiers, 3-D systems can roughly be classified into three primary categories: *System-in-Package* (SiP), *System-on-Package* (SoP), and fine-grain 3-D ICs [7].

1.2.1 System-in-Package

An SiP is a system assembling either bare or packaged dies along the vertical direction. The individual dies of an SiP are separately designed. Interconnections among dies are primarily implemented with four methods [7]:

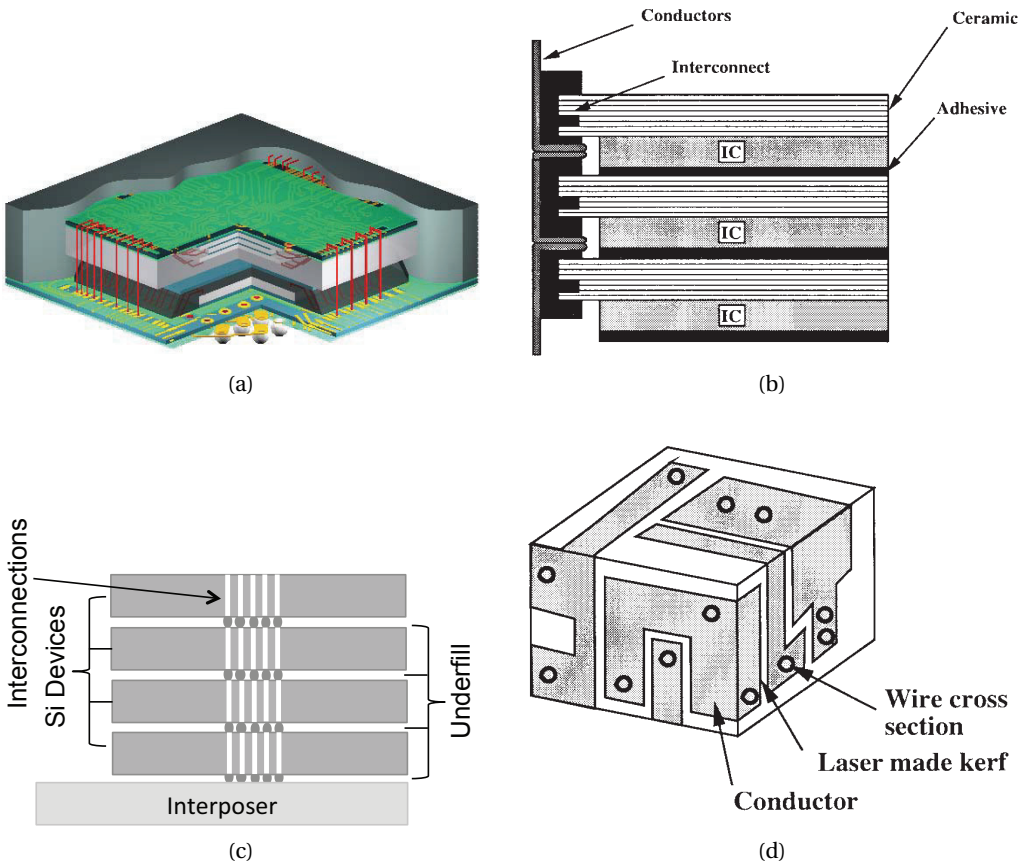


Figure 1.6: System-in-Packages implemented with different methods: (a) wire bonding [8], (b) interconnects on the periphery of dies [9], (c) vertical interconnect array [10], and (d) interconnects on the faces of a 3-D stack [9].

- Wire bonding, as shown in Fig. 1.6(a)
- Vertical interconnects on the periphery of dies/packages, as illustrated in Fig. 1.6(b)
- Low aspect ratio (length over diameter) and low density vertical interconnects arranged in an array, as depicted in Fig. 1.6(c)
- Metallization on the faces of a 3-D stack, as illustrated in Fig. 1.6(d)

Due to the ease of design and fabrication, wire bonding is the most common inter-die communication mechanism used in SiPs [7]. SiPs significantly enhance the packaging efficiency and reduce the form factor. The dies within an SiP are integrated at the die or package level, where only coarse-grain interconnections can be achieved among the circuitries in different tiers. As shown in Fig. 1.6, due to the limited locations and low density of vertical interconnects, the advantage of 3-D integration to support shorter interconnects cannot be fully utilized in 3-D SiPs.

1.2.2 System-on-Package

In SiPs, inter-die communication can only be implemented at specific locations. To facilitate the communication among dies, System-on-Packages have been proposed [50]. As illustrated in Fig. 1.7, a silicon interposer is used to host all the dies. Metal wires can be used in this interposer to interconnect different dies. TSVs are used to connect all the chips through the interposer with the package. SoPs are also referred as 2.5-D ICs, since multiple dies are located on a planar interposer. In 2.5-D SoPs, all the chips can be designed similar to conventional 2-D circuits. These chips are flipped down and bonded to the interposer. Consequently, 2.5-D SoPs are easier to design and fabricate as compared with fine-grain 3-D ICs. Industrial products of 2.5-D FPGAs have recently been released by Xilinx Inc. [11].

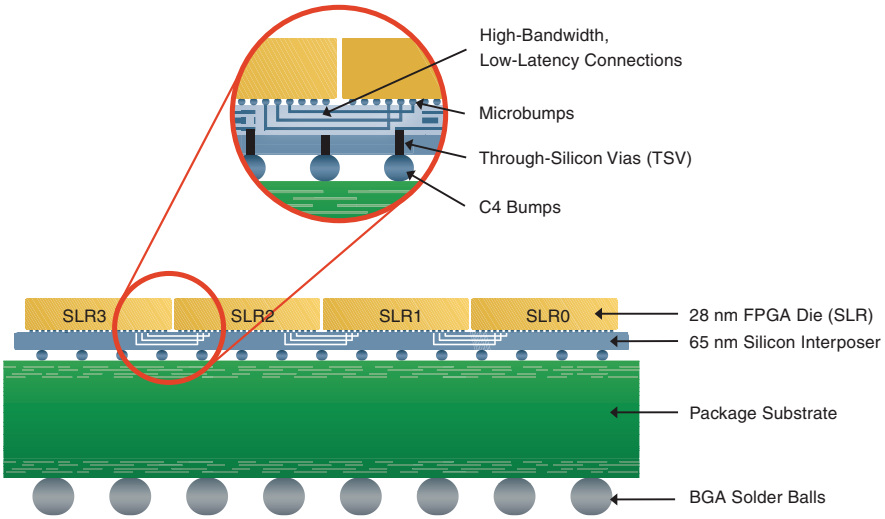


Figure 1.7: Xilinx Virtex-7 FPGA based on 2.5-D System-on-Package [11].

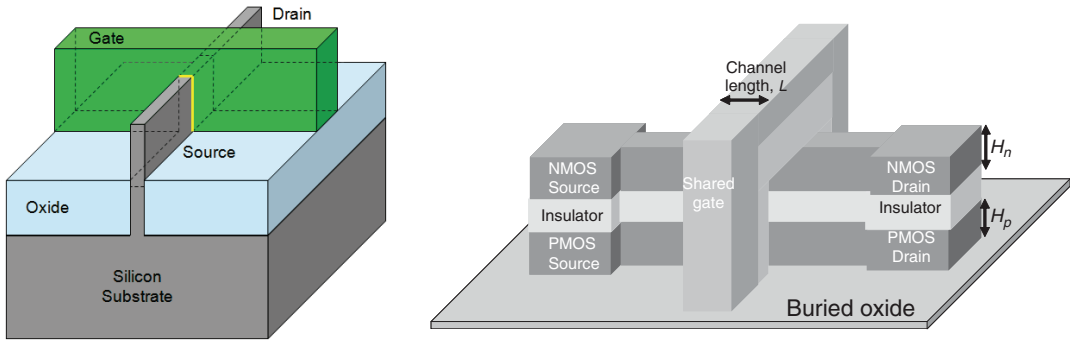
1.2.3 Fine-grain 3-D ICs

To further decrease the interconnect length and delay, fine-grain 3-D ICs have been proposed. In general, fine-grain 3-D ICs refer to the circuits where the devices, gates, and circuit blocks can be vertically distributed among physical planes [51]. The interconnection among planes can be implemented (*e.g.*, by TSVs) at any "legal" location. To avoid confusion, the term "3-D IC" refers to fine-grain 3-D ICs in the remainder of this dissertation. Differently from the package-level integration in SiPs, 3-D ICs are integrated at lower levels. Depending on the fabrication process, 3-D ICs are classified into two main categories: monolithic and polylithic 3-D circuits [7].

Monolithic 3-D ICs

Monolithic 3-D ICs are fabricated in a batch process. Different planes of a monolithic 3-D circuit are successively fabricated. The devices in the upper planes are grown above the lower planes. Monolithic 3-D ICs can be further divided into two types: stacked 3-D ICs and 3-D *fin field effect transistors* (fin-FETs).

- In stacked monolithic 3-D ICs, multiple layers of planar transistors are successively grown on top of conventional CMOS or *Silicon-on-Insulator* (SOI) planes. As illustrated in Fig. 1.3(b), the transistor-level integration is achieved by stacking conventional transistors [52].
- A fin-FET is a nonplanar and multi-gate transistor built on an SOI substrate [53]. The conducting channel is surrounded by a silicon fin. Fin-FET-like transistors are successfully utilized in industrial products, *e.g.*, Intel Tri-Gate transistors [12] as illustrated in Fig. 1.8(a). The advantages of fin-FETs include the high driving current and the significant reduction in both the gate area and routing within one gate. 3-D fin-FETs are a novel transistor structure based on conventional fin-FETs. In 3-D fin-FETs, devices stacked together share the same gate [13, 54], as illustrated in Fig. 1.8(b). The density of devices is further increased as compared with conventional fin-FETs.



(a) (b)

Figure 1.8: Different types of fin-FETs, where (a) and (b) are an Intel Tri-Gate [12] and a 3-D fin-FET [13], respectively.

Polylithic 3-D ICs

In polylithic 3-D ICs, different planes of a circuit are separately fabricated and then bonded together. In contrast to SiPs, in polylithic 3-D ICs, vertical interconnections are not limited to the periphery or in a fixed area array arrangement. These interconnections can be implemented in all possible locations (not occupied by transistors) by *Through Silicon Vias* (TSVs) [55], inductive coupling [56], or capacitive coupling [57], as illustrated in Fig. 1.9.

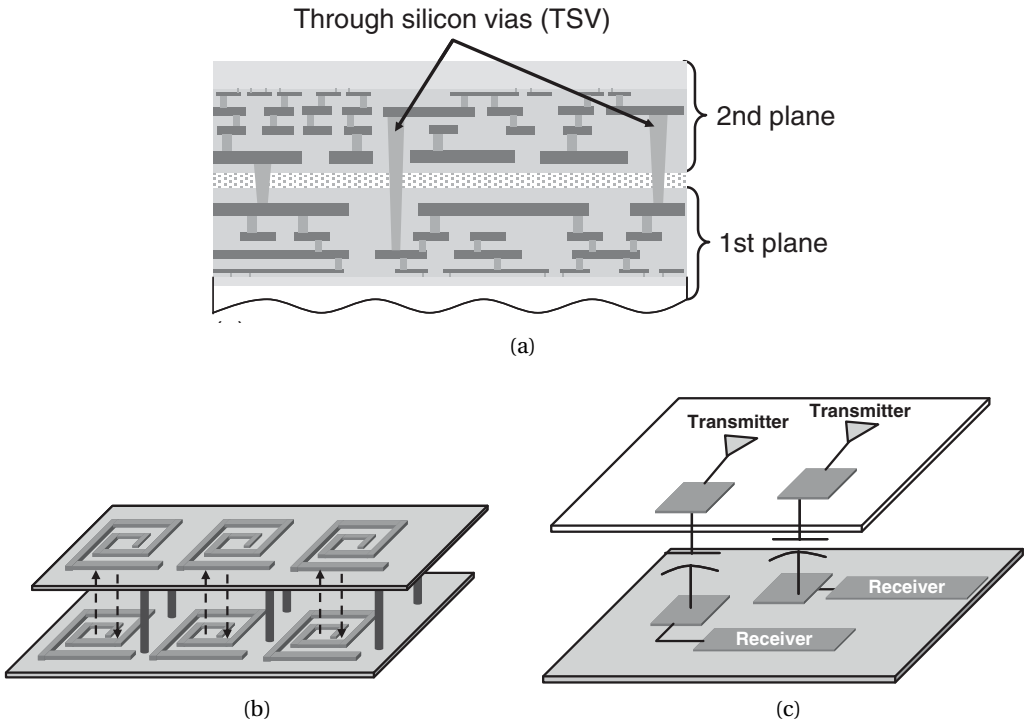


Figure 1.9: Communication mechanisms in different fine-grain 3-D ICs [7]: (a) TSVs, (b) inductive coupling, and (c) capacitive coupling.

Wafer or die-level 3-D integration utilizing TSVs is the most attractive solution for fine-grain 3-D ICs [7,55]. First, TSVs can be inserted at any available location where vertical interconnection is required, fully exploiting the advantage of 3-D ICs in reducing interconnect length and delay. Second, different tiers of TSV-based 3-D ICs are separately fabricated, which shortens the manufacturing time as compared to monolithic 3-D ICs and allows integration of different technologies in different tiers. Consequently, TSV-based 3-D ICs are investigated in this dissertation. In the following context, 3-D ICs directly imply 3-D circuits using TSVs to communicate among tiers.

1.3 Manufacturing Technologies for 3-D ICs

Manufacturing technologies and processes for 3-D ICs are introduced in this section. Different fabrication processes of TSVs are introduced in Section 1.3.1. The resulting physical and electrical characteristics of TSVs are presented in Section 1.3.2.

1.3.1 TSV-based 3-D ICs

Although no standardized fabrication technique has yet been established, several types of fabrication processes have been used to manufacture TSV-based 3-D ICs [58, 59]. These processes, actually, share a similar sequence of fabrication stages [7], as depicted in Fig. 1.10. As the first step, CMOS or SOI wafers are separately fabricated, which provide the physical

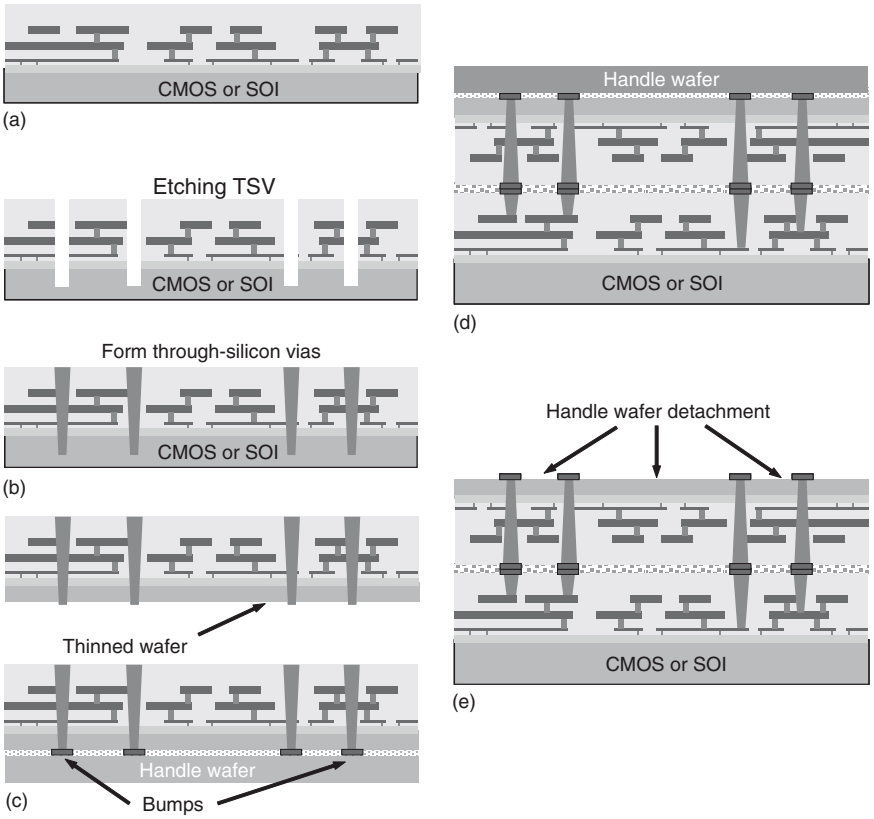


Figure 1.10: Typical fabrication steps for 3-D ICs [7]: (a) wafer preparation, (b) TSV etching, (c) wafer thinning, bumping, and handle wafer attachment, (d) wafer bonding, and (e) handle wafer removal.

planes (tiers) for subsequent bonding, as shown in Fig. 1.10(a). In Fig. 1.10(b), vertical TSVs are etched and filled with a conductive material, such as tungsten (W), copper (Cu), or low-resistance polysilicon. To reduce the length of TSVs, wafers need to be thinned to different thickness depending on the techniques. To mechanically support the thinned wafers which are difficult to handle and bond, these wafers are attached to "handle" wafers, as shown in Fig. 1.10(c). The alignment and bonding between dies or wafers follow successively as in Fig. 1.10(d). Afterwards, the handle wafer is removed and the corresponding side of wafer is processed and bonded to another physical plane, if any.

Since all the inter-plane communication is implemented through TSVs, the fabrication of TSVs with high density is crucial to 3-D ICs. A fabrication technology for TSVs should provide reliable, inexpensive, low-impedance, and area-efficient vertical interconnections. Depending on which stage TSVs are fabricated at, there are presently three types of fabrication processes: via-first, via-middle, and via-last [43, 60–63]. In some other works, via-first and via-middle technologies are both designated as via-first technology [7]. To avoid confusion, via-first and via-middle are differentiated from each other in this dissertation.

- In via-first process, TSVs are fabricated before the silicon *front-end of line* (FEOL) device fabrication. Polysilicon TSVs are usually employed in via-first processes to avoid metal contamination. Via-first TSVs usually have small aspect ratio with a diameter below 5 μm [63].
- In via-middle process, TSVs are fabricated after the silicon FEOL device fabrication processing and before the *back-end of line* (BEOL) interconnect process.
- In via-last process, TSVs are fabricated after or during BEOL interconnect process [43]. Via-last TSVs can have an aspect ratio between 3 and 20 and the diameter is about 10-50 μm [63]. An important advantage of via-last process is the possibility for the foundries without TSV manufacturing capabilities to fabricate the individual tiers separately. Tungsten and copper TSVs can be used in both via-middle and via-last processes.

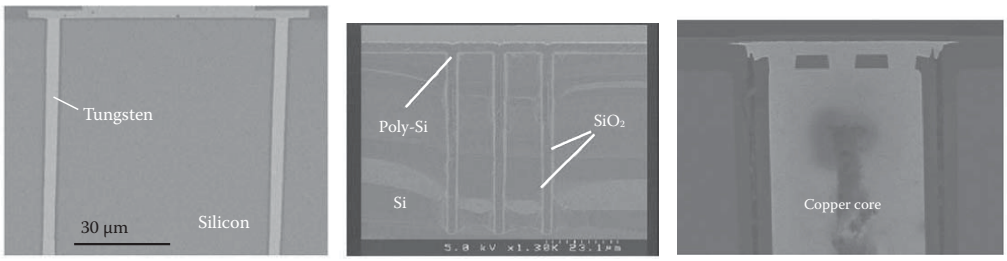
TSVs exhibit different geometric sizes and electrical characteristics in different fabrication processes. The physical and electrical characteristics of TSVs are introduced in the following subsection.

1.3.2 Physical and electrical characteristics of TSVs

Typically, a TSV consists of a dielectric liner, a barrier layer (for copper TSVs to prevent Cu diffusion), and the filling material. Examples of TSVs with different filling materials are illustrated in Fig. 1.11. The shape and geometric sizes differ among different types of TSVs. In general, TSVs have either straight or tapered shape with various aspect ratios [7]. TSVs with a wide range of diameters and depths have been manufactured. For instance, a comparison in the diameter and depth of TSVs among different technologies is listed in Table 1.1 [7].

As shown in this table, the diameter of TSVs can range from a few micrometers to nearly 100 μm. The effect of the physical parameters of TSVs on the timing, power, and temperature of circuits needs to be investigated. Several electrical models have been proposed to model the electrical behavior of TSVs [67–73]. For instance, an early electrical model of TSVs is illustrated in Fig. 1.12. The *RLC* characteristics of TSVs can be abstracted based on the structure and material of the vias.

The total resistance, capacitance, and inductance are commonly used to concisely describe the electrical characteristics of TSVs. These electrical characteristics vary with the physical



(a) (b) (c)

Figure 1.11: Examples of TSVs using different filling materials: (a) IBM tungsten TSV [14], (b) Tohoku University polysilicon TSV [15], and (c) Cu TSV [16].

Table 1.1: Dimensions and resistances of TSVs from different technologies [7].

Process	Depth [μm]	Diameter [μm]	Total resistance [mΩ]
[64]	25	4	140
[65]	30	2×12	230
[66]	80	5/15	9.4/2.6
[66]	150	5/15	2.7/1.9
[17]	90	75	2.4

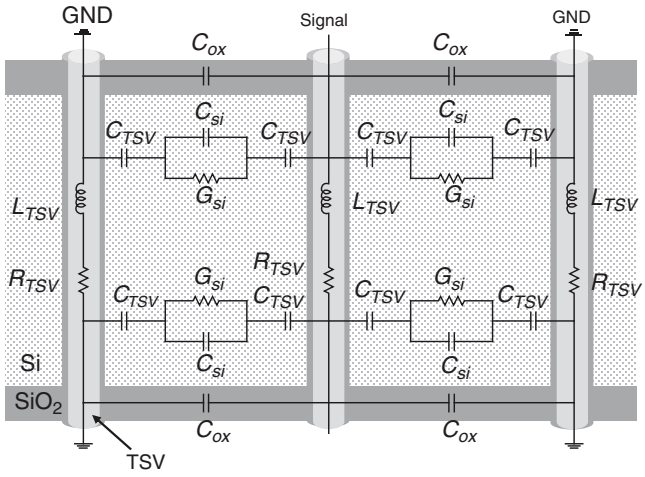


Figure 1.12: Electrical model of a TSV [7, 17].

Table 1.2: Electrical characteristics of different TSVs.

Reference	Diameter [μm]	Length [μm]	R [mΩ]	C [fF]	L [pH]
[68]	2	20	119.3	52.4	13.8
[70]	1.5	18	152.4	2.1	4.7
[17]	55	165	12	922	35

parameters of TSVs. As reported in Table 1.1, the total resistance can vary from a few milliohms to 200 mΩ. As another example, the electrical characteristics of other TSV processes are listed in Table 1.2. As reported in this table, the resistance, capacitance, and inductance differ significantly among fabrication technologies. In general, TSVs with larger diameter produce lower resistance, but higher capacitance and inductance. These electrical parameters all increase with the length of TSVs. To compare the electrical characteristics of TSVs with horizontal wires, a set of typical *RLC* parameters of global interconnects in 2-D circuits for different technology nodes is listed in Table 1.3 [41]. As listed in this table, the *RC* delay of

Table 1.3: Electrical characteristics of horizontal global interconnects [41].

Tech. [nm]	Width [μm]	Space [μm]	Height [μm]	ILD thickness [μm]	R [Ω/mm]	C [fF/mm]	L [nH/mm]
32	0.23	0.23	0.39	0.25	245.3	141.9	1.3
65	0.45	0.45	1.20	0.20	40.7	205.9	1.1
90	0.50	0.50	1.20	0.30	36.7	234.2	1.1

horizontal interconnects significantly increases as technology scales due to the large resistance. For a wire at 32 nm technology with a length longer than 18 μm, the *RC* constant is already larger than the largest *RC* constant of TSVs in Table 1.2. Consequently, TSVs can provide fast vertical interconnection in 3-D ICs, thereby decreasing the interconnect delay.

Although 3-D ICs exhibit various advantages over conventional 2-D circuits, new challenges also come with 3-D ICs that need to be addressed. These challenges include thermal issues, fabrication difficulties, manufacturing cost, test flow, and physical design issues [7, 49].

Among the physical design issues, synchronization among the tiers is a predominant problem. As a significantly larger number of devices can be vertically integrated within one 3-D circuit, synchronizing these devices in different tiers is even more challenging than in 2-D ICs. Different sources of variations from different tiers also complicate the synchronization among tiers. A short discussion on the synchronization approaches for digital circuits is provided in the following section.

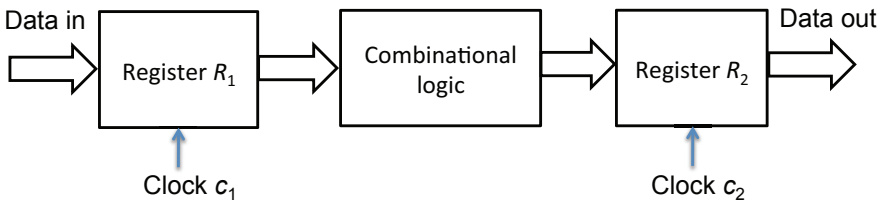


Figure 1.13: A data path including combinational and sequential circuits.

1.4 Clock Distribution Networks in 3-D ICs

The clock distribution networks in 3-D ICs are introduced in this section. Clock distribution networks are used to synchronize data transfer in digital circuits. The synchronization mechanism in digital ICs is introduced. Different clock distribution topologies for both 2-D and 3-D ICs are then presented.

A digital circuit consists of combinational and sequential parts, as illustrated in Fig. 1.13. The sequential elements R_1 and R_2 are usually implemented with flip-flops controlled by a clock signal. Depending on the relationship between c_1 and c_2, the synchronization approaches can be classified into five categories [2, 74]:

1. Synchronous. Clocks have the same frequency and phase. The data signal can be directly sampled with the clock.

2. Mesochronous. Clocks have the same frequency but different phases. The data signal can be sampled with a specified delay.

3. Plesiochronous. Clocks have nearly the same frequency. The difference in phase shifts slowly with time. The data signal can be sampled with a variable but predictable amount of delay. The difference in frequency may lead to missed or duplicated data.

4. Periodic. The data generated from the sender (R_1) is periodic at an arbitrary frequency. The data can be sampled with a predicted varying delay based on the periodic property.

5. Asynchronous. Clock signals have unknown difference in both the frequency and phase. Synchronizers are required to correctly sample the data at the receiver (R_2).

Synchronous circuits are the most widely used synchronization approach for on-chip communication due to its simplicity and robustness. In addition, no extra synchronizing element is required for the receiver. Consequently, synchronous circuits are investigated in this dissertation. The corresponding synchronization mechanism is introduced in the following subsection.

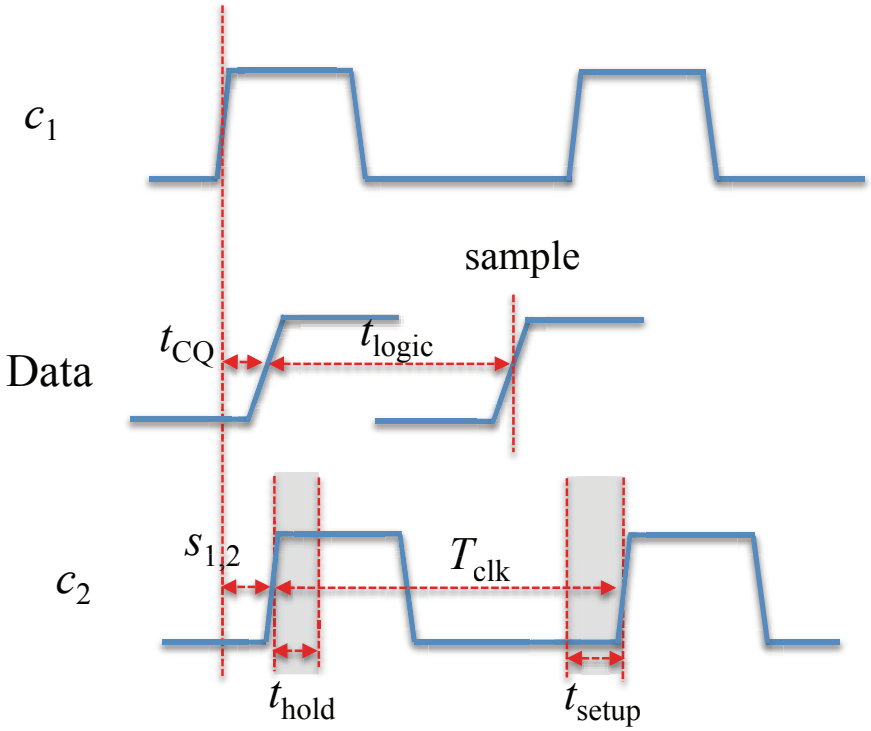


Figure 1.14: The waveforms of clock and data signals of the components in Fig. 1.13.

1.4.1 Synchronous circuits

As previously mentioned, all the registers in a synchronous circuit are synchronized by the same clock. Assuming the circuit in Fig. 1.13 is a synchronous circuit, the waveforms of the data and clock signals are drawn in Fig. 1.14. The data is sent from R_1 to R_2, which are synchronized by clocks c_1 and c_2, respectively. To correctly sample the data at R_2, the time when the data arrives at R_2 should not fall into the shadowed area. This requirement can be formulated by the following expressions [75],

$$\min(t_{\text{logic}}) + t_{\text{CQ}} \geq s_{1,2} + t_{\text{hold}}, \tag{1.1}$$

$$\max(t_{\text{logic}}) + t_{\text{CQ}} \leq s_{1,2} + T_{\text{clk}} - t_{\text{setup}}. \tag{1.2}$$

The clock period is denoted by T_{clk}. The delay from the clock pin to the data output of R_1 is t_{CQ}. The delay of the combinational logic is t_{logic}. The hold and setup times of R_2 are denoted by t_{hold} and t_{setup}, respectively. These two terms are determined by the design of the flip-flops. The difference between the delay of c_2 and c_1 is defined as clock skew $s_{1,2}$. Expressions (1.1) and (1.2) describe hold and setup time constraints, respectively.

As shown by (1.1) and (1.2), correct data transfer is determined by the data propagation delay, the traits of the flip-flops, and the clock distribution. For high-speed digital circuits with a high clock frequency, clock period T_{clk} is relatively short. The hold and setup slacks, therefore,

are highly sensitive to $s_{1,2}$. In synchronous circuits, clock c_1 and c_2 ideally have the same clock frequency and phase. If the delay from the clock source to the clock pins of R_1 and R_2 is the same, skew $s_{1,2}$ ideally should be zero (except for intentional skew scheduling [76]). Nevertheless, due to the large number of flip-flops in a circuit, clock is propagated to all these sinks through a large clock distribution network. The clock delay to different sinks, consequently, is significantly affected by the structure of clock distribution networks. A careful design of the clock distribution is important for the correct operation of circuits.

1.4.2 Clock signal distribution

The design of modern clock distribution networks faces four primary challenges [18]: 1) the strict constraint on clock skew and jitter due to the high clock frequency, 2) the large capacitive load of clock networks and long clock paths due to the large number of devices and large circuit area, 3) the increase in on-chip variations due to technology scaling, and 4) the strict power envelope imposed by the system specifications. To address these challenges, different types of clock distribution networks have been proposed.

Unbalanced clock tree

Clock trees are a widely used structure to propagate the clock signal from a unique clock source to different clock sinks. Clock trees can be classified into two categories: unbalanced (or asymmetric) and balanced (or symmetric) clock trees. An unbalanced clock tree is illustrated in Fig. 1.15. Unbalanced clock trees are usually generated from clock tree synthesis flows [77–79]. In unbalanced clock trees, clock sinks can be routed with an asymmetric topology. Different numbers and sizes of buffers are inserted along different clock paths. Clock skew among clock sinks is minimized or scheduled as desired. The clock synthesis algorithms can be designed to optimize different objectives, such as clock skew, total wire length, power, *etc.* Nevertheless, due to the unmatched clock buffers and paths, unbalanced clock trees are highly sensitive to *process, voltage, and temperature* (PVT) variations. Consequently, unbalanced clock trees are usually used in small circuits or within several blocks of a circuit.

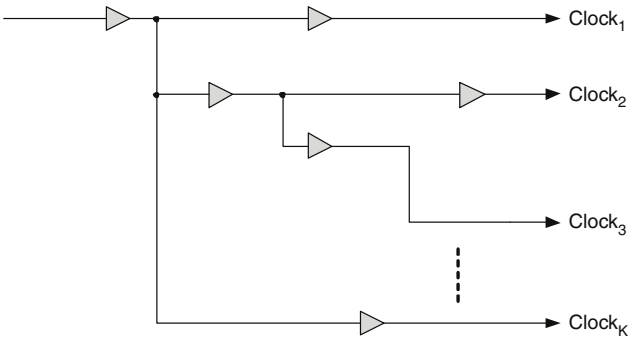


Figure 1.15: An unbalanced clock tree [18].

Balanced clock tree

Typical structures of balanced clock trees are illustrated in Fig. 1.16. A symmetric topology (H-tree or X-tree, as illustrated in Figs. 1.16(a) and 1.16(b), respectively) is used to connect clock sinks. Identical number and size of buffers are inserted along different clock paths. Consequently, balanced clock trees are more robust than unbalanced clock trees for PVT variations. The routing resources required by symmetric trees, however, are relatively high. In addition, since clock buffers are located across the chip, these buffers are non-uniformly affected by on-chip variations. To further improve the robustness, binary trees are used, as illustrated in Fig. 1.16(c). Different clock paths can be shorted at different levels. Clock buffers can be placed close in a clock trunk to ensure the proximity of devices and the robustness of clock trees [80].

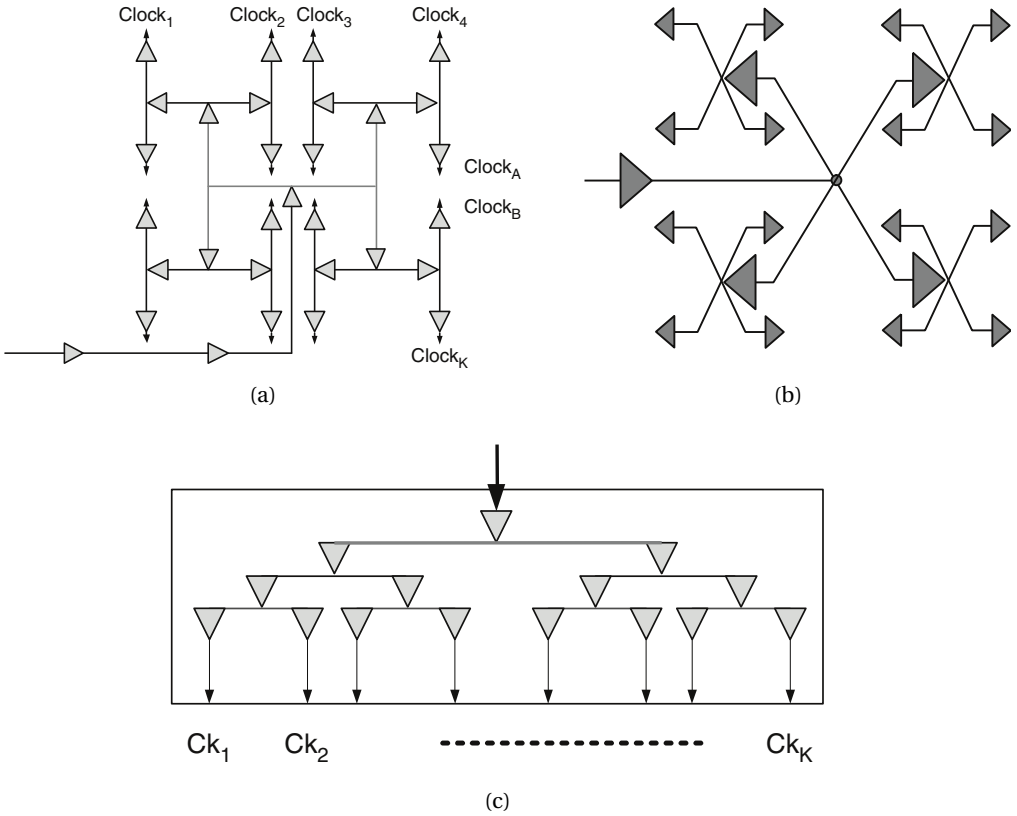


Figure 1.16: Balanced clock trees, where (a), (b), and (c) are an H-tree, an X-tree, and a binary tree, respectively [18].

Clock spine

A clock spine is a special implementation of binary trees [18]. A clock distribution network consisting of three clock spines is illustrated in Fig. 1.17. The clock signal is propagated from

the output of clock spines through individual branches to clock sinks. These ending branches of clock spines are matched to provide similar clock delays to clock sinks. Within a clock spine, branches are shorted at different levels of the clock trees to decrease skew.

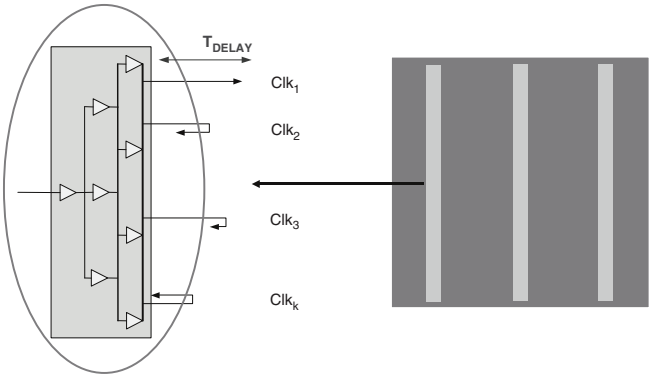


Figure 1.17: A clock distribution network consisting of three clock spines [18].

Clock grid

Clock distribution networks can be implemented based on grid or mesh structures. A clock grid is illustrated in Fig. 1.18. The clock signal is fed into the grid through the drivers on each side. The clock skew among clock sinks is significantly decreased as compared with tree structures. This skew is affected by the location of clock drivers and the pitch of clock grids. The main disadvantages of clock grids are the high requirement on routing resources and the high power consumption [75].

Hybrid distribution

To combine different advantages of clock distribution networks, hybrid clock distribution can be used. For instance, a global H-tree can be used to propagate the clock signal to different parts of a circuit. Local unbalanced clock trees or clock grids are then used to connect the global clock tree to local clock sinks [38]. Consequently, a tradeoff between clock skew, wire resources, and power is achieved.

All these clock distribution networks are fully investigated and optimized for 2-D ICs. In 3-D ICs, since clock sinks are distributed across different tiers, traditional 2-D clock distribution networks cannot be directly applied to a 3-D circuit. Alternatively, several 3-D clock trees have been proposed to support the robust and power-efficient clock signal distribution within 3-D ICs.

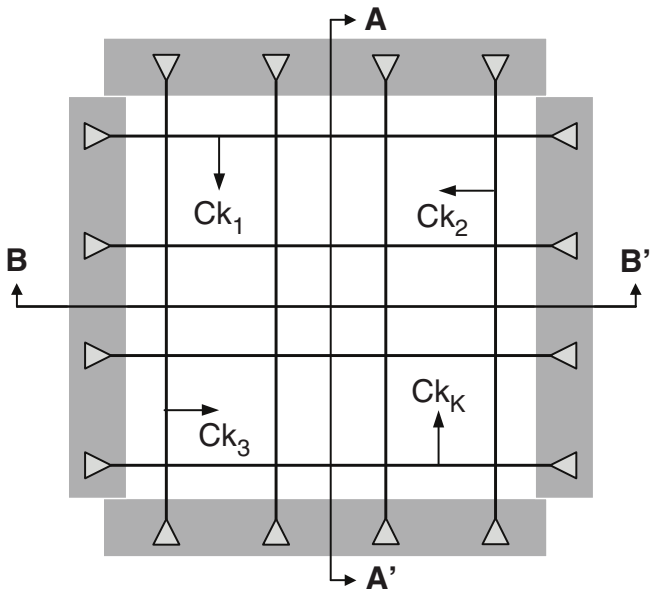


Figure 1.18: A clock grid with clock drivers on four sides [18].

1.4.3 3-D clock trees

Both balanced and unbalanced 3-D clock trees have been proposed. 3-D H-trees with different distributions across multiple tiers have been discussed in [19, 81]. Three different types of 3-D H-trees are illustrated in Fig. 1.19. In Fig. 1.19(a), 2-D H-trees are replicated in each tier. A large TSV or a group of TSVs are used to propagate the clock signal among tiers from the clock source. This topology only requires a limited number of TSVs and is easy to implement. Nevertheless, due to the replicated clock paths and buffers, this topology consumes high power and introduces large skew variation among tiers.

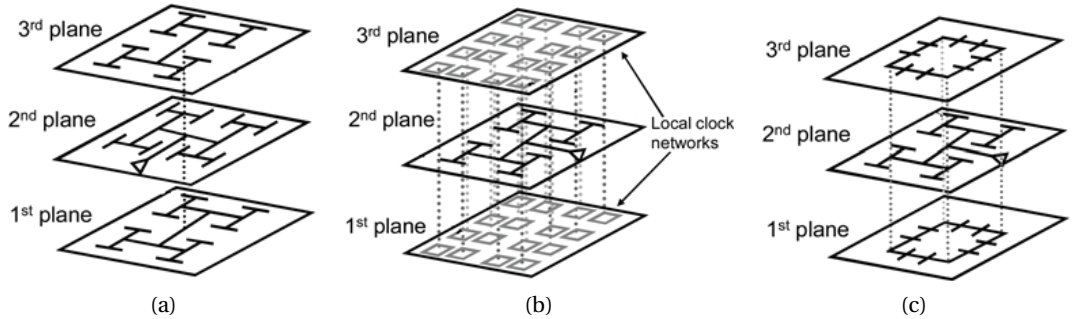
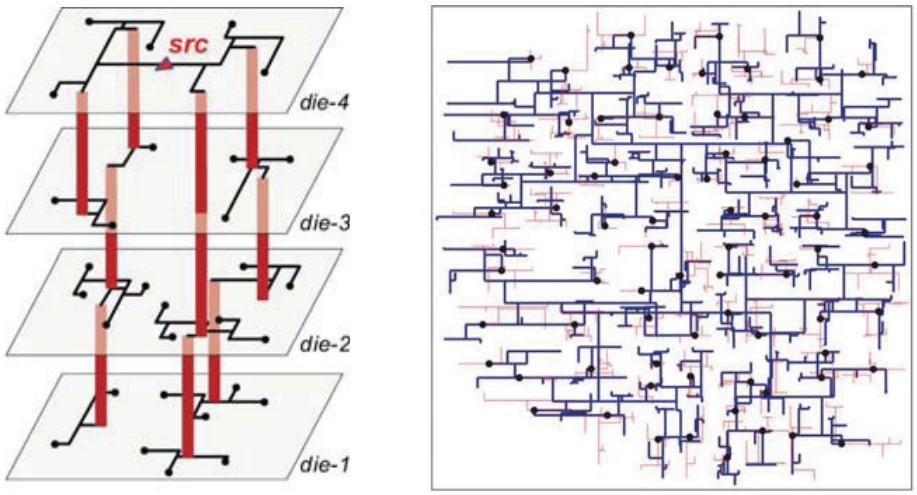


Figure 1.19: 3-D H-trees with different topologies across tiers, where (a) is a 3-D H-tree with replicated 2-D H-trees on each tier [19], (b) is a 2-D H-tree with local rings in other tiers, and (c) is an H-tree with global rings in other tiers.

Alternatively, only one 2-D H-tree is implemented in Fig. 1.19(b), the clock signal is propagated to other tiers through TSVs at the leaves of this H-tree. Local rings are used in these tiers to propagate the clock signal to the flip-flops. Global rings instead of local rings are used to propagate the clock signal in other tiers in Fig. 1.19(c). Both these topologies produce lower skew variation and power than replicated 3-D H-trees [19]. The 3-D clock tree using global rings generates lower clock skew but higher power consumption as compared to local rings. The number of TSVs of these two topologies, however, is significantly higher than the replicated 3-D H-trees.

Unbalanced 3-D clock trees are generated from automated *clock tree synthesis* (CTS) algorithms. 3-D clock tree synthesis algorithms have been proposed in [20, 82–85]. These algorithms focus on different optimization objectives. Low power clock tree synthesis algorithms considering pre-bond test problems are developed in [83, 84]. The CTS algorithm in [85] focuses on enhancing the tolerance of 3-D clock trees to TSV faults. A low power synthesis algorithm minimizing the number of TSVs is proposed in [20]. CTS considering temperature variations is proposed in [82]. Examples of these synthesized unbalanced 3-D clock trees are shown in Fig. 1.20 [20]. As shown in this figure, the resulting 3-D clock trees are asymmetric in both the horizontal and vertical directions.



(a) (b)

Figure 1.20: 3-D clock trees from the CTS algorithm [20], where (a) is a clock tree in a four-tier circuit and (b) is the top view of a clock tree in a two-tier circuit. TSVs are denoted by dots in (b).

Although different topologies of 3-D clock trees have been proposed, the analysis on these clock trees is mostly based on the deterministic behavior of clock buffers and wires. The variation of the resulting clock skew and jitter has not been thoroughly investigated. Since multiple tiers are vertically stacked in 3-D ICs, the effect of different sources of variations, such as process variations and power supply noise, on clock uncertainty differs from planar

circuits. Consequently, different sources of variations in 3-D ICs and their effect on 3-D clock distribution networks are carefully modeled and analyzed in this dissertation.

1.5 Contributions

Modeling and design techniques for 3-D ICs under process, voltage, and temperature variations are investigated in this dissertation. The proposed models and design techniques primarily focus on mitigating the effect of these variations on 3-D clock distribution networks. The main contributions of this thesis are:

1. The effect of process variations on 3-D ICs, especially 3-D clock trees, is investigated. A statistical model is proposed to describe the process-induced clock skew in 3-D clock trees. A comparison in skew variations among different topologies of 3-D clock distribution networks is presented. A set of design guidelines is proposed to mitigate skew variation.
2. Power supply noise in 3-D ICs is studied. The effect of power supply noise and process variations on 3-D clock trees is simultaneously modeled to more accurately estimate clock uncertainty. The resulting clock skew and jitter are statistically modeled in terms of skitter. Based on the proposed statistical model, a set of guidelines is proposed to improve the robustness of both 2-D and 3-D clock trees.
3. Two analytic heat transfer models are proposed to model *Thermal TSVs* (TTSVs) in 3-D ICs, with different tradeoffs in accuracy and run time. The temperature of 3-D ICs is shown to significantly vary with the physical characteristics of TTSVs. The relation between the temperature of circuits and geometric parameters of TTSVs is investigated to facilitate the design and placement of TTSVs in 3-D circuits.

1.6 Assumptions and Limitations

As mentioned before, fine-grain polylithic 3-D ICs based on TSVs are investigated in this dissertation. The behavior of other types of 3-D circuits, such as SiPs, SoPs, and monolithic 3-D circuits, is not discussed. Other assumptions and limitations are listed below:

- Process variations are approximated by normal (Gaussian) distributions. In the following chapters, the variation of different parameters of devices and interconnects (TSVs and wires) is modeled by a normal distribution. This approximation is based on the common assumption validated in related works [86–89]. In addition, process variations with non-Gaussian distribution can be converted to a combination of variables following Gaussian distribution [90].
- The die-to-die process variations are assumed to be independent from within-die variations [23, 91, 92]. Die-to-die process variations are assumed to be independent among

tiers in a 3-D IC. The correlation among tiers is neglected according to the analysis and observations in [88, 89, 93].

- The first droop of the resonant power supply noise is assumed to uniformly affect the devices across a physical plane (tier) of 3-D ICs. This assumption is based on the observation in [25, 29, 30, 34, 94] and on the fact that the resonant supply noise is determined by the total decoupling capacitance and package inductance.
- Clock distribution networks based on tree structures are modeled in this dissertation. Other topologies of clock distribution networks, such as clock grids and clock spines, are not modeled. The clock uncertainty in these clock distribution networks, however, is compared with clock trees through simulations.
- Only the steady-state heat transfer is modeled for thermal TSVs. The transient thermal behavior of TSVs and 3-D ICs is not investigated. The analysis on the transient heat transfer is typically implemented with *finite element analysis* (FEA) or *finite difference analysis* (FDA). Related works are described in Section 2.3.2.

1.7 Organization of the Dissertation

The remainder of this thesis is organized as follows. The background of PVT variations and the related modeling techniques for 2-D ICs are introduced in Chapter 2. The sources of process variations, the effect of process variations on timing and power, and the delay models used to describe process variations are introduced in Section 2.1. Power distribution networks and power supply noise are introduced in Section 2.2, where *IR*-drop and resonant supply noise are discussed. The thermal issues in integrated circuits and the related modeling techniques for 2-D ICs are presented in Section 2.3. Note that all these models are suitable for 2-D circuits.

For 3-D ICs, process variations significantly differ from 2-D circuits, since the stacked dies can be fabricated separately. Consequently, describing the resulting die-to-die variations throughout a 3-D stack is a greatly complex task as compared to 2-D circuits. The concept of statistical timing analysis considering process variations is introduced in Section 3.1, where related works on process variations in 3-D ICs are also reviewed. The effect of process variations on clock distribution networks is discussed in Section 3.2. A novel model for process-induced skew in 3-D clock trees is then proposed in Section 3.3. Based on this model, different topologies of 3-D clock distribution networks are compared with each other in Section 3.4. Consequently, a set of design guidelines is proposed to mitigate skew variation in 3-D ICs.

Voltage variation, or power supply noise, in 3-D ICs is investigated in Chapter 4. Potential structures of 3-D power distribution networks are introduced in Section 4.1. Due to the large number of TSVs in 3-D power distribution networks, the conventional methods used to analyze *IR*-drop cannot be directly applied to 3-D ICs. Consequently, a novel method for fast *IR*-drop analysis is proposed in Section 4.2. On the other hand, the resonant supply noise in 3-D ICs is discussed in Section 4.3. It is shown that the resonant supply noise varies among

tiers according to the switching current, the resistance of TSVs, and the number of tiers. In Section 4.4, the effect of resonant supply noise on clock distribution networks is discussed, where the analytic model of clock jitter is presented.

Since a circuit is simultaneously affected by process variations and power supply noise, the combined effect of these variations on 3-D clock trees is investigated in Chapter 5. Conventionally, the effect of process variations on clock trees is denoted by clock skew, while the clock uncertainty caused by power supply noise is described by clock jitter. In Section 5.1, a unified treatment, based on clock skitter, is used to describe both clock skew and jitter. A simplified model for skitter in 2-D ICs is proposed in Section 5.2, where methods used to decrease skitter in 2-D ICs are discussed. This model is extended to accurately model skitter in 3-D ICs in Section 5.3. The effect of skitter on both setup and hold time slacks is investigated. A set of guidelines is proposed to mitigate skitter in 3-D ICs in Section 5.4. To illustrate the efficiency of these guidelines, a case study on the skitter of synthesized 3-D clock trees is presented in Section 5.5. In this case study, clock buffers are inserted under the constraint in capacitive load. Alternatively, a fast buffer insertion algorithm used to decrease the total and maximum path delay in 3-D ICs is proposed in Section 5.6.

Thermal issues in 3-D ICs are investigated in Chapter 6. The significant increase in the temperature of 3-D circuits is introduced in Section 6.1. The exacerbated temperature highly affects the timing and power consumption of circuits. To decrease the temperature of 3-D ICs, TTSVs can be used to improve the heat transfer across tiers. The structure of these TSVs is introduced in Section 6.2. Two steady-state analytical models for TTSVs are proposed in Section 6.3. Based on these models, the effect of the physical characteristics of TTSVs (the diameter, the depth, the density, *etc.*) on temperature is investigated in Section 6.4. A case study on 3-D DRAM-Microprocessor structure is presented to show the efficiency of the proposed models.

The conclusions of this dissertation and the potential future research directions are drawn in Chapter 7. The models proposed in this thesis can facilitate designers to understand and quantitatively evaluate the effect of PVT variations on the performance of 3-D ICs. Since PVT variations become increasingly important as the technology scales, the robustness of circuits becomes a main challenge. The provided design guidelines help to significantly improve the robustness of 3-D ICs under these variations.

2 Process, Voltage, and Temperature Variations in Integrated Circuits

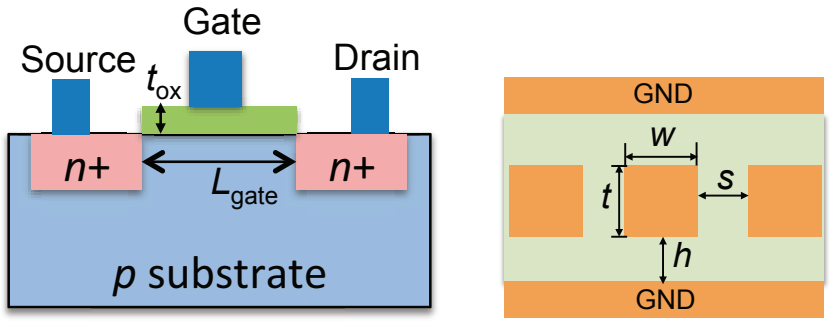
Both the physical and electrical characteristics of integrated circuits are subjected to fluctuations. Different sources of variability in integrated circuits are introduced in this chapter. These sources of fluctuations include *process variations, supply voltage noise, and temperature/thermal variations* (PVT variations). The effect and conventional models of process variations are introduced in the following section. Power supply noise and temperature variation are introduced in Sections 2.2 and 2.3, respectively.

2.1 Process Variations

Technology scaling has been the driving force to increase integration density for the past decades. In very deep sub-micrometer technologies, process variations significantly complicate the IC design process [21]. The different sources of process variations are introduced in the following subsection. The effect of these variations on the timing and power of circuits is briefly presented in Section 2.1.2. The models used to describe the statistical delay of devices and interconnects due to process variations are presented in Section 2.1.3.

2.1.1 Sources of process variability

Process variations are introduced in the manufacturing process and are attributed to the imperfections of the related equipment. Both the devices (transistors) and interconnects (metal wires) are affected by process variations. The primary physical parameters of devices and interconnects are illustrated in Fig. 2.1. For the transistor, the channel length and oxide thickness are denoted by L_{gate} and t_{ox}, respectively. For the interconnect, the width and thickness of metal wires, the space between wires, and the thickness of the *Inter-Layer Dielectric* (ILD) are denoted by w, t, s, and h, respectively. The variability of these parameters increases significantly as the technology scales, as illustrated in Fig. 2.2. The effective channel length, the threshold voltage of transistors, and the dielectric constant of the ILD are denoted by L_{eff}, V_{th}, and ρ, respectively.



(a) (b)

Figure 2.1: Physical parameters of transistors and metal interconnects, where (a) and (b) are the cross-sections of an NMOS transistor and a metal wire, respectively.

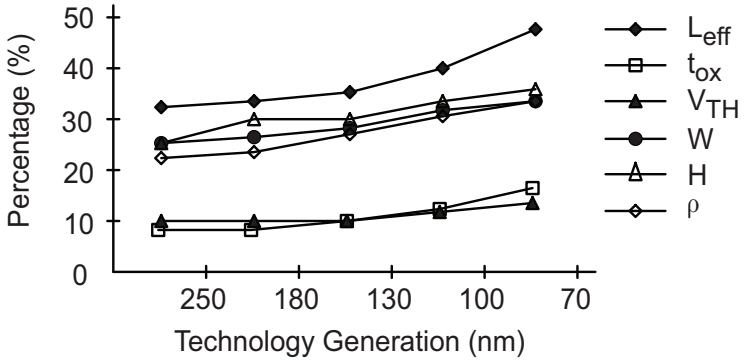


Figure 2.2: The 3σ variation of several parameters *vs.* technology generations [21], [22].

The variability of devices is also named front-end variability [22], which mainly includes the variations in L_{gate}, t_{ox}, transistor width (W_{gate}), and doping density (N_{A}). The gate length L_{gate} and the related parameter L_{eff} strongly affect the drive current and, consequently, the speed of the transistor. L_{eff} can be defined as L_{gate} minus the overlapping parts of the gate with the source and drain. For simplicity, the term L_{gate} is used in this dissertation to denote L_{eff}. The variation of L_{gate} (ΔL_{gate}) is caused by several processing steps and procedures, such as the mask, the exposure stage, etching, and implantation of the source and drain. Gate length variation is one of the most important variations for a circuit [22, 23, 91, 92]. The gate width variation ΔW_{gate} has negligible impact on large-width transistors while a non-negligible impact on minimum-size transistors has been observed [22]. Mask alignment is a major source causing ΔW_{gate}.

An important element in the design of the devices is the thin dielectric layer of the gate terminal. This dielectric film is used to isolate the gate from the channel region. The thickness of this dielectric film t_{ox} significantly affects the drive current, the threshold voltage V_{th}, and the leakage current. The silicon dioxide film can be grown with a thermal oxidation process [2]. As the technology scales, t_{ox} has now reached the atomic-level (*i.e.*, only a few layers of atoms

are grown). Consequently, the interface roughness and the atomic-scale discreteness result in a non-negligible Δt_{ox}.

The doping density N_A is another important factor affecting V_{th} and, consequently, the performance of the transistors. Dopant atoms are placed into the channel of the transistors by ion implantation. Both the ion implantation and the following step, activation through annealing, cause a random distribution of dopant atoms inside the channel. The spread of this random distribution increases with technology scaling.

The variability of interconnects is called back-end variability referring to the back end of the IC fabrication process. This variability includes the variations in metal width (Δw), metal thickness (Δt), the thickness of ILD (Δh), and the material properties (*e.g.*, $\Delta \rho$) [95]. Many of the sources of the front-end variability also result in the back-end variability, such as lithography and etching. In addition, the backend process includes copper electroplating and *chemical-mechanical polishing* (CMP), which also introduce variations in both the physical and material properties of the interconnects. The resulting variations in the resistance and capacitance of interconnects significantly affect the timing and power of circuits as technology scales.

The classification or, the decomposition, of process variations is illustrated in Fig. 2.3 [22, 95, 96]. In general, process variations consist of systematic and random variations. The systematic variations can be modeled deterministically, while the random variation can only be described statistically. At different design stages of a circuit, specific random variations can be treated as systematic variations. For instance, during logic synthesis, the layout-related variation of metal wire thickness Δt is a random variable due to the lack of layout information. After placement and routing, however, this layout-related Δt can be predicted with a fixed value based on the pattern of interconnects. In the early stages of the design flow, most of the systematic and random variations are statistically modeled [96]. The process variations in this dissertation, consequently, refer to random variations.

Since variations originate from any step of the fabrication stage, the effects of these fluctuations apply to different physical scales. The random process variations can be decomposed into lot-to-lot, wafer-to-wafer, across-wafer, across-reticle, and within-die variations [22]. Retaining all these levels of information in the design process would unnecessarily increase the complexity of the process. For this reason, the individual sources are abstracted to two major categories: inter-die and intra-die variations. This abstraction captures the overall effect of the variations but hides the complicated characterization process from the designers.

Inter-die variations

Inter-die (or die-to-die, D2D) variations affect the characteristics of devices and interconnects differently among dice, but the same parameter of different components within one die is uniformly affected. D2D variations include lot-to-lot, wafer-to-wafer, parts of the across-wafer

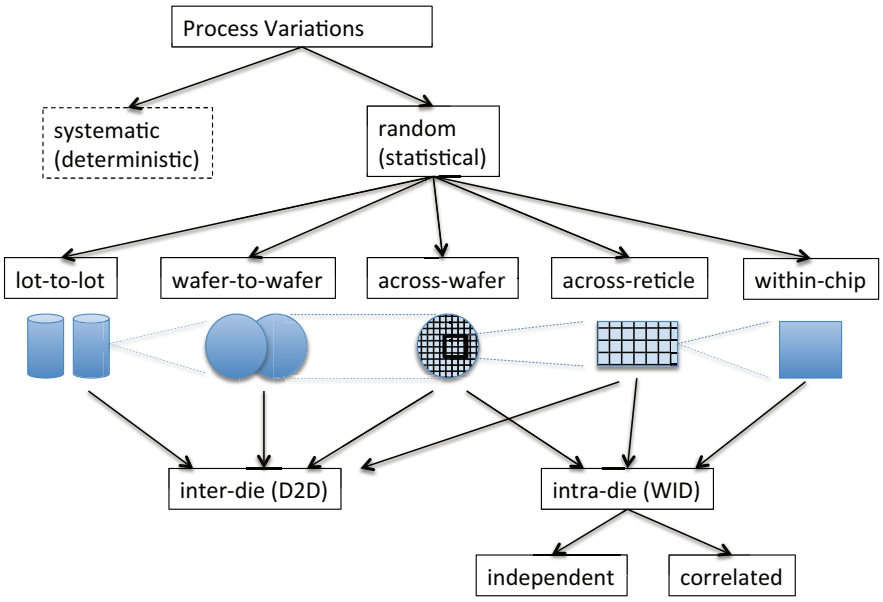


Figure 2.3: The classification of process variations.

and across-reticle variations. Traditionally, D2D variations have been the dominant factor in process variations [95].

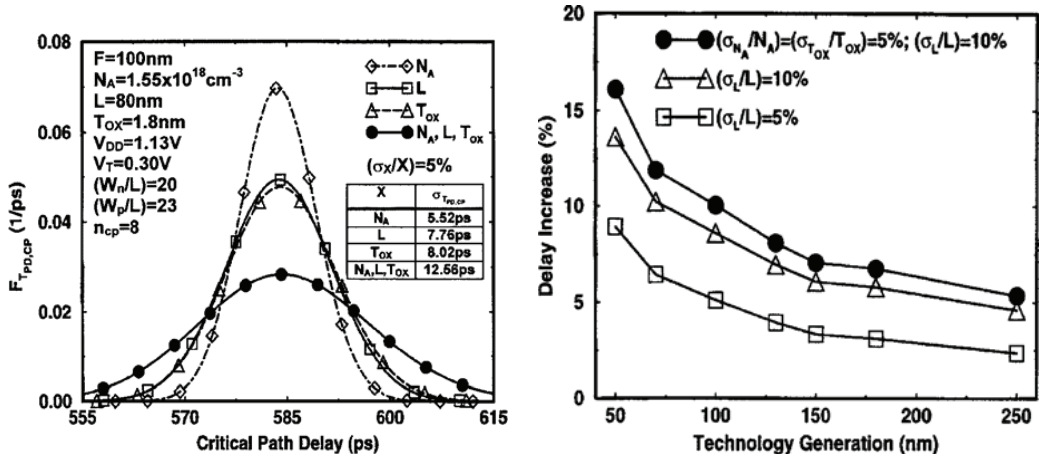
Intra-die variations

Intra-die (or within-die, WID) variations affect the characteristics of devices unequally within one die. For sub-micrometer circuits, WID variations become non-negligible and increasingly important [91, 92, 97]. WID variations can be further divided to correlated and independent components. In some publications, the correlated WID variations are also called systematic variations [91, 92]. To avoid confusion, the systematic variations, herein, only refer to the deterministic process variations.

Most of the physical parameters are simultaneously affected by D2D and WID variations. For instance, ΔL_{gate} can be decomposed to D2D variations and *across-chip linewidth variation* (ACLV). The D2D ΔL_{gate} is caused by variations in the resist bake, the radial variations in the photoresist coating thickness, and the fluctuation of the etch process. The WID component, ACLV, is primarily induced by variations in the stepper, reticle imperfections, *etc* [22]. The D2D and WID parameter variations of devices and interconnects, together, lead to variations in the timing and power of circuits. This effect is discussed in the following subsection.

2.1.2 Effect of process variations on timing and power

The delay and power of transistors highly depend on the drive current and threshold voltage. These electrical characteristics are greatly affected by process variations. Consequently, pro-



(a) (b)

Figure 2.4: The delay variation of critical paths due to process variations, where (a) is the distribution of critical path delay for different parameter variations and (b) is the increase in critical path delay corresponding to a 3σ delay deviation. [23]

cess variations lead to non-negligible variations in delay, transition time, and dynamic and leakage power of transistors. In addition, as the interconnect delay increases with the scaling technology, the variations in the electrical characteristics of the interconnects also introduce significant wire delay variation into circuits [98, 99].

The device and wire variations result in delay fluctuations in both logic and clock paths. An example of the increase in the delay of critical paths due to process variations is illustrated in Fig. 2.4 [23]. As shown in this figure, at advanced technology nodes, the critical delay variation increases significantly for all cases of process variations (ΔN_{A}, ΔL_{gate}, and Δt_{ox}). The accumulating effect of different sources of process variations causes a wide distribution of the critical path delay, as shown in Fig. 2.4(a).

In addition to speed, power consumption has become another critical factor in modern IC design. The power consumed by a circuit includes dynamic and static power. The dynamic power of devices includes two sources: (1) the charging and discharging of the intrinsic and extrinsic parasitic capacitances of the transistors and (2) the short-circuit current during the gate switching. Due to the variation of the drive current, the input transition time, and the capacitive load, dynamic power is also affected by process variations. The increase in power corresponding to the distribution of parameters in Fig. 2.4 is illustrated in Fig. 2.5 [23]. Similar to Fig. 2.4(b), the power variation also increases with technology scaling.

Leakage power is the other power component and becomes increasingly important as the CMOS technology scales. Sub-threshold leakage and gate oxide leakage are two major constituents of leakage power, both of which are highly affected by process variations of devices [87, 100]. An example of the leakage current distribution of microprocessors is illustrated

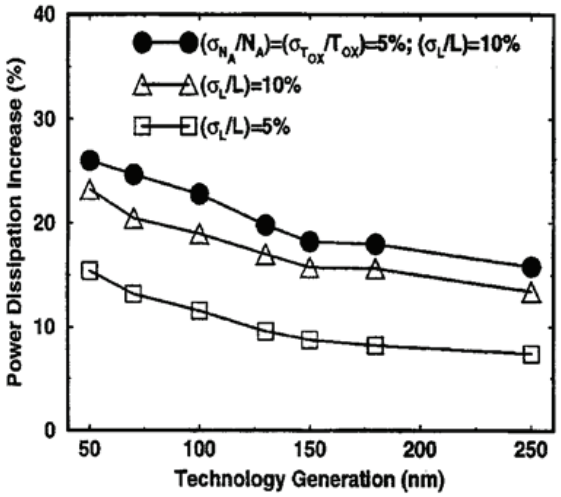


Figure 2.5: The increase in dynamic power due to process variations corresponding to a 3σ critical-path delay deviation [23].

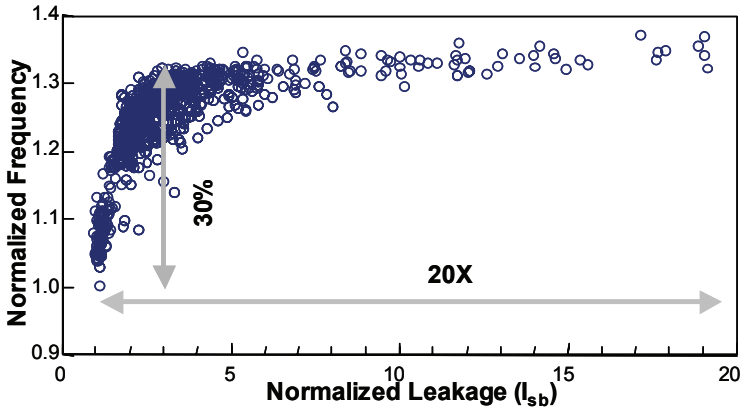


Figure 2.6: Distribution of clock frequency and leakage current due to process variations [24].

in Fig. 2.6 [24]. As shown in this figure, both the clock frequency and the leakage current and consequently, the leakage power, vary significantly with the parameters of transistors.

Since process variations introduce high variation in timing and power of circuits, accurately modeling these variations is necessary to design a robust circuit. This dissertation focuses on timing variability. The models used to describe the delay of devices and interconnects are introduced in the following subsection.

2.1.3 Delay model for devices and interconnects

Multi-corner analysis and statistical models are two common methods to predict the variation in the delay of transistors and interconnects. The corner-based analysis determines the timing information of transistors and interconnects based on a finite combination of process, voltage,

and temperature parameters [101]. For instance, a transistor can be characterized under three corners: fast, typical, and slow. The actual delay of transistors and interconnects is approximated to the closest corner. To include the worst case of PVT variations, the resulting delay is usually pessimistic and can lead to over-designing a circuit.

Statistical models more accurately describe the variations in delay. The statistical delay of transistors and interconnects can be determined through sensitivity analysis. Due to the reasonable accuracy and the ease for computation, the first-order Taylor expansion is widely used to describe the sensitivity of the delay of transistors and interconnects to the variational parameters [95],

$$d = d_0 + \frac{\partial d}{\partial p}(p - p_0), \tag{2.1}$$

where d and p are the delay and process parameter, respectively. The nominal value of d and p are denoted by d_0 and p_0, respectively. The sensitivity of delay d to parameter p is determined by

$$\frac{\Delta d}{d} = \frac{\partial d}{\partial p}\frac{\Delta p}{d}, \tag{2.2}$$

where $\Delta d = d - d_0$ and $\Delta p = p - p_0$. Although (2.1) and (2.2) are only low-order polynomial approximations to the process-induced delay variation, this approximation exhibits considerably high precision and efficiency in practice [22, 86, 95, 96, 98]. The partial derivative in (2.1) and (2.2) can be determined through the deterministic delay model of transistors and interconnects.

Delay model for transistors

A number of analytical models have been developed to describe the delay of transistors [99, 102–105]. Most of these models obtain a simplified expression for the transistor delay by introducing fitting parameters. For instance, one of the most prolific analytic models, Alpha-Power Law Model (α-model) [105], estimates the transistor delay as

$$\tau \propto \frac{C_{\text{load}} V_{\text{dd}}}{I_{\text{dsat}}}, \tag{2.3}$$

where C_{load}, V_{dd}, and I_{dsat} are the capacitive load, supply voltage, and saturation drain current, respectively. The saturation current is approximated with the α-model as,

$$I_{\text{dsat}} = \frac{W_{\text{gate}}}{2L_{\text{gate}}}\mu_{\text{eff}} C_{\text{ox}}(V_{\text{gs}} - V_{\text{th}})^{\alpha}, \tag{2.4}$$

where C_{ox} is the gate oxide capacitance per unit area and μ_{eff} is the effective mobility. The short channel effects are described by α. The accuracy of this model is determined by the proper fitting of "α". Based on these formulas, the partial derivative in (2.1) and (2.2) can be obtained.

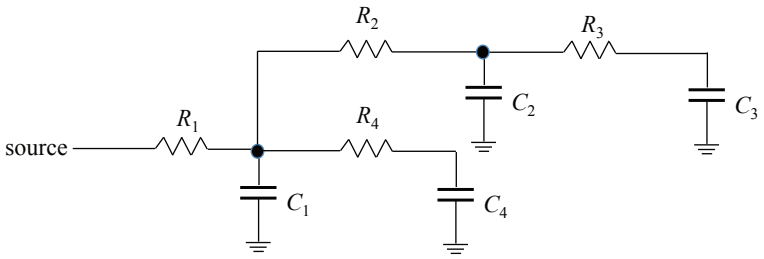


Figure 2.7: An *RC* tree with two sinks.

In industrial standard cell libraries, *non-linear delay models* (NLDMs) have been widely used to characterize the timing information of different cells. NLDMs determine the delay of a cell based on the input slew and output load through a non-linear transformation. Current Source Model (CSM) is another novel delay model which shows many advantages in predicting the delay of a cell [106, 107]. Several academic approaches have been developed to use current source models in statistical timing analysis [108, 109]. Nevertheless, a large number of simulations is required to generate CSMs. In addition, CSM currently lacks the wide support from commercial EDA tools [95].

Delay model for interconnects

There are also a large number of delay models proposed for *RC* and *RLC* interconnects. For instance, for an *RC* wire with a capacitive load C_{load}, the 50% and 90% signal delay corresponding to a step input, respectively, can be estimated by [39, 110]

$$d(50\%) = 0.38RC + 0.69RC_{\text{load}}, \tag{2.5}$$

$$d(90\%) = 1.02RC + 2.3RC_{\text{load}}, \tag{2.6}$$

where R and C are the total resistance and capacitance of this wire, respectively. For tree-like *RC* interconnects, the Elmore model and the extended versions of this model have broadly been used to predict the delay from the source to the sinks [111, 112]. For an *RC* tree shown in Fig. 2.7, the time constant of the signal delay from the source to C_3 and C_4 is modeled as

$$\tau_{C_3} = R_1(C_1 + C_4) + (R_1 + R_2)C_2 + (R_1 + R_2 + R_3)C_3, \tag{2.7}$$

$$\tau_{C_4} = R_1(C_1 + C_2 + C_3) + (R_1 + R_4)C_4. \tag{2.8}$$

Based on (2.7) and (2.8), the delay from the clock source to all the clock sinks of an interconnect tree can be obtained. Closed-form expressions have also been proposed to model the delay of *RLC* interconnects [113, 114].

The electrical characteristics R, C, and L of interconnects can be determined by empirical analytic expressions [41],

$$R = \frac{\rho_m l}{wt}, \tag{2.9}$$

$$L = \frac{\mu_0 l}{2\pi}\left[\ln\left(\frac{2l}{w+t}\right) + 0.5 + \frac{0.22(w+t)}{l}\right], \tag{2.10}$$

$$C = 2C_g + 2C_c, \tag{2.11}$$

$$C_g = \epsilon\left[\frac{w}{h} + 2.04\left(\frac{s}{s+0.54h}\right)^{1.77}\left(\frac{t}{t+4.53h}\right)^{0.07}\right], \tag{2.12}$$

$$C_c = \epsilon\left[1.41\frac{t}{s}e^{-\frac{4s}{s+8.01h}} + 2.37\left(\frac{w}{w+0.31s}\right)^{0.28}\left(\frac{h}{h+8.96s}\right)^{0.76}e^{-\frac{2s}{s+6h}}\right], \tag{2.13}$$

where ρ_m, μ_0, and ϵ are the resistivity of metal, permeability of wire, and permittivity of insulator, respectively. Given the distribution of the physical parameters of interconnects, the variations of electrical characteristics of interconnects can be determined from (2.9) through (2.13). Consequently, the delay variation of the interconnects can also be determined.

The effect of the delay variation of devices and interconnects on the timing of the circuits will be discussed in Chapter 3. In general, process variations are time-invariant for a fabricated circuit. The parameter variation does not vary during operation [95]. The fluctuation in supply voltage, however, changes with time. The impact of power supply noise and the related models are introduced in the following section.

2.2 Power Supply Noise

The fluctuation of supply voltage, called power supply noise, is another important source of variations. This supply noise largely affects the electrical characteristics of transistors. The supply voltage is provided to the transistors through *power distribution networks* (PDNs). Typical structures of power distribution networks are introduced in the following subsection. The sources and effect of power supply noise are presented in Section 2.2.1. The related modeling methods are introduced in Section 2.2.3.

2.2.1 Power distribution networks

Due to the increasing number of devices integrated within a chip, supplying power/ground to a circuit is a challenging task. The design of PDNs significantly affects the supply voltage and, consequently, the timing and power of a circuit. Typically, a PDN consists of interconnect networks with decoupling capacitance on the *printed circuit board* (PCB), the circuit package, and the circuit die. A cross-sectional view of a typical PDN is illustrated in Fig. 2.8 [25]. As shown in this figure, the power is supplied from a switching voltage regulator on the PCB. Decoupling capacitance is placed on the PCB, the circuit package, and the circuit to mitigate

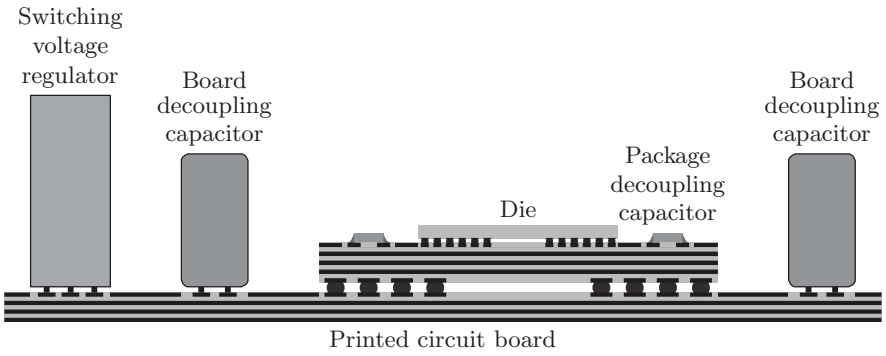


Figure 2.8: Cross-section of the PDN hierarchy with decoupling capacitance [25].

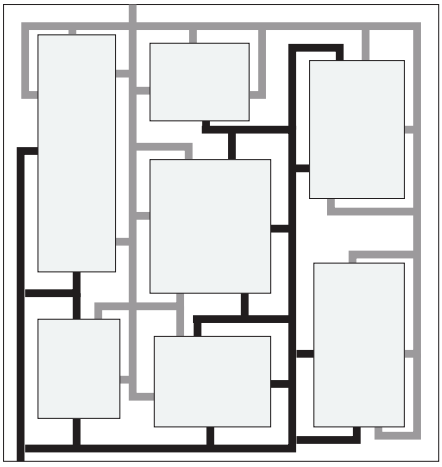


Figure 2.9: Routed P/G networks [25, 26].

the fluctuation of the supply voltage. Within the circuit, on-chip PDNs are used to properly distribute the supply voltage.

Typically, the on-chip PDN consists of *power and ground* (P/G) interconnect networks. Several topologies have been proposed and used to distribute P/G to the entire circuit. Some of these topologies are briefly described in the next paragraphs.

Routed networks

A routed network used to propagate power and ground is illustrated in Fig. 2.9 [25, 26]. Dedicated wire trunks are used to supply P/G to circuit blocks. The primary advantage of routed networks is the savings in routing resources. The main drawback is the low redundancy of the networks due to the limited number of P/G trunks, which decreases the robustness of the PDNs.

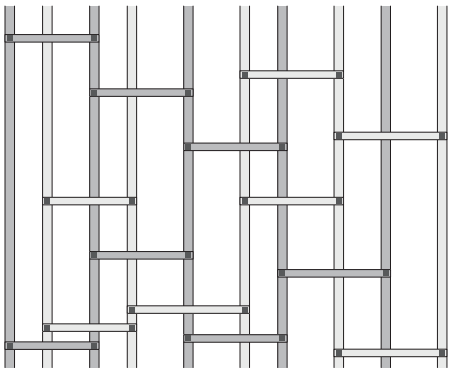
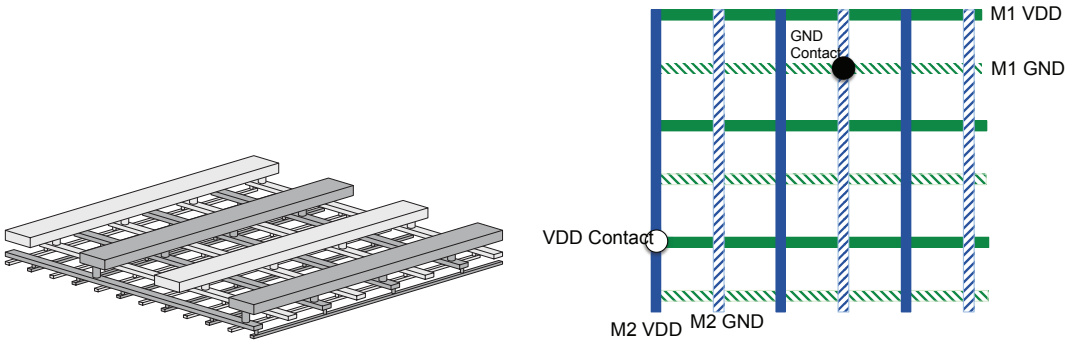


Figure 2.10: A power and ground mesh [25].



(a) (b)

Figure 2.11: A power (VDD) and ground (GND) grid network, where (a) [25,27] and (b) are a 3-D plot and a top-view of power grids, respectively.

Mesh networks

Mesh-based P/G networks are used to add redundancy, thereby increasing the robustness of PDNs, as illustrated in Fig. 2.10. Parallel wide wires are placed on the upper metal layers to globally distribute P/G. Irregular short straps orthogonal to the global wires are used in lower metal layers to supply P/G to the devices and circuit blocks. Power meshes are used in low-power circuits with limited routing resources in the upper metal layers. When the upper metal layers can be utilized to implement a grid structure to distribute P/G, power grids are usually a more robust solution for PDNs.

Grid networks

The grid-based P/G networks are illustrated in Fig. 2.11. Parallel P/G wires are connected by vias to the adjacent metal layers. Power grids are more regular structures than power meshes, which utilize more metal resources in the upper metal layers. For high-performance designs, power grids provide a PDN with high redundancy and robustness.

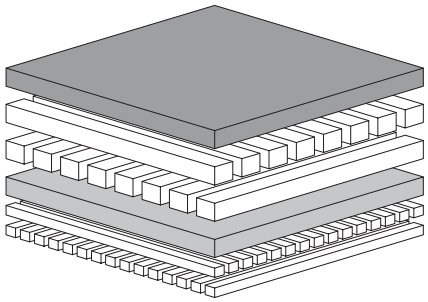


Figure 2.12: A PDN with P/G planes [25], where the power and ground planes are depicted in dark and light gray, respectively.

P/G planes

For the PDN with P/G planes, an entire metal layer is used to distribute power and ground, respectively, as illustrated in Fig. 2.12 [25, 115]. The dedicated power and ground planes provide low-impedance current paths for P/G. The cost in metal resources, however, is prohibitively high, since no signal (or clock) wires can be placed in the P/G planes.

Cascaded P/G rings

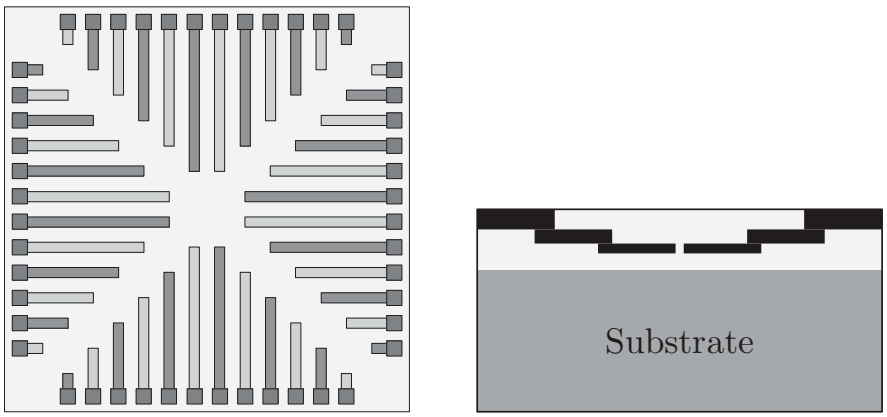
PDNs with cascaded P/G rings are used in the circuits with peripheral *input/output* (I/O) pads. As illustrated in Fig. 2.13 [25, 28], the P/G are supplied to the chip through the peripheral P/G pads towards the center of the chip. Sub-PDNs (*e.g.*, power grids) can be placed under the P/G rings to supply voltage to the devices, where other signals can also be routed. With cascaded P/G rings, more metal resources are saved for signal routing. Nevertheless, this structure is typically used for peripheral I/Os and careful design is required for the cascaded P/G rings to avoid a PDN with large impedance.

Hybrid-structure networks

The boundaries among the listed PDN structures are not strictly defined. For modern VLSI design, a combination of different types for PDNs is widely used for complex circuits. As the technology advances, more devices are integrated into a single chip and an increasingly larger amount of current is required by the circuit. Consequently, power grids are commonly used for global PDNs and other types of local PDNs can be used for the sub-circuits [25].

2.2.2 Sources and effect of power supply noise

Due to the large amount of current and high switching frequency of modern ICs, the devices within a circuit can experience significant power supply noise. Power supply noise is caused by the flow of currents (either DC or transient currents) through PDNs. The supply noise is



(a) (b)

Figure 2.13: PDNs with P/G cascaded rings [25, 28], where (a) and (b) are the top-view and cross-section of cascaded rings, respectively.

primarily determined by the impedance characteristics of PDNs. The sources of the power supply noise include the *IR*-drop (resistive) and $L\frac{di}{dt}$ (inductive) noise.

IR-drop in PDNs

IR-drop is due to the resistance of the PDNs, which causes a voltage drop when currents flow through. The resistance along a PDN includes the resistance of on-chip PDN wires and vias, bond wires or solder bumps to the package, package traces, and PCB planes. Since the copper wires in the package and PCB are much thicker and wider than the on-chip interconnects, *IR*-drop is primarily determined by on-chip PDNs.

Inductive noise

In addition to the resistance, the inductance of PDNs can also cause significant voltage fluctuations for transient currents. The inductance in PDNs includes the partial self inductance and partial mutual inductance of the wires, vias, and connectors from the chip, the package, and PCB. The inductive component of the impedance of PDNs contributes considerable noise due to the increasing switching speed and, consequently, the faster current transients, of circuits [116].

The simplified effect of *IR*-drop and $L\frac{di}{dt}$ drop is illustrated in Fig. 2.14 [25]. As shown in this figure, the supply voltage seen by the power load (devices) deviates from the ideal voltage with the transient current. The fluctuations in the supply voltage cause significant variation in the delay of the transistors. For instance, the delay of a CMOS inverter under different fluctuations of the supply voltage is illustrated in Fig. 2.15. The ideal V_{dd} (VDD) for this inverter is 0.9 V. As

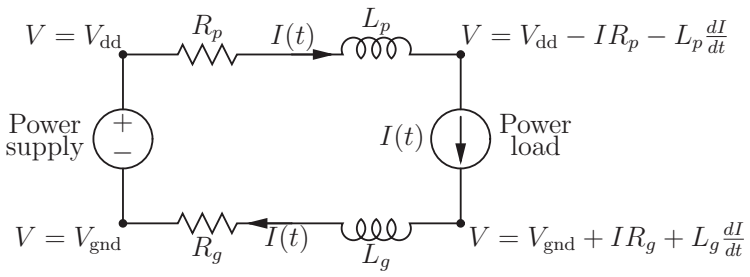
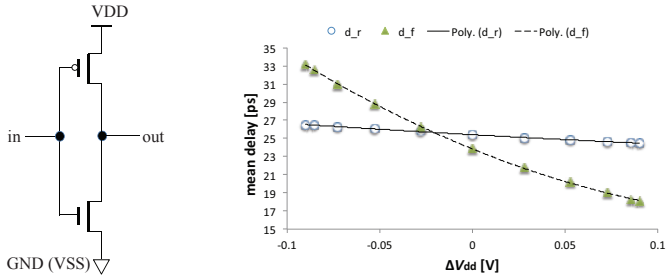


Figure 2.14: Power supply noise caused by *IR*-drop and $L\frac{di}{dt}$ drop in a simplified PDN [25].



(a) (b)

Figure 2.15: Delay of a CMOS inverter *vs.* the fluctuation of supply voltage, where (a) is the schematic of a CMOS inverter. (b) is the change of the inverter delay, where both the rise-fall (d_r) and fall-rise (d_f) delay are shown.

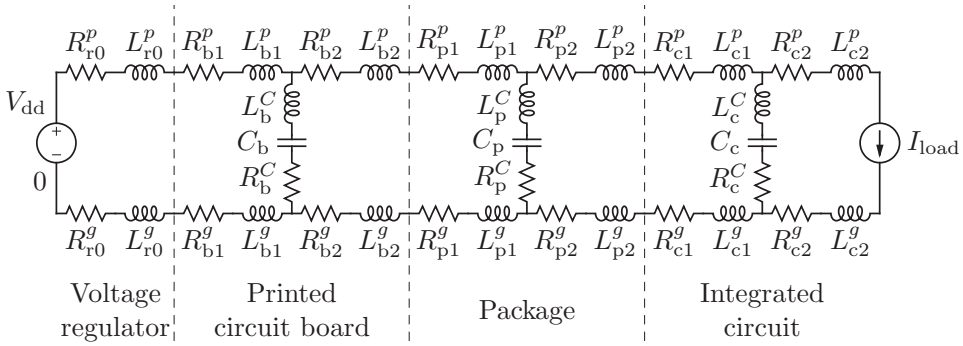
shown in this figure, under a +10% (-10%) variation in $V_{\rm dd}$, the rise-to-fall delay (d_r) can vary by 37% (-24%). Power supply noise, therefore, highly affects the timing of a circuit.

2.2.3 Modeling techniques for power supply noise

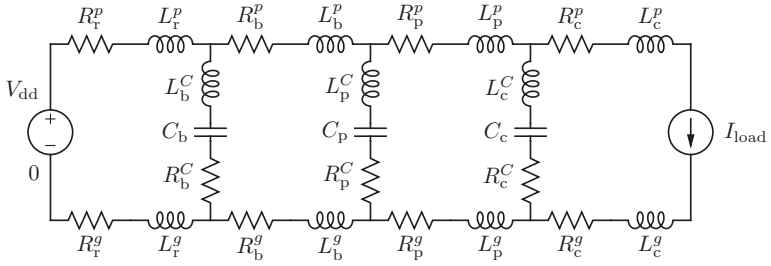
As mentioned above, the power supply noise includes two major components: the inductive component $L\frac{di}{dt}$ and the resistive component *IR*. To investigate the two sources of power supply noise, different models for the PDNs have been proposed. Two important models are introduced in the following subsections.

The one-dimensional model of PDNs

The one-dimensional model has been proposed to model the entire PDN [25, 117]. For the PDN shown in Fig. 2.8, the corresponding model is illustrated in Fig. 2.16 [25]. The electrical characteristics of the voltage regulator, PCB, package, and chip are labeled with the subscripts “r”, “b”, “p”, and “c”, respectively. The decoupling capacitors are modeled by series *RLC* circuits, of which *R* and *L* are the parasitic impedances of the capacitors and labeled with the superscript *“C”*.



(a)



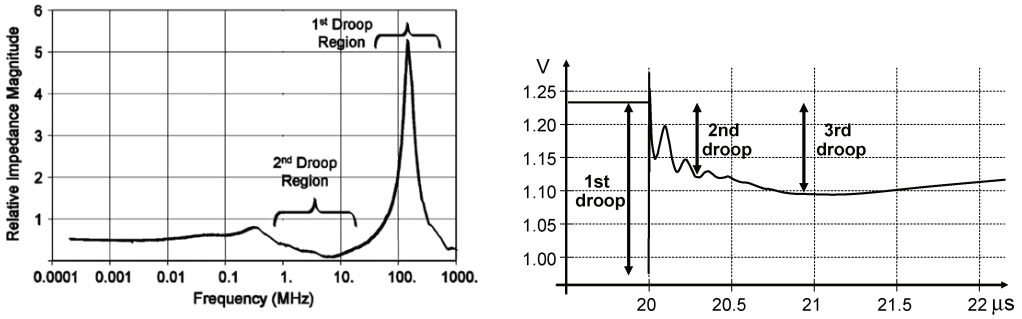
(b)

Figure 2.16: A simplified one-dimensional circuit model for PDNs [25], where (a) models both the upstream and downstream impedance of all levels of a PDN and (b) is a compact version of (a).

With the model in Fig. 2.16, the important impedances of a PDN and the decoupling capacitances are modeled as lumped resistance, inductance, and capacitance. The power supply noise at different frequencies can be estimated efficiently [118]. For instance, the impedance seen at I_{load} at different frequencies is illustrated at Fig. 2.17 [29]. For a current pulse (*e.g.*, circuit wakeup or clock activation), the resulting resonant supply noise is illustrated in Fig. 2.17(b) [30]. As shown in this figure, the first droop of the supply voltage is the deepest voltage drop and significantly affects the performance of the system. Consequently, the resonant supply noise refers to the first droop of supply noise herein [29].

Model for on-chip PDNs

Although the model in Fig. 2.16 provides a fast way to estimate the supply noise in a PDN, the structure of the on-chip PDNs has been abstracted. Consequently, the difference in the supply voltage among circuit blocks and devices are ignored. As mentioned before, *IR*-drop is primarily determined by the on-chip resistances due to the narrow on-chip interconnects. To describe the voltage drop within a chip, resistor networks are used to model the on-chip PDNs.



(a) (b)

Figure 2.17: Power supply noise in PDNs. (a) is the impedance of a PDN at different frequencies [29] and (b) is the waveform of a resonant supply noise [30].

The VDD and GND (VSS) parts of an on-chip PDN are usually modeled separately by two resistor networks. For instance, the GND network of a power grid can be modeled by the resistor network illustrated in Fig. 2.18(a). The GND wires on different metal layers are flattened to a planar network. The circuit blocks or devices are modeled as current sources. Given the resistance of each wire segment, the current through each current source, and the connections to the package, the voltage at each junction node and current source can be obtained [119, 120]. This voltage describes the *IR*-drop of PDNs and the voltage distribution across the chip. An example of this voltage map is illustrated in Fig. 2.18(b). Since only the resistance of PDNs is considered, *IR*-drop analysis focuses on the spatial distribution but not the temporal change of power supply noise.

In addition to the one-dimensional model for the entire PDN and the resistor networks for on-chip PDNs, a large number of models has been proposed with emphasis on different characteristics of the power supply noise [25]. The most detailed models often require a full-chip transient simulation with given circuit events [121].

2.3 Temperature Variations

In addition to process and voltage variations, the devices and interconnects also experience temperature variations. The heat generated by devices and interconnects spatially and temporally changes the temperature distribution across a circuit. As the power consumption increases, thermal issues become increasingly important to integrated circuits. Thermal issues and related modeling methods are introduced in Sections 2.3.1 and 2.3.2, respectively.

2.3.1 Thermal issues in integrated circuits

For advanced technology nodes, on-chip power densities rapidly increase due to the large number of devices integrated within a circuit and the increased switching speed of devices. For instance, the power density of Intel microprocessors is illustrated in Fig. 2.19 [31,32]. As shown

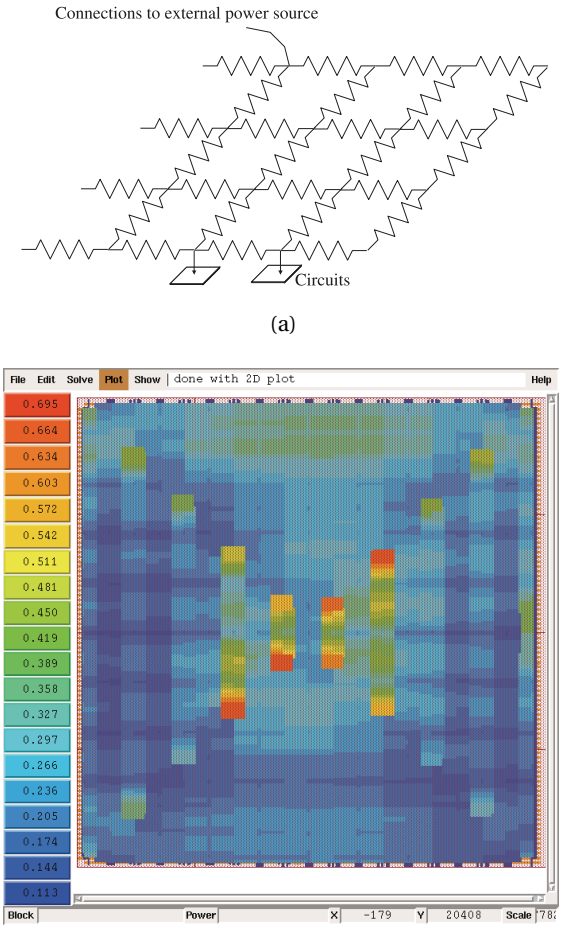


Figure 2.18: Resistor network used to model the on-chip PDNs [27], where (a) and (b) are the topology (not to scale) and voltage drop of the GND network.

in this figure, the power density increases extraordinarily as the feature size of transistors decreases. This increasing power density can generate high temperatures in integrated circuits, as illustrated in Fig. 2.20, which results in non-negligible thermal issues. One of the most important issues is the high junction temperature of transistors. The junction temperature highly affects the speed, leakage power, and long-term reliability of circuits [33].

The high junction temperature decreases the mobility of the carriers, which decreases the driving current of transistors. The delay and transition time of transistors, consequently, increases. Meanwhile, the thermal variation among on-chip devices can lead to spatial and temporal delay variation. The leakage power also increases with the junction temperature. The increased power, in turn, can further raise the temperature. In extreme cases, this positive feedback between temperature and power can destroy the circuit due to excessive heat dissipation. The increasing temperature also reduces the reliability of the circuit. The temperature-related failures of circuits can be caused by gate oxide breakdown, electro-migration of metal intercon-

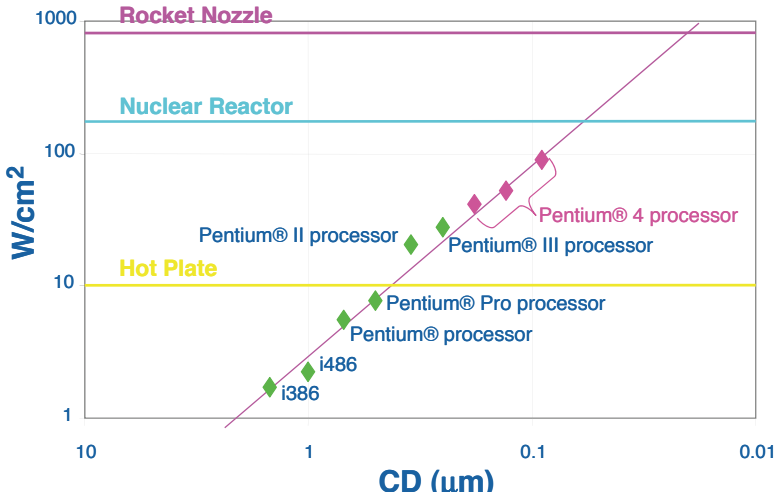


Figure 2.19: The power density of Intel microprocessors *vs.* feature size of transistors [31, 32].

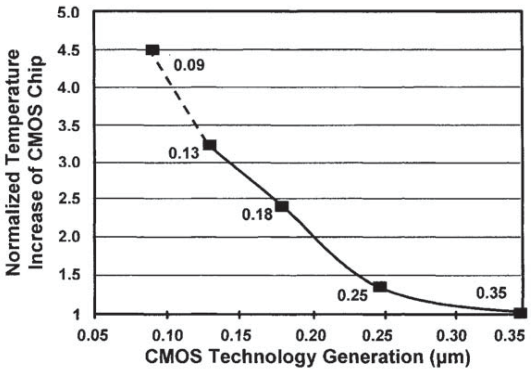


Figure 2.20: The temperature increase in CMOS circuits *vs.* feature size of transistors [33].

nects, hot electron effects, negative bias temperature instability, *etc.* In summary, temperature variations become a challenge for the design of circuits as technology scales and integration density increases.

2.3.2 Thermal modeling methods for integrated circuits

To accurately estimate the timing performance and power consumption of circuits under temperature variations and to mitigate the thermal issues, it is necessary to accurately and efficiently model the temperature (thermal) profile of circuits. Heat transfer models are commonly used to estimate the temperature distribution and variations. The heat transfer models can be divided into two types: steady-state and transient heat transfer models.

Steady-state heat transfer model

In steady-state models, only the thermal conductance of circuits is considered, which is determined by the thermal resistance and is time-invariant [122]. Given a constant power generated in the circuit, the resulting temperature distribution is constant at any time. Based on the thermal-electrical analogy, thermal resistor networks are commonly used to model the steady-state heat transfer [123]. The thermal conductance is obtained based on the geometric parameters and thermal conductivity of different parts of a circuit.

The heat transfer is modeled by Fourier's law of conduction [124]. Due to Fourier's law, the heat flux q (heat generated per unit area) is proportional to the negative temperature gradient ∇T, with a coefficient proportional to the thermal conductivity of the material k_t,

$$q = -k_t \nabla T. \tag{2.14}$$

Different models have been developed to obtain T at different abstraction levels. The simplest model, 1-D thermal model focuses on the vertical heat transfer paths, as illustrated in Fig. 2.21 [34]. Although analytic expressions for temperature can be obtained fast from 1-D model, non-vertical heat transfer is ignored, which introduces high error to the estimated temperature.

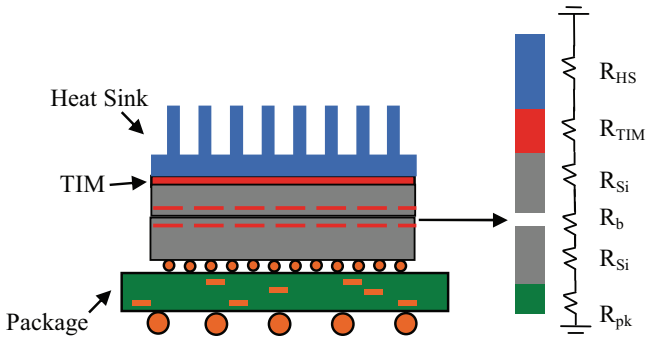


Figure 2.21: One-dimensional steady-state heat transfer model for a two-tier 3-D IC [34], where TIM is the abbreviation for thermal interface material.

To fully describe the heat transfer and accurately estimate the temperature, the divergence of q in different directions within a region needs to be considered. Based on (2.14), the steady-state divergence of q is determined by

$$\nabla \cdot q = g(\mathbf{r}) = -k_t \nabla^2 T(\mathbf{r}), \tag{2.15}$$

where $\mathbf{r}$ is the coordinate vector of the node where the temperature is being investigated and $g(\mathbf{r})$ is the power generated per unit volume. This *partial differential equation* (PDE) is also called Poisson's equation. For integrated circuits, *finite difference method* (FDM) and

finite element method (FEM) are widely used to discretize equation (2.15) to determine the temperature distribution. The key difference between FDM and FEM is that FDM discretizes the differential operator of T in Poisson's equation [125, 126], while FEM discretizes the temperature field [127]. In both methods, the entire circuit is divided into multiple cells to calculate the temperature at a finite number of nodes.

Both FDM and FEM convert the Poisson's equation (2.15) into a large linear equations array. The matrices describing this system is usually large and sparse. Consequently, solving this linear equation array is similar to the *IR*-drop analysis problem for power grids. Several methods can be used to obtain the solutions, *e.g.*, Gaussian elimination, iterative methods, random walk methods, and hierarchical solving methods [126].

Transient heat transfer model

In transient heat transfer model, the time-variant power is considered. The partial differential of T to time t needs to be added to the heat transfer equation (2.15) to include the time domain,

$$-k_t \nabla^2 T(\mathbf{r}) = g(\mathbf{r}, t) - \rho c_p \frac{\partial T(\mathbf{r}, t)}{\partial t}, \tag{2.16}$$

where ρ is the density of the material (kg/m^3) and c_p is heat capacity of the materials employed for the circuit. Instead of the resistor networks in steady-state analysis, *RC* models are used to denote different parts of an integrated system in transient thermal analysis. FDM models are widely used to obtain the transient temperature distribution [126].

At the architecture level, HotSpot is widely used to analyze the transient thermal behavior for a circuit [35, 36, 128]. The objectivity of HotSpot is to obtain a coarsely discretized FDM solution across the chip, which is modeled by *RC* models, as illustrated in Fig. 2.22. The temperature difference within a module is ignored in HotSpot. To obtain a more detailed temperature distribution, fine grain methods such as *alternating-direction-implicit* (ADI) method can be used [129].

2.4 Summary

The background of process variations, power supply noise, and temperature variations is introduced in this chapter. A number of techniques have been proposed to model PVT variations in 2-D circuits. In 3-D integration, the vertically stacked tiers have introduced a new physical dimension and new materials (*e.g.*, TSVs and bonding layers). Consequently, the traditional modeling and design techniques for 2-D ICs need to be adapted to cope with 3-D circuits. The process, voltage, and temperature variations and the related effects on these circuits are investigated in the following chapters.

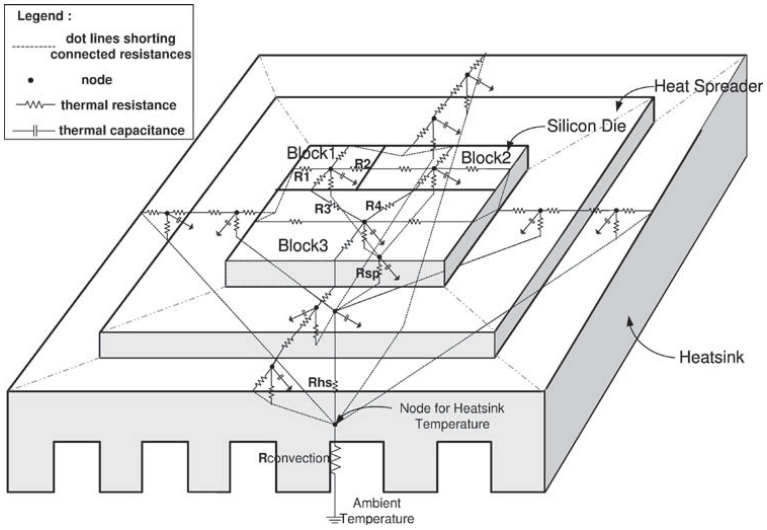


Figure 2.22: A HotSpot *RC* model for a circuit with three architectural modules [35, 36].

3 Process Variations in 3-D ICs

Process variations have been introduced in Section 2.1, where the parameter variations of transistors and wires have been discussed. The effect of process variations on the timing of 3-D ICs is investigated in this chapter. In particular, the focus of this chapter is the process-induced uncertainty in 3-D clock distribution networks.

Techniques to model process variations for the entire circuit are first introduced in Section 3.1. Work related to process variations in 3-D ICs is surveyed in Section 3.1.3. The effect of process variations on clock distribution networks is investigated in Section 3.2. A novel model to capture this effect is proposed in Section 3.3. Based on this analysis, new topologies for 3-D clock trees are proposed in Section 3.4 to mitigate the deleterious effects of process variations.

3.1 Process Variations Modeling for Integrated Circuits

Methods for full-chip process variations modeling are introduced in this section. For the statistical modeling of circuits, process variations are modeled as D2D and WID variations, as discussed in the previous chapter (see Fig. 2.3). The variation models introduced in Section 2.1 are used to describe the behavior of transistors and wires. The objective of full-chip analysis, however, is to describe the delay variations of data and clock paths. Two types of methodologies are commonly used to achieve this goal: corner-based analysis and statistical timing analysis. Corner-based analysis has been introduced in Section 2.1.3, which usually results in pessimistic results [101]. The statistical timing analysis methods, therefore, are discussed in this chapter. Two important statistical timing analysis methods, Monte-Carlo simulations and *Statistical Static Timing Analysis* (SSTA), are described in the following subsections.

3.1.1 Monte-Carlo simulations

Monte-Carlo methods have widely been used in statistical modeling [130]. Multiple iterations of simulations need to be run in Monte-Carlo methods. In each iteration, the random variables

are re-sampled based on the given distribution of these variables. The statistical results, *e.g.*, mean and standard deviation, are obtained after a specified number of iterations.

In general, an accurate estimation of the statistical data can be obtained after a large number of Monte-Carlo simulations. For instance, SPICE-based Monte-Carlo simulations can be used to capture the non-deterministic characteristics of devices [131]. Delay variation can be obtained by applying Monte-Carlo simulations for the entire data or clock paths. Nevertheless, the runtime for Monte-Carlo simulations is excessively high due to the large number of simulations. An example for the obtained standard deviation σ of a path delay with the number of Monte-Carlo simulations is illustrated in Fig. 3.1, where a path with 20 inverters is simulated. As shown in this example, more than 2000 Monte-Carlo simulations are required to obtain a converging σ. Considering the time required by the SPICE-based simulations and the large number of pairs of clock paths in a clock distribution network, this method is prohibitively time-consuming. Statistical static timing analysis, therefore, is usually an efficient alternative to obtain the process-induced delay.

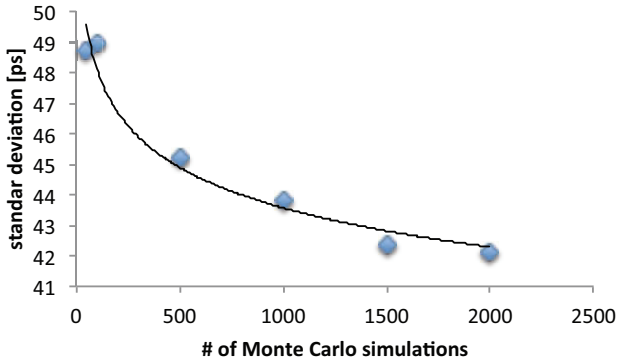


Figure 3.1: The standard deviation of the delay of an inverter chain *vs.* the number of Monte-Carlo simulations.

3.1.2 Statistical static siming analysis

SSTA has been developed as a statistical version of the classic *Static Timing Analysis* (STA) method to analyze the path delay and timing of circuits [96, 132, 133]. Different from the transient state timing analysis (*e.g.*, SPICE), STA determines the path delay based on the topology of the circuit and the given static delay model of devices and interconnects. The topology of the circuit is interpreted as timing graphs, where a node represents a pin of a gate and the edge denotes a connection between a pair of nodes. A timing graph considering the delay of both gates and wires is illustrated in Fig. 3.2.

The weight of the edge denotes the delay between two nodes determined by the fanin and fanout of the related nodes and the given timing models for gates and wires. By traversing

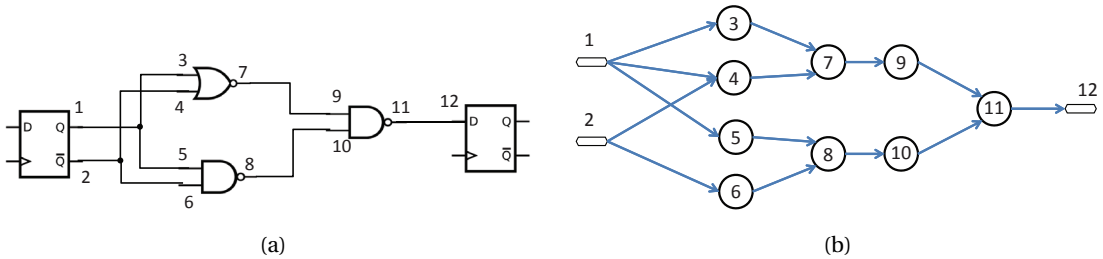


Figure 3.2: An example of timing graph used in STA, where (a) and (b) are the logic gates between two Flip-Flops and the corresponding timing graph, respectively.

the timing graphs, the delay can be determined for all the data and clock paths. Two types of traversal methods have been developed for STA, *i.e.*, path-based and block-based STA. In path-based STA, the timing graph is traversed path by path to obtain the delay for all specified data or clock paths [134, 135]. The maximum (minimum) delay between a source and a sink is determined by the maximum (minimum) delay among all the corresponding paths. Path-based methods are suitable for small designs but not for the large circuits due to the large number of paths.

In block-based STA, however, the delay is calculated node by node, where each node is traversed only once [136, 137]. The maximum (minimum) arrival time to each node is determined after the maximum (minimum) arrival time of all its fanins is determined. For large-scale circuits, block-based STA is more efficient than path-based STA, since the number of nodes is typically lower than the number of paths between Flip-Flops [138].

The delay of gates and wires in STA is considered to be deterministic. When process variations are included, however, the delay becomes a random variable. Consequently, SSTA has been developed to address this situation [96]. The timing graph used in SSTA is similar to STA, except that the weight of each edge is denoted by a random variable. The target of SSTA is, therefore, to obtain the distribution of the maximum path delay between the given source and sink. SSTA can also be classified into path-based and block-based methods. Typically, block-based methods are preferred due to their higher efficiency, similar to block-based STA. SSTA methods have been used to model the effect of process variations on the delay of both data and clock paths for 2-D ICs [96, 139]. For 3-D circuits, SSTA has been used to investigate the effect of process variations on critical datapaths, which is discussed in the following subsection.

3.1.3 Related works on process variations in 3-D ICs

Recent works analyzing the effect of process variations on the speed of 3-D ICs are presented in [88, 89, 140], where the impact of process variations on the delay of datapaths is investigated. An analytical model, 3D-GCP, for the impact of process variations on the critical path delay distribution of 3-D ICs is proposed in [88]. The model is used to describe the distribution of

the maximum delay among a given set of critical paths. The investigated critical paths are illustrated in Fig. 3.3. In this model, two types of critical paths are modeled: WID paths (shown

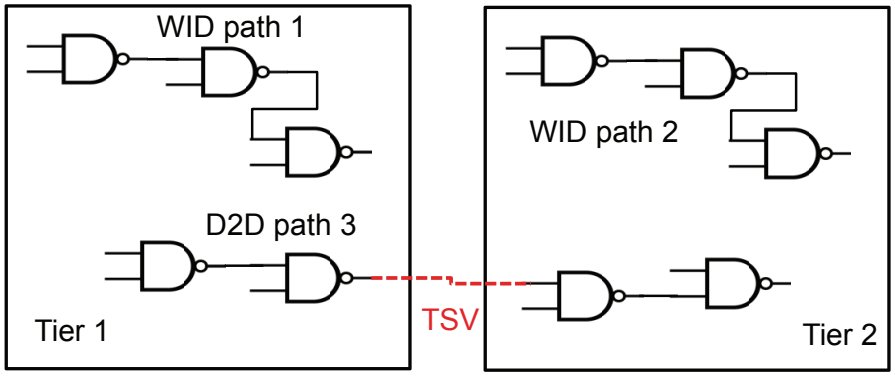


Figure 3.3: The critical paths modeled by 3D-GCP.

by paths 1 and 2 in Fig. 3.3) and D2D paths (shown by path 3). As with these datapaths, all the critical paths are assumed to have no common segments. Based on this assumption, it has been shown that 3-D circuits behave very differently under the impact of process variations as compared to 2-D circuits [88]. The distribution of datapaths across planes significantly affects the timing yield (the distribution of the highest clock frequency).

Based on 3D-GCP, a system-level process variations analysis has been presented in [89], where the timing yield in 3-D ICs with multiple clock domains has also been investigated. Again, only the effect of process variations on datapaths, which share no parts, is discussed. Furthermore, it is assumed that for multiple-clock designs, the different clock domains are separately located in different planes. Another comparison between process variations in 2-D and 3-D ICs is presented in [140]. The difference in process-induced timing variability has, again, been observed between 2-D and 3-D circuits. Statistical timing analysis has been implemented for datapaths to show the advantage of 3-D ICs over 2-D circuits in reducing the variation of critical path delay due to process variations.

These prior works focus on the timing uncertainty in 3-D ICs caused by the varying delay of the datapaths. Nevertheless, timing also depends on the clock uncertainty, due to the variations of the clock distribution. For instance, the setup slack between two flip-flops can be determined by [75]

$$\text{slack}_{\text{setup}} = \text{clock period} + \text{clock skew} - \text{Data delay} - \text{setup time}, \tag{3.1}$$

where clock skew is the difference between the clock delay to the sink and source flip-flops and the setup time is determined by the sink flip-flop. Consequently, the timing of a circuit is not only determined by the delay variation of critical datapaths but is also highly affected by skew variation within the clock distribution networks. Accurately modeling the skew variability due to process variations is, therefore, necessary to evaluate the timing of a 3-D circuit.

The aforementioned modeling methods for the process-induced delay variation in 3-D ICs, however, cannot be used to analyze clock distribution networks. In these works, the datapaths are assumed to have no common segments. This assumption cannot be used to determine the skew of clock distribution networks, since the clock paths to different sinks in clock distribution networks can (actually are designed to) share segments, as illustrated in Fig. 3.4. This "structure correlation", therefore, should be considered when calculating the distribution of skew variation. In addition, the WID variation is considered either independent or fully

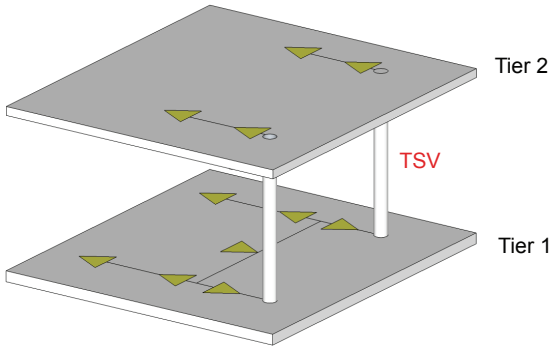


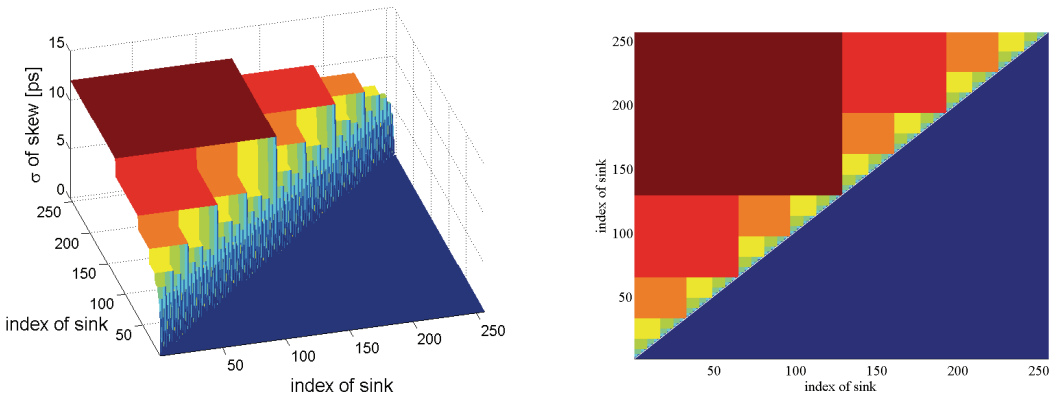
Figure 3.4: Clock paths sharing different branches within a 3-D circuit.

correlated among the devices within the same plane in [88, 89, 140]. The spatial correlation of WID variations (between zero and one), however, has been shown to be non-negligible [37,86,97,139,141]. Consequently, clock delay and skew variation in 3-D ICs should be modeled considering both the shared branches among clock paths and the spatial WID correlation.

The traditional methods employed to model skew variation in 2-D clock trees cannot be directly applied to 3-D clock trees either. For the same parameter in 2-D ICs, the D2D variation remains uniform for all the devices or interconnects. Consequently, D2D variations are often ignored or simplified in skew analysis [37, 139, 142]. In 3-D ICs, however, the D2D variations differ among tiers (planes). The WID variations are independent among tiers. Both the clock and data paths can span more than one tier (see Figs. 3.3 and 3.4), which complicates the statistical timing analysis for 3-D circuits. These observations are considered in the following sections to accurately model the clock uncertainty, where the effect of process variations on 3-D clock distribution networks is investigated and a novel model is proposed to describe this effect.

3.2 The Effect of Process Variations on Clock Distribution Networks

As introduced in Section 1.4, the maximum clock frequency of a circuit is significantly affected by the clock skew between clock sinks. Both the setup and hold slacks vary with clock skew. Clock skew is introduced at the design and fabrication stages and during the operation of ICs. There is a plethora of methods to manage the excessive clock skew in the design phase [39,



(a) (b)

Figure 3.5: The standard deviation of skew between each pair of clock sinks in a 2-D H-tree, where (b) is the top-view of (a).

75, 79] for 2-D circuits. Careful physical design, however, does not guarantee the elimination of the undesirable skew since some skew can be introduced in the fabrication phase. This unwanted skew is due to the process variations of the clock buffers and wires.

3.2.1 Process-induced statistical skew

The delay variations of clock buffers and clock wires introduce differences in the delay among clock paths. Even for a symmetric clock tree, this difference in the path delay can be significant. Since the origins of the difference are statistically described, the resulting delay difference also exhibits a statistical behavior. For instance, for a 2-D H-tree with 256 clock sinks, the standard deviation of clock skew between each pair of clock sinks is illustrated in Fig. 3.5. In this figure, $\sigma_{i,j}$ denotes the standard deviation of skew between sinks i and j. Since $\sigma_{i,j} = \sigma_{j,i}$, only half of the skew array is shown in this figure for clarity. The skew variation changes significantly with the location of the corresponding pairs. Accurately modeling process-induced skew variation in a clock tree is, consequently, important to precisely estimate the timing performance of a circuit.

Several techniques for analyzing the effect of process variation on clock skew have been developed for 2-D circuits emphasizing intra-die variations. As introduced in Section 3.1, process-induced timing uncertainty can be modeled by corner-based analysis or statistical timing analysis. A method for statistical clock skew analysis based on Monte Carlo simulations is described in [143]. The computational time of this method is, however, prohibitively high for large scale ICs. Based on statistical timing analysis [133], other statistical skew modeling methods considering intra-die variations are presented in [37, 139, 142, 144, 145] to efficiently analyze skew variations. In 2-D ICs, since the inter-die process variations uniformly affect the devices within a circuit, the majority of the skew analysis methods emphasizes the

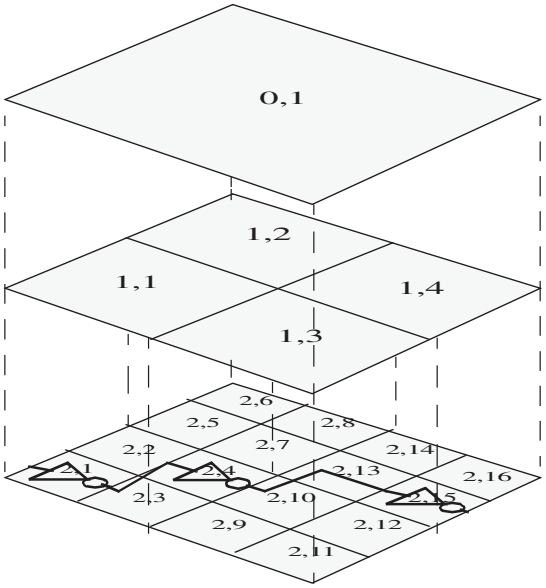


Figure 3.6: Modeling spatial correlations using quad-tree partitioning [37].

intra-die (WID) variations. For WID variations, one of the most important and tedious tasks is to determine the spatial correlation of the same parameter among different devices (or interconnects).

3.2.2 Spatial correlation

The WID variations typically exhibit a spatial correlation [37, 92, 97, 141]. In general, the correlation between a pair of devices (or interconnects) located within the same plane decreases with their distance. Several methods have been proposed to model this spatial correlation [37, 86, 141].

The spatial correlation model (multi-level correlation) used in this dissertation is based on the statistical timing analysis method proposed in [37]. A multi-level quad-tree partitioning is used and the WID variation of a parameter is divided into l levels, as illustrated in Fig. 3.6 [37]. At the k^{th} level, there are 4^{k-1} regions.

An independent variable is assigned to each region to represent a component of the WID variation of a parameter. The overall WID variation of a parameter of a device or interconnect is composed by the sum of these independent components at different levels. For instance, the WID variation of the channel length of transistor i is described by

$$\Delta L_{\text{gate}(i)} = \sum_{\text{region } r \text{ intersects } k}^{1 \le k \le l} \Delta L_{\text{gate}}(k, r), \tag{3.2}$$

where $\Delta L_{\text{gate}}(k,r)$ is the random variable associated with the quad-tree at level k, region (k,r), as shown in Fig. 3.6. This distribution is obtained by dividing the total WID variability among the different levels. All the random variables $\Delta L_{\text{gate}}(k,r)$ associated with a particular level k are assigned to identical and independent probability distributions. Consequently, the spatial correlation between devices in the same plane can be modeled. Devices located close to each other are highly correlated, while the devices separated by a large physical distance exhibit low correlation.

Given that the total WID variation is equally divided into different levels, the total correlation between the WID variation of the same parameter of devices (or interconnects) i and j is described by the sum of the correlations at all the levels,

$$\text{corr}(i,j) = \frac{1}{l}\sum_{k=1}^{l} \text{corr}_k(i,j), \tag{3.3}$$

where $\text{corr}_k(i,j)$ is the correlation between buffers i and j at the k^{th} level. As illustrated in Fig. 3.6, assuming buffers i and j are located in the zones (k,region_i) and (k,region_j), respectively,

$$\text{corr}_k(i,j) = \begin{cases} 1, & \text{if } (k,\text{region}_i) = (k,\text{region}_j) \\ 0, & \text{if } (k,\text{region}_i) \neq (k,\text{region}_j) \end{cases}. \tag{3.4}$$

WID variations are well modeled for 2-D ICs. Nevertheless, D2D variations are usually neglected when analyzing the variation of skew in 2-D clock trees. In 3-D ICs, however, both WID and D2D variations need to be included in the statistical model for skew, as discussed in the following section.

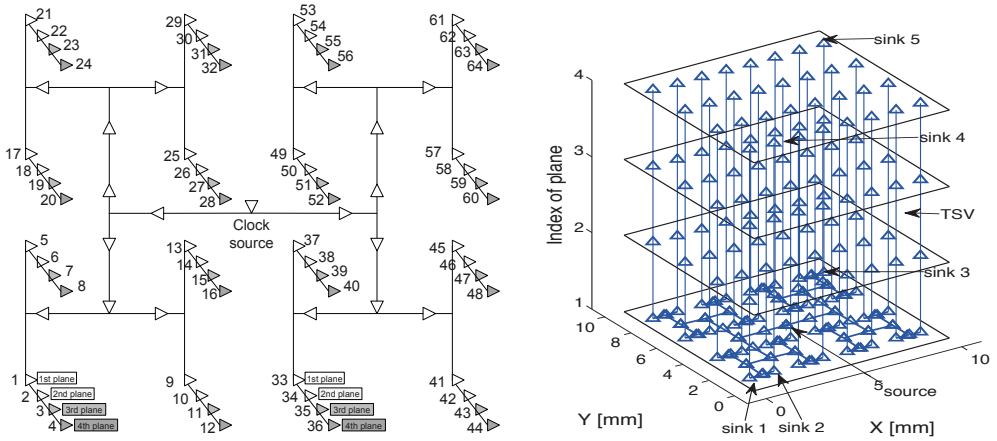
3.3 A Novel Model for Process-Induced Skew in 3-D ICs

Although statistical skew analysis has been studied in 2-D ICs, the resulting methods cannot be directly applied to 3-D systems. As mentioned in the previous section, the skew analysis methods for 2-D clock distribution networks focus on the effect of WID variations, since D2D variations uniformly affect the devices (wires) within a circuit. For well-designed clock distribution networks, this uniform effect can be neglected due to the balanced clock paths. In 3-D ICs, however, the D2D variation cannot be neglected, since clock paths can span multiple dies and these paths are affected by different D2D variations. The problem of modeling skew variation is formulated in the following subsection. A new method to characterize the statistical electrical characteristics of clock buffers is presented in Section 3.3.2. This method is employed to provide the new statistical skew model for 3-D ICs, which is proposed in Section 3.3.3.

3.3.1 Problem formulation for modeling skew variation

The problem of skew analysis for 3-D clock distribution networks considering process variations is formulated in this subsection. As discussed in [75], only the clock skew between the sequential elements which transfer data between each other (data-related or "sequentially-adjacent" registers) affects the performance of a circuit. Consequently, in addition to global skew, appropriate pairwise skew distributions are used to evaluate the performance of clock distribution networks [145].

H-tree is a common topology used to globally distribute the clock signal within a circuit [39,75]. A typical buffered 3-D H-tree is illustrated in Fig. 3.7. The pairwise clock skew is defined as the skew between every pair of sinks in 3-D clock distribution networks, $S_{\text{skew}} = \{s_{i,j} \mid s_{i,j} = D_i - D_j, i \neq j \text{ and } 1 \leq i, j \leq n_{\text{sink}}\}$. Sinks i and j can be located in any plane of the 3-D circuit and $s_{i,j}$ denotes the skew between sinks i and j. The clock delay to sinks i and j is denoted by D_i and D_j, respectively. The number of clock sinks is n_{sink}.



(a) (b)

Figure 3.7: 3-D H-trees spanning four planes, where (a) is the topology of a 3-D H-tree and (b) is the 3-D view of a 3-D H-tree.

The number of buffers and the length of the interconnects in clock trees significantly affect the distribution of clock skew. These quantities depend, in turn, on the area and number of the physical planes comprising a 3-D IC, affecting the highest clock frequency that can be supported by a 3-D IC. By investigating the effect of process variations on S_{skew}, several guidelines for the design of 3-D global clock trees with a low ΔS_{skew} are offered.

3.3.2 Modeling the statistical delay of a buffer stage

Typical methods to model the process-induced delay of transistors have been introduced in Section 2.1.3. These methods require the detailed information about the distribution of all the parameters of transistors. For some industrial device libraries, however, only the Monte-Carlo

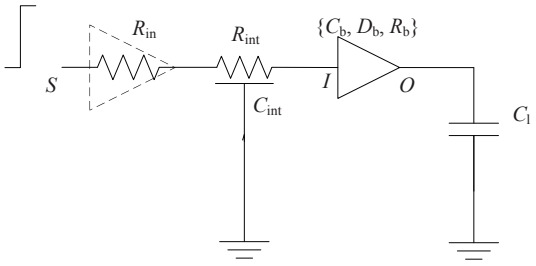


Figure 3.8: An elemental circuit used to measure the variations in the buffer characteristics.

model of transistors is provided, where the parameter variation is unknown [146]. A fast method to extract the distribution of electrical characteristics of clock buffers is proposed in this subsection, where the detailed information of the distribution of parameters is not required.

The fluctuation of the buffer delay is typically approximated as being linear in the device parameter variations [145,147]. Alternatively, the variation in delay can be determined through the variations of the input capacitance and output resistance [142]. The second method is enhanced by considering the input slew rate to more accurately model the distribution of the buffer delay. The interconnects are modeled as distributed *RC* wires. The circuit illustrated in Fig. 3.8 is utilized to obtain the variation of buffer delay for different slew rates of the input signal.

Let R_{in} denote the output resistance of a buffer driving the buffer under consideration. The load capacitance of the buffer under consideration is denoted by C_{l}. Interconnects with diverse impedance characteristics are modeled by employing different R_{int} and C_{int}, where R_{int} and C_{int} denote the resistance and capacitance of the interconnects, respectively. The interconnect R_{int} and C_{int} can also be adjusted to produce different slew rates for the input signal of the buffer in Fig. 3.8.

For a step input signal, the Elmore delay [111] from source S to nodes I and O in Fig. 3.8, respectively, is

$$D_{SI} = 0.69R_{\text{in}}C_{\text{int}} + 0.38R_{\text{int}}C_{\text{int}} + 0.69(R_{\text{in}} + R_{\text{int}})C_{\text{b}}, \tag{3.5}$$

$$\Delta D_{SI} = 0.69(R_{\text{in}} + R_{\text{int}})\Delta C_{\text{b}}, \tag{3.6}$$

$$D_{SO} = D_{SI} + D_{\text{b}} + 0.69R_{\text{b}}C_{\text{l}}, \tag{3.7}$$

$$\Delta D_{SO} = \Delta D_{SI} + \Delta D_{\text{b}} + 0.69C_{\text{l}}\Delta R_{\text{b}}, \tag{3.8}$$

where C_{b}, R_{b}, and D_{b} are the input capacitance, the output resistance, and the intrinsic delay of the buffer, respectively. The variations of C_{b}, R_{b}, and D_{b} are denoted by ΔC_{b}, ΔR_{b}, and ΔD_{b}, respectively. While investigating the buffer as shown in Fig. 3.8, the R_{in} is presumed constant (for the moment).

The delay variation at nodes I and O are evaluated with Monte-Carlo simulations. The standard deviation $\sigma_{D_{SI}}$ and $\sigma_{D_{SO}}$ can be obtained from these simulations. The distribution of ΔD_{SO} is measured in two cases:

1. setting C_l to zero (corresponding to ΔD_{SO_0} and $\sigma_{D_{SO_0}}$)
2. setting C_l to a value close to the actual capacitive load of a buffer stage in a 3-D clock tree (*e.g.*, 200 fF, corresponding to ΔD_{SO_1} and $\sigma_{D_{SO_1}}$).

The mean and standard deviation of ΔC_b, ΔR_b, and ΔD_b can, consequently, be obtained by (3.6) and (3.8). Assuming the sources of the process variations can be described by Gaussian distribution, the characteristics of a buffer can also be approximated as Gaussian distribution [86],

$$\Delta C_\mathrm{b} \sim \mathcal{N}(0, \sigma^2_{C_\mathrm{b}}), \quad \Delta R_\mathrm{b} \sim \mathcal{N}(0, \sigma^2_{R_\mathrm{b}}), \quad \Delta D_\mathrm{b} \sim \mathcal{N}(0, \sigma^2_{D_\mathrm{b}}). \tag{3.9}$$

Given $\sigma_{D_{SI}}$, the standard deviation of the input capacitance σ_{C_b} is directly obtained through (3.6),

$$\sigma_{C_\mathrm{b}} = \frac{\sigma_{D_{SI}}}{0.69(R_\mathrm{in} + R_\mathrm{int})}. \tag{3.10}$$

According to (3.6) and (3.8), $\sigma_{D_{SO}}$ is determined by σ_{C_b}, σ_{D_b}, σ_{R_b}, and the covariance among these variables,

$$\sigma^2_{D_{SO_0}} = (0.69(R_\mathrm{in} + R_\mathrm{int})\sigma_{C_\mathrm{b}})^2 + \sigma^2_{D_\mathrm{b}} + 1.38(R_\mathrm{in} + R_\mathrm{int}) \cdot \mathrm{cov}(D_\mathrm{b}, C_\mathrm{b}), \tag{3.11}$$

$$\begin{aligned}\sigma^2_{D_{SO_1}} = {} & (0.69(R_\mathrm{in} + R_\mathrm{int})\sigma_{C_\mathrm{b}})^2 + \sigma^2_{D_\mathrm{b}} + (0.69\sigma_{R_\mathrm{b}} C_\mathrm{l})^2 + 1.38(R_\mathrm{in} + R_\mathrm{int}) \cdot \mathrm{cov}(D_\mathrm{b}, C_\mathrm{b}) \\ & + 1.38 C_\mathrm{l} \cdot \mathrm{cov}(D_\mathrm{b}, R_\mathrm{b}) + 0.952 C_\mathrm{l}(R_\mathrm{in} + R_\mathrm{int}) \cdot \mathrm{cov}(C_\mathrm{b}, R_\mathrm{b}).\end{aligned} \tag{3.12}$$

Recalling that $\mathrm{cov}(a,b) = \sigma_a \sigma_b \mathrm{corr}(a,b)$ and (3.10), the above expressions can be rewritten as

$$\sigma^2_{D_{SO_0}} = \sigma^2_{D_{SI}} + \sigma^2_{D_\mathrm{b}} + 2\sigma_{D_{SI}}\sigma_{D_\mathrm{b}} \cdot \mathrm{corr}(D_\mathrm{b}, C_\mathrm{b}), \tag{3.13}$$

$$\begin{aligned}\sigma^2_{D_{SO_1}} = {} & \sigma^2_{D_{SO_0}} + (0.69 C_\mathrm{l})^2 \sigma^2_{R_\mathrm{b}} + 1.38 C_\mathrm{l} \sigma_{D_\mathrm{b}} \sigma_{R_\mathrm{b}} \cdot \mathrm{corr}(D_\mathrm{b}, R_\mathrm{b}) \\ & + 1.38 C_\mathrm{l} \sigma_{D_{SI}} \sigma_{R_\mathrm{b}} \cdot \mathrm{corr}(C_\mathrm{b}, R_\mathrm{b}).\end{aligned} \tag{3.14}$$

Given the measured $\sigma_{D_{SI}}$ and $\sigma_{D_{SO_0}}$, the standard deviation of the intrinsic delay of buffers σ_{D_b} can be obtained based on a quadratic equation transformed from (3.13),

$$\sigma^2_{D_\mathrm{b}} + \left(2\sigma_{D_{SI}} \cdot \mathrm{corr}(D_\mathrm{b}, C_\mathrm{b})\right)\sigma_{D_\mathrm{b}} + (\sigma^2_{D_{SI}} - \sigma^2_{D_{SO_0}}) = 0. \tag{3.15}$$

With the obtained σ_{C_b} and σ_{D_b}, the standard deviation of the output resistance of buffers σ_{R_b} is calculated based on the quadratic equation transformed from (3.14),

$$(0.69 C_\mathrm{l})^2 \sigma^2_{R_\mathrm{b}} + A \cdot \sigma_{R_\mathrm{b}} + (\sigma^2_{D_{SO_0}} - \sigma^2_{D_{SO_1}}) = 0,$$

$$A = 1.38C_{\mathrm{l}}\sigma_{D_{\mathrm{b}}} \cdot \mathrm{corr}(D_{\mathrm{b}}, R_{\mathrm{b}}) + 1.38C_{\mathrm{l}}\sigma_{D_{SI}} \cdot \mathrm{corr}(C_{\mathrm{b}}, R_{\mathrm{b}}). \tag{3.16}$$

As shown in (3.15) and (3.16), the obtained $\sigma_{R_{\mathrm{b}}}$ and $\sigma_{D_{\mathrm{b}}}$ depend on the correlation $\mathrm{corr}(D_{\mathrm{b}}, C_{\mathrm{b}})$, $\mathrm{corr}(D_{\mathrm{b}}, R_{\mathrm{b}})$, and $\mathrm{corr}(C_{\mathrm{b}}, R_{\mathrm{b}})$. In the proposed model of skew variation, $\sigma_{C_{\mathrm{b}}}$, $\sigma_{R_{\mathrm{b}}}$, and $\sigma_{D_{\mathrm{b}}}$ are used to obtain the delay variation of each buffer stage Δd_i, which is similar to ΔD_{SO_1}. The process to determine Δd_i is illustrated in Fig. 3.9. When calculating σ_{d_i}, the pre-calculated

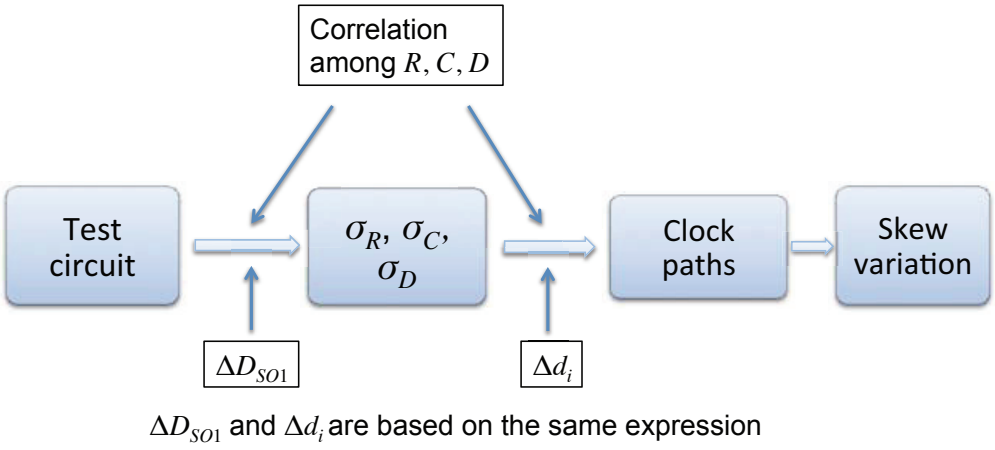


Figure 3.9: The flow to determine clock skew variation by using the parameters extracted from the test circuit.

$\sigma_{C_{\mathrm{b}}}$, $\sigma_{R_{\mathrm{b}}}$, $\sigma_{D_{\mathrm{b}}}$, and the correlation among them are substituted into (3.14) again. Note that these terms are extracted based on the simulation of a test circuit similar to a buffer stage (see Fig. 3.8). Consequently, the correlation among ΔC_{b}, ΔR_{b}, and ΔD_{b} does not affect Δd_i significantly, as long as Δd_i is calculated based on the same correlation. Since ΔC_{b}, ΔR_{b}, and ΔD_{b} are due to the same process variation sources, these variables are assumed to be fully correlated herein.

3.3.3 Modeling the statistical skew in 3-D circuits

An example of a 3-D clock path is illustrated in Fig. 3.10. Note that this path is general and can be applied to any 3-D clock tree in addition to the 3-D topologies investigated herein. The devices in different physical planes are connected by TSVs [68], which, in turn, are modeled as *RC* wires of different resistance and capacitance as compared to the horizontal wires (*e.g.*, R_{TSV} and C_{TSV} in Fig. 3.10). R_{TSV} and C_{TSV} are considered fixed.

Consider the clock path consisting of buffers $i-1$, i, and $i+1$. From (3.6) and (3.8), the delay variation Δd_i attributed to the variation of buffer i along the investigated path is

$$\begin{aligned}\Delta d_i = {} & 0.69(\overline{R}_{\mathrm{in}(i)} + \Delta R_{\mathrm{b}(i-1)})\Delta C_{\mathrm{b}(i)} + 0.69\Delta R_{\mathrm{b}(i)}(\overline{C}_{\mathrm{l}(i)} + \Delta C_{\mathrm{b}(i+1)} + \Delta C_{\mathrm{b}(j)}) \\ & + 0.69\overline{R}_{\mathrm{b}(i)}\Delta C_{\mathrm{b}(j)} + \Delta D_{\mathrm{b}(i)},\end{aligned} \tag{3.17}$$

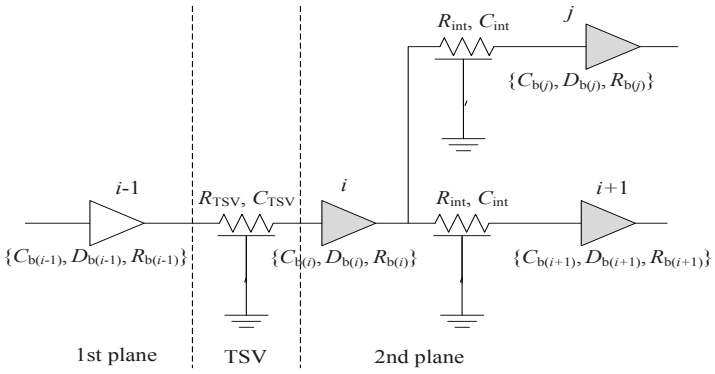


Figure 3.10: The electrical model of a segment of a clock path.

$$R_{\text{in}(i)} = R_{\text{b}(i-1)} + R_{\text{TSV}}, \tag{3.18}$$

$$C_{\text{l}(i)} = 2C_{\text{int}} + C_{\text{b}(i+1)} + C_{\text{b}(j)}, \tag{3.19}$$

where the bar ($\bar{}$) denotes the nominal value. For buffer i, the $\Delta R_{\text{b}(i-1)}$ of the upstream buffer and $\Delta C_{\text{b}(i+1)}$ of the downstream buffer are both included in (3.17). To determine the delay of a clock path, Δd_i for all the buffers along this path is summed up. In this case, $\Delta R_{\text{b}(i-1)}\Delta C_{\text{b}(i)}$ and $\Delta R_{\text{b}(i)}\Delta C_{\text{b}(i+1)}$ are duplicated. Therefore, one of these two terms needs to be removed. Consequently, Δd_i is rewritten as

$$\begin{aligned}\Delta d_i =& 0.69\left(\overline{R}_{\text{in}(i)}\Delta C_{\text{b}(i)} + \Delta R_{\text{b}(i)}(\overline{C}_{\text{l}(i)} + \Delta C_{\text{b}(i+1)} + \Delta C_{\text{b}(j)}) + \overline{R}_{\text{b}(i)}\Delta C_{\text{b}(j)}\right) + \Delta D_{\text{b}(i)} \\ =& 0.69\left(\overline{R}_{\text{in}(i)}\Delta C_{\text{b}(i)} + \overline{R}_{\text{b}(i)}\Delta C_{\text{b}(j)}\right) + \Delta D_{\text{b}(i)} + \delta_i,\end{aligned} \tag{3.20}$$

where $\delta_i = 0.69\Delta R_{\text{b}(i)}(\overline{C}_{\text{l}(i)} + \Delta C_{\text{b}(i+1)} + \Delta C_{\text{b}(j)})$.

The variation of ΔC_{b} is relatively low as compared with the nominal C_{b} ($\sigma/\mu < 3\%$ for both D2D and WID variations in the simulations). The observed delay variation of buffers in other works is also much lower than the nominal value (*e.g.*, $\sigma/\mu \leq 5\%$ for both D2D and WID variations as reported in [92]). δ_i can, therefore, be approximated linearly using a first-order Taylor expansion around the nominal value [86],

$$\begin{aligned}\delta_i \approx& \left[\frac{\partial \delta_i}{\partial \Delta R_{\text{b}(i)}}\right]_0 \Delta R_{\text{b}(i)} + \left[\frac{\partial \delta_i}{\partial \Delta C_{\text{b}(i+1)}}\right]_0 \Delta C_{\text{b}(i+1)} + \left[\frac{\partial \delta_i}{\partial \Delta C_{\text{b}(j)}}\right]_0 \Delta C_{\text{b}(j)} \\ =& 0.69\overline{C}_{\text{l}(i)}\Delta R_{\text{b}(i)}.\end{aligned} \tag{3.21}$$

As reported in [88, 89, 92], the σ/μ of the transistor characteristics is typically considered $\leq 5\%$. For instance, for transistors and interconnects based on PTM 45 nm CMOS and global wire models [41] and the ITRS reports [43], the $\sigma_{R_{\text{b}}}/\mu_{R_{\text{b}}}$ and $\sigma_{C_{\text{b}}}/\mu_{C_{\text{b}}}$ are below 5.1% and 2.3%, respectively (as will be shown in Table 3.2). The 3σ variation is smaller than 15% of the nominal value for R_{b} and 10% for ΔC_{b}. Since ΔC_{b} and ΔR_{b} are modeled by Gaussian distributions,

for more than 99.7% buffers, $\Delta C_\mathrm{b}\Delta R_\mathrm{b}$ is lower than $1.5\%C_\mathrm{b}R_\mathrm{b}$. Moreover, from the nominal value and standard deviation of C_b, R_b, and D_b reported in the simulations, $0.69\Delta C_\mathrm{b}\Delta R_\mathrm{b}$ and $0.69C_\mathrm{b}R_\mathrm{b}$ are much lower than ΔD_b and D_b, respectively. Consequently, approximating δ_i with (3.21) does not introduce significant loss of accuracy.

As mentioned previously, $\Delta R_{\mathrm{b}(i)}$, $\Delta C_{\mathrm{b}(i)}$, and $\Delta D_{\mathrm{b}(i)}$ are approximated as Gaussian distributions and can be assumed to be fully correlated. According to (3.20) and (3.21), Δd_i can be approximated as a Gaussian distribution,

$$\Delta d_i \sim \mathcal{N}(0, \sigma^2_{d_i^{\mathrm{D2D}}} + \sigma^2_{d_i^{\mathrm{WID}}}), \tag{3.22}$$

$$\sigma^2_{d_i^{\mathrm{D2D}}} = \begin{cases} (\sigma_1+\sigma_2+\sigma_3)^2+\sigma_4^2 & \text{if buffers } i \text{ and } j \text{ are in different planes} \\ (\sigma_1+\sigma_2+\sigma_3+\sigma_4)^2 & \text{if buffers } i \text{ and } j \text{ are in the same plane} \end{cases}, \tag{3.23}$$

$$\sigma^2_{d_i^{\mathrm{WID}}} = (\sigma_5+\sigma_6+\sigma_7)^2+\sigma_8^2+2\mathrm{corr}(i,j)(\sigma_5+\sigma_6+\sigma_7)\sigma_8, \tag{3.24}$$

$$\sigma_1 = 0.69\overline{R}_{\mathrm{in}(i)}\sigma_{C_{\mathrm{b}(i)}^{\mathrm{D2D}}}, \quad \sigma_2 = 0.69\overline{C}_{\mathrm{l}(i)}\sigma_{R_{\mathrm{b}(i)}^{\mathrm{D2D}}}, \quad \sigma_3 = \sigma_{D_{\mathrm{b}(i)}^{\mathrm{D2D}}}, \quad \sigma_4 = 0.69\overline{R}_{\mathrm{b}(i)}\sigma_{C_{\mathrm{b}(j)}^{\mathrm{D2D}}},$$

$$\sigma_5 = 0.69\overline{R}_{\mathrm{in}(i)}\sigma_{C_{\mathrm{b}(i)}^{\mathrm{WID}}}, \quad \sigma_6 = 0.69\overline{C}_{\mathrm{l}(i)}\sigma_{R_{\mathrm{b}(i)}^{\mathrm{WID}}}, \quad \sigma_7 = \sigma_{D_{\mathrm{b}(i)}^{\mathrm{WID}}}, \quad \sigma_8 = 0.69\overline{R}_{\mathrm{b}(i)}\sigma_{C_{\mathrm{b}(j)}^{\mathrm{WID}}}.$$

The correlation between buffers i and j is denoted by $\mathrm{corr}(i,j)$, the model of which has been discussed in Section 3.2.2.

Consequently, for a 3-D clock path to a sink u which includes n_u clock buffers, the variation of the delay is expressed as the summation of (3.20) applied to each buffer along the path. The variance of the distribution of a 3-D clock path is a Gaussian distribution consisting of the WID and D2D variations of the buffers,

$$\Delta D_u = \sum_{i=1}^{n_u} \Delta d_i, \tag{3.25}$$

$$\Delta D_u \sim \mathcal{N}(0, \sigma^2_{D_u^{\mathrm{D2D}}} + \sigma^2_{D_u^{\mathrm{WID}}}). \tag{3.26}$$

The D2D and WID sources of delay variation along a 3-D clock path are, respectively, discussed in the following paragraphs.

D2D Variation Model for the Delay of 3-D Clock Paths

The variation of the delay of 3-D clock paths due to the D2D process variations is the sum of the D2D variations of the buffer delay in all the planes,

$$\Delta D_u^{\mathrm{D2D}} = \sum_{j=1}^{N_\mathrm{p}} \Delta D_{u(j)}^{\mathrm{D2D}}, \tag{3.27}$$

$$\Delta D_{u(j)}^{\mathrm{D2D}} = \sum_{i=1}^{n_{u(j)}} \Delta D_{u(j,i)}^{\mathrm{D2D}}, \tag{3.28}$$

where N_p is the number of the planes that the clock tree spans. $\Delta D^{D2D}_{u(j)}$ is the variation of the delay of the clock path from the clock source to sink u in plane j. The number of buffers located in plane j along this clock path is denoted by $n_{u(j)}$. The variation of the delay related to the i^{th} buffer in plane j is denoted by $\Delta D_{u(j,i)}$.

Since the D2D variations affect the buffers in the same plane equally, according to (3.22), (3.23), and (3.27), the distribution of $\Delta D^{D2D}_{u(j)}$ is a Gaussian distribution. The D2D variations affect the buffers in different planes independently and, therefore, $\Delta D^{D2D}_{u(j)}$ is independent from $\Delta D^{D2D}_{u(k)}$ for any $j \neq k$. Consequently, according to (3.27), the distribution of ΔD^{D2D}_{u} is also a Gaussian distribution,

$$\Delta D^{D2D}_{u} \sim \mathcal{N}(0, \sigma^2_{\Delta D^{D2D}_{u}}), \tag{3.29}$$

$$\sigma^2_{D^{D2D}_{u}} = \sum_{j=1}^{N_p} \sigma^2_{D^{D2D}_{u(j)}} = \sum_{j=1}^{N_p} (\sum_{i=1}^{n_{u(j)}} \sigma_{D^{D2D}_{u(j,i)}})^2. \tag{3.30}$$

WID variation model for the delay of 3-D clock paths

The delay of a 3-D clock path affected by WID variations is the sum of WID variations of all the buffers along this path. Consequently, according to (3.24), the distribution of ΔD^{WID}_{u} is also a Gaussian distribution. The resulting variance of the delay of sink u due to WID variations is

$$\Delta D^{WID}_{u} \sim \mathcal{N}(0, \sigma^2_{D^{WID}_{u}}), \tag{3.31}$$

$$\sigma^2_{D^{WID}_{u}} = \sum_{i=1}^{n_u} \sigma^2_{d_i^{WID}} + 2 \sum_{1 \leq i < j \leq n_u} \mathrm{corr}(i,j) \sigma_{d_i^{WID}} \sigma_{d_j^{WID}}, \tag{3.32}$$

where $\mathrm{corr}(i,j)$ is the correlation between the WID variations of buffers i and j. If buffers i and j are located in different planes, $\mathrm{corr}(i,j) = 0$. The spatial correlation of WID variations of different buffers within the same plane has been discussed in Section 3.2.2.

The clock skew between any pair of sinks in a 3-D clock tree is the difference of the clock delay between these sinks. For a 3-D clock tree with n_{sink} sinks distributed in N_p planes, the nominal value and the variation of clock skew $s_{u,v}$ between sinks u and v, respectively, are

$$\overline{s}_{u,v} = \overline{D}_u - \overline{D}_v, \tag{3.33}$$

$$\Delta s_{u,v} = \Delta s^{WID}_{u,v} + \Delta s^{D2D}_{u,v} = \Delta D^{WID}_{u} - \Delta D^{WID}_{v} + \Delta D^{D2D}_{u} - \Delta D^{D2D}_{v}. \tag{3.34}$$

The mean of $\Delta s_{u,v}$ is $E(\Delta s_{u,v}) = E(\Delta s^{WID}_{u,v}) - E(\Delta s^{D2D}_{u,v}) = 0$. The WID part $\left(\Delta D^{WID}_{u} - \Delta D^{WID}_{v}\right)$ and D2D part $\left(\Delta D^{D2D}_{u} - \Delta D^{D2D}_{v}\right)$ are independent from each other. Consequently, $\Delta s^{D2D}_{u,v}$ and $\Delta s^{WID}_{u,v}$ are discussed separately in the following subsections.

Skew model of 3-D clock trees with D2D variations

The correlation between each pair of terms in the expression of $\Delta s_{u,v}^{\mathrm{D2D}}$ can be one or zero (*i.e.*, fully correlated or uncorrelated, respectively). According to (3.27), $\Delta s_{u,v}^{\mathrm{D2D}}$ can be written as the sum of the terms in different planes,

$$\Delta s_{u,v}^{\mathrm{D2D}} = \sum_{j=1}^{N_{\mathrm{p}}} \Delta s_{(u,v)_j}^{\mathrm{D2D}} \tag{3.35}$$

$$\Delta s_{(u,v)_j}^{\mathrm{D2D}} = \sum_{i=1}^{n_{u(j)}} \Delta D_{u(j,i)}^{\mathrm{D2D}} - \sum_{i=1}^{n_{v(j)}} \Delta D_{v(j,i)}^{\mathrm{D2D}}, \tag{3.36}$$

where $\Delta D_{u(j,i)}^{\mathrm{D2D}}$ is the D2D delay variation related to the i^{th} buffer in the j^{th} plane along the clock path ending at sink u. The number of buffers in the j^{th} plane along this path is denoted as $n_{u(j)}$.

All the buffers in the same plane are equally affected by the D2D variations, which means that the correlation between each pair of variables in (3.36) is one. Since $\Delta D_{u(j,i)}^{\mathrm{D2D}}$ and $\Delta D_{v(j,i)}^{\mathrm{D2D}}$ are both modeled as Gaussian distributions, $\Delta s_{(u,v)_j}^{\mathrm{D2D}}$ is also a Gaussian distribution. In (3.36), $\forall\, j_1 \neq j_2 (1 \leq j_1, j_2 \leq N_{\mathrm{p}})$, $\Delta s_{(u,v)_{j_1}}^{\mathrm{D2D}}$ is independent from $\Delta s_{(u,v)_{j_2}}^{\mathrm{D2D}}$. Consequently, $\Delta s_{u,v}^{\mathrm{D2D}}$ is also described by a Gaussian distribution,

$$\Delta s_{u,v}^{\mathrm{D2D}} \sim \mathcal{N}(0, \sigma^2_{s_{u,v}^{\mathrm{D2D}}}), \tag{3.37}$$

$$\sigma^2_{s_{u,v}^{\mathrm{D2D}}} = \sum_{j=1}^{N_{\mathrm{p}}} \sigma^2_{s_{u,v(j)}^{\mathrm{D2D}}} = \sum_{j=1}^{N_{\mathrm{p}}} \Big(\sum_{i=1}^{n_{u(j)}} \sigma_{D_{u(j,i)}^{\mathrm{D2D}}} - \sum_{i=1}^{n_{v(j)}} \sigma_{D_{v(j,i)}^{\mathrm{D2D}}} \Big)^2. \tag{3.38}$$

Skew model of 3-D clock trees with WID variations

According to (3.32), the distribution of $\Delta s_{u,v}^{\mathrm{WID}}$ is also a Gaussian distribution,

$$\Delta s_{u,v}^{\mathrm{WID}} \sim \mathcal{N}(0, \sigma^2_{s_{u,v}^{\mathrm{WID}}}), \tag{3.39}$$

$$\begin{aligned} \sigma^2_{s_{u,v}^{\mathrm{WID}}} = & \sum_{i=n_{u,v}+1}^{n_u} \sigma^2_{D_{u(i)}^{\mathrm{WID}}} + \sum_{j=n_{u,v}+1}^{n_v} \sigma^2_{D_{v(j)}^{\mathrm{WID}}} + 2 \sum_{\substack{i,j=n_{u,v}+1 \\ i<j}}^{n_u} \mathrm{corr}(i,j) \sigma_{D_{u(i)}^{\mathrm{WID}}} \sigma_{D_{u(j)}^{\mathrm{WID}}} \\ & + 2 \sum_{\substack{i,j=n_{u,v}+1 \\ i<j}}^{n_v} \mathrm{corr}(i,j) \sigma_{D_{v(i)}^{\mathrm{WID}}} \sigma_{D_{v(j)}^{\mathrm{WID}}} - 2 \sum_{\substack{n_{u,v}+1 \leq i \leq n_u \\ n_{u,v}+1 \leq j \leq n_v}} \mathrm{corr}(i,j) \sigma_{D_{u(i)}^{\mathrm{WID}}} \sigma_{D_{v(j)}^{\mathrm{WID}}}, \end{aligned} \tag{3.40}$$

where $n_{u,v}$ is the number of the buffers shared by the clock paths ending at sinks u and v, as depicted in Fig. 3.11. After buffer $n_{u,v}$, the sub-paths to u and v do not share any buffer. The correlation between the variation of buffers has been introduced in Section 3.2.2.

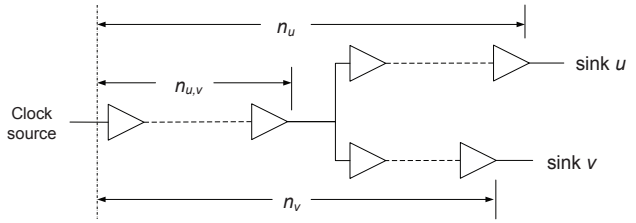


Figure 3.11: The clock paths to sinks u and v where the paths share $n_{u,v}$ buffers.

According to (3.34) through (3.40), the variation of the clock skew $\Delta s_{u,v}$ between sinks u and v in a 3-D clock tree is modeled as a Gaussian distribution,

$$\Delta s_{u,v} \sim \mathcal{N}(0, \sigma^2_{s^{\mathrm{WID}}_{u,v}} + \sigma^2_{s^{\mathrm{D2D}}_{u,v}}). \tag{3.41}$$

If the maximum tolerant skew variation is $\Delta S \geq 0$, the probability that a 3-D clock tree satisfies this constraint is

$$P(|s_{u,v}| \leq \Delta S) = \int_{-\Delta S - \bar{s}_{u,v}}^{\Delta S - \bar{s}_{u,v}} f_{\Delta s_{u,v}}(t)\, dt, \tag{3.42}$$

$$f_{\Delta s_{u,v}}(t) = \frac{1}{\sqrt{2\pi\sigma^2_{s_{u,v}}}} e^{-t^2/(2\sigma^2_{s_{u,v}})}. \tag{3.43}$$

The model of skew variations is used to analyze the effect of process variations in various 3-D clock trees. This model can be extended to include the variations of horizontal interconnects, as analyzed in Section 3.3.5. The investigated 3-D clock distribution networks and simulation results are presented in the following section.

3.3.4 Accuracy of the proposed model

The skew variation model is compared with Monte-Carlo simulations in this section. The structure used for this purpose is an H-tree clock distribution network. This H-tree is placed in a circuit with total area 10 mm × 10 mm.

The circuit is assumed to be implemented at a 45 nm CMOS technology. The parameters of the transistors and the interconnects are extracted from the PTM 45 nm CMOS and global interconnect models [41] and the ITRS reports [43]. The clock buffers consist of two inverters connected in series. The circuit parameters used in the following sections are listed in Table 3.1. The ratio of the width to the channel length is denoted by $W_{\mathrm{n}}/L_{\mathrm{n}}$ and $W_{\mathrm{p}}/L_{\mathrm{p}}$ for NMOS and PMOS, respectively. The interconnect resistance and capacitance per unit length are denoted by r_{int} and c_{int}, respectively. The physical and electrical characteristics of TSVs are also listed in Table 3.1 and are based on the data reported in [68, 71]. The diameter and length of the TSVs are notated as $\varnothing_{\mathrm{TSV}}$ and l_{TSV}, respectively.

Table 3.1: Device and Interconnect Parameters of the Investigated Circuit.

Parameter	W_n/L_n	W_p/L_p	V_{dd} [V]	R_b [Ω]	C_b [fF]	D_b [ps]
Value	30	60	1.0	349.0	5.7	24.8
Parameter	r_{int} [Ω/mm]	c_{int} [fF/mm]	$\varnothing_{TSV}$ [μm]	l_{TSV} [μm]	R_{TSV} [mΩ]	C_{TSV} [fF]
Value	51.2	230.2	2	20	133	52

Table 3.2: Variations of the Electrical Characteristics of the Buffers.

Input Slew	R_b [Ω]			C_b [fF]			D_b [ps]		
	μ	σ_{WID}	σ_{D2D}	μ	σ_{WID}	σ_{D2D}	μ	σ_{WID}	σ_{D2D}
47 [mV/ps]	371	18.8	15.3	4.9	0.04	0.03	19.9	1.04	0.85
	σ/μ	5.1%	4.1%	σ/μ	0.8%	0.7%	σ/μ	5.2%	4.3%
16 [mV/ps]	349	17.8	14.7	5.7	0.31	0.16	24.8	1.49	1.21
	σ/μ	5.1%	4.2%	σ/μ	2.3%	2.1%	σ/μ	6.0%	4.9%
6 [mV/ps]	345	16.7	13.7	7.2	0.08	0.06	30.1	2.19	1.79
	σ/μ	4.8%	4.0%	σ/μ	1.1%	0.9%	σ/μ	7.3%	5.9%

The variation of the effective channel length of transistors, L_{eff}, is considered in this section, which has been identified as the most significant component of device variations [21,37,91,92]. Note that the effect of other sources of process variations can also be determined by the circuit illustrated in Fig. 3.8 and described with the proposed model. The corresponding nominal L_{eff}, D2D variation ($3\sigma_{L_{eff}}^{D2D}$), and WID variation ($3\sigma_{L_{eff}}^{WID}$) are 27 nm, 2.2 nm, and 2.7 nm, respectively [43]. Cadence Spectre is used for the Monte-Carlo simulations [148]. The resulting variations of R_b, C_b, and D_b are listed in Table 3.2, which are obtained based on different input transition times (1.7%, 5.0%, and 13.3% of the clock period). The corresponding input slew rates are 47 mV/ps, 16 mV/ps, and 6 mV/ps, respectively. The mean value and standard deviation are denoted by μ and σ, respectively. The ratio σ/μ usually indicates the importance of variations [91]. The Monte-Carlo simulation is repeated 1500 times. As reported in Table 3.2, the σ of R_b, C_b, and D_b also depends on the input slew rate. To consider the slew rate and the load is, therefore, necessary while evaluating the variations of the buffer delay. The method presented in Section 3.3.2 is applicable to accurately consider this dependence.

Two H-tree topologies are used to verify the accuracy of the skew variation model. The first topology (multi-via) is illustrated in Fig. 3.7. The second topology (single-via) is illustrated in Fig. 3.13. Both these topologies are discussed in the following section. The H-tree spans four planes. The clock source is located at the center of the first plane. There are 128 clock sinks in total, 32 in each plane. Clock buffers, which are marked with $\triangle$, are inserted following the technique described in [77]. The clock frequency is 1 GHz and the constraint on the input slew rate is 16 mV/ps (the transition time is 5% of the clock period). The numbers of the inserted buffers in the multi-via and single-via topologies are 168 and 540, respectively. Only few of these buffers are illustrated in Figs. 3.7 and 3.13 for improved readability. The wire segments between two buffers are simulated using a standard π model.

Table 3.3: σ of Skew Variation of the 3-D Circuits Shown in Figs. 3.7 and 3.13.

Correlation		Independent				Multi-Level				CPU
Skew variation		$\sigma_{s_{1,2}}$	$\sigma_{s_{1,3}}$	$\sigma_{s_{1,4}}$	$\sigma_{s_{1,5}}$	$\sigma_{s_{1,2}}$	$\sigma_{s_{1,3}}$	$\sigma_{s_{1,4}}$	$\sigma_{s_{1,5}}$	time
Multi-via	Model [ps]	7.0	14.8	7.4	15.0	7.0	33.6	7.1	32.9	26 sec.
	Spectre [ps]	7.3	15.7	7.6	16.0	7.4	34.9	7.3	35.1	48 min.
	Error [%]	-4	-6	-2	-6	-5	-4	-3	-6	-
Single-via	Model [ps]	3.9	13.6	51.3	51.3	2.4	29.0	71.1	69.6	39 sec.
	Spectre [ps]	3.8	13.1	50.1	50.1	2.3	28.1	68.2	67.9	55 min.
	Error [%]	2	3	2	2	4	3	4	3	-

As shown in Fig. 3.7(b), sinks 1, 2, and 3 are located in the first plane. Sinks 4 and 5 are located in the topmost plane. Skews $s_{1,2}$, $s_{1,3}$, $s_{1,4}$, and $s_{1,5}$ are considered to demonstrate the accuracy of the developed model. The difference between the resulting standard deviation σ produced by Spectre simulations and the skew variation model is reported in Table 3.3. The skew variations with uncorrelated (independent) WID variations are reported as "Independent". The variations modeled by the multi-level spatial correlation are reported as "Multi-Level", where five levels are assumed ($l = 5$). The error of the skew variation model is below 6% between any pair of sinks in the investigated clock tree. As listed in Table 3.3, the distribution of the clock skew determined by the skew variation model exhibits reasonable accuracy as compared with Monte-Carlo simulations. The *cumulative distribution functions* (CDF) from Spectre and the proposed skew variation model for the 3-D tree with independent WID variations are shown in Fig. 3.12.

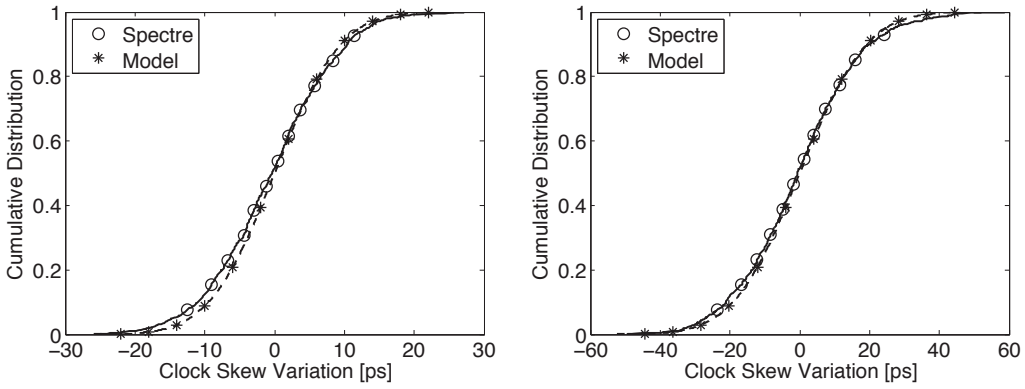


Figure 3.12: Comparison of skew variation between Spectre simulations and the analytic skew model, where (a) and (b) are the CDFs of $\Delta s_{1,4}$ and $\Delta s_{1,5}$, respectively.

The computational time for the proposed model and the SPICE-based Monte-Carlo simulations is also listed in Table 3.3. The proposed statistical model is implemented in Matlab and run on a PC with Intel i5 M540 CPU@2.53 GHz, 4 GB DDR2 memory, and 64 bit Windows 7 system. The Monte-Carlo simulations are run with Cadence Virtuoso 6.1.4 on a server with

Intel Xeon X5650 CPU@2.67GHz, 12 GB DDR3 memory, and 64 bit Scientific Linux 6.0 system. The run time is reported as the average time for independent and spatially correlated WID variations. To decrease the run time, only the clock paths related to the reported σ are simulated. Although only a part of the 3-D clock trees is simulated, using the proposed model, the run time is reduced by 85×. The efficiency of the variability-aware design of 3-D clock distribution networks can significantly be improved by the proposed model. When comparing different topologies of clock trees in terms of skew variation, the proposed model helps to fast select the best solution. When iterative design modifications to the clock tree are required to reduce skew variation, the proposed model also helps to decrease the iteration time. A comparison of the run time for the entire 3-D clock distribution networks is provided in Section 3.4.2.

3.3.5 Extension of the model to include interconnect variations

The proposed model can be extended to include the variations of interconnects. Consider the 3-D clock tree shown in Fig. 3.10, where the delay variation of a buffer stage $\Delta d_{\text{stage}(i)}$ includes the variation of the capacitance ΔC_{int} and resistance ΔR_{int} of the wires,

$$\begin{aligned}\Delta d_{\text{stage}(i)} = \Delta d_i + 0.69(\overline{R}_{\text{b}(i)} + \Delta R_{\text{b}(i)})\Delta C_{\text{int}} + 0.38(\overline{R}_{\text{int}}\Delta C_{\text{int}} + \Delta R_{\text{int}}\overline{C}_{\text{int}} + \Delta R_{\text{int}}\Delta C_{\text{int}}) \\ + 0.69(\overline{R}_{\text{int}}\Delta C_{\text{b}(i+1)} + \Delta R_{\text{int}}\overline{C}_{\text{b}(i+1)} + \Delta R_{\text{int}}\Delta C_{\text{b}(i+1)}).\end{aligned} \tag{3.44}$$

According to the definition of Δd_i in (3.20), the term $0.69\overline{R}_{\text{int}}\Delta C_{\text{b}(i+1)}$ is included in Δd_{i+1}. Consequently, $\Delta d_{\text{stage}(i)}$ is rewritten as

$$\begin{aligned}\Delta d_{\text{stage}(i)} = \Delta d_i + 0.69(\overline{R}_{\text{b}(i)} + \Delta R_{\text{b}(i)})\Delta C_{\text{int}} + 0.38(\overline{R}_{\text{int}}\Delta C_{\text{int}} + \Delta R_{\text{int}}\overline{C}_{\text{int}} + \Delta R_{\text{int}}\Delta C_{\text{int}}) \\ + 0.69(\Delta R_{\text{int}}\overline{C}_{\text{b}(i+1)} + \Delta R_{\text{int}}\Delta C_{\text{b}(i+1)}) = \Delta d_i + \Delta d_{\text{int}(i)},\end{aligned} \tag{3.45}$$

where the delay variation due to the wires is denoted by $\Delta d_{\text{int}(i)}$.

As discussed in [86], since the variation of the characteristics of metal wires is relatively low as compared with the nominal value, the variation of the wire delay can be approximated by the first order Taylor series expansion without significant loss of accuracy. Similar to expression (3.21), $\Delta d_{\text{int}(i)}$ can be approximated as

$$\Delta d_{\text{int}(i)} \approx \sum_{p_j \in \vec{P}} \left(\left[\frac{\partial \Delta d_{\text{int}(i)}}{\partial \Delta R_{\text{int}}} \frac{\partial \Delta R_{\text{int}}}{\partial \Delta p_j} \right]_0 \Delta p_j + \left[\frac{\partial \Delta d_{\text{int}(i)}}{\partial \Delta C_{\text{int}}} \frac{\partial \Delta C_{\text{int}}}{\partial \Delta p_j} \right]_0 \Delta p_j \right), \tag{3.46}$$

where p_j is the j^{th} parameter of the wire and $\vec{P}$ is the vector of parameters of wires affected by process variations. For example, consider the variation of the width and the thickness of the metal and the thickness of ILD [86, 92], $\vec{P} = (W_{\text{m}}, t_{\text{m}}, t_{\text{ILD}})$. Assuming these parameters are modeled by Gaussian distributions and independent from each other [86], the distribution of $\Delta d_{\text{int}(i)}$ can be approximated by a Gaussian distribution,

$$\Delta d_{\text{int}(i)} \sim \mathcal{N}(0, \sigma^2_{d_{\text{int}(i)}}), \tag{3.47}$$

$$\sigma^2_{d_{\text{int}(i)}} = \sum_{p_j \in \vec{P}} \left(\left[\frac{\partial \Delta d_{\text{int}(i)}}{\partial \Delta R_{\text{int}}} \frac{\partial \Delta R_{\text{int}}}{\partial \Delta p_j} \right]_0 + \left[\frac{\partial \Delta d_{\text{int}(i)}}{\partial \Delta C_{\text{int}}} \frac{\partial \Delta C_{\text{int}}}{\partial \Delta p_j} \right]_0 \right)^2 \sigma^2_{p_j}, \tag{3.48}$$

$$R_{\text{int}} = \frac{\rho l}{t_{\text{m}} W_{\text{m}}}, \tag{3.49}$$

$$C_{\text{int}} = 2(C_{\text{g}} + C_{\text{c}}) l, \tag{3.50}$$

where C_{g} includes the ground and fringe capacitances and C_{c} is the coupling capacitance. The expressions of C_{g} and C_{c} are obtained from [41].

Considering the delay variation caused by both the clock buffers and wires in (3.45), the skew variation $\Delta s_{u,v}$ includes two terms, $\Delta s_{u,v} = \Delta s_{\text{b}(u,v)} + \Delta s_{\text{int}(u,v)}$. The distribution of $\Delta s_{\text{b}(u,v)}$ is obtained through (3.20) to (3.41). The distribution of $\Delta s_{\text{int}(u,v)}$ can be obtained through (3.25) to (3.41) by substituting $\Delta d_{\text{int}(i)}$ for Δd_i. Consequently, $\Delta s_{u,v}$ can be described by a Gaussian distribution,

$$\Delta s_{u,v} \sim \mathcal{N}(0, \sigma^2_{s_{\text{b}(u,v)}} + \sigma^2_{s_{\text{int}(u,v)}}). \tag{3.51}$$

The extended model is compared with Monte-Carlo simulations including the variations of r_{int} and c_{int} in the π model of interconnects. Based on the parameters used in [86], the nominal value and standard deviation of the parameters of wires are listed in Table 3.4. Two types of 3-D clock trees, multi-via and single-via trees, are used to verify the accuracy of the extended model. The topologies of these clock trees will be introduced in the following section. The results for the independent WID variations are reported in Table 3.5. As reported in this table, the accuracy of the model including the variation of wires is reasonably high.

Table 3.4: Parameters of Horizontal Interconnects.

Parameters	W_{m} [nm]	t_{m} [nm]	t_{ILD} [nm]
Nominal	430	1000	160
$3\sigma_{\text{D2D}}$	43	50	12
$3\sigma_{\text{WID}}$	21.5	25	6

Table 3.5: Skew Variation of the 3-D Circuits Considering Wire Variations.

Topology	multi-via				single-via			
Skew variation	$\sigma_{s_{1,2}}$	$\sigma_{s_{1,3}}$	$\sigma_{s_{1,4}}$	$\sigma_{s_{1,5}}$	$\sigma_{s_{1,2}}$	$\sigma_{s_{1,3}}$	$\sigma_{s_{1,4}}$	$\sigma_{s_{1,5}}$
Model [ps]	7.01	15.09	7.45	15.3	3.99	13.94	56.46	56.46
Spectre [ps]	7.19	16.44	7.64	16.55	4.00	13.77	56.38	56.3
Error [%]	-3	-8	-2	-8	0	1	0	0

3.4 Process Variations Tolerant 3-D Clock Distribution Networks

The variation of process-induced skew in different 3-D clock trees is investigated in this section. The skew variation is first compared among the conventional 3-D clock trees in Section 3.4.1. Combining the advantages of the conventional clock trees, a novel multi-group topology for 3-D clock trees is proposed in Section 3.4.2. Another type of low-skew 3-D clock distribution networks, clock grids, are presented and compared with 3-D clock trees in Section 3.4.3. The skew variation in 3-D ICs with multiple clock domains is investigated in Section 3.4.4.

3.4.1 Skew variation of conventional 3-D clock trees

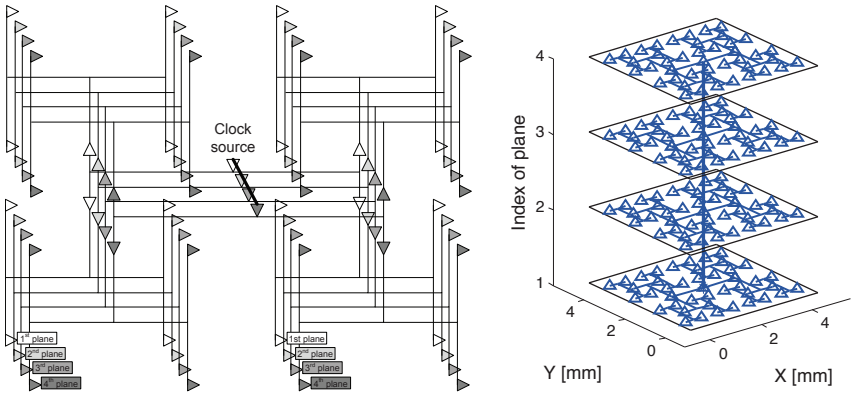
The skew variation for two types of conventional 3-D global H-trees is investigated in this subsection. Both of these networks have been utilized in the design of a prototype 3-D circuit [81] and other case studies of multiplane circuits [149]. H-trees are typically used to globally deliver the clock signal to large scale circuits [75]. The regularity of these topologies facilitates the investigation of WID and D2D variations as compared to synthesized clock trees which exhibit significantly different wire length and TSV density characteristics. In other words, the main objective is to demonstrate the physical behavior of a 3-D system under these variations and the related tradeoffs, rather than the decrease in the wire length of a clock tree produced by an efficient 3-D clock tree synthesis technique. A number of local clock networks, such as local meshes, clock trees [150], and rings can be used to distribute the clock signal in the vicinity of each leaf of the H-tree. Although H-trees are considered, the analysis also applies to other global clock architectures, such as X-trees.

The first topology (multi-via topology) has been shown in Fig. 3.7, where the clock source and buffers (except for the buffers at the last level) of a 3-D H-tree are located in a single physical plane (*e.g.*, the first plane). In this topology, the clock signal is propagated to the sinks in other planes by multiple TSVs. The vertical lines at each leaf correspond to a cluster of TSVs.

The second clock tree topology (single-via topology) is illustrated in Fig. 3.13, where a 2-D H-tree is replicated in each plane. The clock signal is propagated by a single-via (or a group of TSVs to prevent TSV failures and to lower the resistance of this vertical path) connecting the clock source to each H-tree replica.

Skew variations between the clock sinks in the same plane

In 3-D H-trees and for intra-plane paths, the number, size, and location of the buffers along these paths are equal for a single plane, since the multi-via and single-via topologies are both symmetric topologies (at least within the x and y directions). The D2D variations in each plane, therefore, affect these clock paths equally. Consequently, according to (3.36), for both the multi-via and single-via topologies, only WID variations affect the variation of skew



(a) (b)

Figure 3.13: A single-via 3-D clock H-tree, where 2-D view (a) and 3-D view (b) are illustrated.

between sinks located in the same plane. For both topologies, the variation of skew between the buffers located in the same plane exhibits the same behavior as in 2-D circuits.

For the considered topologies, the clock buffers are inserted by uniform buffer insertion techniques under the same constraints of skew and slew rate [77]. R_{in} and C_{l} of each buffer, therefore, are approximately equal. For a 3-D clock tree, as described by (3.40), if R_{in} and C_{l} of each buffer remain unchanged, the $\sigma_{s_{(u,v)}^{\text{WID}}}$ between two sinks decreases as the number of the non-shared clock buffers (*e.g.*, the buffers after the $n_{u,v}$ buffers in Fig. 3.11) decreases. For a 3-D IC with total area A, the side length of each plane is $L \propto \sqrt{\frac{A}{N_{\text{p}}}}$. Consequently, the number of buffers in one plane decreases as L decreases for an increasing number of planes forming the 3-D circuit.

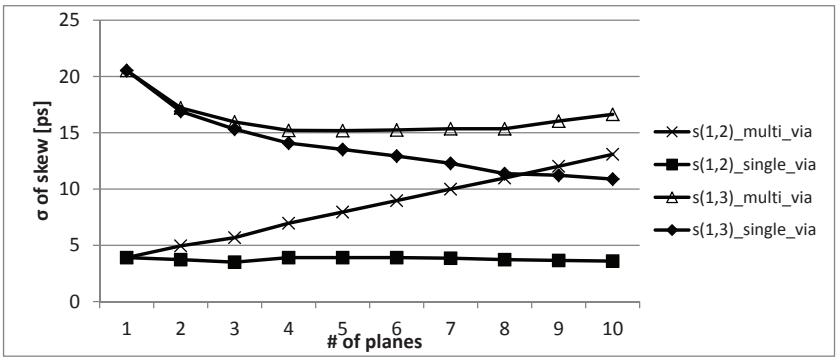
For the single-via topology, all the clock sinks within a plane are connected to the clock source by the same TSV. The length of this TSV and the increasing number of buffers vertically connected to this TSV do not affect the intra-plane skew. Consequently, based on the proposed model and the above analysis, it is concluded that

Observation 3.1. *For the single-via topology, the distribution of the skew between the clock sinks in the same plane becomes narrower as the number of planes increases.*

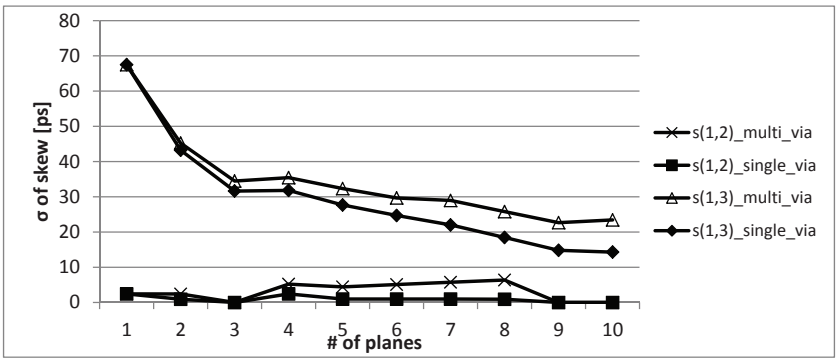
For the multi-via topology, however, the clock sinks in the same plane connect to different TSVs. As the number of planes increases, both the number of buffers connecting to a TSV and the length of the TSVs increase. The input slew rate decreases since an increasing load is driven. As reported in Table 3.2, the resulting delay variation of the buffers after the TSVs increases. Moreover, the load of the buffers driving the TSVs increases. These changes of the topological characteristics result in the increase of $\sigma_{d(i)}$, as described by (3.24). This increase, consequently, counteracts and can surmount the decrease in variations due to the decreasing number of clock buffers along the clock paths.

Observation 3.2. *For the multi-via topology, the distribution of the skew between the clock sinks in the same plane changes non-monotonically as the number of planes increases.*

Example 3.1 : Simulation results exhibiting the different behavior of the single-via and multi-via topologies are shown in Fig. 3.14. In this example, a global clock tree with 256 sinks is placed in a 3-D IC with increasing number of planes. A sink can be either a sub-tree, a local clock mesh, or a cluster of registers (or a buffer driving any of these structures). The total area of the circuit is 100 mm^2. The impact of process variations on the skew between pairs of sinks within the same plane is demonstrated by skews $s_{1,2}$ and $s_{1,3}$. The physical location of these sinks is illustrated in Fig. 3.7(b). The results of the simulations with the independent and multi-level correlated WID variations ($l = 5$) are illustrated in Figs. 3.14(a) and 3.14(b), respectively.



(a)



(b)

Figure 3.14: σ of skew within the first plane for increasing number of planes, where the WID variations are considered (a) independent and (b) multi-level correlated.

The buffers inserted into the 3-D clock trees are reported in Table 3.6. The number of buffers inserted within one plane in the single-via topology is lower than the multi-via topology, which introduces a lower skew variation than the multi-via topology. The total number of buffers in a single-via topology is, however, much higher than the multi-via topology.

Table 3.6: The Number of Buffers Inserted into the 3-D Clock Trees.

# of planes	1	2	3	4	5	6	7	8	9	10
multi-via	981	588	558	296	264	242	234	138	134	134
single-via (per plane)	981	460	430	231	199	177	169	105	101	101
single-via (total)	981	920	1290	924	995	1062	1183	840	909	1010

Observation 3.3. *The variance of the skew between two intra-plane sinks of the single-via 3-D H-tree is smaller than the corresponding variance of the skew in a multi-via 3-D H-tree.*

The numbers of buffers in each plane for both topologies decrease as the number of planes increases. The increasing number of buffers connected to TSVs (due to the greater number of planes) increases $\sigma_{s_{u,v}}$ in the multi-via topology but does not affect $\sigma_{s_{u,v}}$ in the single-via topology. Consequently, the decrease in the number of buffers leads to a reduction in skew variation within the same plane for the single-via topology, as shown by the ♦ and ■ curves in Fig. 3.14. Nevertheless, for the multi-via topology, as shown by the △ and × curves, $\sigma_{s_{u,v}}$ within the same plane changes non-monotonically with the number of planes. For the sinks with short distance, the $\sigma_{s_{1,2}}$ even increases with the number of planes. As a result, for multi-via 3-D H-trees, simply increasing the number of planes does not necessarily improve skew variation. By employing the proposed skew variation model, the number of planes that produces the lowest skew variation is determined.

The maximum supported clock frequency $f_{\max}$ of a circuit is constrained by skew [75]. Although $f_{\max}$ is typically determined by the critical path delay, the skew criterion is used here to offer a tangible explanation of the effect of process variations on the performance of circuits. The maximum allowed skew is assumed to be $10\%\frac{1}{f_{\max}}$ for the simulated 3-D clock trees. To achieve a timing yield higher than 99%, $f_{\max}$ should be smaller than $10\%\frac{1}{3\sigma_{s_{u,v}}}$, where $3\sigma_{s_{u,v}}$ is the skew at the 3σ point from the mean value. Assuming that the clock frequency is limited by the largest skew variation, the $f_{\max}$ corresponding to $\sigma_{s_{1,3}}$ shown in Fig. 3.14, is illustrated in Fig. 3.15. The results with independent WID variations are illustrated by "multi-via (I)" and "single-via (I)". The results with multi-level WID correlations are illustrated by "multi-via (II)" and "single-via (II)".

As illustrated in Fig. 3.15, the single-via topology can produce an up to 53% and 64% higher clock frequency for independent and multi-level correlated WID variations, respectively, as compared to the multi-via topology. This improvement increases as the number of planes increases. The $f_{\max}$ in the multi-via topology changes non-monotonically with the number of planes.

Guideline 3.1. *In a 3-D circuit, if the data-related sinks are located mostly within the same plane, the single-via topology is more efficient in reducing the skew variations and can support a higher clock frequency.*

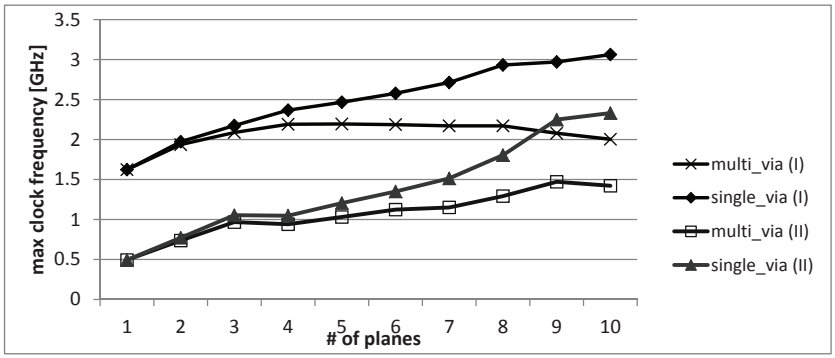
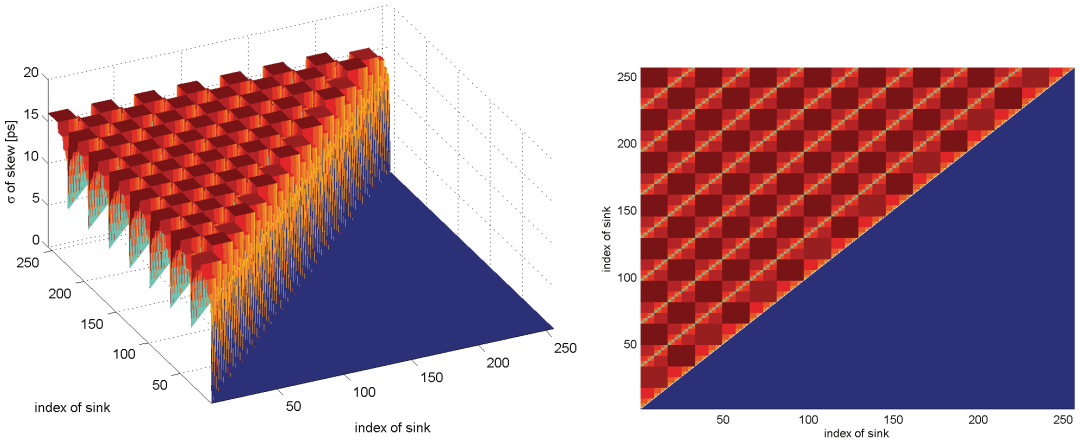


Figure 3.15: The maximum supported clock frequency determined by the skew variation within one plane.



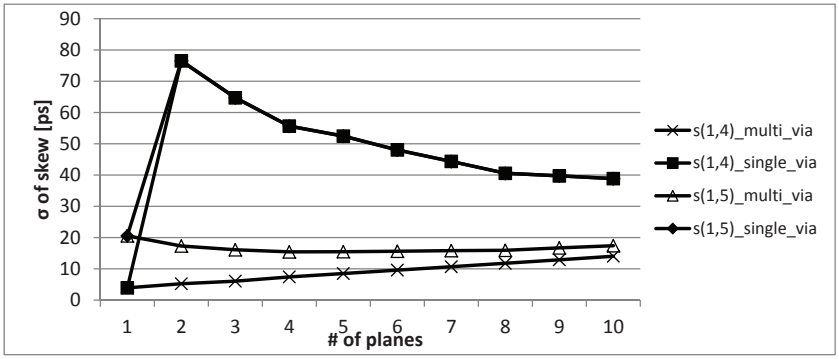
(a) (b)

Figure 3.16: The σ of skew between each pair of clock sinks under the multi-via topology where (a) is the 3-D view and (b) is the top view.

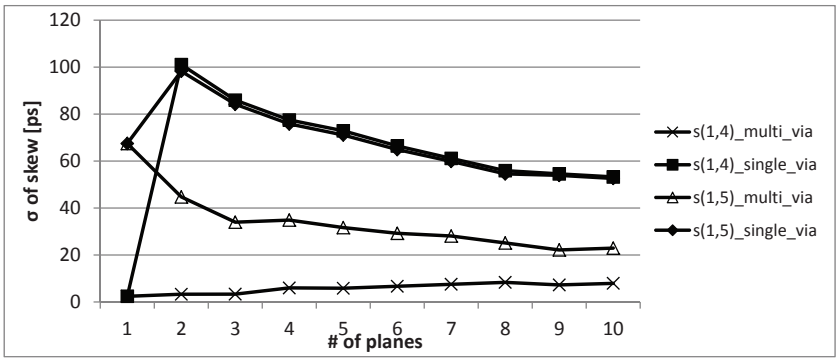
The $f_{\max}$ produced by a 3-D tree with independent WID variations, not surprisingly, is higher than a 3-D tree with multi-level correlated WID variations. According to (3.40), this situation is due to the larger spatial correlation between devices which introduce higher skew variations into a 3-D clock tree.

Skew variations between the clock sinks in different planes

As described by (3.36)-(3.38), when the investigated clock sinks are located in different planes, the corresponding clock skew is also affected by D2D variations. As a result, the skew variations between inter-plane sinks vary from the intra-plane skew variation. To demonstrate this difference, the $\sigma_{s_{u,v}}$ between each pair of sinks of a multi-via 3-D clock tree is illustrated in Fig. 3.16. Since $\sigma_{s_{u,v}} = \sigma_{s_{v,u}}$, only half of the skew array is shown in Fig. 3.16 for clarity.



(a)



(b)

Figure 3.17: σ of skew between the sinks in the first and the topmost plane for the single-via and multi-via topologies. The locations of the pairs of sinks defining $s_{1,4}$ and $s_{1,5}$ are shown in Fig. 3.7(a). (a) is based on independent WID variations and (b) is based on multi-level correlated WID variations.

Example 3.2 : In this example, a 3-D clock tree spanning eight planes is implemented using the single- and multi-via topologies. The resulting skew of the multi-via topology is illustrated in Fig 3.16. The electrical parameters of the wires are given in Section 3.3.4. There are 32 sinks in each plane and 256 sinks in total. Sinks 1 to 32 are located in the first plane and sinks 33 to 64 are located in the second plane, *etc.* As an example, consider the σ of the skew between sinks 3 and 4. This standard deviation is determined by the z value of the point $x = 3$ and $y = 4$. From these figures, the σ_s of inter-plane skew is larger than the σ_s of intra-plane skew. The change of skew variation between inter-plane sinks with the number of planes is illustrated in Fig. 3.17.

For the single-via topology, the skew variation for the inter-plane pairs of sinks remains approximately the same irrespective of the planes that the sinks belong. This behavior is because the paths to sinks located in different planes do not share any segment (see Fig. 3.13).

When the number of planes is greater than two, the skew variation decreases as the number of planes increases, as also shown in Fig. 3.17. Since the paths lay in different planes, according

Table 3.7: The Maximum Clock Frequency Supported by Multi-Via and Single-Via Topologies.

	# of planes	1	2	3	4	5	6	7	8	9	10
I	multi-via [GHz]	1.62	1.93	2.07	2.16	2.15	2.13	2.11	2.09	2.00	1.92
	single-via [GHz]	1.62	0.44	0.52	0.60	0.64	0.69	0.75	0.82	0.84	0.86
	multi / single	1	4.4	4.0	3.6	3.4	3.1	2.8	2.5	2.4	2.2
II	multi-via [GHz]	0.49	0.74	0.98	0.96	1.05	1.14	1.19	1.33	1.50	1.45
	single-via [GHz]	0.49	0.34	0.40	0.44	0.47	0.51	0.56	0.61	0.62	0.63
	multi / single	1.0	2.2	2.5	2.2	2.2	2.2	2.1	2.2	2.4	2.3

to (3.38), the effect of D2D variations on the 3-D single-via topology is much larger than in planar H-trees.

The skew variation under the multi-via topology varies significantly from the single-via topology, as illustrated in Fig. 3.17. The skew variation between planes significantly depends on the location of the related sinks. According to (3.36)-(3.38), the impact of D2D variations increases as the number of buffers located in different planes increases. For the multi-via topology, all clock paths preceding the TSVs are in the first plane. The effect of D2D variations on the multi-via topology, therefore, is much smaller than the single-via topology, as shown in Fig. 3.17. Nevertheless, as shown in Fig. 3.17, the skew variation of the multi-via topology changes non-monotonically with the number of planes. The reason is similar to the skew variation within the same plane as discussed previously.

Guideline 3.2. *In a 3-D circuit, if the data-related sinks are widely distributed in several planes, the multi-via topology is more efficient in reducing the skew variation and supports a higher clock frequency.*

Assuming the $f_{\max}$ of a 3-D IC is limited by the inter-plane skew variation, the maximum operating frequency supported by the single-via and multi-via topologies is reported in Table 3.7. The results based on independent and multi-level correlated WID variations are reported after "I" and "II", respectively. As listed in this table, for a circuit with a different number of planes and different clock tree topology, the $f_{\max}$ can vary from 440 MHz to 2.16 GHz. The corresponding largest $\sigma_{s_{u,v}}$ varies from 77 ps to 15 ps. For the same number of planes, the $f_{\max}$ of the multi-via topology is up to 4.4 times higher than the single-via topology. As reported by "I", the single-via topology produces lower $f_{\max}$ than a planar tree when the WID variation is completely random. The single-via topology, however, can produce a higher $f_{\max}$ as compared to a 2-D tree when the systematic WID variation is considered as reported by "II".

3-D integration is considered to significantly reduce the interconnect delay and enhance the clock frequency of circuits [7, 151]. This enhancement, however, is shown to not grow directly proportionally with the number of planes where process variations (both WID and D2D) are considered in the design process.

Results indicate that the performance improvement in a 3-D clock network depends significantly on the distribution of the sinks (and consequently the clock paths) among the planes. As reported in Table 3.7, when the data-related sinks are distributed in different planes, the skew of single-via 3-D clock trees is affected more by process variations than the corresponding 2-D clock trees. This behavior is consistent with the conclusions made in [88, 93]. The effect of process variations on 3-D clock distribution networks can be mitigated by employing a multi-via topology in this case. This topology can better exploit the traits of vertical integration (*i.e.*, shorter wires) to significantly increase the operating frequency.

3.4.2 A novel multi-group 3-D clock tree

As stated in Guidelines 3.1 and 3.2, the single-via 3-D clock H-tree topology is more efficient in reducing the skew variation within a single plane, while the multi-via topology is more efficient in reducing the skew variation between planes. To exploit these advantages, a hybrid H-tree topology (multi-group topology) combining the features of these topologies is proposed in this section.

The new multi-group topology is illustrated in Fig. 3.18. The key idea is that the N_p planes forming a 3-D circuit are divided in G groups of "data-related planes". The data-related planes are the physical planes containing data-related registers. The i^{th} group of data-related planes consists of $h_i (\leq N_p)$ physical planes. The clock signal is distributed within these h_i planes by a multi-via topology.

An example of this H-tree topology is illustrated in Fig. 3.18. This H-tree includes two groups of data-related planes ($G = 2$). Each group spans three ($h_1 = 3$) and two physical planes ($h_2 = 2$), respectively. The buffers contained in each group of data-related planes are denoted by $\triangle$ and $\circ$. The TSVs connecting these buffers are called "sink-TSVs". The roots of the multi-via topologies are connected with a "root-TSV" (or a cluster of TSVs) as illustrated by the segment at the center of the planes.

For a 3-D IC, if all the data-related clock sinks cannot be located within the same plane but in adjacent planes, the multi-group topology is more efficient in reducing the skew variation than the aforementioned topologies. Compared with the single-via topology, using G instead of N_p H-trees, the multi-group topology significantly reduces the skew variation between data-related planes. Compared with the multi-via topology, the buffers connected to the sink-TSVs for the multi-group topology are fewer than the buffers connected to the TSVs of the multi-via topology. Therefore, both the skew variation within a single plane and the skew variation between data-related planes are reduced.

Example 3.3 : A 3-D circuit with eight planes is simulated for the three topologies to investigate the efficiency of the multi-group 3-D clock tree. The physical and electrical characteristics of the circuit are reported in Section 3.3.4. Two variants of the multi-group topology are

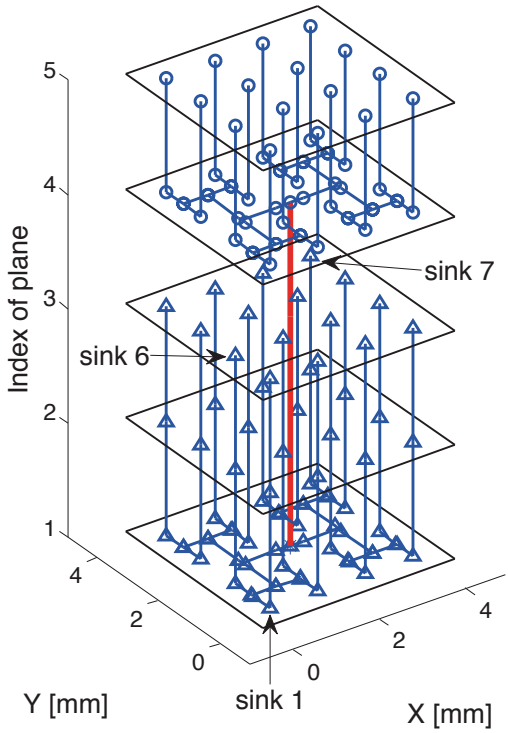


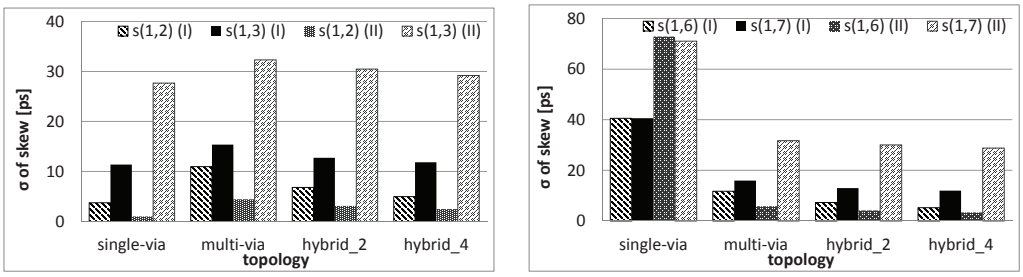
Figure 3.18: An example of the multi-group 3-D clock H-tree topology.

simulated, including two groups (hybrid_2, $G = 2$, $h_i = 4$) and four groups (hybrid_4, $G = 4$, $h_i = 2$) of data-related planes, respectively. Simulation results are shown in Fig. 3.19.

In Fig. 3.19, skews $s_{1,2}$, $s_{1,3}$ (defined in Fig. 3.7(a)), $s_{1,6}$, and $s_{1,7}$ (defined in Fig. 3.18) are depicted showing the skew variation between the nearest and the farthest sinks. The results based on independent and multi-level correlated WID variations are denoted by (I) and (II), respectively. The $\sigma_{s_{1,2}}$ and $\sigma_{s_{1,3}}$ produced by the multi-group topology are lower than the multi-via topology and decrease as the number of sub-H-trees increases. For the topology with four sub-H-trees (hybrid_4), $s_{1,2}$(I), $s_{1,3}$(I), $s_{1,2}$(II), and $s_{1,3}$(II) are reduced by 55%, 23%, 44%, and 10 % respectively, as compared with the multi-via topology.

Although the $\sigma_{s_{1,3}}$ within the same plane of the multi-group topology is still greater (4% for hybrid_4) than the single-via topology, the inter-plane skews $\sigma_{s_{1,6}}$ and $\sigma_{s_{1,7}}$ within a group of data-related planes of the multi-group topology are significantly reduced as shown in Fig. 3.19(b). This reduction is also greater than the multi-via topology. The number of sub-H-trees within a multi-group 3-D topology can be determined by the distribution of the data-related sinks.

Guideline 3.3. *When the data-related sinks are located in adjacent planes of a 3-D circuit, the multi-group 3-D clock tree topology is more efficient in reducing the skew variation than both the single- and multi-via topologies.*



(a) (b)

Figure 3.19: σ of skew for three 3-D clock tree topologies. (a) Intra-plane skews $s_{1,2}$ and $s_{1,3}$. (b) Inter-plane skews $s_{1,6}$ and $s_{1,7}$ within a group of data-related planes.

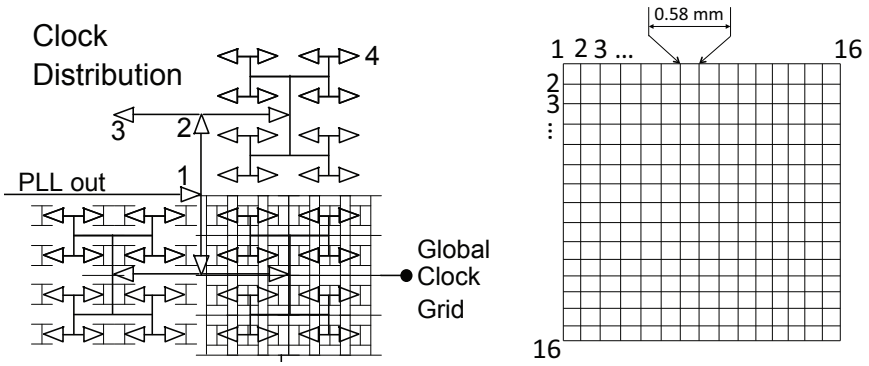
Table 3.8: $\sigma_{s_{1,7}}$ and Computational Time of three 3-D Clock Tree Topologies.

Topology		single-via	multi-via	hybrid_2	hybrid_4
$\sigma_{1,7}$	Model [ps]	40.6	15.9	13.0	11.9
	Spectre [ps]	41.6	16.8	13.7	12.3
	Error [%]	-2	-5	-5	-3
CPU time	Model [min.]	29	28	25	27
	Spectre [h.]	500	173	221	265
	Spec./Model	1034	364	535	582

Furthermore, as illustrated in Fig. 3.19, for the sinks with a short horizontal distance in a multi-group topology, the multi-level correlated WID variations (denoted by (II)) introduce lower skew variation than the random WID variations (denoted by (I)), *e.g.*, $s_{1,2}$ and $s_{1,6}$. For the sinks with a large horizontal distance (*e.g.*, $s_{1,3}$ and $s_{1,7}$), the skew variation produced by the multi-level correlated WID variations is higher.

The results illustrated in Fig. 3.19 are compared with Monte-Carlo simulations. The setup of the Monte-Carlo simulation environment is listed in Section 3.3.4. The σ of $s_{1,6}$ and $s_{1,7}$ within a group of data-related planes is reported for the independent WID variations in Table 3.8. As reported in this table, the above analysis on the multi-group 3-D H-trees is consistent with the results of Monte-Carlo simulations. The error of the skew variation model is typically smaller than 5% as compared with the Monte-Carlo simulations.

The computational time is also listed for different topologies in Table 3.8. Since this run time is for the entire 3-D clock trees, the computational time is significantly higher than that reported in Table 3.3 for both the proposed model and Monte-Carlo simulations. As the complexity of the 3-D clock tree increases, the time savings by the proposed model significantly increases, up to 1000×. Consequently, the efficiency of the variability-aware design of 3-D clock distribution networks can considerably be improved by estimating the skew variation with the proposed model.



(a) (b)

Figure 3.20: An example of combining clock trees and grids, where (a) is the topology of a tree-grid structure [38] and (b) is the investigated global grid.

3.4.3 Mitigating skew variations with clock grids

To compare the skew variation in 3-D clock trees with other clock distribution topologies, the skew of clock grids is discussed in this section. A typical hybrid structure of clock trees and grids is simulated and compared with the 3-D clock trees in terms of process-induced clock skew.

A pregrid clock distribution network is required to drive a grid structure. A combination of clock trees and grids (tree-grid structure) can meet this requirement [38], as illustrated in Fig. 3.20(a). The sinks of the clock tree are connected to the global clock grid. Buffers are inserted into the clock tree to meet the constraint on the slew rate.

A clock tree-grid structure with 256 sinks is compared with the previous 3-D clock trees in terms of skew variation. The 2-D global clock grid has 256 nodes and the area of each cell is 0.58 mm × 0.58 mm, as illustrated in Fig. 3.20(b). When the grid is embedded to a 3-D IC, the area of the grid is shrunk proportionally to the area of each plane. The grid is located in the first plane and the clock signal is propagated to other planes through TSVs at each node similar to the multi-via 3-D tree. The electrical and physical characteristics of the circuit are reported in Tables 3.1 and 3.2. The Monte-Carlo simulation results of the tree-grid are reported in Table 3.9, where the independent WID variations are considered. The results correspond to the circuits with one, two, four, and eight planes, respectively. The maximum mean skew and the standard deviation within each clock distribution network are denoted by $\mu_{\max}$ and $\sigma_{\max}$, respectively. The average power for each clock distribution network for the nominal device parameters is also reported at the clock frequency of 1 GHz.

As shown in Table 3.9, the tree-grid structure produces the lowest skew variation ($\sigma_{\max}$) compared with the other topologies. Nevertheless, the tree-grid produces the largest mean skew, which is significantly higher than the 3-D H-trees. This situation is due to the considerable delay and capacitance of the wires in the global clock grid. Furthermore, the average power

Table 3.9: Monte-Carlo Results of Different Clock Distribution Networks.

# of planes	μ_{max} [ps]			σ_{max} [ps]			Power@1 GHz [mW]		
	grid	multi	single	grid	multi	single	grid	multi	single
1	9.77	0.00		12.00	21.53		189.60	125.90	
2	8.83	0.04	0.04	11.25	18.30	76.10	105.10	72.01	110.00
4	4.15	0.15	0.18	6.15	16.22	56.65	67.86	51.70	103.10
8	3.10	0.34	0.38	6.39	16.82	41.58	47.68	39.84	89.71

consumed by the tree-grid is the highest among all the three topologies. Extending the grid to multiple planes can, however, improve the power consumption and mean skew. In conclusion, clock grids reduce the skew variations compared with clock trees but increase the mean skew and the power consumption. The multi-via 3-D H-trees can significantly reduce the power consumption and maintain a sufficiently low mean skew while reducing the skew variation. The single-via 3-D H-trees produce the highest power consumption and skew variations due to the large number of buffers, as reported in Table 3.6.

3.4.4 3-D clock trees with multiple domains

Multi-domain clock distribution networks have widely been used to improve the performance of complex large circuits [18, 152]. The effect of process variations on the clock skew of potential 3-D synchronization architectures with multiple clock domains is discussed in this subsection. The case studies include regular clock networks based on single-domain clock trees in a 3-D stack [81, 149]. The analysis methods can also be used to analyze synthesized 3-D clock trees [83, 150]. The resulting skew variations in synthesized clock trees also depend on the efficiency of the synthesis technique. Since the intention is to investigate the effect of process variations rather than the efficiency of a 3-D clock tree synthesizer, regular structures, such as H-trees are explored.

Topologies of 3-D clock trees with multiple domains

In 3-D circuits, the clock trees belonging to different clock domains can be located in the same or different planes. A straightforward idea is to assign each clock domain to a single tier, as illustrated in Fig. 3.21. For each clock domain, a PLL is assumed to generate the clock signal for the corresponding clock network. In this scenario, excluding any synchronization requirement between different clock domains, the impact of D2D process variations expressed by (3.37) can be eliminated. Only WID variations need to be considered.

As illustrated in Fig. 3.21, the sinks of a clock domain are distributed across the entire plane. Long interconnects and a large number of buffers can, consequently, be required. Each clock tree can significantly be affected by WID variations. An approach to mitigate this problem is to decrease the total wire length of the tree, by distributing the clock registers to other planes.

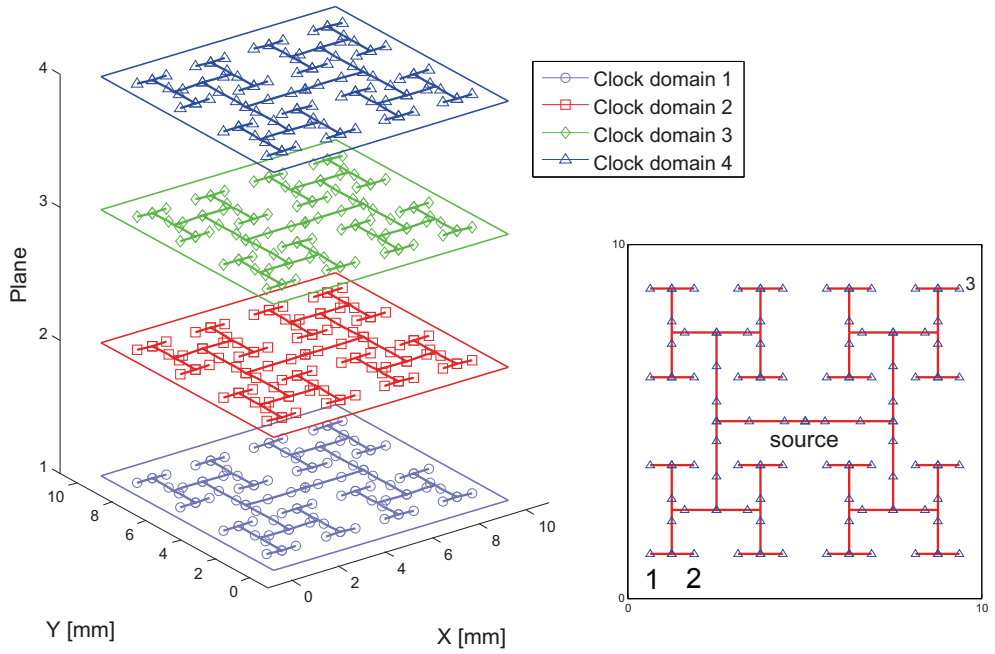


Figure 3.21: A four-plane 3-D IC with four clock domains. A PLL and an H-tree are used to generate and distribute, respectively, the clock signal within each domain (plane). The clock sources are located at the center of each plane.

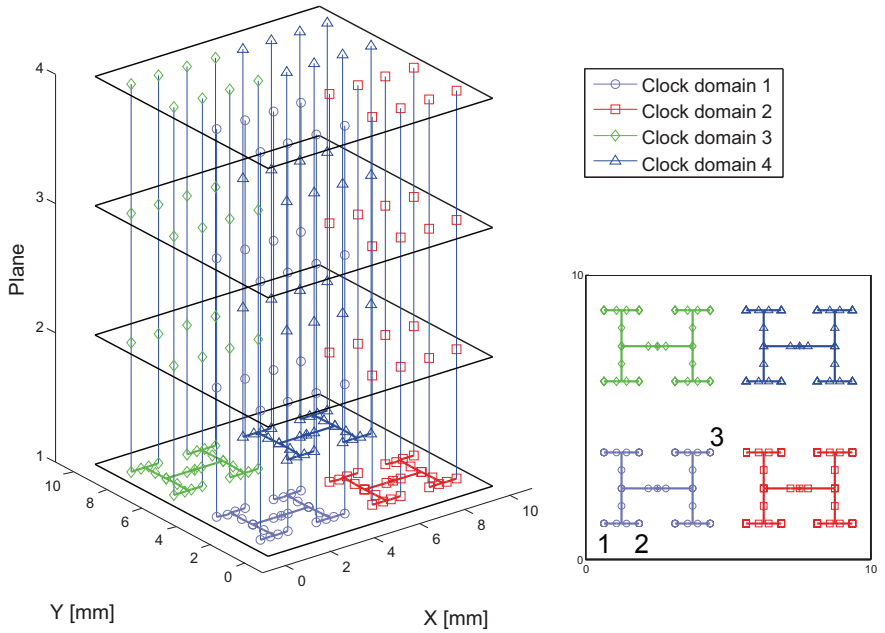
In this case, several clock domains are integrated in one plane, as illustrated in Fig. 3.22(a). The design of the 3-D clock H-tree within each domain is based on [81].

In Fig. 3.22(a), each clock tree spans four planes through TSVs. The skew variation within each clock domain is affected by the D2D variations in all the four planes. The topology illustrated in Fig. 3.22(b) produced by combining the topologies in Figs. 3.21 and 3.22(a) provides another approach to manage the effect of D2D and WID variations. A comparison of different D2D and WID variation scenarios for the investigated 3-D circuits with multiple clock domains is presented in the following section.

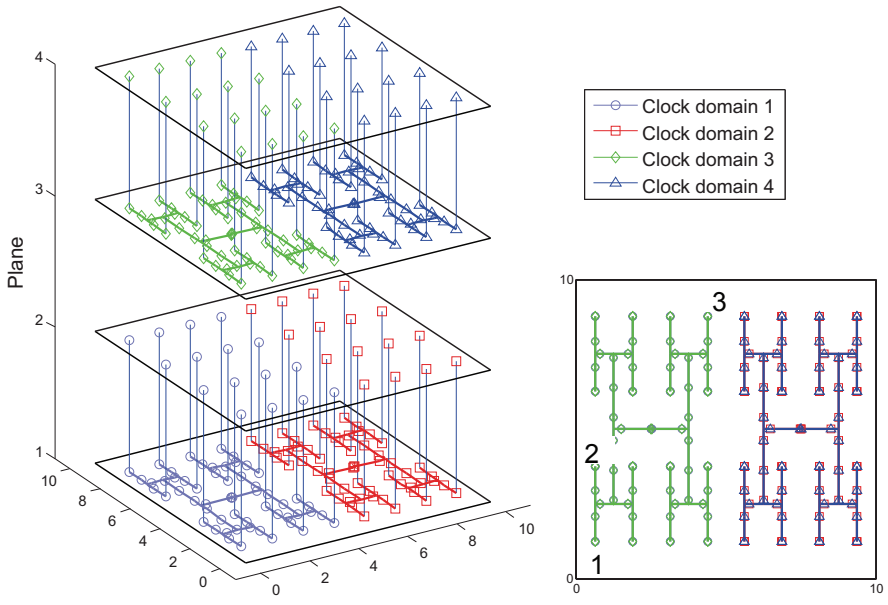
Skew variation in different multi-domain 3-D clock trees

The skew variation is compared among different multi-domain 3-D clock trees. Several combinations of D2D and WID process variations are simulated to investigate the efficiency of different allocations of the clock domains within a 3-D stack.

The PTM model for a 90 nm technology node is used [41]. The characteristics of TSVs are extracted based on [68]. An eight-plane 3-D IC (10 mm × 10 mm per plane), envisioning highly complex 3-D systems, with eight clock domains is simulated. There are 128 clock sinks within each clock domain *i.e.*, 1024 sinks in total. A clock buffer is inserted at each sink driving the downstream circuitry (*e.g.*, a cluster of flip-flops or a local clock mesh). Clock buffers are



(a)



(b)

Figure 3.22: Different assignments of clock domains in a four-plane 3-D IC. (a) Four clock domains within each plane. (b) Two clock domains within each plane (a total of four clock domains).

Table 3.10: Electrical Characteristics of the Investigated Circuits.

Buffer		Interconnect		TSV		Clock	
R_b	536 [Ω]	r	244.44 [Ω/mm]	R_v	0.13 [Ω]	V_{dd}	1.2 [V]
C_b	15.7 [fF]	c	225.039 [fF/mm]	C_v	50 [fF]	f_{clk}	1 [GHz]

inserted into the clock trees after [77], where the constraint on the slew rate is 8.8 mV/ps. The electrical characteristics of the clock networks are listed in Table 3.10. The output resistance and input capacitance of the buffers, the resistance and capacitance per unit length of the interconnects, and the resistance and capacitance of the TSVs are denoted by R_b, C_b, r, c, R_v, and C_v respectively. Four schemes of multiple clock domains are investigated:

(A) One clock domain per plane (see Fig. 3.21).

(B) Two clock domains per plane, each spanning two planes.

(C) Four clock domains per plane each traversing four planes (similar to Fig. 3.22(b)).

(D) Eight clock domains each extending in all of the planes (similar to Fig. 3.22(a)).

Note that the total number of clock domains remains the same for all four schemes; the distribution of these domains among and within the planes, however, changes. The objective is to determine the scheme with the lowest skew variations within each domain. The sinks located the farthest within one domain demonstrate the largest skew variation $s_{\max}$, *e.g.*, $s_{\max} = s_{1,3}$ between sinks 1 and 3 in Fig. 3.21. The smallest skew variation $s_{\min}$ is $s_{1,2}$, a typical trait of an H-tree.

The variations of the gate length (L_{gate}) of both the NMOS and PMOS are considered [88]. Other sources of variations can also be described by the proposed model. The resulting variations in R_b, C_b, and the intrinsic delay of the buffers are extracted by SPICE simulations. Three different scenarios for D2D and WID process variations are investigated:

1. D2D variations are assumed to be higher than the WID variations (D2D > WID). The $\sigma_{L_{\text{gate}}}$ due to D2D and WID variations is assumed to be $\sigma_{L_{\text{gate}}}^{\text{D2D}} = 6\%$ and $\sigma_{L_{\text{gate}}}^{\text{WID}} = 2\%$, respectively.

2. The WID variations are dominant (D2D < WID), $\sigma_{L_{\text{gate}}}^{\text{D2D}} = 2\%$ and $\sigma_{L_{\text{gate}}}^{\text{WID}} = 6\%$.

3. The D2D and WID variations are equivalent (D2D = WID), $\sigma_{L_{\text{gate}}}^{\text{D2D}} = \sigma_{L_{\text{gate}}}^{\text{WID}} = 5\%$.

A clock distribution network based on Scheme (A) is simulated through SPICE. The waveform of the clock signal at sink 1 is illustrated in Fig. 3.23. The slew rate at the sinks is well constrained by the buffer insertion. The delay variation due to the process variations, however,

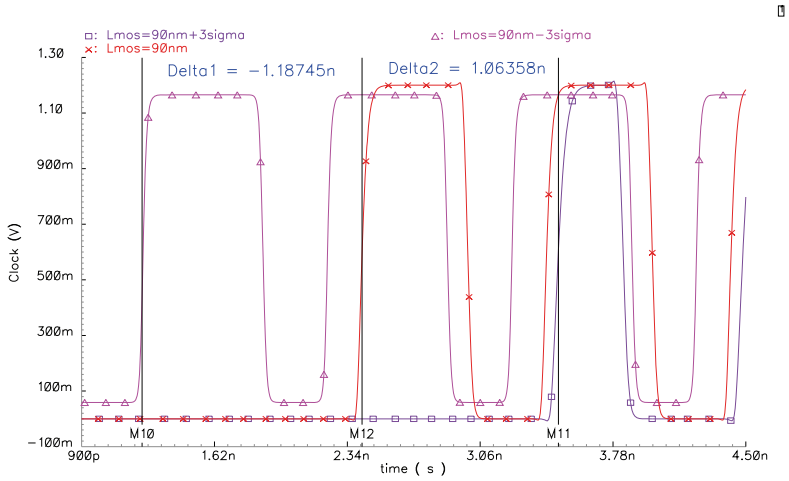


Figure 3.23: Waveform of the signal at s_1 with different gate lengths of MOSFET.

Table 3.11: Skew Variation Analysis of an Eight-Plane 3-D IC with Eight Clock Domains.

σ_{skew} [ps]		Uncorrelated WID				Multi-correlation			
		A	B	C	D	A	B	C	D
D2D > WID	$\sigma_{s_{\min}}$	**5.0**	8.1	9.5	16.1	**4.5**	9.1	10.2	16.0
	$\sigma_{s_{\max}}$	25.8	23.6	**20.0**	21.8	50.6	41.3	28.3	**28.1**
D2D < WID	$\sigma_{s_{\min}}$	**15.1**	24.8	29.2	49.5	**13.7**	27.7	31.1	49.3
	$\sigma_{s_{\max}}$	79.0	69.2	**57.7**	63.5	154.7	126.1	**85.3**	85.8
D2D = WID	$\sigma_{s_{\min}}$	**12.6**	20.6	24.2	41.0	**11.3**	23.0	25.8	40.8
	$\sigma_{s_{\max}}$	65.4	57.3	**48.1**	52.9	128.2	104.5	**70.7**	71.1

is significant. The delay variation with $L_{\text{gate}} - 3\sigma_{L_{\text{gate}}}$ and $L_{\text{gate}} + 3\sigma_{L_{\text{gate}}}$ are -1.2 ns and 1.1 ns, respectively.

The largest and smallest skew variations within a clock domain are reported for the four clock schemes (Scheme A, B, C, and D) and the three variation scenarios (Scenario 1, 2, and 3) in Table 3.11. The numbers of buffers and TSVs per clock domain are listed in Table 3.12, where the mean skew is also reported.

Table 3.12: Statistics of the Eight-Plane 3-D IC with Eight Clock Domains.

Domain Scheme	A	B	C	D
Number of clock buffers	716	533	291	203
Number of clock signal TSVs	0	512	256	128
$\mu_{s_{\max}}$ [fs]	0	4	29	151

Uncorrelated WID variations

In this case, the WID variations are assumed to be independent among the devices within one plane. As reported in Table 3.11, Scheme A produces the highest $\sigma_{S_{\max}}$ for all the three scenarios of variations. This behavior is because the horizontal area (total wire length) occupied by each tree in scheme A is the greatest among the four schemes, requiring the largest number of buffers (see Table 3.12). As described by (3.37) and (3.39), the skew variations of scheme A are higher than the other schemes.

For clock schemes B, C, and D, $\sigma_{s_{\max}}$ varies significantly with the allocation of 3-D clock trees to the planes. Note that although reducing the horizontal area of a tree helps to decrease the WID variations, scheme D does not produce the smallest $\sigma_{S_{\max}}$. The reason is that scheme D introduces a larger number of buffers connected to a TSV in different planes. The effect of D2D variations, therefore, increases. As reported in Table 3.11, scheme C produces the smallest $\sigma_{S_{\max}}$ in all the three variation scenarios.

As the number of planes that a clock tree spans increases, the load capacitance connected to a TSV increases. Consequently, more buffers are inserted along the path from the last branching point to the TSV. For these pairs of sinks which are in short distance, this increase in the number of buffers along this specific path has a greater effect than the decreasing number of buffers for the entire tree. Consequently, skew variations between the nearest sinks increase with the number of planes a tree spans and scheme A produces the smallest $\sigma_{s_{\min}}$. In the three investigated variation scenarios, extending a clock tree of a domain to multiple planes decreases $\sigma_{s_{\max}}$ up to 26% as compared with Scheme A. $\sigma_{s_{\min}}$ increases, however, by 3.3 times.

Guideline 3.4. *For independent WID process variations, extending a clock domain to multiple planes of a 3-D circuit decreases the maximum skew variation. Extending the clock tree to the greatest supported number of planes, however, does not necessarily produce the smallest skew variations. If most of the data-related sinks are distributed close to each other, having one domain within each plane can decrease the skew variations.*

Spatially correlated WID variations

In this case, the correlation of WID variations is modeled by (3.3). As reported in the column "Multi-correlation" in Table 3.11, the behavior of the investigated clocking schemes differs from the uncorrelated WID variations.

For all the three variation cases, extending a clock domain to multiple planes produces a smaller $\sigma_{s_{\max}}$, as compared with scheme A. For "D2D > WID", this decrease in $\sigma_{s_{\max}}$ increases as the number of planes that a tree spans increases. For "D2D < WID" and "D2D = WID", extending the clock tree to all the planes does not, however, produces the smallest $\sigma_{s_{\max}}$. Consequently, the efficiency of extending a clock tree to multiple planes depends on the relation between D2D and WID variations. For $\sigma_{s_{\min}}$, in this correlation model, the behavior of

the four clocking schemes is similar to the independent WID correlation. $\sigma_{s_{\min}}$ increases as the number of planes that a clock tree spans increases.

Guideline 3.5. *For multi-level WID process variations, increasing the number of planes a clock domain spans increases the skew variation between sinks located within a short distance. The change in the maximum skew variation depends on the relation between D2D and WID variations.*

3.5 Summary

The effect of process variations in 3-D ICs is investigated in this chapter. The focus of this chapter is to model and analyze the effect of process variations in 3-D clock distribution networks. The contributions and the major points of this chapter are:

- A novel model to describe the distribution of process-induced skew in 3-D clock trees, which exhibits reasonably high accuracy, is proposed.
- Typical 3-D clock distribution networks are compared among each other in terms of clock skew variation. 3-D clock grids exhibit the lowest skew variation but with a significant cost in power consumption.
- For 3-D clock trees, the multi-via topology outperforms the single-via topology in terms of the maximum skew variation and power consumption, since the single-via topology requires a larger number of buffers. For clock sinks within the same plane, however, single-via 3-D clock trees usually produce a lower skew variation due to the smaller number of buffers per plane.
- A new 3-D clock tree topology is proposed to combine the advantages of both multi-via and single-via topologies, which produces a low skew variation for the clock sinks within the same group.
- The skew variation in multi-domain clock trees is also investigated. It is shown that placing different clock domains in different tiers does not necessarily produce the lowest skew. Skew variation can be decreased by locating different clock domains within the same plane and vertically extending these domains.
- For multi-level WID process variations, increasing the number of planes a clock domain spans increases the skew variation between the sinks located within a short distance. The maximum skew variation, however, is determined by the sinks with the farthest distance. The change of the maximum skew depends on the relation between D2D and WID variations.

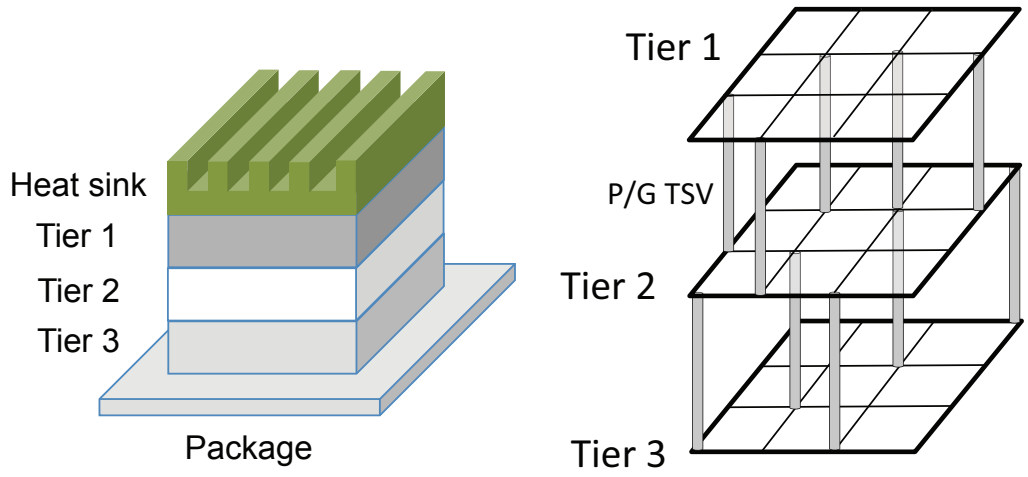
4 Power Supply Noise in 3-D ICs

As introduced in Chapter 2, in addition to process variations, integrated circuits are affected by the voltage variations across the power distribution systems. The power supply noise in 3-D ICs is investigated in this chapter, where the effect of power supply noise on clock jitter is also discussed. Power distribution networks for 3-D ICs are first presented in Section 4.1. An approach for fast *IR*-drop analysis of 3-D power distribution networks is proposed in Section 4.2. The resonant supply noise determined by the inductance and capacitance of potential 3-D PDNs is discussed in Section 4.3. The clock jitter caused by this resonant supply noise is introduced in Section 4.4. The conclusions are drawn in Section 4.5.

4.1 3-D Power Distribution Networks

Some PDN structures for 2-D circuits have been introduced in Section 2.2.1. For 3-D circuits, the off-chip parts of a PDN, *i.e.*, the voltage regulator, PCB, package, and the related decoupling capacitance are similar to a 2-D system. The on-chip PDN, however, varies from 2-D circuits, since the power and ground need to be distributed across multiple planes/tier.

In TSV-based 3-D ICs, power/ground TSVs are used to distribute P/G from one tier to another. A 3-D circuit and the corresponding PDN are illustrated in Fig. 4.1. As shown in this figure, each tier has its own planar PDN. The planar PDNs are connected with each other at several nodes by P/G TSVs to form the entire on-chip PDN. Consequently, the voltage variation in different tiers interacts with each other. Several 3-D PDNs have been investigated in [34, 153]. These 3-D PDNs mainly differ from each other in the electrical characteristics of P/G TSVs. A 3-D PDN similar to a 2-D network is implemented in [153], while 3-D PDNs with different characteristics among tiers are discussed in [34]. The resulting supply noise differs among tiers, which substantially differentiates the behavior of power supply noise from that observed in planar circuits. To model the supply noise in different 3-D PDNs, a fast method to analyze *IR*-drop is proposed in the following section. The method to describe the resonant noise is presented in Section 4.3.



(a) (b)

Figure 4.1: An example of a 3-D circuit where (a) is a schematic of a three-tier circuit and (b) is the corresponding PDN.

4.2 A Method for Fast *IR*-Drop Analysis of 3-D ICs

As introduced in Section 2.2.2, *IR*-drop is an important component of supply noise. The voltage drop due to the resistance of the interconnects of PDNs is dominated by the on-chip PDNs. Consequently, a resistor network is used to model a PDN. Power grids are investigated in this section due to their wide application [25, 27]. A resistor network abstraction of a power grid is illustrated in Fig. 4.2.

As shown in Fig. 4.2, P/G TSVs are modeled as vertical resistors connecting adjacent tiers. The voltage at the nodes of the resistor network is the variable to be determined. For instance, voltage $v_{x,y,1}$ denotes the voltage of the node in the x^{th} row and the y^{th} column of the first tier. The devices of the circuit are modeled as constant current sources, as shown by the current source connecting to $v_{x,y,1}$. VDD and GND grids are modeled as two independent networks [27]. Only one network is illustrated in Fig. 4.2. In a VDD grid, the current flows from the grid through current sources to the ideal ground. In a GND grid, the current flows from the ideal VDD through current sources to the grid. The voltage at the package is considered as ideal VDD (or GND). The resistance of package P/G pins and the corresponding vias is modeled as r_{pkg}. Note that multiple P/G pins are used to connect the on-chip PDN to the package at different nodes.

4.2.1 Problem formulation

As shown in Fig. 4.2, devices (modeled by the constant current sources) are connected to different nodes of a power grid. The voltage at these nodes needs to be determined for *IR*-drop

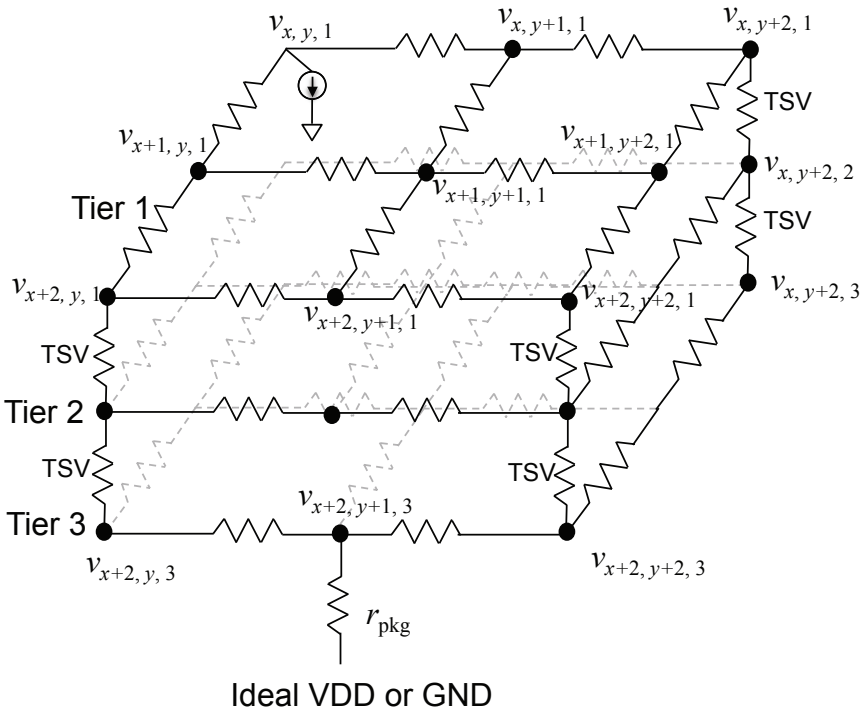


Figure 4.2: A resistor network used to model a 3-D PDN for *IR*-drop analysis.

analysis. A node connected with four resistors in the same tier and two TSVs is illustrated in Fig. 4.3. This node is denoted by $v_{x,y,z}$, which is located in the x^{th} row and y^{th} column of the VDD grid in the z^{th} tier. The four resistors in the same tier connected to this node are denoted by $r_{x,(y\pm1),z}$ and $r_{(x\pm1),y,z}$, indicating the resistors on the left, right, upper, and lower side in the topview, respectively. The TSVs connecting $v_{x,y,z}$ to the $(z-1)^{\text{th}}$ and $(z+1)^{\text{th}}$ tiers are denoted by $r_{x,y,(z,z-1)}$ and $r_{x,y,(z,z+1)}$, respectively. The voltage $v_{x,y,z}$ is determined by the adjacent resistance and nodes. According to *Kirchholff's Current Law* (KCL),

$$\begin{aligned}(v_{x,y,z}-v_{x,y-1,z})g_{x,(y,y-1),z}+(v_{x,y,z}-v_{x,y+1,z})g_{x,(y,y+1),z}+\\ (v_{x,y,z}-v_{x-1,y,z})g_{(x,x-1),y,z}+(v_{x,y,z}-v_{x+1,y,z})g_{(x,x+1),y,z}+\\ (v_{x,y,z}-v_{x,y,z-1})g_{x,y,(z,z-1)}+(v_{x,y,z}-v_{x,y,z+1})g_{x,y,(z,z+1)}=-I_{x,y,z},\end{aligned} \tag{4.1}$$

where $g_i = 1/r_i$.

For a 3-D power grid with Z tiers, the number of rows and columns in each tier can be different from other tiers. The numbers of rows and columns in the z^{th} tier are denoted by X_z and Y_z, respectively. Consequently, similar to the *IR*-drop analysis in 2-D power grids [120], the voltage in 3-D power grids can be solved by a linear system,

$$\mathbf{A}\cdot\mathbf{v}=\mathbf{b}, \tag{4.2}$$

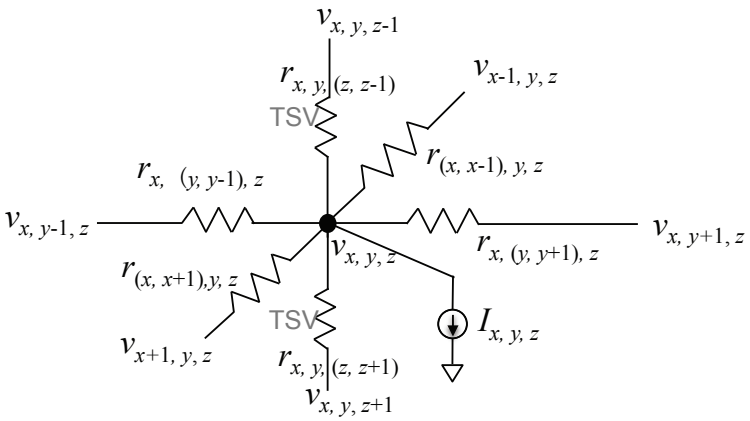


Figure 4.3: A node of a power grid connected with four resistors in the same tier and two TSVs.

where **A** is the conductance matrix. **v** and **b** are the voltage and current vectors, respectively,

$$\mathbf{v} = [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_Z]^{\mathrm{T}}, \tag{4.3}$$

$$\mathbf{v}_z = [\mathbf{v}_{1,z}, \mathbf{v}_{2,z}, \cdots, \mathbf{v}_{X_z,z}],\ (1 \le z \le Z) \tag{4.4}$$

$$\mathbf{v}_{x,z} = [v_{x,1,z},\ v_{x,2,z},\ \cdots,\ v_{x,Y_z,z}],\ (1 \le x \le X_z) \tag{4.5}$$

$$\mathbf{b} = [\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_Z]^{\mathrm{T}}, \tag{4.6}$$

$$\mathbf{b}_z = [\mathbf{b}_{1,z}, \mathbf{b}_{2,z}, \cdots, \mathbf{b}_{X_z,z}],\ (1 \le z \le Z) \tag{4.7}$$

$$\mathbf{b}_{x,z} = -[I_{x,1,z},\ I_{x,2,z},\ \cdots,\ I_{x,Y_z,z}],\ (1 \le x \le X_z) \tag{4.8}$$

where $\mathbf{v}_z$ and $\mathbf{b}_z$ are the voltage and current in the z^{th} tier, respectively. The conductance matrix **A**, accordingly, consists of the conductance of the power grid in each tier and the TSVs,

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{B}_1 & & \\ \mathbf{B}_1 & \mathbf{A}_2 & \ddots & \\ & \ddots & \ddots & \mathbf{B}_{Z-1} \\ & & \mathbf{B}_{Z-1} & \mathbf{A}_Z \end{bmatrix}. \tag{4.9}$$

The sub-matrix $\mathbf{A}_z\ (1 \le z \le Z)$ is the conductance of the power grid in the z^{th} tier and $\mathbf{B}_i$ is the conductance of the TSVs connecting the z^{th} and $(z+1)^{\text{th}}$ tiers. Matrix $\mathbf{A}_z$ is determined by the conductance of each row and the conductance between adjacent rows,

$$\mathbf{A}_z = \begin{bmatrix} \mathbf{A}_{1,z} & \mathbf{D}_{1,z} & & \\ \mathbf{D}_{1,z} & \mathbf{A}_{2,z} & \ddots & \\ & \ddots & \ddots & \mathbf{D}_{X_z-1,z} \\ & & \mathbf{D}_{X_z-1,z} & \mathbf{A}_{X_z,z} \end{bmatrix}, \tag{4.10}$$

where $\mathbf{A}_{x,z}$ and $\mathbf{D}_{x,z}$ are the conductance of the x^{th} row and the conductance between the x^{th} and $(x+1)^{\text{th}}$ rows in the z^{th} tier, respectively. These two sub-matrices are determined by

$$\mathbf{A}_{x,z} = \begin{bmatrix} G_{x,1,z} & -g_{x,(1,2),z} & & & \\ -g_{x,(2,1),z} & G_{x,1,z} & \ddots & & \\ & \ddots & \ddots & -g_{x,(Y_z-1,Y_z),z} \\ & & -g_{x,(Y_z,Y_z-1),z} & G_{x,Y_z,z} \end{bmatrix}, \tag{4.11}$$

$$\mathbf{D}_{x,z} = \begin{bmatrix} -g_{(x,x+1),1,z} & & & \\ & -g_{(x,x+1),2,z} & & \\ & & \ddots & \\ & & & -g_{(x,x+1),Y_z,z} \end{bmatrix}, \tag{4.12}$$

where, $G_{x,y,z}$ is the sum of all the conductances related to the node at (x, y, z),

$$G_{x,y,z} = g_{x,(y,y-1),z} + g_{x,(y,y+1),z} + g_{(x,x-1),y,z} + g_{(x,x+1),y,z} + g_{x,y,(z,z-1)} + g_{x,y,(z,z+1)}, \tag{4.13}$$

The node connectivity of the power grid within the xy dimensions of the z^{th} tier can be described by A_z through (4.10) to (4.13). Similarly, the vertical connection between the z^{th} and $(z+1)^{\text{th}}$ tiers is described by B_z,

$$\mathbf{B}_z = \begin{bmatrix} \mathbf{B}_{1,z} & & & \\ & \mathbf{B}_{2,z} & & \\ & & \ddots & \\ & & & \mathbf{B}_{X_z,z} \end{bmatrix}, \tag{4.14}$$

$$\mathbf{B}_{x,z} = \begin{bmatrix} -g_{x,1,(z,z+1)} & & & \\ & -g_{x,2,(z,z+1)} & & \\ & & \ddots & \\ & & & -g_{x,Y_z,(z,z+1)} \end{bmatrix}. \tag{4.15}$$

Consequently, the voltage at all the nodes is determined through (4.2) to (4.15). The objective of steady-state *IR*-drop analysis is to determine the voltage of all the nodes in a 3-D power grid satisfying (4.2).

Note that the total number of nodes, $N = \sum_{z=1}^{Z} X_z \cdot Y_z$, is relatively large (≥ tens of millions for contemporary circuits) [27], which means **A** is a large sparse matrix. Similar to the conclusion in [120], **A** is also a symmetric semi-positive matrix. Special methods, consequently, can be used to solve this large linear system.

For 2-D ICs, different techniques have been developed to efficiently analyze a power grid. The grid-reduction [154] and hierarchical analysis [155] methods convert the large power grid to coarser or smaller blocks to reduce the size of the linear system. The random walk method can be utilized to solve a subset of a large power grid [156], with a large number of

iterations. Iterative methods, such as the row-based [120] and multi-grid preconditioned conjugated gradients [157, 158] algorithms have been developed to overcome the drawbacks of the previous methods. For a 3-D power grid, since **A** is also semi-positive, the efficient iterative methods can be extended to solve this linear system.

4.2.2 Row-based algorithm for 3-D PDNs

Since the resistance of P/G TSVs is introduced into 3-D PDNs, the traditional *IR*-drop analysis methods for 2-D PDNs, where only four adjacent resistors are considered, cannot be directly applied to 3-D cases. For 3-D ICs, a voltage propagation algorithm has been proposed in [159] to solve a specific type of 3-D PDNs, where the resistance of P/G TSVs is assumed to be much smaller than the resistance within a planar grid. If these two resistances are comparable to each other, the voltage propagation algorithm may not converge. Nevertheless, as reported in the industrial benchmarks in [27], the resistors in a planar grid can be lower than 10 mΩ. This resistance is smaller than the resistance of different proposed TSV technologies [34, 68, 160]. In this case, the algorithm proposed in [159] cannot be utilized. A novel *IR*-drop analysis algorithm, therefore, is proposed herein to model different types of 3-D PDNs.

The novel *IR*-drop analysis algorithm extends the traditional row-based algorithm [120] to 3-D PDNs, which is named by “RB3D”. The key idea is illustrated in Fig. 4.4. The algorithm iteratively traverses from the first tier (farthest from the package) to the bottommost tier to calculate the voltage of each node row by row. The pseudocode of RB3D is listed in Algorithm 4.1.

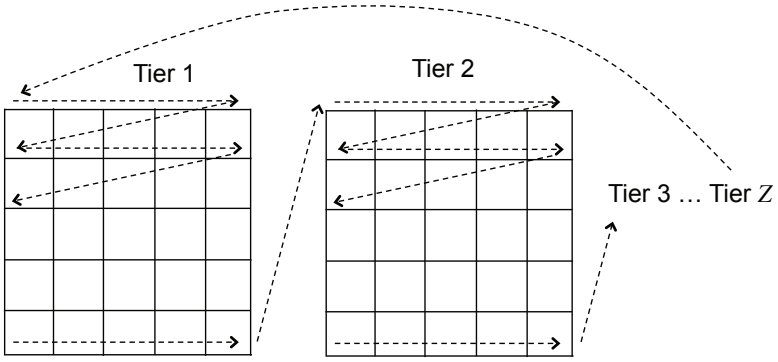


Figure 4.4: The traversal direction of the row-based algorithm for a 3-D PDN with Z tiers.

The function CALC_ROW_VOLTAGE(x, z) is used to calculate the voltage of the x^{th} row in the z^{th} tier. This function is adapted from the "Solving-one-row" algorithm in [120] by considering the resistance of TSVs. For the x^{th} row in the z^{th} tier, the variables a, b, c, and d are determined

Algorithm 4.1 RB3D algorithm

Input: The topology and current sources of a 3-D power grid.
Output: Voltage at all the nodes **v**.

Initialize **v**.
$dV \leftarrow$ max difference in **v** between two iterations ▷ specified by users
$\Delta v \leftarrow \infty$
while $\Delta v > dV$ **do**
  $\mathbf{v}_{old} \leftarrow \mathbf{v}$
  for $z = 1 \rightarrow Z$ **do** ▷ Z is the number of tiers
    for $x = 1 \rightarrow X_z$ **do**
      $\mathbf{v}(x, 1 \rightarrow Y_z, z) \leftarrow$ CALC_ROW_VOLTAGE(x, z)
    end for
  end for
  $\Delta v \leftarrow \max(\mathbf{v} - \mathbf{v}_{old})$
end while
return v

function CALC_ROW_VOLTAGE(x, z)
  $u_1 \leftarrow G_{x,1,z}$
  for $y = 2 \rightarrow Y_z$ **do**
    $l_y \leftarrow \frac{a_y}{u_{y-1}}$ ▷ a, b, c, and d are defined in (4.16) and (4.17)
    $u_y \leftarrow b_y - l_y c_{y-1}$
  end for
  $w_1 \leftarrow d_1$
  for $y = 2 \rightarrow Y_z$ **do**
    $w_y \leftarrow d_y - l_y w_{y-1}$
  end for
  $V_{Y_z} \leftarrow \frac{w_{Y_z}}{u_{Y_z}}$
  for $y = Y_z - 1 \rightarrow 1$ **do**
    $V_y \leftarrow \frac{w_y - c_y V_{y+1}}{u_y}$
  end for
  $\mathbf{v}_{x,z} \leftarrow [V_1, \cdots, V_{Y_z}]$
  return $\mathbf{v}_{x,z}^{\mathrm{T}}$
end function

by

$$\mathbf{A}_{x,z} = \begin{bmatrix} b_1 & c_1 & & & \\ a_2 & b_2 & c_2 & & \\ & a_3 & \ddots & \ddots & \\ & & \ddots & \ddots & c_{Y_z-1} \\ & & & a_{Y_z} & b_{Y_z} \end{bmatrix}, \tag{4.16}$$

$$\begin{bmatrix} d_1 \\ \vdots \\ d_{Y_z} \end{bmatrix} = \begin{bmatrix} -I_{x,1,z} \\ \vdots \\ -I_{x,Y_z,z} \end{bmatrix} + \mathbf{D}_{x-1,z} \cdot \mathbf{v}^{\mathrm{T}}_{x-1,z} + \mathbf{D}_{x,z} \cdot \mathbf{v}^{\mathrm{T}}_{x+1,z} + \mathbf{B}_{x,z-1} \cdot \mathbf{v}^{\mathrm{T}}_{x,z-1} + \mathbf{B}_{x,z} \cdot \mathbf{v}^{\mathrm{T}}_{x,z+1}, \tag{4.17}$$

where $\mathbf{v}$ is the voltage currently obtained. The voltage in different tiers is iteratively calculated until the difference between two successive iterations is lower than the specified threshold dV. Since $\mathbf{A}$ in (4.2) is positive semi-definite, the algorithm will converge for any initial $\mathbf{v}$ [120]. For a 3-D PDN, the algorithm RB3D is applied to the VDD and GND grids independently. The *IR*-drop at each node is obtained by adding up the voltage drop on the VDD grid and voltage rise on the GND grid.

4.2.3 Simulation results

The proposed algorithm "RB3D" is compared with SPICE-based DC analysis to verify the accuracy of RB3D. The benchmark 3-D power grids are generated from industrial 2-D power grids [27] by replicating the planar grids in multiple tiers. The statistics of the 2-D benchmarks are reported in Table 4.1. Among tiers, the P/G TSVs are uniformly inserted based on a user-specified interval.

Table 4.1: IBM power grid benchmarks for IR-drop analysis.

Benchmark	# current sources	# nodes	# resistors	# P/G pads	# metal layers
ibmpg1	10774	30638	30027	277	2
ibmpg2	37926	127238	208325	330	5
ibmpg3	201054	851584	1401572	955	5
ibmpg4	276976	953583	1560645	962	6
ibmpg5	540800	1079310	1076848	277	3
ibmpg6	761484	1670494	1649002	381	3

The "RB3D" is compared with HSPICE [161] for the original 2-D benchmarks and the generated 3-D power grids with two and three tiers. To provide enough current, P/G TSVs (50 mΩ each) are inserted every four nodes. The worst voltage drop for VDD and GND, the maximum error, and the acceleration in CPU time as compared to HSPICE are reported in Table 4.2.

Table 4.2: Simulation results for 2-D and 3-D power grids based on IBM benchmarks.

Circuit	max ΔVDD [V][1]			max ΔGND [V] [2]			error[3]	time[4]	#TSVs/Tier
	1p	2p	3p	1p	2p	3p			
ibmpg1	0.811	1.391	1.953	0.694	0.983	1.260	4%	3×	1510
ibmpg2	0.500	0.866	1.222	0.370	0.566	0.756	0%	22×	5489
ibmpg4	0.012	0.017	0.021	0.004	0.006	0.008	20%	4×	38363
ibmpg5	0.067	0.118	0.161	0.050	0.067	0.090	15%	7×	62543
ibmpg6	0.211	0.308	0.403	0.113	0.167	0.219	6%	2×	97714

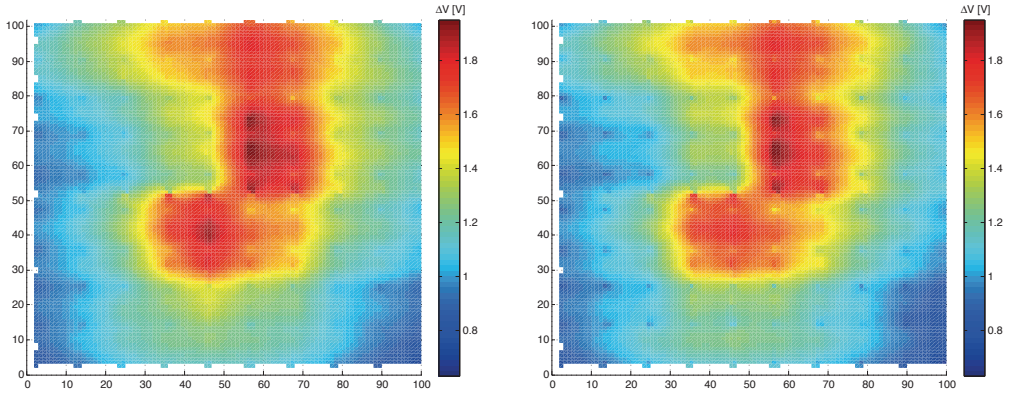
[1, 2] The maximum *IR*-drop in VDD (GND) grids among all tiers.
[3] The maximum error for both VDD and GND grids as compared with HSPICE.
[4] The maximum acceleration in CPU time as compared with HSPICE simulations.

As shown in Table 4.2, for all the benchmark power grids, the proposed algorithm "RB3D" achieves a reasonably high accuracy as compared to HSPICE simulations. For the circuits ibmpg4 and ibmpg5, although the percentage is over 10%, the maximum error is actually below 7 mV since the IR-drop is relatively low. The savings in CPU time depends on the specified accuracy. The memory required by RB3D is much lower than HSPICE. HSPICE is not able to simulate large circuits (*e.g.*, ibmpg5 and ibmpg6) at 32 bits platforms, while RB3D is capable to run at both 32 and 64 bits platforms.

The *IR*-drop among different tiers of a three-tier VDD grid based on ibmpg1 is illustrated in Fig. 4.5. As shown in this figure, the *IR*-drop is similar among tiers, since the same planar circuit is replicated in each tier and the resistance of TSVs is relatively low (50 mΩ). Nevertheless, it is still clear that the voltage drop of tier 3 (closer to the P/G pads) is lower than tier 1 (farther to the P/G pads). The maximum difference in V_{dd} between the first and third tiers is 35.5 mV. This difference in the voltage drop among tiers significantly increases with the resistance of TSVs and the number of tiers.

4.3 Modeling Resonant Supply Noise in 3-D ICs

In addition to the *IR*-drop, the resonant supply noise caused by the inductance and capacitance of PDNs also highly affects the timing of circuits, as introduced in Section 2.2.3. To investigate the resonant supply noise in 3-D PDNs, the 1-D model for 2-D PDNs (see Fig. 2.16) can be extended to include the electrical characteristics of the TSVs and the on-chip PDNs in different tiers [34]. The resonant noise is mainly determined by the *LC* tank formed between the package inductance and the on-die capacitance. This noise is typically stimulated by a large current spike due to the clock edge or simultaneous switching of transistors. A simplified circuit used to simulate the resonant (the first droop) supply noise in 3-D PDNs is illustrated in Fig. 4.6. A three-tier 3-D IC is shown in this figure. The P/G signal is supplied from the package to tier 3, tier 2, and to the topmost tier 1. The on-die decoupling capacitance is denoted by C_3,



(a) (b)

Figure 4.5: *IR*-drop in a three-tier circuit based on ibmpg1, where (a) and (b) are the top-views of tiers 1 and 3 with 50 mΩ TSVs, respectively.

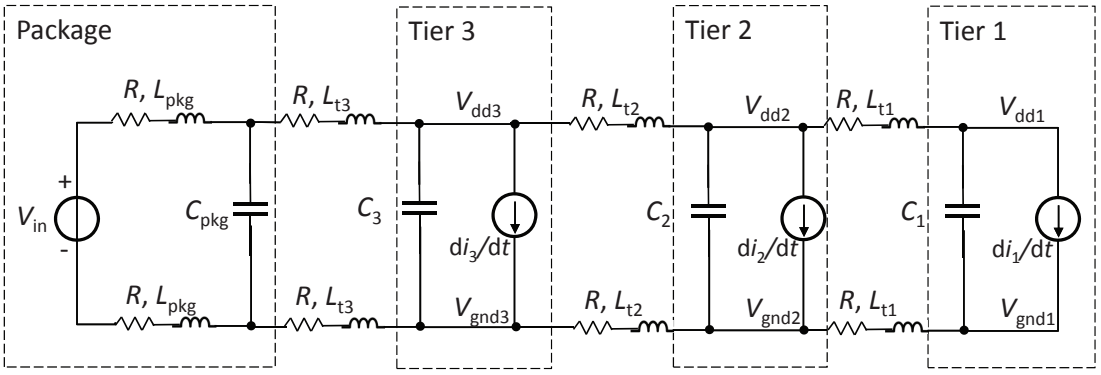


Figure 4.6: A simplified circuit used to simulate the first-droop of the power supply noise.

Table 4.3: Electrical Characteristics of the Simplified Circuit.

	R_t [mΩ]	L_t [pH]	$C_{1,2,3}$ [nF]	I_3 [A]	I_2 [A]	I_1 [A]
Base	50	100	1	0.03	0.04	0.05
Diff_I1	50	100	1	0.02	0.04	0.06
Diff_I2	50	100	1	0.04	0.04	0.04
Diff_I3	50	100	1	0.06	0.04	0.02

C_2, and C_1, respectively. The resonant noise in different tiers is investigated under different scenarios of 3-D PDNs in the following subsections.

4.3.1 Resonant Noise *vs.* On-Chip Current

In 3-D ICs, the current dissipated by the tiers can differ due to the different numbers and sizes of devices. Moreover, due to the different clock delay and wakeup time among tiers, the current peaks can temporally vary among tiers. The waveform of the resonant supply noise due to different turn-on times is illustrated in Fig. 4.7. The time instances where the circuits in the different tiers switch are separated by 0.1 ns, 1 ns, and 10 ns, respectively, for Switch1, Switch2, and Switch3 scenarios. The electrical characteristics of the 3-D circuit are listed in Table 4.3.

As shown in Fig. 4.7, the amplitude of the supply noise decreases with the separation among the switching instances of the tiers. When the separation is lower than 1 ns (Switch1 and Switch2), the resonant noise can be approximated by a damped sinusoidal signal. Since the supply noise of Switch1 and Switch2 is much higher than Switch3, the sinusoidal waveform can be used to describe the worst resonant noise of 3-D ICs.

Consequently, the resonant supply noise seen by the circuit at time t can be described by a damped sinusoidal waveform

$$v(t) = V_n e^{-\sigma t} \sin(2\pi f_n t + \phi). \tag{4.18}$$

The clock frequency is much higher than the supply noise frequency and the clock delay is typically lower than the period of the supply noise. As shown in Fig. 4.7, the worst supply noise occurs in the first period of the resonant noise. Consequently, to investigate the effect of the worst supply noise on clock distribution networks, (4.18) can be approximated by an undamped sinusoidal waveform [94],

$$v(t) \approx V_n \sin(2\pi f_n t + \phi). \tag{4.19}$$

The amplitude V_n, frequency f_n, and the initial phase ϕ are determined by the switching current and the characteristics of the circuits. The relation between resonant noise and the current in different tiers is illustrated in Fig. 4.8. The load current peak in different tiers is

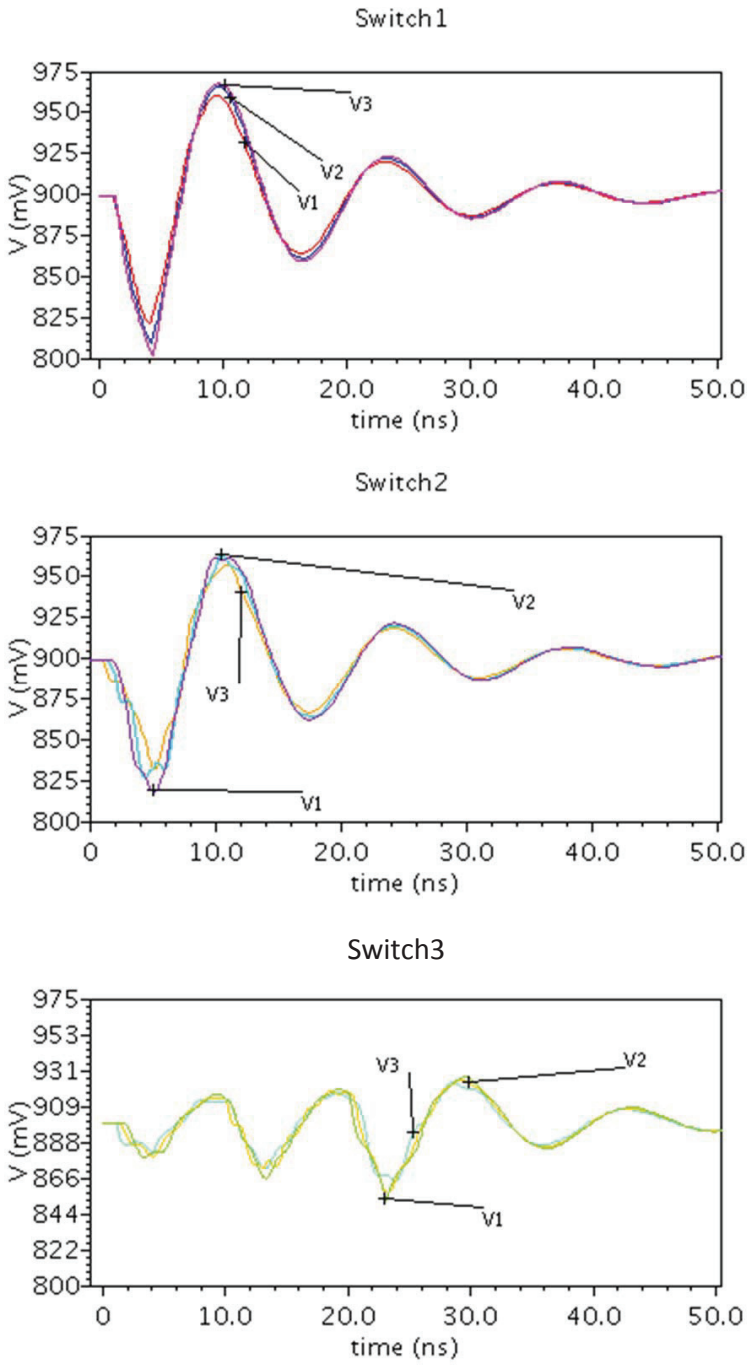
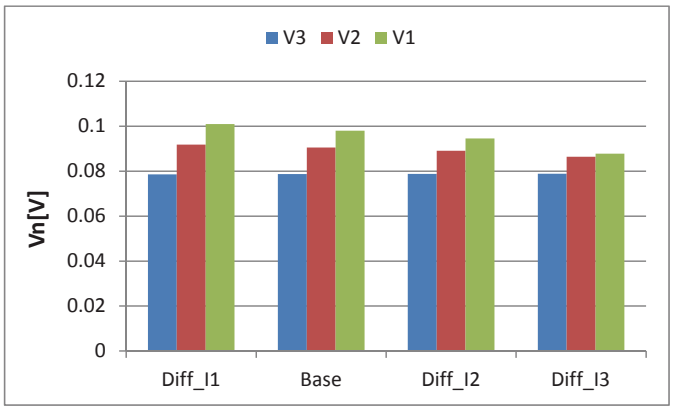
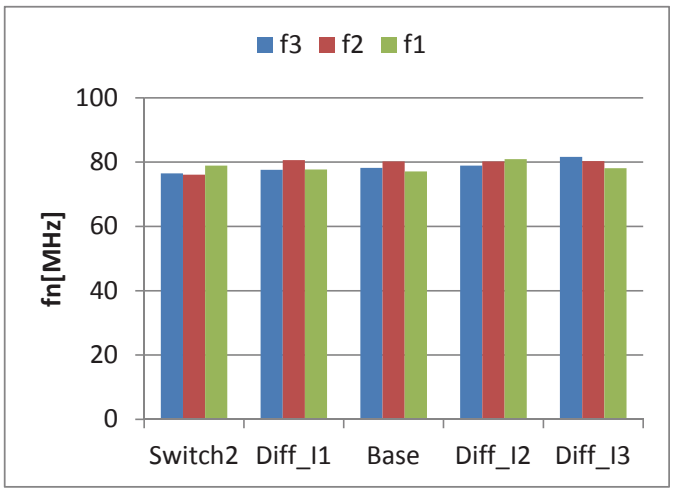


Figure 4.7: Resonant noise in 3-D ICs due to different temporal separation of circuits switching within the three tiers.



(a)



(b)

Figure 4.8: Resonant noise *vs.* switching current in different tiers. The change of V_n and f_n are illustrated in (a) and (b), respectively.

reported in Table 4.3, where Diff_I1, Diff_I2, and Diff_I3 are given. As shown in Fig. 4.8(a), the difference in V_n among tiers decreases as I_3 increases and I_1 decreases. Considering the severe thermal issues in 3-D ICs, the majority of switching components need to be placed in the tiers close to the heat sink (Tiers 1 and 2) to reduce temperature [34]. Consequently, the switching current in Tiers 1 and 2 can be significantly higher than Tier 3, which is similar to Diff_I1 in Table 4.3. As illustrated in Fig. 4.8(a), this current distribution introduces non-negligible ΔV_n among tiers.

The change of f_n with the switch current is shown in Fig. 4.8(b). The frequency of resonant noise is similar among tiers ($\Delta f_n \leq 5$ MHz) and does not significantly change with the current. As illustrated in Fig. 4.7, assuming all the components are idle at $t = 0$, the initial supply noise is the same for all tiers ($v(0) = 0$). Due to the same $v(0)$ and similar f_n among tiers, the initial phase ϕ of the supply noise is similar among tiers.

4.3.2 Resonant Noise *vs.* Resistance of TSVs

The electrical characteristics of TSVs differ with the manufacturing technology [68, 71]. The change of the power supply noise with the resistance of TSVs is illustrated in Fig. 4.9. Larger

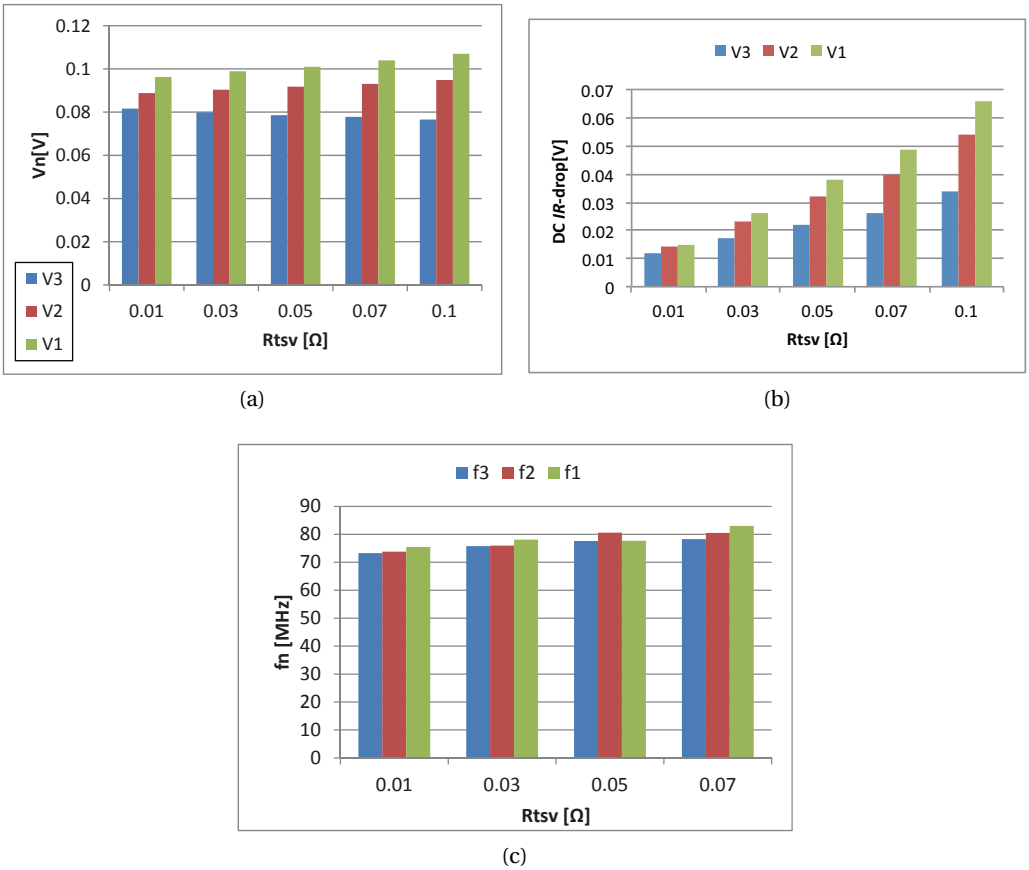


Figure 4.9: Resonant noise *vs.* resistance of TSVs. The change of V_n, *IR*-drop, and f_n is illustrated in (a) to (c), respectively.

R_{tsv} results in both higher DC *IR*-drop (Fig. 4.9(b)) and higher ΔV_n (Fig. 4.9(a)). Moreover, the difference in *IR*-drop and ΔV_n among tiers increases. The difference in frequency Δf_n also differs slightly with R_{tsv}. Consequently, the resonant supply noise is sensitive to the resistance of TSVs.

4.3.3 Resonant Noise *vs.* Number of Tiers

The resonant noise for different number of tiers in a 3-D IC is plotted in Fig. 4.10. The total switch current and on-die decoupling capacitance are assumed identical for all the four cases and are evenly distributed among tiers. As shown in Fig. 4.10, both ΔV_n and Δf_n increase with the number of tiers. As more dies are vertically stacked, the difference in resonant noise among tiers increases.

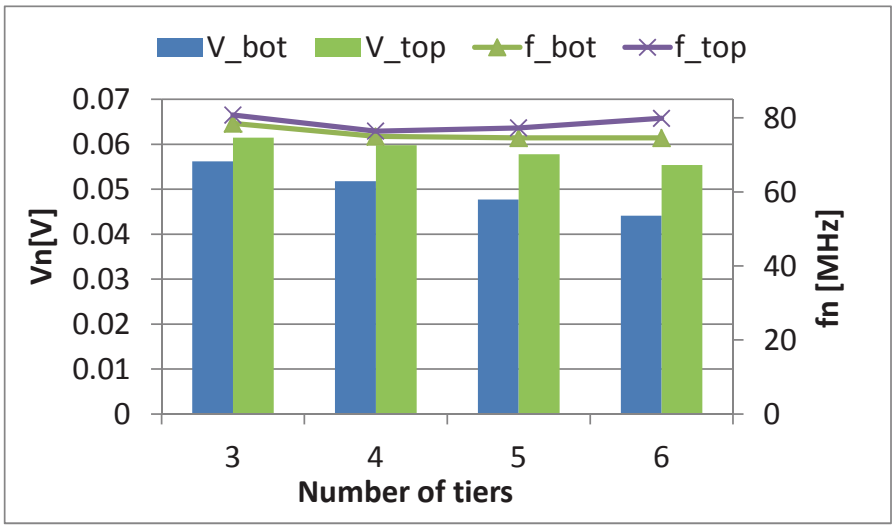


Figure 4.10: Resonant noise *vs.* number of tiers. The changes in both V_n and f_n are depicted.

4.3.4 High-Frequency Power Supply Noise

Other faster and more mutable droops of power supply noise ($f_n \geq 400$ MHz) can be directly described as random variables with probabilistic formulations, as modeled in [121]. To obtain the distribution of these variables, the switching activity of all the cells is required. Afterwards, a full-chip transient simulation of the PDN is performed to determine the temporal and spatial change of the power supply noise. The high- and mid-frequency supply noise, however, can be greatly reduced by *RC* filters [25] and the first-droop of the resonant supply noise is usually the deepest droop. Consequently, the focus of the remaining part of this dissertation is mainly on the first-droop (resonant) supply noise.

4.4 Clock Jitter due to the First Droop of Power Supply Noise

The effect of power supply noise on clock distribution networks is discussed in this section. As presented in [29, 30], the first droop of the power supply noise, or the resonant noise, is typically the worst (largest) supply noise in a circuit. Consequently, the effect of resonant noise on the clock uncertainty is the focus of this section. Since the resonant noise is determined by the *LC* tank formed by the package/bonding inductance and the on-chip capacitance, the entire chip is uniformly affected by this noise [29]. A clock path affected by the first-droop noise is illustrated in Fig. 4.11(a). As introduced in Section 4.3.1, the first-drop power supply noise can be modeled by a sinusoidal waveform. The path delay of two consecutive clock edges under this supply noise is illustrated in Fig. 4.11(b).

As shown in Fig. 4.11(b), two consecutive rising clock edges arrive at the source of the clock path at time t_0 and t_1, respectively. The ideal clock period is assumed to be $t_1 - t_0$. The path delay of these two edges is denoted by pd_0 and pd_1, respectively. Since these edges arrive

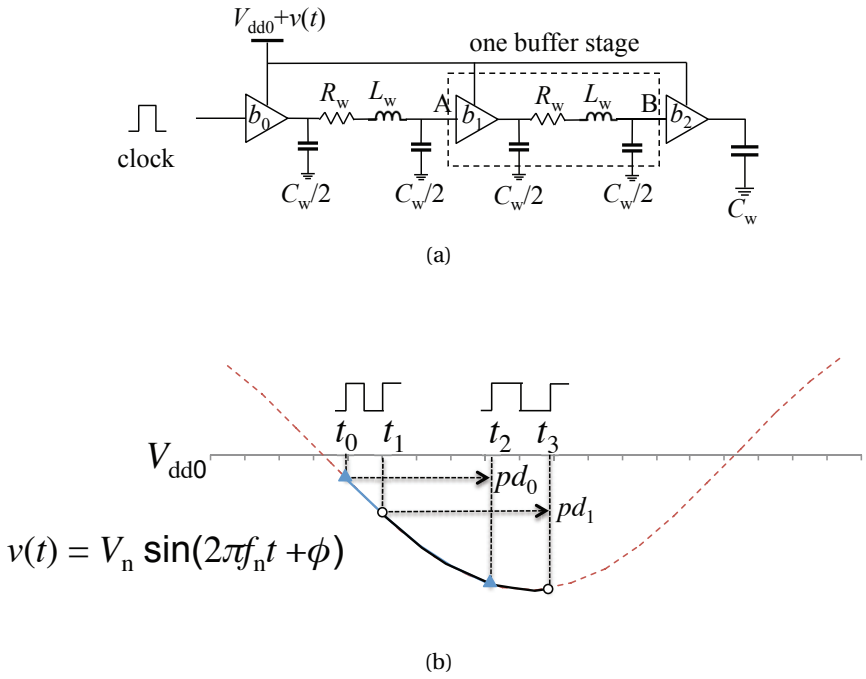


Figure 4.11: A clock path affected by the first-droop supply noise, where (a) and (b) are the clock path and path delay, respectively.

at the clock buffers at different time, the supply noise affecting the buffers is different. The resulting path delay pd_0 is, consequently, different from pd_1, which suggests $t_3 - t_2 \neq t_1 - t_0$. Therefore, the clock period seen at the end of the clock path is different from the ideal clock period. This type of variation of the clock signal is described by clock jitter. Clock jitter is defined as the deviation of the clock signal from the ideal temporal occurrences [18]. Clock jitter can be described in three ways: period jitter, cycle-to-cycle jitter, and phase jitter (time interval error). The corresponding definitions are illustrated in Fig. 4.12, where the clock edges deviate from the ideal occurrence to a different extend.

- The period jitter can be defined as either the difference between the actual and the ideal clock periods [18] or the actual clock period itself [94]. The former is used in this dissertation to avoid the confusion between period jitter and clock period. In Fig. 4.12, the ideal clock period is denoted by T_0. The period jitter for the first and the second clock periods is determined by $J_1 = T_1 - T_0$ and $J_2 = T_2 - T_0$, respectively.

- The cycle-to-cycle jitter is the difference between different clock cycles of the actual clock signal. As shown in Fig. 4.12, the cycle-to-cycle jitter between the first and the second clock cycles is determined by $CJ_{1,2} = T_2 - T_1$.

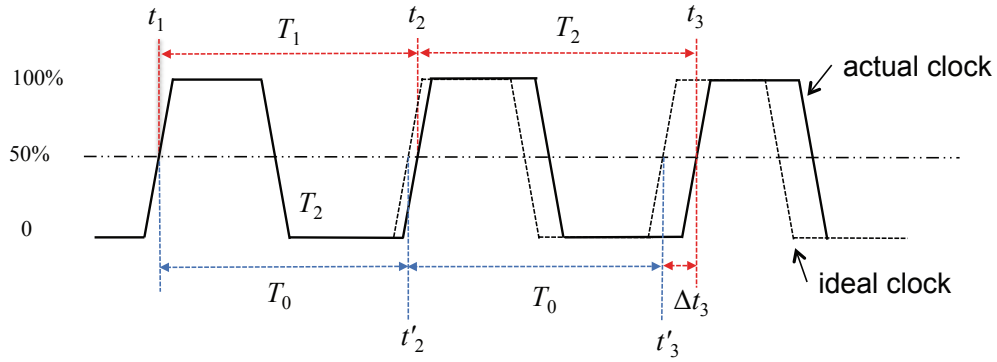


Figure 4.12: Different definitions of clock jitter.

- The phase jitter, or the time interval error, is the difference between an actual clock edge and the corresponding ideal clock edge. In Fig. 4.12, the phase jitter for the third clock edge is determined by $\Delta t_3 = t_3 - t'_3$.

These three definitions of clock jitter are essentially similar to each other. The period jitter is the most explicit description of the clock jitter within a circuit [18]. Therefore, the clock jitter directly refers to period jitter in the remainder of this dissertation. Note that clock jitter can be generated by both the *Phase-Locked Loop* (PLL) and clock distribution network. PLL jitter can be mitigated by careful PLL design [162]. The clock jitter generated by clock distribution networks, however, is affected by the power supply noise, as shown in Fig. 4.11(b). In the following context, the clock jitter implicitly refers to the jitter generated in clock distribution networks.

In 2-D ICs, closed-form formulas of period jitter are proposed in [94]. Based on the approximation of the power supply noise, non-recursive and recursive expressions of supply noise can be used to model period jitter. For a clock path consisting of clock inverters, the non-recursive period jitter is based on the non-recursive supply noise. When a clock edge arrives at the i^{th} inverter at time $t(i)$, the supply noise is approximated by

$$v(t(i)) \approx v\left(t_0 + \frac{(i-1)(\overline{d}_\text{r} + \overline{d}_\text{f})}{2}\right), \tag{4.20}$$

where $\overline{d}_\text{r}$ and $\overline{d}_\text{f}$ are the nominal delay of an inverter stage for the rising and falling input, respectively. The time where the investigated clock edge arrives at the input of the clock path is denoted by t_0. When $i-1$ is an odd number, an additional $\overline{d}_\text{r}$ or $\overline{d}_\text{f}$ is added to (4.20). The period jitter J_1 in Fig. 4.12 can be determined by

$$J_1 = t_2 - t_1 - T_0, \tag{4.21}$$

$$t_1 = \sum_{i=1}^{k} d_p\Big(v\big(t_1(i)\big)\Big) + t_0, \tag{4.22}$$

$$t_2 = \sum_{i=1}^{k} d_p\Big(v\big(t_2(i)\big)\Big) + t_0 + T_0, \tag{4.23}$$

where $v(t_1(i))$ and $v(t_2(i))$ are the supply noise seen by the i^{th} inverter when the first and the second clock edges arrive, respectively. These supply noises are determined by (4.20). When the input is a rising (falling) edge, d_p equals d_r (d_f). Although the non-recursive expression provides a concise formula to estimate the period jitter, it determines the supply voltage for each buffer stage based on the nominal clock delay. This assumption introduces some error, however, since the actual clock delay is affected by the supply noise.

To further improve the accuracy, a recursive expression for period jitter is developed in [94]. The clock jitter and delay are still determined through (4.21) to (4.23). The supply noise $v(t)$ seen by the i^{th} inverter, however, is determined based on the actual clock delay $t(i)$, instead of the approximation in (4.20). For instance, when the first clock edge arrives at the k^{th} buffer, the supply noise $v(t_1(k))$ is, recursively, determined through (4.19) by

$$v(t_1(k)) = V_\text{n} \sin(2\pi f_\text{n} \cdot t_1(k) + \phi), \tag{4.24}$$

$$t_1(i) = \sum_{j=1}^{i-1} d_p\Big(v\big(t_1(j)\big)\Big). \tag{4.25}$$

Given the characteristics of the power supply noise, clock jitter can be determined through (4.21) to (4.25). Nevertheless, as shown in Fig. 4.11, this clock jitter applies to a single clock path. Considering that the setup and hold slacks are determined by a pair of clock paths, clock jitter and skew need to be simultaneously modeled to accurately describe the clock uncertainty of a clock distribution network. This combined model is investigated in the following chapter.

4.5 Summary

The power supply noise in 3-D PDNs is investigated in this chapter. A fast steady-state *IR*-drop analysis method is developed for 3-D power grids. In this method, the row-based algorithm for 2-D power grids is extended to consider the influence of P/G TSVs and the interaction among tiers. Compared to SPICE-based simulations, the proposed method achieves reasonably high accuracy and savings in the computing resources.

The resonant noise in 3-D PDNs is investigated based on the one-dimensional model. Under different scenarios of 3-D PDNs, the resonant noise exhibits different characteristics among tiers.

- For various schemes of switching current and turn-on time, the tier adjacent to the package and the heat sink experience the lowest and highest amplitude of resonant noise, respectively.
- The difference in the amplitude of resonant noise increases with the resistance of P/G TSVs.
- The difference in the amplitude of resonant noise increases with the number of tiers.
- In all scenarios, the frequency of resonant noise slightly differs among tiers, with a difference lower than 10% in the simulations.

The clock jitter under power supply noise is introduced. A non-recursive model of power supply noise provides a concise method to determine clock jitter, while a recursive model provides a more accurate and complex method to obtain clock jitter. The combined effect of the power supply noise and process variations on the timing uncertainty of clock distribution networks is discussed in the next chapter.

5 Combined Effect of Process and Voltage Variations on Clock Uncertainty

The combined effect of process variations and power supply noise on clock distribution networks is investigated in this chapter. As shown in Chapters 3 and 4, the process variations and power supply noise introduce clock skew and jitter, respectively, into clock distribution networks. Since a circuit is simultaneously affected by these variation sources, the resulting clock uncertainty needs to be modeled considering process and voltage variations at the same time.

The term "skitter" [163] is used to denote this combined clock uncertainty in the following section. A simplified model for skitter in 2-D ICs is proposed in Section 5.2, where the skitter for different buffer insertion methodologies is also discussed. A more accurate model for skitter in 3-D ICs is proposed in Section 5.3, where different process variations and supply noise among tiers are considered. Methods to mitigate skitter are presented in Section 5.4. A case study for the skitter in synthesized clock trees is presented in Section 5.5, where a 3-D clock tree synthesis algorithm is also described. To insert buffers along 3-D interconnect trees, a timing-driven fast buffer insertion algorithm is proposed in Section 5.6. The conclusions are drawn in Section 5.7.

5.1 Skitter: A Unified Treatment of Skew and Jitter

Clock distribution networks are simultaneously affected by different sources of variations, such as static process variations and dynamic voltage noise [18]. The resulting clock uncertainty due to these variations consists of clock skew and jitter. This uncertainty can severely constrain the highest clock frequency of a circuit. In addition, the design of robust clock distribution networks requires a comprehensive analysis and proper mitigation of these variations.

3-D integration emerges as a promising solution to alleviate the increasing interconnect delay and to enhance the density of devices in modern integrated circuits [7]. Multiple planar circuits (tiers) with different technologies can be vertically stacked, which complicates the variation analysis for 3-D circuits. The effect of process variations in 3-D ICs has been

discussed in Chapter 3. The resulting skew variation is modeled by considering the time-invariant variability in the characteristics of devices and wires. The effect of power supply noise, especially the resonant noises, on 2-D clock distribution networks has been investigated in [29, 30, 94]. The effect of power supply noise on 3-D clock distribution networks, however, has not been adequately explored.

For 3-D ICs, different PDNs have been investigated in Chapter 4. A 3-D PDN similar to a 2-D network is implemented in [153], while 3-D PDNs with different characteristics among tiers are discussed in [34]. The resulting amplitude of supply noise differs among tiers substantially differentiating the behavior of power supply noise from that observed in planar circuits. The effect of this different power supply noise on the timing uncertainty of 3-D clock distribution networks, however, remains unclear. The change of the clock uncertainty with both the different characteristics of supply noise and process variations is investigated in this chapter.

In most of the previous works, the effect of process variations and power supply noise is discussed separately in terms of skew and jitter. Clock skew, the difference in delay among clock paths, is considered to be significantly affected by process variations and is well modeled for 2-D ICs [139, 144]. A model of process-induced skew in 3-D clock trees is investigated in Chapter 3. The other constituent of clock uncertainty is clock jitter, as introduced in Section 4.4. Period jitter is the difference between the measured clock period and the ideal period. The period jitter generated in clock distribution networks is mainly due to the power supply noise on the clock buffers [164]. The effect of the power supply noise on period jitter in 2-D ICs is analyzed in [29, 30, 94], while to the best of the author's knowledge, this dissertation discusses period jitter in 3-D ICs for the first time.

Clock distribution networks are simultaneously affected by process variations and power supply noise. For 2-D ICs, a statistical timing analysis method considering process variations and power supply noise is proposed in [165], where full-chip simulations are required to obtain the distribution of power supply noise. Moreover, the effect of these variations on clock distribution networks is not adequately explored. The combination of skew and jitter, "skitter", is introduced in [163] to model the co-effect of all sources of variations on clock distribution networks, while no closed-form expression is given to model the distribution of skitter. A subcircuit is designed to measure the skitter in [163], which can be utilized to mitigate undesired skitter during operation [166]. If the skitter is high, frequent recovery and adaptation procedures have to be executed to correctly transfer data. Moreover, these architectural procedures cannot be used for each pair of clock sinks. Consequently, by better understanding the behavior of skitter, this component of clock uncertainty can be mitigated through the proper design of clock distribution networks. In addition, the overhead of the adaptive circuits and architectural procedures can be reduced.

The combined effect of process variations and dynamic power supply noise on 3-D clock distribution networks has not been explored, although clock skew and jitter need to be treated cohesively. As mentioned before, these variations in 3-D ICs are more complex due to the

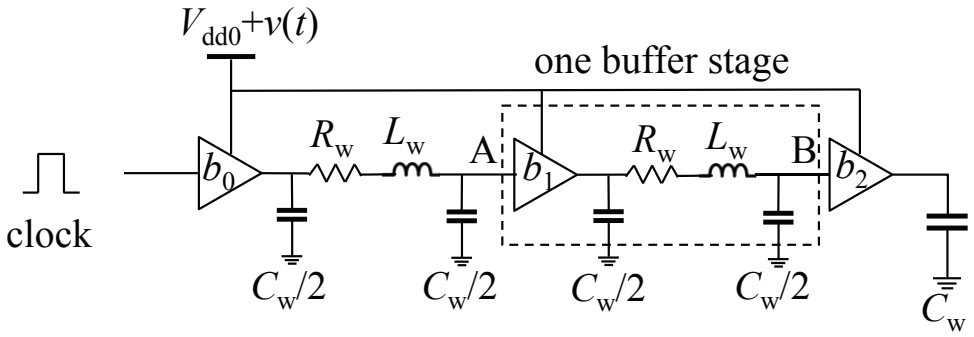


Figure 5.1: A circuit used to measure the delay variation of one buffer stage due to process variations and power supply noise.

stacked tiers. The effect of these variations on clock distribution networks is investigated in terms of skitter in this chapter.

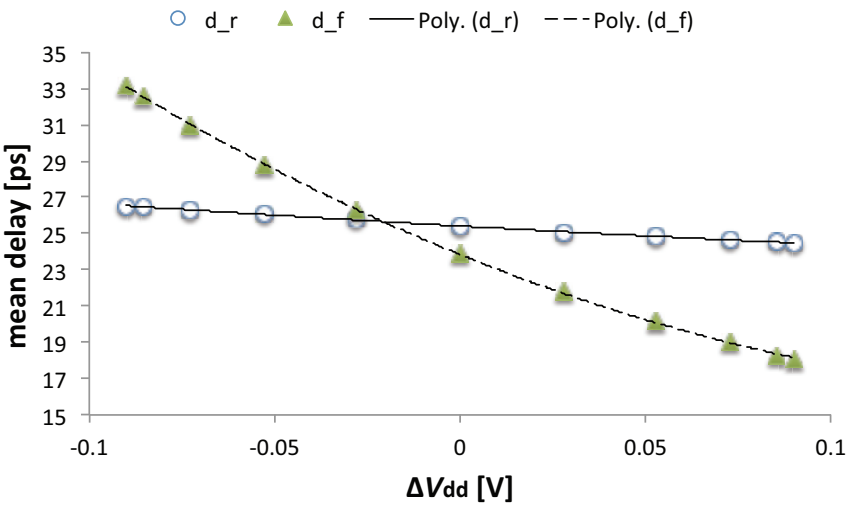
5.2 Modeling Skitter in 2-D Clock Distribution Networks

Skitter in 2-D clock distribution networks is investigated in this section. Considering the uniform D2D process variations and resonant supply noise in 2-D circuits, a simplified model is proposed to fast describe skitter. The delay model of a buffer stage considering both process variations and power supply noise is proposed in the following subsection. Based on this model, the skitter of clock trees is investigated in Section 5.2.2. The skitter between clock paths with different numbers and sizes of buffers is discussed in Section 5.2.3. Several methods to mitigate skitter in 2-D circuits is presented in Section 5.2.4.

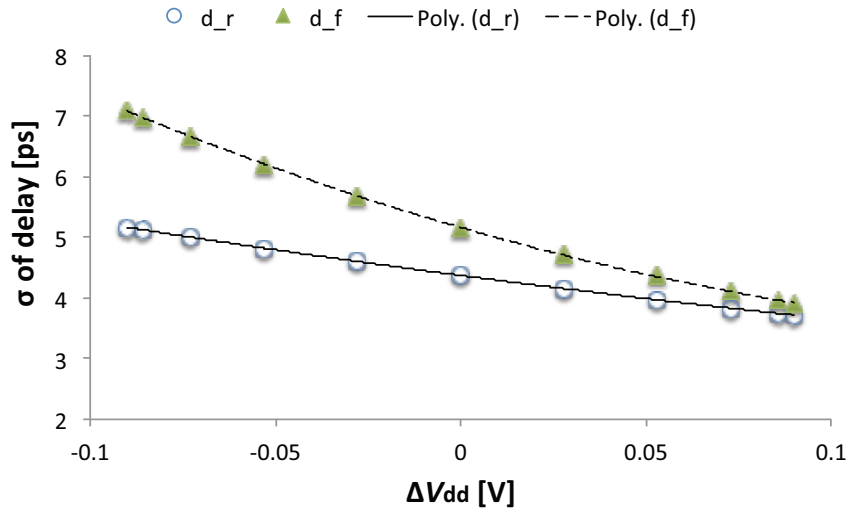
5.2.1 Delay distribution of a buffer stage

A simplified method to model the delay variation of a buffer stage simultaneously considering process variations and power supply noise is presented in this subsection. For a clock tree with uniform clock buffer insertion, the input slew rate and load of each buffer stage are similar. Consequently, the delay variation of a buffer stage can be evaluated by a parameterized lumped circuit, as illustrated in Fig. 5.1. The investigated buffer stage is depicted with a dashed rectangle in Fig. 5.1. The interconnect between two buffers is modeled as an *RLC* network comprising one π-section. The power supply to buffers b_0, b_1, and b_2 can be adapted to model the power supply noise v. By measuring the delay variation from pin A to pin B, the effect of process variations under different power supply noise can be described.

An example of the delay variation of a buffer stage with the supply noise is illustrated in Fig. 5.2. The mean and standard deviation of the delay of a buffer stage are shown in this figure. In this example, a clock buffer is an inverter, based on a PTM 32 nm CMOS model [41]. The supply voltage is $V_{dd} + \Delta V_{dd}$, where $V_{dd} = 0.9$ V is the nominal supply voltage. As shown in Fig.



(a)



(b)

Figure 5.2: The mean and standard deviation of the delay of a buffer stage.

5.2(a), the delay of a buffer stage for a rising and falling input edges is denoted by, d_{r} and d_{f}, respectively. The mean delay $\mu_{d_{\mathrm{f}}}$ decreases with v much faster than $\mu_{d_{\mathrm{r}}}$. In Fig. 5.2(b), $\sigma_{d_{\mathrm{r}}}$ and $\sigma_{d_{\mathrm{f}}}$ also decrease with ΔV_{dd}. Consequently, a higher V_{dd} can produce lower mean and standard deviation of the delay of a clock buffer stage.

Both $\mu_{d_{\mathrm{r}}}$ ($\mu_{d_{\mathrm{f}}}$) and $\sigma_{d_{\mathrm{r}}}$ ($\sigma_{d_{\mathrm{f}}}$) under different power supply noise can be obtained by polynomial fitting from SPICE based Monte-Carlo simulations [94]. Considering $\Delta V_{\mathrm{dd}} = v$, the delay variation of a buffer stage is approximated by a second-order polynomial,

$$y = a_2 v^2 + a_1 v + a_0, \tag{5.1}$$

where y denotes the mean or standard deviation of the delay of a buffer stage. With the expressions for the delay variation of one buffer stage, the skitter $J_{1,2}$ can be obtained by traversing all the clock buffer stages along the investigated pair of clock paths and estimating the supply noise at different time instances.

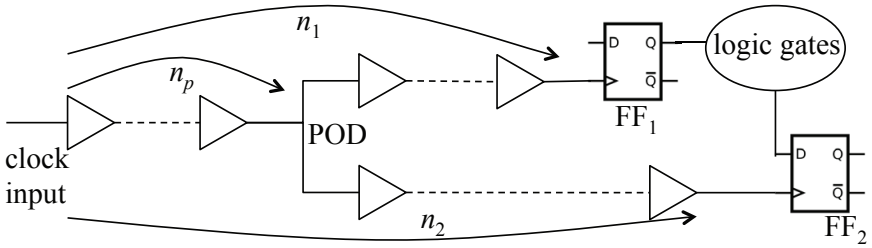
5.2.2 Skitter considering process variations and power supply noise

The skitter in 2-D clock trees can be determined with the delay model of a buffer stage by estimating the power supply noise in either non-recursive or recursive formulas, as introduced in Section 4.4. To fast model skitter, the non-recursive formula of supply noise is used in this section.

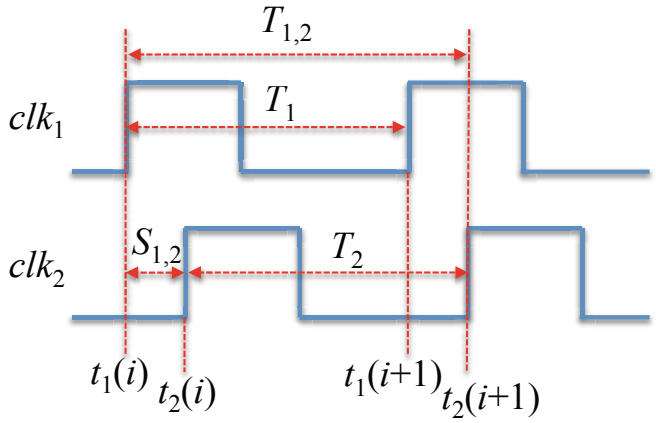
The definition of the clock skew, period jitter, and skitter between a pair of clock paths in 2-D ICs is illustrated in Fig. 5.3. The clock signal is fed into the clock tree from the primary clock driver. Two flip-flops are driven by this clock signal, denoted as FF_1 and FF_2 in Fig. 5.3(a). The corresponding waveforms are illustrated in Fig. 5.3(b). The waveforms clk_1 and clk_2 denote the clock signal driving FF_1 and FF_2, respectively. Assuming the time where the i^{th} rising edge arrives at clock input is zero, the time where this edge arrives at FF_1 and FF_2 is, respectively, denoted by $t_{1,i}$ and $t_{2,i}$. The number of buffers before the *point of divergence* (POD) is n_p. The numbers of buffers from the clock input to FF_1 and FF_2 are denoted by n_1 and n_2, respectively.

The skew between the i^{th} edge of clk_1 and clk_2 is $S_{1,2}(i)$. The ideal clock period is T_{clk}. The measured clock periods after the i^{th} edge for FF_1 and FF_2 are T_1 and T_2, respectively. The corresponding period jitters are $J_1 = T_1 - T_{\mathrm{clk}}$ and $J_2 = T_2 - T_{\mathrm{clk}}$. Assuming the data is propagated from FF_1 to FF_2 within one clock cycle, $T_{1,2}$ is the resulting time interval that determines the maximum data transfer speed. Consequently, the variation of $T_{1,2}$ is denoted as skitter $J_{1,2}$,

$$\begin{aligned} J_{1,2} &= T_{1,2} - T_{\mathrm{clk}} \\ &= t_2(i+1) - t_1(i) - T_{\mathrm{clk}} \\ &= S_{1,2}(i) + J_2. \end{aligned} \tag{5.2}$$



(a)



(b)

Figure 5.3: Clock period jitter and skew between two clock paths. The clock paths and FFs are illustrated in (a). The corresponding waveforms of the clock signal are illustrated in (b).

As shown in (5.2), the effective time window $T_{1,2}$ is determined by $J_{1,2}$, which is the sum of the skew $S_{1,2}(i)$ and the period jitter J_2 along clock path 2. Simultaneously modeling the skew and jitter can, therefore, more accurately determine delay uncertainty.

Skitter $J_{1,2}$ is the difference in the delay between two clock paths,

$$J_{1,2} = \sum_{k=1}^{n_2} d_{2,k}(i+1) - \sum_{k=1}^{n_1} d_{1,k}(i), \tag{5.3}$$

where $d_{1,k}(i)$ is the delay of the k^{th} buffer stage along the path to FF_1 for the i^{th} clock edge. As modeled in [94] and introduced in Section 4.4, $d_{1,k}(i)$ can be approximated by a non-recursive formula for the first-droop power supply noise,

$$d_{1,k}(i) = \mathscr{F}\left(v_k(i), \vec{P}\right), \tag{5.4}$$

$$v_k(i) \approx v\left(t_0(i) + \frac{k\left(d_{\text{r}}(v_1(i)) + d_{\text{f}}(v_1(i))\right)}{2}\right), \tag{5.5}$$

$$v(t) = V_{\text{n}} \sin(2\pi f_{\text{n}} t + \phi), \tag{5.6}$$

where $v_k(i)$ is the voltage noise which affects the k^{th} buffer stage when the i^{th} clock edge arrives at this stage. The power supply noise v is modeled as a sinusoidal waveform with amplitude V_{n}, frequency f_{n}, and initial phase ϕ. This deterministic model is widely used to describe the first droop of the power supply noise, which is considered as the worst supply noise in a circuit [29, 94]. Other faster and more erratic droops of the supply noise can be included as random variables with probabilistic formulations, similar to process variations.

The delay of a buffer stage under $v_1(i)$ for a rising and falling input is denoted by $d_{\text{r}}(v_1(i))$ and $d_{\text{f}}(v_1(i))$, respectively. The set of parameters affected by process variations is denoted by $\vec{P}$. For instance, if the variations in channel length and threshold voltage are considered, $\vec{P} = \{L_{\text{eff}}, V_{\text{th}}\}$. The expression for $\mathscr{F}\left(v_k(i), \vec{P}\right)$ is obtained by quadratic fitting as presented in the previous subsection. This expression can be approximated as a Gaussian distribution if the parameters in $\vec{P}$ are also described by a Gaussian distribution.

The distribution of $J_{1,2}$ is, therefore, approximated as a Gaussian distribution from (5.3) and (5.4). The mean value and the standard deviation of $J_{1,2}$ are discussed separately.

- Mean value of skitter $\mu_{J_{1,2}}$. The term $J_{1,2}$ can be expressed as the difference of the delay of the $i+1^{\text{th}}$ and i^{th} clock edges,

$$J_{1,2} \sim \mathcal{N}(\mu_{J_{1,2}}, \sigma^2_{J_{1,2}}), \tag{5.7}$$

$$\mu_{J_{1,2}} = \sum_{k=1}^{n_2} \mu_{d_{2,k}(i+1)} - \sum_{k=1}^{n_1} \mu_{d_{1,k}(i)}. \tag{5.8}$$

- Standard deviation of the skitter $\sigma_{J_{1,2}}$. The variation on $J_{1,2}$ is determined by both the D2D and WID variations, which are independent from each other. All the devices are affected by D2D variations uniformly. The WID variations on different devices consist of random and spatially correlated components [37, 92, 144]

$$\sigma^2_{J_{1,2}} = \sigma^2_{J^{\text{D2D}}_{1,2}} + \sigma^2_{J^{\text{WID}}_{1,2}}, \tag{5.9}$$

$$\sigma_{J^{\text{D2D}}_{1,2}} = \sum_{k=1}^{n_2} \sigma_{d^{\text{D2D}}_{2,k}(i+1)} - \sum_{k=1}^{n_1} \sigma_{d^{\text{D2D}}_{1,k}(i)}, \tag{5.10}$$

$$\begin{aligned}\sigma^2_{J^{\text{WID}}_{1,2}} = &\sum_{k=1}^{n_2} \sigma^2_{d^{\text{WID}}_{2,k}(i+1)} + \sum_{k=1}^{n_1} \sigma^2_{d^{\text{WID}}_{1,k}(i)} + 2\sum_{k=1}^{n_2-1}\sum_{h=k+1}^{n_2} \text{Cov}\left[d^{\text{WID}}_{2,k}(i+1), d^{\text{WID}}_{2,h}(i+1)\right] \\ &+2\sum_{k=1}^{n_1-1}\sum_{h=k+1}^{n_1} \text{Cov}\left[d^{\text{WID}}_{1,k}(i), d^{\text{WID}}_{1,h}(i)\right] - 2\sum_{k=1}^{n_2}\sum_{h=1}^{n_1} \text{Cov}\left[d^{\text{WID}}_{2,k}(i+1), d^{\text{WID}}_{1,h}(i)\right],\end{aligned} \tag{5.11}$$

$$\text{Cov}(a,b) = \text{corr}(a,b)\sigma_a\sigma_b. \tag{5.12}$$

  In this model, it is assumed that the covariance between the delay variation is determined by the spatial correlation. Assuming the number of buffers before POD is n_p, for $k \le n_p$, $\text{corr}\left[d^{\text{WID}}_{2,k}(i+1), d^{\text{WID}}_{1,k}(i)\right] = 1$. Other correlations in (5.12) are determined

Table 5.1: Different Buffer Insertion Strategies for an Interconnect.

# Buffers	10	14	20	30	40	50	60
length [μm]	1000	714	500	333	250	200	167
min W_n [μm]	1.8	1.5	1.2	0.9	0.9	0.9	0.6

based on the existing models. For instance, WID variations can be modeled as independent [91] or spatially-correlated [37, 86], as introduced in Section 3.2.2.

With the delay model of a buffer stage fitted through (5.1) and the skitter model described through (5.2) to (5.12), the skitter between different paths in a clock tree can be estimated quickly. Consequently, the effect of process variations and power supply noise can, simultaneously, be modeled.

5.2.3 Skitter for different buffer insertions

The skitter is determined by the varying delay of all buffer stages due to process variations and power supply noise. For the same pair of clock paths, the effect of the number and size of buffers on skitter is investigated in this subsection. The electrical parameters of the transistors are based on a 32 nm PTM model [41]. The variation in channel length ($\sigma^{\text{D2D}} = 3\%\mu$ and $\sigma^{\text{WID}} = 5\%\mu$ based on ITRS data [43]) is considered in the simulations. Note that different sources of variations can also be modeled by the proposed modeling approach. The parameters of the interconnects are based on an Intel 32 nm interconnect technology [94]. The resistance, inductance, and capacitance of the interconnects per unit length are 388.007 Ω/mm, 68.683 fF/mm, and 1.768 nH/mm, respectively.

The skitter including skew and period jitter between two paths of a clock tree are investigated. Considering two clock paths with a length of 10 mm, seven cases of buffer insertion are investigated, as listed in Table 5.1. The maximum size of the investigated nMOS transistors is assumed to be 22.5 μm. The size of the pMOS transistors is twice the W_n to produce close to equal rise and fall times. The accuracy of the proposed methodology to estimate $J_{1,2}$ is verified in the following subsection. The case, where the paths contain 20 buffers with $W_n = 3\,\mu$m, is taken as an example. The result is compared with SPICE based Monte-Carlo simulations [148].

The efficiency of different buffer insertion cases in reducing $J_{1,2}$ is then discussed. The buffer insertion can be driven by considering 1) only process variations, 2) only the power supply noise, and 3) both of these sources of variations, respectively. The skitter of the interconnects with different length, the tradeoff between power consumption and skitter, and the effect of recombining clock paths and dynamic voltage scaling in reducing skitter are presented in this section.

Table 5.2: Comparison between the Proposed Modeling Method and Monte-Carlo Simulations.

n_p	μ_M [ps]	μ_{MC} [ps]	μ%	σ_M [ps]	σ_{MC} [ps]	σ%
0	-57.63	-54.2	5.4%	32.9	29.5	11.4%
10	-57.63	-55.0	3.8%	22.4	22.5	-0.2%
18	-57.63	-54.6	4.6%	12.4	11.9	3.9%

Accuracy of the proposed methodology

The accuracy of $J_{1,2}$ obtained from (5.2) through (5.12) is verified by comparing with SPICE-based Monte-Carlo simulations. Twenty buffers are inserted along one interconnect. For a fixed $\phi = \frac{3}{2}\pi$ in (5.6), three cases of n_p are examined, $n_p = 0, 10, 18$. The estimated $\mu_{J_{1,2}}$ and $\sigma_{J_{1,2}}$ and the results from Monte-Carlo simulations are listed in Table 5.2. The mean delay from the proposed model and the Monte-Carlo simulations are denoted by μ_M and μ_{MC}, respectively. As reported in Table 5.2, for all the three cases of n_p, the proposed model exhibits reasonably high accuracy (below 5.4% for μ and below 11.4% for σ).

For $n_p = 0$, different initial noise phases ϕ are also examined. The $\mu_{J_{1,2}}$ and $\sigma_{J_{1,2}}$ from the proposed model and the Monte-Carlo simulations (MC) are illustrated in Fig. 5.4. Since $J_{1,2}$ is approximated as a Gaussian distribution based on (5.2) - (5.12), the probability for $J_{1,2}$ to lie within the range $[\mu - 3\sigma, \mu + 3\sigma]$ is 99.7%. The negative $J_{1,2}$ with the maximum absolute value can be expressed as $\max(J_{1,2}) = \mu_{J_{1,2}} - 3\sigma_{J_{1,2}}$, which results in the shortest time period for data transfer. The $\max(J_{1,2})$ from the proposed model and the Monte-Carlo simulations is also illustrated in Fig. 5.4.

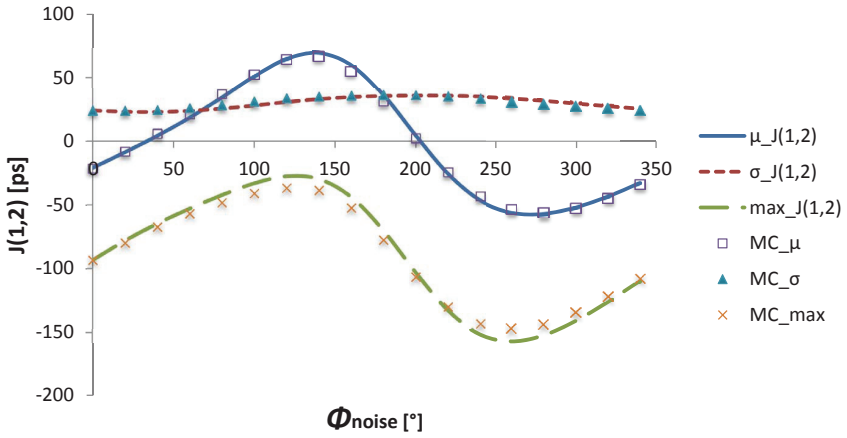


Figure 5.4: $\mu_{J_{1,2}}$ and $\sigma_{J_{1,2}}$ from the proposed modeling method and Monte-Carlo simulations (notated by "MC").

As shown in this figure, the proposed modeling method produces reasonable accuracy for different ϕ. The worst $\max(J_{1,2})$ or the *worst case period jitter* (WJ), occurs where $\phi = \frac{3}{2}\pi$ (270°). This behavior is consistent with the conclusion made in [94], when $f_n \ll f_{clk}$. Consequently,

$\phi = \frac{3}{2}\pi$ is utilized and $J_{1,2}$ implies WJ in the remainder of this section. In this case, $\mu_{J_{1,2}}$ and $\max(J_{1,2})$ are both negative and are described with absolute values for clarity.

Different objectives for buffer insertion

The three previously-mentioned objectives for performing buffer insertion are compared in this subsection. The resulting number and size of buffers are also presented. The slew rate (rise time) for different buffer insertions is investigated, as shown in Fig. 5.5. Since the rise time for 10 inverters is greater than 75 ps, these solutions are not considered in the following analysis.

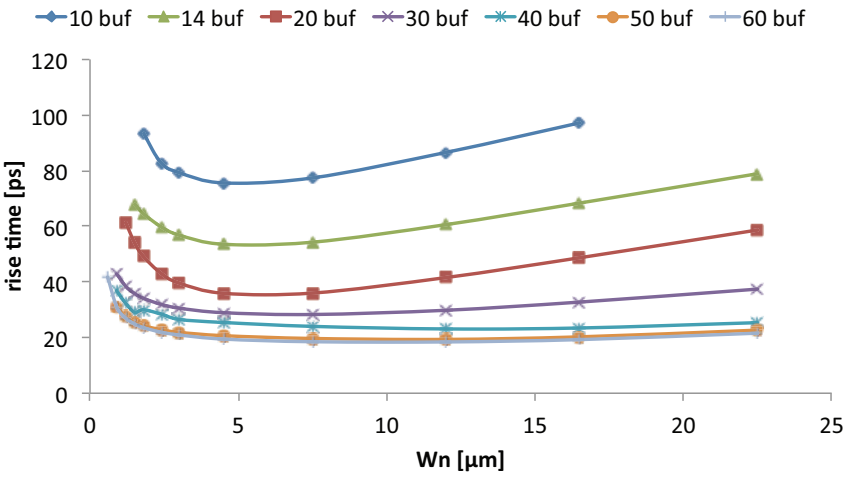


Figure 5.5: Mean slew rate for different buffer insertion under process variations and power supply noise.

Buffer insertion under process variations

There are plenty of works focusing on buffer insertion considering process variations [167, 168]. In these methodologies, the buffers are inserted to reduce both the delay and power while alleviating the delay uncertainty due to process variations. All the buffer stages are considered to be supplied with a constant V_{dd} (instead of using (5.1) to determine the distribution of the delay and skew). Consequently, the period jitter J_1 and J_2 in Fig. 5.3(b) are neglected. The variation of skew $S_{1,2}$ determines $J_{1,2}$.

The buffer insertion cases and the resulting $\sigma_{J_{1,2}}$ from Monte-Carlo simulations are illustrated in Fig. 5.6, where only process variations are considered. The lowest $\sigma_{1,2}$ is achieved for 14 buffers with $W_n = 12\,\mu$m. The resulting minimum $\sigma_{1,2}$ is 21.17 ps.

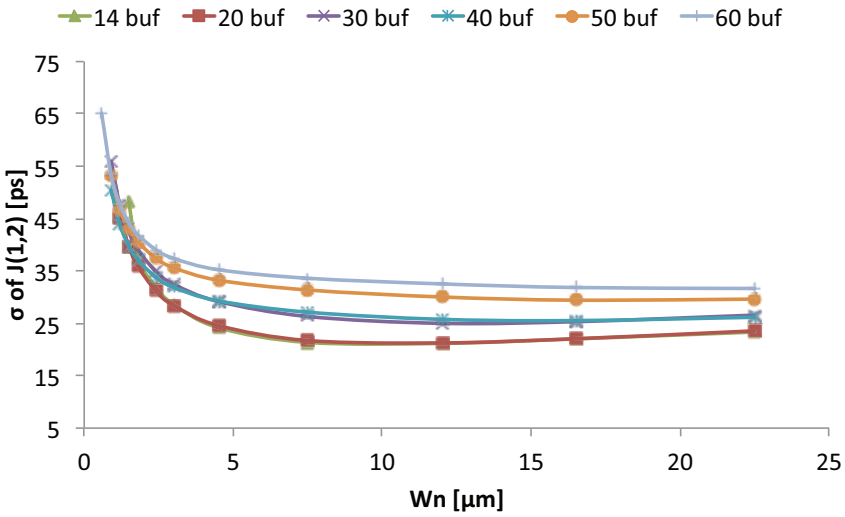


Figure 5.6: $\sigma_{J_{1,2}}$ for different buffer insertion under process variations.

Buffer insertion under power supply noise

There are also existing works on buffer insertion considering the power supply noise [94]. In these works, the effect of the power supply noise on clock jitter is modeled and buffers are inserted such that this effect is suppressed. In this case, process variations are not considered. Consequently, $S_{1,2}$ and $J_{1,2}$ are constant for a given power supply noise scenario.

The WJ from SPICE based simulations for different numbers and sizes of buffers is illustrated in Fig. 5.7. The lowest WJ is achieved by 14 buffers but with $W_n = 7.5\,\mu$m. The resulting minimum $\mu_{J_{1,2}}$ is 36.2 ps. Solutions with fewer buffers produce lower WJ.

Buffer insertion under both process variations and power supply noise

Since the process variations and power supply noise coexist in a real circuit, investigating the combined effect of these variations is necessary. Skitter $J_{1,2}$ combining $S_{1,2}$ and J_2 can be obtained from (5.2) to (5.12). In this case, both the effect of process and voltage variations are considered to determine the size and number of buffers.

The max($J_{1,2}$) from Monte-Carlo simulations for different buffer solutions is illustrated in Fig. 5.8(a). In this example, the minimum $\mu_{J_{1,2}}$, $\sigma_{J_{1,2}}$, and max($J_{1,2}$) from different buffer insertions are 35.7 ps, 22.36 ps, and 102.98 ps, respectively. The corresponding solutions are 14 buffers with $W_n = 7.5\,\mu$m, $12\,\mu$m, $12\,\mu$m, respectively. The solution with fewer buffers, again, produces lower $J_{1,2}$. The comparison in $\mu_{J_{1,2}}$ and $\sigma_{J_{1,2}}$ between the proposed model and Monte-Carlo simulations for different numbers of buffers ($W_n = 7.5\,\mu$m) is reported in Table 5.3. As reported in this table, for the clock paths with different numbers of buffers, the proposed model exhibits

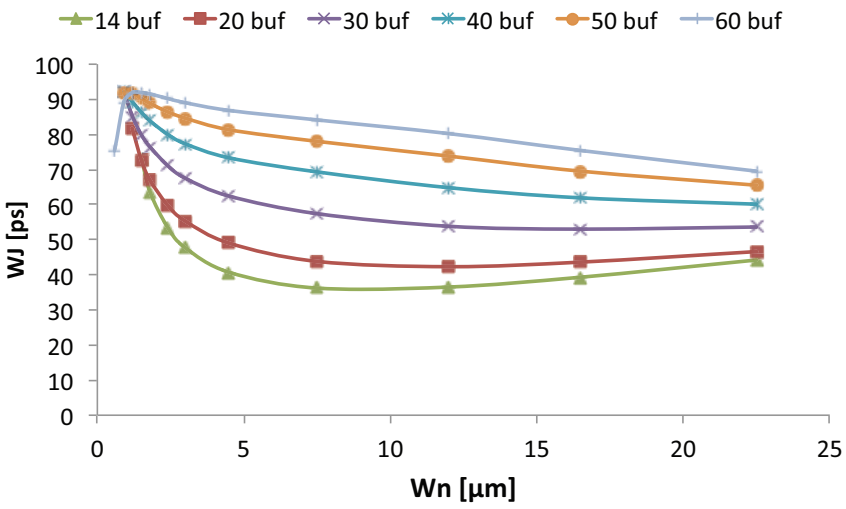
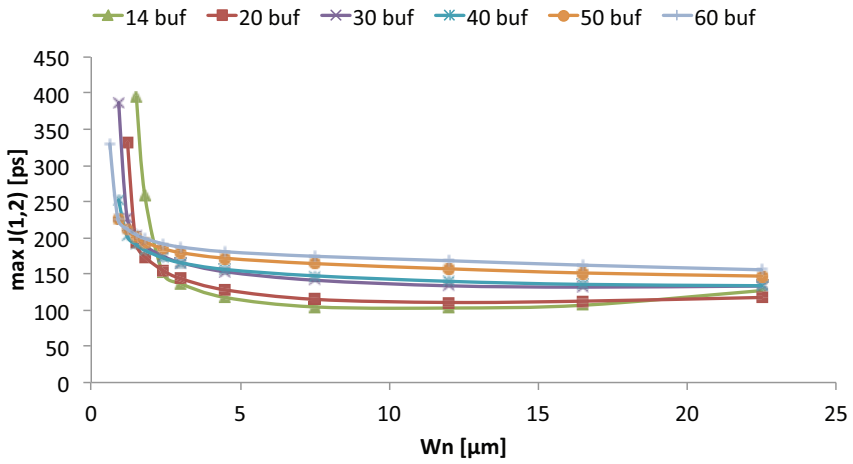


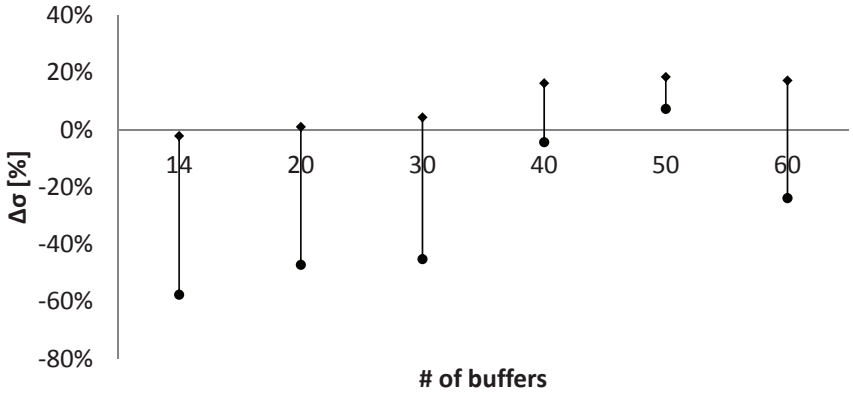
Figure 5.7: $WJ_{1,2}$ for different buffer insertions under power supply noise.

Table 5.3: Comparison between the proposed modeling method and Monte-Carlo simulations for different numbers of buffers.

# buf	μ_M [ps]	μ_{MC} [ps]	μ%	σ_M [ps]	σ_{MC} [ps]	σ%
14	-33.3	-35.7	-6.9%	22.5	22.9	-1.8%
20	-39.4	-43.3	-9.1%	22.5	23.9	-5.8%
30	-51.2	-57.1	-10.3%	25.7	28.1	-8.4%
40	-64.0	-69.0	-7.3%	29.8	26.2	14.0%
50	-73.9	-77.7	-4.8%	33.0	28.9	14.0%
60	-80.8	-82.7	-2.3%	34.7	30.6	13.4%



(a)



(b)

Figure 5.8: $J_{1,2}$ for different buffer insertion under process variations and power supply noise. (a) is the maximum $J_{1,2}$. The max and min difference on $\sigma_{J_{1,2}}$ between PV only and PV&PSN is shown in (b).

reasonable accuracy (below 10% for μ and below 14% for σ). For clarity, the skitter is described by the results from Monte-Carlo simulations in the remainder of this section.

Comparing the results of the three considerations for buffer insertion, it is shown that under process and voltage variations, the mean of the resulting $J_{1,2}$ is dominated by power supply noise (the difference in $\mu_{J_{1,2}}$ between considering power supply noise only (PSN) and considering both process variations and power supply noise (PV&PSN) is typically below 2%). This behavior is because $\mu_{J_{1,2}}$ is the linear combination of the mean delay of each buffer stage as expressed by (5.8), which is mainly determined by the power supply noise.

The difference between the $\sigma_{J_{1,2}}$ considering process variations only (PV) and PV&PSN is reported in Fig. 5.8(b). The non-negligible $\Delta\sigma_{J_{1,2}}$ is reported for the clock paths with different

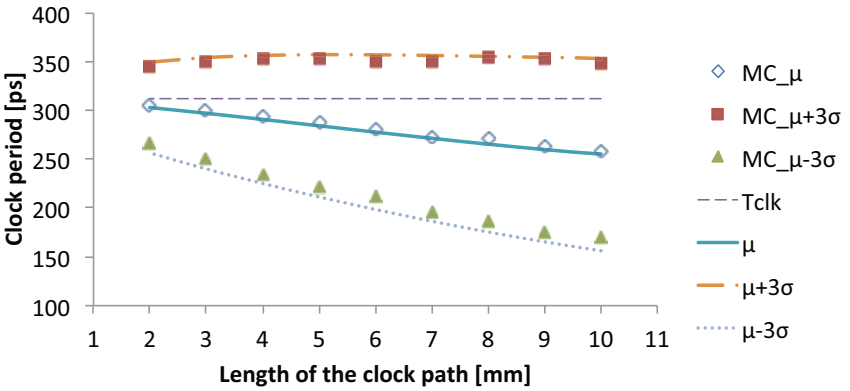


Figure 5.9: Skew and jitter with different length of clock paths.

numbers of buffers. The $\Delta\sigma_{J_{1,2}}$ for 14 buffers is the highest, although this number of buffers actually produces the lowest skitter. Modeling PV and PSN simultaneously is, therefore, necessary to estimate the variation of $J_{1,2}$.

Skitter for various lengths of the clock path

As the length of the clock path changes, the skitter also differ. An example of clock skew and jitter for different interconnect length is drawn in Fig. 5.9. The same buffers ($W_n = 3\,\mu$m) are inserted at the same distance (500 μm) for all the clock paths. The ideal clock period ($T_{\text{clk}} = 312.5$ ps) is denoted by the dashed line. The actual mean ($T_{\text{clk}} - \mu_{J_{1,2}}$), the highest ($T_{\text{clk}} - \mu_{J_{1,2}} + 3\sigma_{J_{1,2}}$), and the lowest ($T_{\text{clk}} - \mu_{J_{1,2}} - 3\sigma_{J_{1,2}}$) periods within 99.7% confidence range are denoted by ◊, ■, and ▲, respectively.

As shown in Fig. 5.9, the skitter increases with the length of the clock path, given the same buffer insertion. The mean and the variation of the period jitter increase with the interconnect length. The largest clock period, however, remains nearly constant as the interconnect length varies, since the increase in period jitter and skew counteract each other. The results from the proposed model are also illustrated in Fig. 5.9, which fit well with Monte-Carlo results.

Power consumption with constraints on skitter

The power consumed by clock distribution networks still constitutes a significant portion of the total power consumed by a circuit [18, 169]. The power consumption of the clock network under different constraints on skitter is investigated in this subsection.

For the investigated clock paths, the total power consumption under different constraints on $\max(J_{1,2})$ is illustrated in Fig. 5.10. As shown in this figure, when $\max(J_{1,2}) \geq 220$ ps, all the buffer insertions approximately consume the same power. As the constraint becomes stricter ($\max(J_{1,2})$ decreases), the power increases and the solutions with fewer buffers are

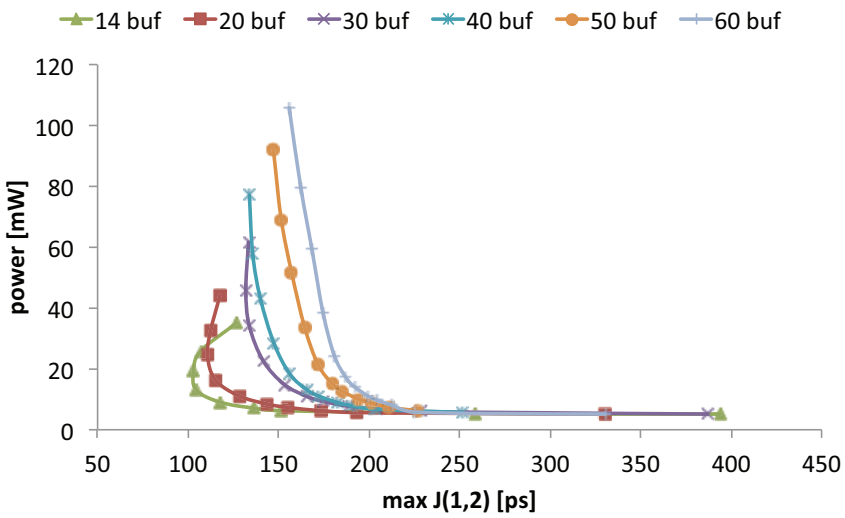


Figure 5.10: Power consumption vs. max($J_{1,2}$) for different buffer insertions.

more power-efficient. The solution with 14 buffers consumes the lowest power and meets the constraint on max($J_{1,2}$).

The constraint on max($J_{1,2}$) ≥ 115 ps can be met with low power overhead. Nevertheless, as the constraint becomes lower than 115 ps, significant power overhead is shown. For example, to decrease the max($J_{1,2}$) from 118 ps to 103 ps (13% improvement), the 14 buffers inserted along each clock path are sized up from 4.5 μm to 12 μm. The resulting power consumption increases from 9.1 mW to 19.2 mW (110% increase). In conclusion, pursuing extreme constraints on clock skew and period jitter results in buffer insertions with high power consumption.

Power consumption with constraints on slew rate

The power consumed by a clock path under different constraints on the slew rate is investigated in this subsection. The output slew is denoted by the rise time at the clock sinks.

The power consumption under different constraints on the rise time is illustrated in Fig. 5.11. In contrast with the buffer insertion solutions minimizing max($J_{1,2}$), the clock path with more and smaller buffers produces a lower output slew (higher slew rate). As shown in Figs. 5.5 and 5.11, the minimum slew rate of the clock path with 14 buffers is much higher than other solutions. Consequently, the solution with 14 buffers can not be used for the clock paths with the strict slew constraint, although this solution produces the lowest max($J_{1,2}$). Among the other approaches, a clock path with 60 buffers consumes the lowest power under the same constraint on slew rate.

Similar to the results in Fig. 5.10, the increase in power becomes severe as the slew constraint becomes stricter (slew rate decreases). For example, as the slew constraint decreases from 21 ps

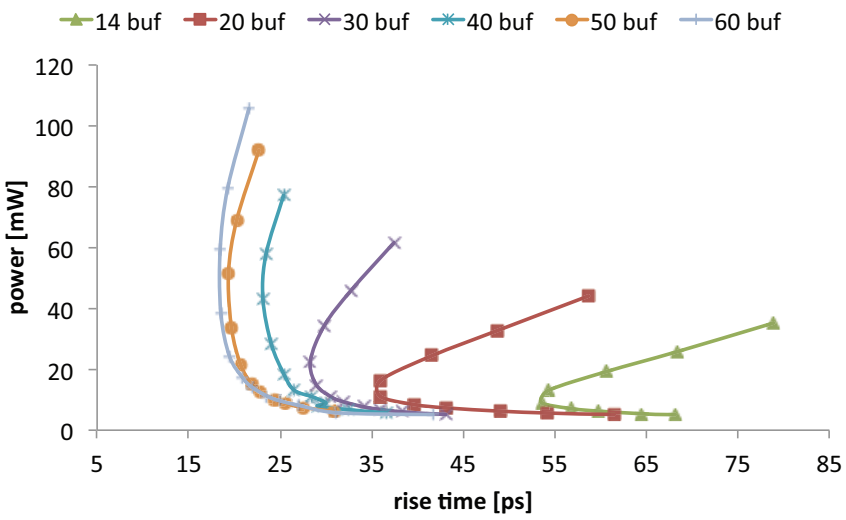


Figure 5.11: Power consumption vs. output slew for different buffer insertions.

to 18.5 ps (12% decrease), the size of the buffers increases from 3 μm to 7.5 μm. The resulting power consumption increases from 17 mW to 38.5 mW (126% increase). In conclusion, pursuing extreme constraints on slew rate also results in high power overhead.

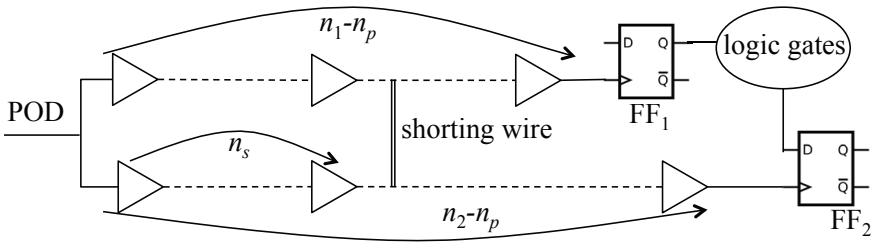
5.2.4 Decreasing skitter in 2-D circuits

Different methods can be utilized to mitigate the skitter by reducing skew and jitter. Two typical methods, recombinant trees and dynamic voltage scaling, are investigated in this subsection.

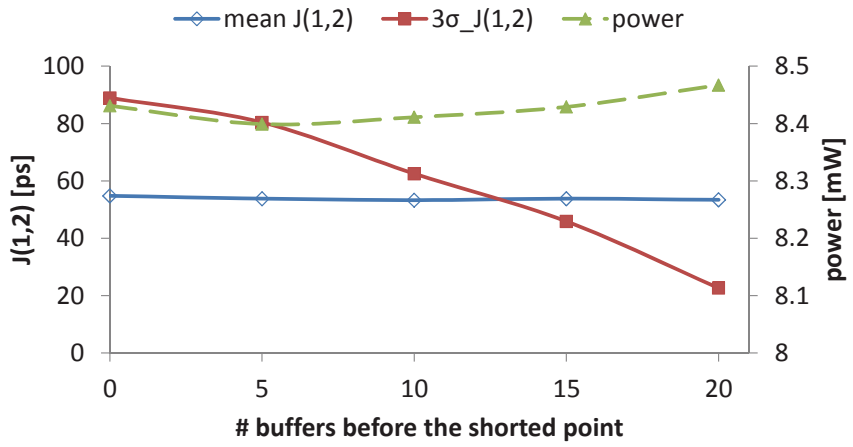
Mitigating skitter with recombinant clock paths

Recombining clock paths (*e.g.*, in binary trees and clock spines) can mitigate skew by shorting different paths at the output of the clock buffers [18, 94]. The interconnects can be shorted at different levels along the clock path, as depicted in Fig. 5.12(a). By shorting the interconnects at different positions along the clock path, the number of shorted clock buffers n_s varies from 0 to $\max(n_1, n_2) - n_p$. The skew and jitter for the clock paths with different n_s are illustrated in Fig. 5.12(b), where $n_1 = n_2 = 20, n_p = 0, n_s = \{0, 5, 10, 15, 20\}, W_n = 3\,\mu$m.

As illustrated in Fig. 5.12(b), $3\sigma_{J_{1,2}}$ significantly decreases with n_s. The mean skitter $\mu_{J_{1,2}}$ between two clock paths is, however, not affected by the position of the shorted interconnect. This situation is due to the symmetry between the clock paths. The mean skitter is mainly determined by the supply noise. Since all the buffers are uniformly affected by resonant noise, the mean clock delay of these clock paths is similar with each other. Consequently, shorting clock paths cannot decrease mean skitter in this case. In other words, the variation of skitter is



(a)



(b)

Figure 5.12: Skitter and power with the shorted wire at different levels of clock paths. The number of buffers before the shorted point is denoted by n_s.

highly reduced by shorting the clock paths at the clock sinks, while the mean skitter cannot be reduced by the shorted interconnect. As n_s increases, $\mu_{J_{1,2}}$ becomes higher than $3\sigma_{J_{1,2}}$. This behavior shows that the period jitter caused by the power supply noise becomes dominant as the skew variation is reduced by recombinant trees. The power consumed by clock buffers increases slightly with n_s, which indicates that the power does not vary a lot while shorting the buffers at different levels between two branches.

Effect of DVS on skitter

The effect of *dynamic voltage scaling* (DVS) on skitter is discussed in this subsection. DVS is an efficient method to mitigate the impact of PVT variations on data transfer [166, 170]. Since DVS is commonly applied to the circuit block by block, the supply voltage for the data paths and the clock distribution networks are both tuned. For example, consider the setup slack t_{slack} between FF_1 and FF_2 in Fig. 5.3(a),

$$t_{\text{slack}} = T_{1,2} - D_{1,2} - t_{\text{setup}} = T_{\text{clk}} - t_{\text{setup}} + J_{1,2} - D_{1,2}. \tag{5.13}$$

The delay $D_{1,2}$ is the propagation time of data from the clock input pin of FF_1, through the logic gates between FF_1 and FF_2, to the data input pin of FF_2. The setup time of FF_2 is denoted by t_{setup}, which is constant.

A positive t_{slack} is required for the data to be successfully latched by FF_2. Since both $J_{1,2}$ and $D_{1,2}$ are affected by process variations and power supply noise, DVS can be used to ensure a positive t_{slack} by voltage scaling [166]. The voltage (consequently, the delay) of logic gates is adjusted according to the measured delay variation. The clock buffers within the adjusted circuit block are also affected by the scheduled supply voltage. An example of the skitter due to different V_{dd} is illustrated in Fig. 5.13.

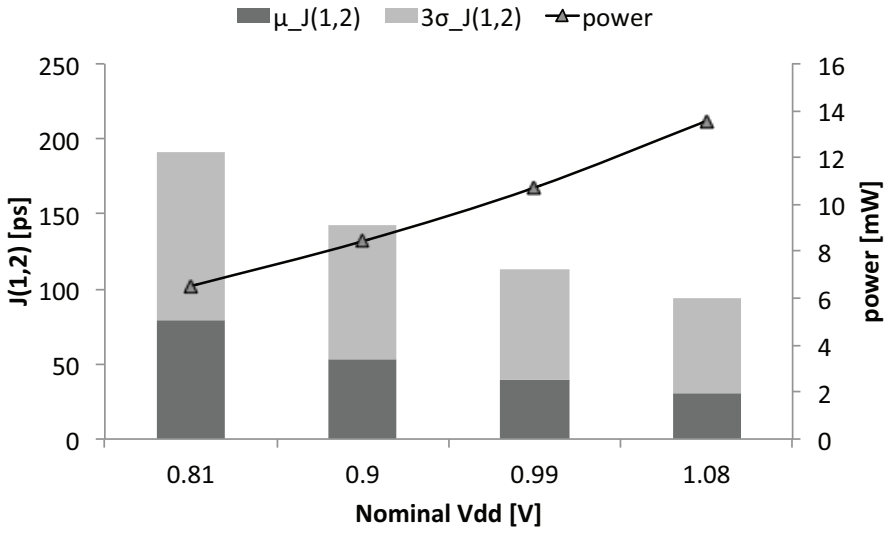


Figure 5.13: Skitter between two branches vs. supply voltage.

The skitter between two clock branches with 20 clock buffers ($W_n = 3\,\mu$m) along each branch is shown in Fig. 5.13. By increasing V_{dd}, both the mean and variation of the skitter decrease. The maximum skitter is, therefore, reduced. Regarding the delay variation of a buffer stage shown in Fig. 5.2, both the mean and variation of the delay decrease with V_{dd}. As a result, the induced skitter decreases.

Since $J_{1,2}$ is negative in this example, decreasing $|J_{1,2}|$ facilitates satisfying (5.13). Increasing V_{dd} can improve the performance of the circuit by both speeding up the logic gates and reducing the skitter of the clock distribution networks. The power consumed by the clock buffers, however, increases quadratically with V_{dd}.

5.3 Extending the Skitter Model to 3-D Clock Distribution Networks

The simplified model of skitter proposed in the previous section provides a fast method to estimate skitter in 2-D ICs. In 3-D ICs, however, high inaccuracy can be introduced due to

more complicated process variations and different supply noise among tiers. To accurately describe the skitter in 3-D clock distribution networks, an extended skitter model is proposed in this section. The delay models for buffers and interconnects are presented in the following subsection. Both the setup and hold skitter are modeled in Section 5.3.2. The skitter $J_{1,2}$ is compared for clock paths with different lengths in Section 5.3.3. Skitter under different scenarios of power supply noise is discussed in Sections 5.3.4 - 5.3.6. In addition, the skitter for different buffer insertions along 3-D clock paths is discussed. The tradeoff between skitter and power is also presented.

5.3.1 Linear statistical model for buffers and interconnects

The distribution of the delay of a buffer stage is modeled in this subsection. The delay of a buffer stage d consists of the delay of the buffer d_{b} and the interconnect (horizontal wire and/or TSV) d_{I}. The variation of d is a random variable affected by both process variations and power supply noise.

Delay variation due to process variations

Since the variation of parameters due to process variations is typically within a small range, the delay of a buffer stage considering the parameter variations can be approximated by the first order Taylor expansion [86],

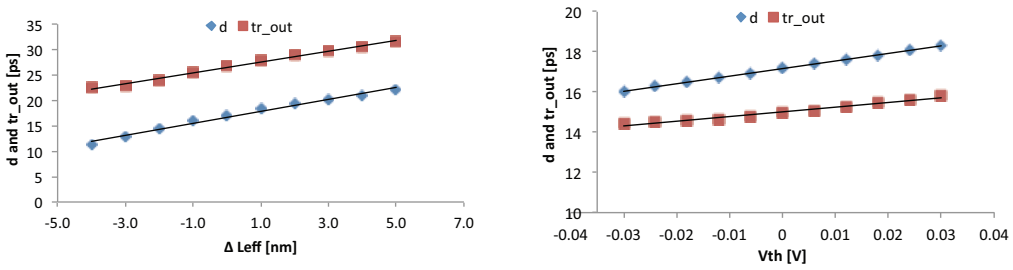
$$\begin{aligned} d(tr,\vec{P},C_{\mathrm{lw}}) &= d_{\mathrm{b}}(tr,\vec{P},C_{\mathrm{lb}}) + d_{\mathrm{I}}(\vec{P},C_{\mathrm{lw}}) \\ &\approx \overline{d} + \sum_{p\in\vec{P}} \left(\frac{\partial d}{\partial p}\Big|_0 \Delta p \right). \end{aligned} \tag{5.14}$$

The input slew of this buffer stage is denoted by tr. The capacitive load seen at the output of the buffer and wire is denoted by C_{lb} and C_{lw}, respectively. The nominal delay is $\overline{d}$ and the subscript "0" denotes the partial derivative with nominal parameters. The set of parameters affected by process variations is denoted by $\vec{P}$. Each parameter is modeled by a random variable. For instance, if the variation of channel length of three buffers is considered, $\vec{P}$ is $\{L_{\mathrm{b},1}, L_{\mathrm{b},2}, L_{\mathrm{b},3}\}$. The variation of a parameter Δp consists of WID and D2D variations,

$$\Delta p = \Delta p_{\mathrm{WID}} + \Delta p_{\mathrm{D2D}}, \tag{5.15}$$

where Δp_{D2D} is consistent among buffers (interconnects) within the same die, while Δp_{WID} varies among the components within the same die [92]. The partial derivatives in (5.14) are determined by

$$\begin{aligned} \frac{\partial d}{\partial p} = &\frac{\partial d_{\mathrm{b}}}{\partial tr}\frac{\partial tr}{\partial p} + \frac{\partial d_{\mathrm{b}}}{\partial C_{\mathrm{lb}}}\frac{\partial C_{\mathrm{lb}}}{\partial p} + \frac{\partial d_{\mathrm{b}}}{\partial p} \\ &+ \frac{\partial d_{\mathrm{I}}}{\partial C_{\mathrm{lw}}}\frac{\partial C_{\mathrm{lw}}}{\partial p} + \frac{\partial d_{\mathrm{I}}}{\partial p}. \end{aligned} \tag{5.16}$$



(a) (b)

Figure 5.14: Change of the delay and output transition time with (a) effective channel length and (b) threshold voltage.

The partial derivatives in (5.16) are determined by the expressions of d_{b} and d_{I}. The expression of $d_{\mathrm{b}}(tr, \vec{P}, C_{\mathrm{lb}})$ can be obtained through analytic formulas [103] or adjoint sensitivity analysis with SPICE-based simulations. To achieve higher accuracy, SPICE-based sensitivity analysis is used in this dissertation. For instance, considering a PTM 32 nm CMOS model [41], for a rising input, the change of d_{b} and the output transition time $t_{\mathrm{r,out}}$ with ΔL_{eff} and the threshold voltage ΔV_{th} is shown in Fig. 5.14. For horizontal wires, the expression of $d_{\mathrm{I}}(\vec{P}, C_{\mathrm{lw}})$ is determined by the *RLC* interconnect model proposed in [113, 171].

The variations introduced by TSVs have been discussed in [172, 173], where the TSV stress-induced delay variation of buffers is well modeled. In the following sections, the keep-out-zone of TSVs is assumed to be large enough ($\leq 10\mu$m [67, 172]) to mitigate the effect of TSV stress. Consequently, TSVs are modeled as *RLC* wires with different electrical characteristics from the horizontal interconnects. The effect of the variation of TSVs on the skitter is discussed in Section 5.3.3.

Delay variation simultaneously considering process variations and power supply noise

Similar to the transistor parameters affected by process variations, supply voltage can be treated as an additional random variable in (5.14). The first step is to correctly model supply noise $v(t)$, as introduced in Section 4.4. The non-recursive expression of supply noise is used in Section 5.2. To improve the accuracy of the skitter model, the recursive expression is used in the extended model.

Assuming a clock edge arrives at the source of a clock path at time zero, t_j is the time when this clock edge arrives at buffer j. The supply noise to buffer j at time t_j can be expressed as $v(t_j)$. As aforementioned, to investigate the effect of worst supply noise on clock distribution networks, (4.18) can be approximated by an undamped sinusoidal waveform [94],

$$v(t_j) \approx V_{\mathrm{n}} \sin(2\pi f_{\mathrm{n}} t_j + \phi), \tag{5.17}$$

$$t_j = \sum_{i=1}^{j-1} d_i. \tag{5.18}$$

According to (5.14), d_i, t_j, and $v(t_j)$ are all random variables. Since Δt_j is low as compared with $\overline{t}_j$, $v(t_j)$ can also be approximated by the first order Taylor expansion,

$$v(t_j) = \overline{v}(t_j) + \Delta v(t_j) \approx \overline{v}(t_j) + \frac{\partial v(t_j)}{\partial t_j}\Big|_0 \Delta t_j, \tag{5.19}$$

$$\Delta v(t_j) \approx 2\pi V_{\rm n} f_{\rm n} \cos(2\pi f_{\rm n} \overline{t}_j + \phi) \sum_{i=1}^{j-1} \Delta d_i. \tag{5.20}$$

In 3-D ICs, different resonant supply noise among tiers is discussed in Section 4.3. Based on the electrical characteristics of different PDNs, $V_{\rm n}$, $f_{\rm n}$, and ϕ can differ among tiers. Within each tier, the corresponding $V_{\rm n}$, $f_{\rm n}$, and ϕ are used in (5.19) to describe the supply noise at different time instances.

According to (5.14), the delay variation Δd is also affected by the input slew Δtr, which is determined by the previous buffer stage. Considering the effect of Δv and Δtr on Δd, the delay variation of the $j^{\rm th}$ buffer stage can be modeled as

$$\Delta d_j \approx \sum_{p \in \vec{P}_j} \left(\frac{\partial d_j}{\partial p}\Big|_0 \Delta p \right) + \frac{\partial d_j}{\partial v}\Big|_0 \Delta v(t_j) + \frac{\partial d_j}{\partial tr}\Big|_0 \Delta tr_j. \tag{5.21}$$

The set of statistical parameters of the $j^{\rm th}$ buffer stage is denoted by $\vec{P}_j$, which is a subset of the entire parameter set, $\vec{P}_j \subseteq \vec{P}$. The input slew of the $j^{\rm th}$ buffer stage Δtr_j can be determined similar to (5.21),

$$\Delta tr_{\rm j} \approx \sum_{p \in \vec{P}_j} \left(\frac{\partial tr_j}{\partial p}\Big|_0 \Delta p \right) + \frac{\partial tr_j}{\partial v}\Big|_0 \Delta v(t_{j-1}) + \frac{\partial tr_j}{\partial tr_{j-1}}\Big|_0 \Delta tr_{j-1}. \tag{5.22}$$

Substituting (5.20) and (5.22) into (5.21), Δd_j can be recursively determined considering both process variations and power supply noise. The coefficients in (5.21) and (5.22) are obtained through adjoint sensitivity analysis as previously mentioned. The resulting expression (5.21) is used to determine skitter in the following subsection.

5.3.2 Modeling setup and hold skitter

For a pair of clock paths in 3-D ICs, the definition of the clock skew, period jitter, and skitter are redrawn in Fig. 5.15 to better follow the discussion in the remainder of the chapter. The clock signal is fed into the 3-D clock tree from the primary clock driver. Two flip-flops are driven by this clock signal, denoted as $\rm FF_1$ and $\rm FF_2$, respectively. The numbers of buffers from the clock input to $\rm FF_1$ and $\rm FF_2$ are denoted by $n_1 + n_2$ and $n_3 + n_4$, respectively.

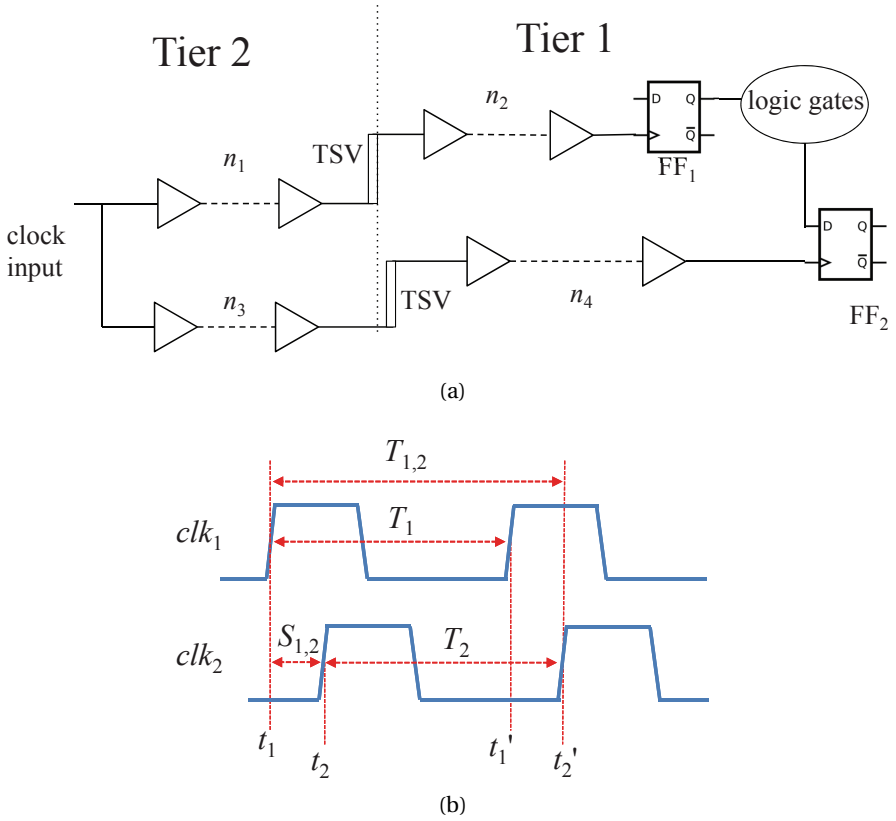


Figure 5.15: Clock uncertainty between 3-D clock paths. Two paths and flip-flops are illustrated in (a). The corresponding clock signals are shown in (b).

The waveforms clk_1 and clk_2 in Fig. 5.15(b) correspond to the clock signal driving FF_1 and FF_2, respectively. The time where the first rising edge in Fig. 5.15(b) arrives at the clock input is defined as the origin. The time when this edge arrives at FF_1 and FF_2 is, respectively, denoted by t_1 and t_2. The arrival time of the next rising edge at FF_1 and FF_2 is t_1' and t_2', respectively. The skew between the first edge of clk_1 and clk_2 is $S_{1,2}$. The measured clock periods after the first edge for FF_1 and FF_2 are T_1 and T_2, respectively. The ideal clock period is T_{clk}. The corresponding period jitters are $J_1 = T_1 - T_{\text{clk}}$ and $J_2 = T_2 - T_{\text{clk}}$.

The effect of skitter on setup time slack

Assuming the data is transferred from FF_1 to FF_2 within one clock cycle, $T_{1,2}$ is the time interval that affects the highest clock frequency of the circuit. The setup time requirement needs to be satisfied for the system to work correctly [18]. The setup time slack $slack_{\text{setup}}$ is defined as

$$slack_{\text{setup}} = T_{1,2} - \max(D_{1,2}) - t_{\text{setup}}, \tag{5.23}$$

$$T_{1,2} = (t_2 - t_1) + T_2 = S_{1,2} + J_2 + T_{\text{clk}}, \tag{5.24}$$

where $\max(D_{1,2})$ denotes the longest data transfer time from FF_1 to FF_2. The setup time for FF_2 is t_{setup}, specified in the cell library. Consequently, the variation of $slack_{\text{setup}}$ is affected by the variation of $T_{1,2}$, called "setup skitter" $J_{1,2}$, which is similar to the skitter defined in the previous section,

$$J_{1,2} = S_{1,2} + J_2 = t_2' - t_1 - T_{\text{clk}}. \tag{5.25}$$

To avoid setup time violations, $slack_{\text{setup}} \geq 0$ is required in any operating condition. This requirement means that sufficient $T_{1,2}$ should be provided for data transfer. For this purpose, as shown in (5.24), T_{clk} should be large enough to compensate the worst $J_{1,2}$. Consequently, clock skew and period jitter should be simultaneously modeled to determine the minimum T_{clk} (the highest clock frequency) of a circuit.

According to (5.18) and (5.25), skitter $J_{1,2}$ is the linear combination of the delay of buffer stages,

$$J_{1,2} = \sum_{k=1}^{n_3+n_4} d'_{2,k} - \sum_{k=1}^{n_1+n_2} d_{1,k}, \tag{5.26}$$

$$\overline{J}_{1,2} = \sum_{k=1}^{n_3+n_4} \overline{d'}_{2,k} - \sum_{k=1}^{n_1+n_2} \overline{d}_{1,k}, \tag{5.27}$$

$$\Delta J_{1,2} = \sum_{k=1}^{n_3+n_4} \Delta d'_{2,k} - \sum_{k=1}^{n_1+n_2} \Delta d_{1,k} \approx \sum_{p\in\vec{P}} \left(\frac{\partial J_{1,2}}{\partial p}\Big|_0 \Delta p \right), \tag{5.28}$$

where $d'_{2,k}$ is the delay of the k^{th} buffer stage along the path to FF_2 for the second clock edge. The mean skitter $\overline{J}_{1,2}$ is determined by the nominal delay of all the buffer stages considering the nominal voltage supply noise (without process variations). Substituting (5.21) into (5.28), the partial derivatives $\frac{\partial J_{1,2}}{\partial p}|_0$ are obtained. Consequently, skitter $J_{1,2}$ is approximated by the first order Taylor expansion.

Assuming all the parameters are described by Gaussian distributions, $\Delta J_{1,2}$ can also be approximated by a Gaussian distribution,

$$\Delta J_{1,2} \sim \mathcal{N}(0, \sigma^2_{J_{1,2}}), \tag{5.29}$$

$$\sigma^2_{J_{1,2}} = \sum_{p\in\vec{P}} \left(\frac{\partial J_{1,2}}{\partial p}\Big|_0^2 \sigma_p^2 \right) + 2 \sum_{p,q\in\vec{P}} \left(\frac{\partial J_{1,2}}{\partial p}\Big|_0 \frac{\partial J_{1,2}}{\partial q}\Big|_0 \text{cov}(p,q) \right), \tag{5.30}$$

where $\text{cov}(p,q)$ denotes the covariance between two parameters. Assuming D2D variations are independent from WID variations [86,88], $\sigma_p^2 = \sigma^2_{p(\text{D2D})} + \sigma^2_{p(\text{WID})}$. The covariance between two parameters is determined according to the tiers where these parameters are located and the spatial correlation between these parameters,

$$\text{cov}(p,q) = \begin{cases} 0, \text{ if } p,q \text{ are of different type or} \\ \quad \text{belong to different tiers,} \\ \text{cov}(p,q)_{\text{WID}} + \sigma_{p(\text{D2D})}\sigma_{q(\text{D2D})}, \text{otherwise,} \end{cases} \tag{5.31}$$

where the WID covariance $\text{cov}(p,q)_{\text{WID}}$ is determined by the spatial correlation between parameters p and q within the same tier. Statistically, the devices (wires) close to each other have higher correlation than those far from each other. This spatial correlation can be obtained from fabricated wafers [174] or through a spatial correlation model [86, 139].

As shown in (5.30) and (5.31), the variance of setup skitter $\sigma^2_{J_{1,2}}$ highly depends on the covariance between process-induced parameters. In 2-D ICs, the change of $\text{cov}(p,q)$ is mainly determined by $\text{cov}(p,q)_{\text{WID}}$, since the parameters of the same type are affected by the same D2D variations. Therefore, the distribution of clock paths only affects $\sigma^2_{J_{1,2}}$ by changing the WID covariance. In 3-D circuits, however, D2D variations vary among tiers and WID covariance among tiers is zero. Consequently, the distribution of clock paths will affect the skitter variation in a more complicated way.

The effect of skitter on hold time slack

In addition to the setup time slack, hold time slack also significantly affects the design of ICs. Hold time violations can also cause the failure of the entire system [18]. Moreover, this type of failure cannot be removed by lowering the clock frequency of the system. As illustrated in Fig. 5.15(b), the hold time slack is modeled as

$$slack_{\text{hold}} = \min(D_{1,2}) - S_{1,2} - t_{\text{hold}}, \tag{5.32}$$

where the hold time requirement t_{hold} is also specified in the cell library. The "hold skitter" affecting $slack_{\text{hold}}$ is determined by $S_{1,2}$, which is the skew between clk_1 and clk_2. Note that $S_{1,2}$ is affected by both process variations and power supply noise.

To correctly latch the data in FF_2, $slack_{\text{hold}} \geq 0$ is required to avoid hold time violations in any operating condition. From Fig. 5.15(b), $S_{1,2}$ can be determined as

$$\begin{aligned} S_{1,2} &= t_2 - t_1 = \sum_{k=1}^{n_3+n_4} d_{2,k} - \sum_{k=1}^{n_1+n_2} d_{1,k} \\ &\approx \sum_{k=1}^{n_3+n_4} \overline{d}_{2,k} - \sum_{k=1}^{n_1+n_2} \overline{d}_{1,k} + \sum_{p\in\vec{P}} \left(\frac{\partial S_{1,2}}{\partial p}\Big|_0 \Delta p \right). \end{aligned} \tag{5.33}$$

Similarly to (5.29) and (5.30), the distribution of $\Delta S_{1,2}$ can be modeled as

$$\Delta S_{1,2} \sim \mathcal{N}(0, \sigma^2_{S_{1,2}}), \tag{5.34}$$

$$\begin{aligned} \sigma^2_{S_{1,2}} = &\sum_{p\in\vec{P}} \left(\frac{\partial S_{1,2}}{\partial p}\Big|_0^2 \sigma_p^2 \right) + \\ &2 \sum_{p,q\in\vec{P}} \left(\frac{\partial S_{1,2}}{\partial p}\Big|_0 \frac{\partial S_{1,2}}{\partial q}\Big|_0 \text{cov}(p,q) \right), \end{aligned} \tag{5.35}$$

Table 5.4: Variations of Devices, Horizontal Wires, and TSVs

Parameters	Nominal	3σ (D2D)	3σ (WID)
Channel length [nm]	32	1.5	2.5
Threshold voltage [mV]	242	24.2	24.2
Wire width [nm]	225	22.5	11.3
Wire height [nm]	388	19.4	9.7
ILD thickness [nm]	252	18.9	9.5
TSV resistance [mΩ]	133	39.9	39.9
TSV capacitance [fF]	52	15.6	15.6

where the partial derivatives are obtained similar to the coefficients in (5.28). As shown through (5.14) to (5.35), both the setup and hold time violations are simultaneously affected by the process variations and power supply noise. Similar to the 2-D circuits shown in Section 5.2.3, it is also necessary to model process variations and power supply noise at the same time for 3-D ICs to accurately capture the clock uncertainty. The accuracy of the proposed model is verified for different lengths of clock paths in the following subsection. In 3-D ICs, in addition to different process variations among tiers, power supply noise can also differ from tier to tier, as discussed in Section 4.3. Consequently, setup and hold skitters under different scenarios of power supply noise in 3-D ICs are investigated in Sections 5.3.4 to 5.3.6.

5.3.3 Skitter *vs.* length of clock paths

The paths of a 3-D clock tree with different lengths are simulated. The electrical parameters of the transistors are based on a 32 nm PTM model [41]. The parameters of the interconnects are based on an Intel 32 nm interconnect technology [94]. The parameters of TSVs are based on data from [68]. Both the horizontal wires and TSVs are modeled by π segments in SPICE-based simulations. The proposed model is implemented in Matlab. All the simulations are performed in a Scientific Linux server (Intel Xeon 2.67 GHz, 24 Gb memory).

The variations considered in the simulations are listed in Table 5.4. The D2D and WID ΔL_{b} are extracted based on ITRS data [43]. The wire variations and ΔV_{th} are based on [86]. The variations of TSVs are based on [173]. Note that other sources of variations can also be described by the proposed modeling approach. For example, the TSV stress-induced delay variation in [172] can be included. In this case, the distribution of d_{B} in (5.14) is adapted based on the distance between the buffer and TSVs and the given expression of stress-induced buffer delay.

In the simulations, the length of clock paths ranges from 0.5 mm to 12.5 mm within 2- and 3-tier circuits. Buffers are inserted to produce a 10% T_{clk} input slew for the next stage. To emphasize the relation between skitter and the length of clock paths, all tiers are assumed to experience similar supply noise ($V_{\mathrm{n}} = 90$ mV, $f_{\mathrm{n}} = 400$ MHz, $\phi = 270^{\circ}$ [94]). Each pair of paths is averagely distributed across different tiers, as shown in Fig. 5.15(a). The resulting $\mu_{J_{1,2}}$ and

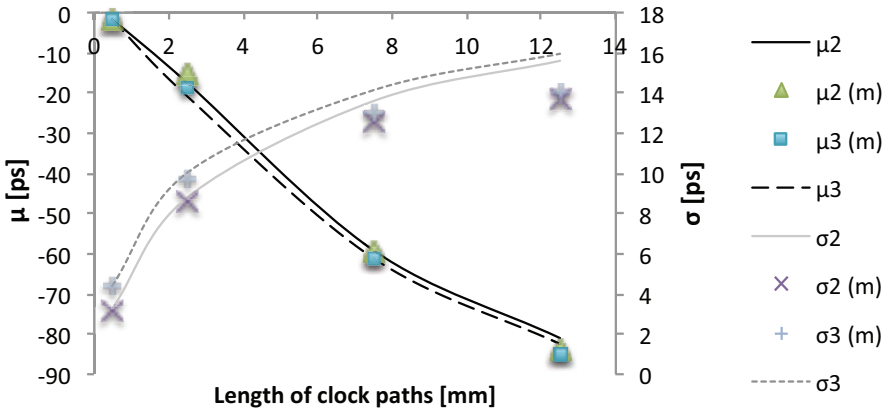


Figure 5.16: Skitter *vs.* length of 3-D clock paths.

Table 5.5: Effect of TSV Variations on Skitter

Length [mm]	0.5	2.5	7.5	12.5
$\sigma_{J_{1,2}}$ (0 ΔTSV) [ps]	4.40	10.03	14.14	15.95
$\sigma_{J_{1,2}}$ (5% ΔTSV) [ps]	4.35	10.36	13.50	16.78
$\sigma_{J_{1,2}}$ (15% ΔTSV) [ps]	4.95	10.96	13.93	17.00

$\sigma_{J_{1,2}}$ are illustrated in Fig. 5.16, where the suffixes "2" and "3" denote the results for 2- and 3-tier circuits, respectively.

The data from SPICE-based Monte-Carlo simulations and the proposed model (labeled with the (m)) are both depicted in Fig. 5.16. As shown in this figure, both $\mu_{J_{1,2}}$ and $\sigma_{J_{1,2}}$ deteriorate with the length of clock paths. This behavior can be described by the proposed model with a reasonable accuracy. The error of the proposed model is below 11% for $\mu_{J_{1,2}}$ and 12% for $\sigma_{J_{1,2}}$, respectively. Not surprisingly, long clock paths introduce high skitter in 3-D clock trees.

Observation 5.1. *Both the mean and standard deviation of skitter increase with the length of clock paths.*

The effect of TSV variations in the 3-tier clock paths is reported in Table 5.5. The skitter is listed for no TSV variation, 5% TSV variation ($\sigma/\mu = 5\%$), and 15% TSV variation, respectively. The difference in $\sigma_{J_{1,2}}$ among these three cases is around 1 ps for all the clock paths. This situation shows that TSV variations are a second-order effect, consistent with the results reported in [173].

5.3.4 Skitter *vs.* V_n in different tiers

3-D PDNs with different amplitudes of power supply noise among tiers are investigated in this subsection. Due to the different switching current in power supply networks and the vertical resistance of P/G TSVs among tiers, the devices in different tiers can be subjected to different

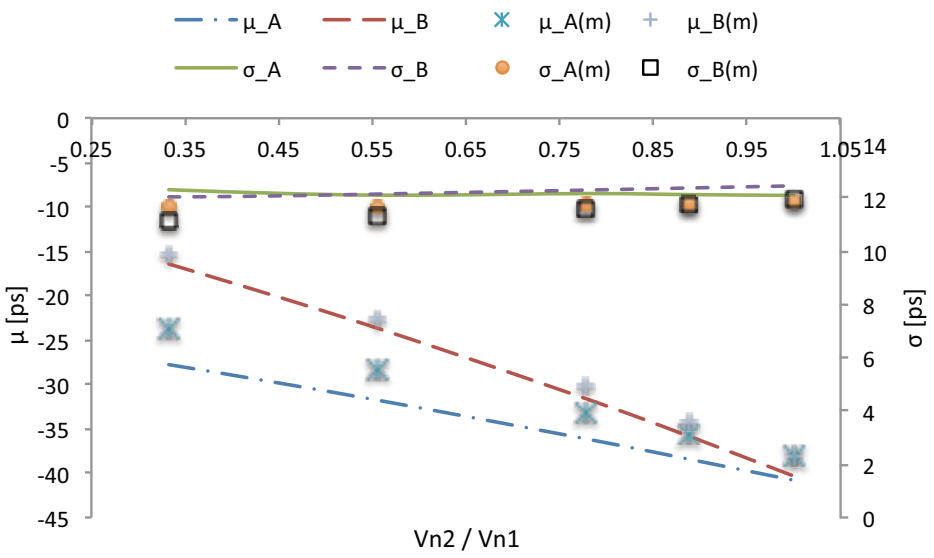


Figure 5.17: Skitter for $V_{n1} = 90$ mV and different V_{n2}.

ΔV_n, as shown in Figs. 4.8 and 4.9. The tier closer to the P/G pads experiences lower supply noise [34].

The clock paths spanning two tiers with 20 buffers ($n_1 + n_2 = n_3 + n_4 = 20$, see Fig. 5.3(a)) are taken as an example. The clock source is located in Tier 2. The total length of each path is 5 mm. The initial phase ϕ (270°) and frequency f_n (400 MHz) are assumed to be the same for both tiers. Two distributions of clock paths are discussed: (A) $n_1 = n_2 = n_3 = n_4 = 10$ and (B) $n_1 = n_3 = 15, n_2 = n_4 = 5$. Distribution (A) denotes the equally-divided 3-D clock paths. Distribution (B) represents placing the longest segment of clock paths in Tier 2. To depict the accuracy of the model, the simulation results of the setup skitter $J_{1,2}$ for $V_{n1} = 90$ mV and different V_{n2} are shown in Fig. 5.17. As shown in this figure, $\mu_{J_{1,2}}$ changes significantly with V_{n2}, while $\sigma_{J_{1,2}}$ does not vary a lot with V_{n2}. This behavior is accurately described by the proposed model. The change of setup and hold skitter with both V_{n2} and V_{n1} is discussed in the following paragraphs.

Setup skitter $J_{1,2}$ *vs.* V_n

The change of $J_{1,2}$ with (V_{n2}, V_{n1}) is illustrated in Fig. 5.18. As shown in Figs. 5.18(a) and 5.18(b), for distribution (A), $\mu_{J_{1,2}}$ increases significantly with both V_{n2} and V_{n1}, since higher supply noise introduces greater period jitter. The clock paths of (A) are equally distributed among tiers. As a result, $\mu_{J_{1,2}}$ is affected by V_{n1} and V_{n2} in the same way. For distribution (B), however, the situation is different. As shown in Figs. 5.18(c) and 5.18(d), $\mu_{J_{1,2}}$ is mainly determined by V_{n2}, since the longest segment of the clock paths in (B) is placed in Tier 2.

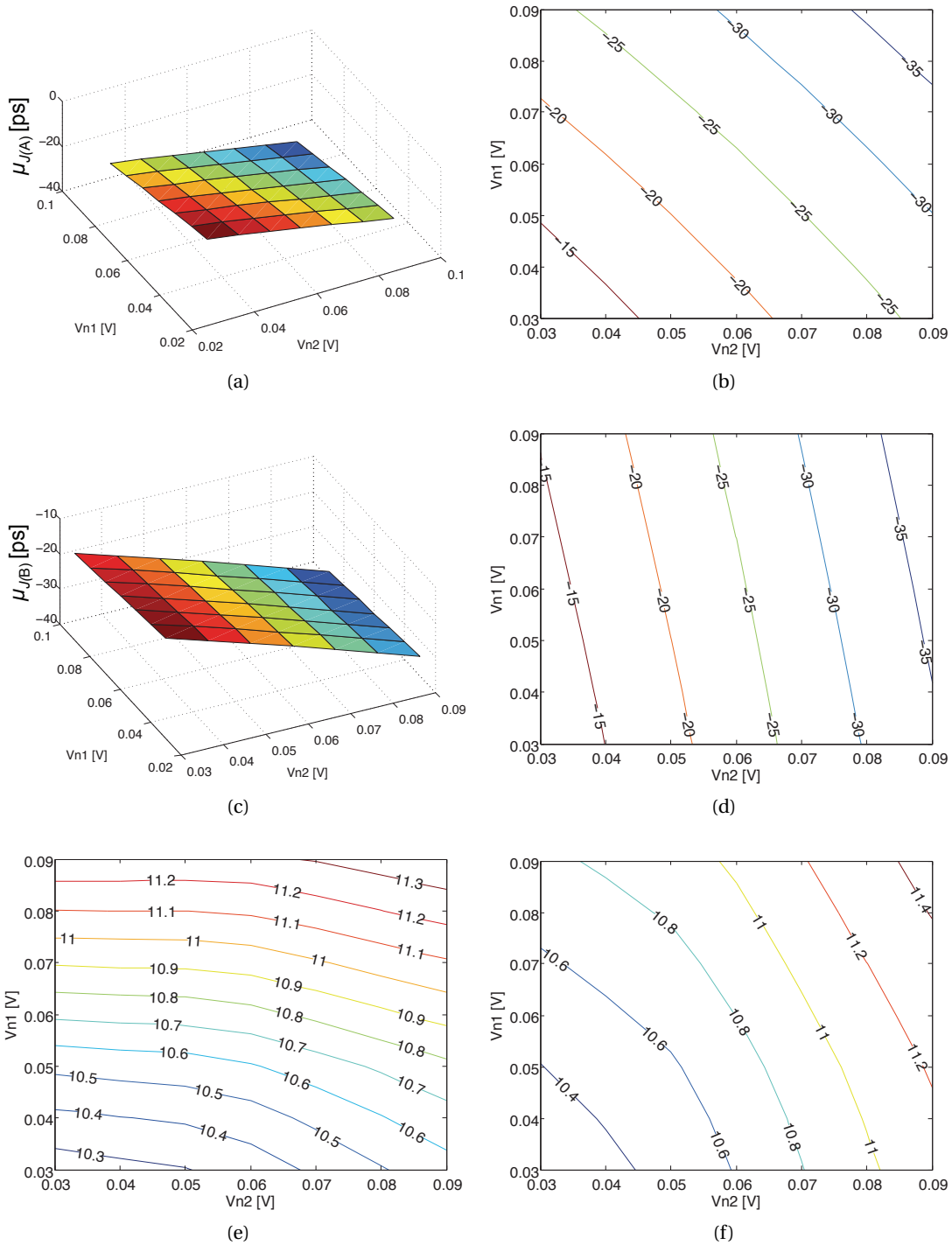


Figure 5.18: Setup skitter *vs.* (V_{n2}, V_{n1}), where (a) and (b) are the 3-D plot and contour for μ_{J_A} for distribution (A), respectively. (c) and (d) are the 3-D plot and contour for μ_{J_B} for distribution (B), respectively. (e) and (f) are the contours of σ_{J_A} and σ_{J_B}, respectively.

Observation 5.2. *For unequally-distributed clock paths, the mean skitter is mainly determined by the tier where the longest part of the clock paths is placed.*

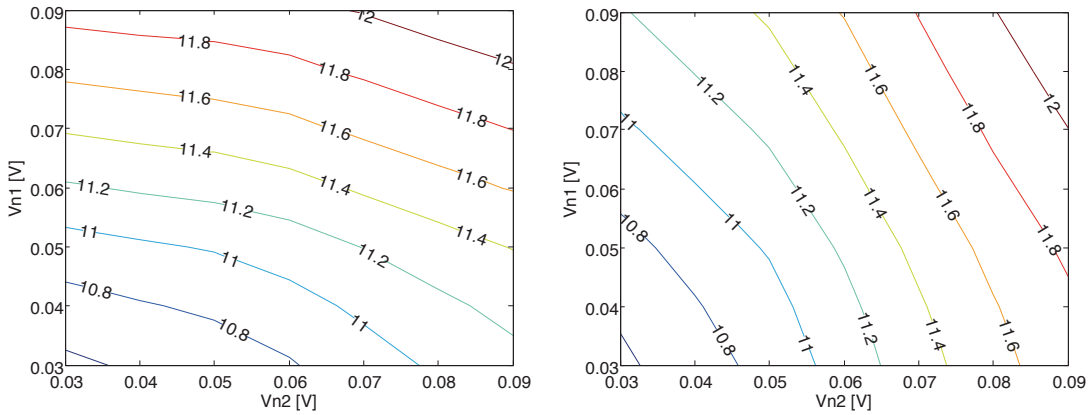
As shown in Figs. 5.18(b) and 5.18(d), assuming $V_{n1} = 0.09$ mV, distribution (A) produces higher $\mu_{J_{1,2}}$ than (B) for different V_{n2}. This difference in $\mu_{J_{1,2}}$ increases with ΔV_n ($\Delta V_n = V_{n1} - V_{n2}$), from 1% to 42% of μ_{J_A}. The reason is that the majority of buffers in (B) is located in Tier 2, which is more susceptible to V_{n2}. More generally, given $V_{n1} > V_{n2}$, the mean skitter of (B) is always lower than (A).

Consequently, the distribution of clock paths in 3-D ICs significantly affects the mean skitter due to the different V_n among tiers. However, in 2-D circuits, this mean skitter does not vary significantly with the distribution of clock paths due to the common effect of the global resonant noise at low frequencies [118].

The standard deviation $\sigma_{J_{1,2}}$ of (A) and (B) is illustrated in Figs. 5.18(e) and 5.18(f), respectively. Similar to $\mu_{J_{1,2}}$, $\sigma_{J_{1,2}}$ also increases with V_{n1} and V_{n2}. Nevertheless, $\Delta\sigma_{J_{1,2}}$ is relatively low as compared with $\Delta\mu_{J_{1,2}}$.

Hold skitter $S_{1,2}$ *vs.* V_n

The mean value of $S_{1,2}$ is relatively low (≤ 0.5 ps), since the two clock paths have the same number, size, and distribution of buffers. Nevertheless, $\sigma_{S_{1,2}}$ is non-negligible for both distributions (A) and (B), as illustrated in Figs. 5.19(a) and 5.19(b), respectively. Similar to $\sigma_{J_{1,2}}$, $\sigma_{S_{1,2}}$ increases with V_{n1} and V_{n2} but $\Delta\sigma_{S_{1,2}}$ is lower than 1.5 ps.



(a) (b)

Figure 5.19: Hold skitter *vs.* (V_{n2}, V_{n1}), where (a) and (b) are the contours for σ_{S_A} and σ_{S_B}, respectively.

Observation 5.3. *The standard deviation of the setup and hold skitter increases slightly with the amplitude of resonant supply noise.*

5.3.5 The effect of ϕ on skitter

The skitter under the power supply noise with different ϕ is investigated in this subsection. As shown in Fig. 4.7, the initial phase ϕ of the supply noise is similar among tiers ($\phi_1 = \phi_2$). The change of $J_{1,2}$ and $S_{1,2}$ with ϕ is illustrated in Fig. 5.20, where $V_{n1} = 0.09$ V and $V_{n2} = 0.07$ V.

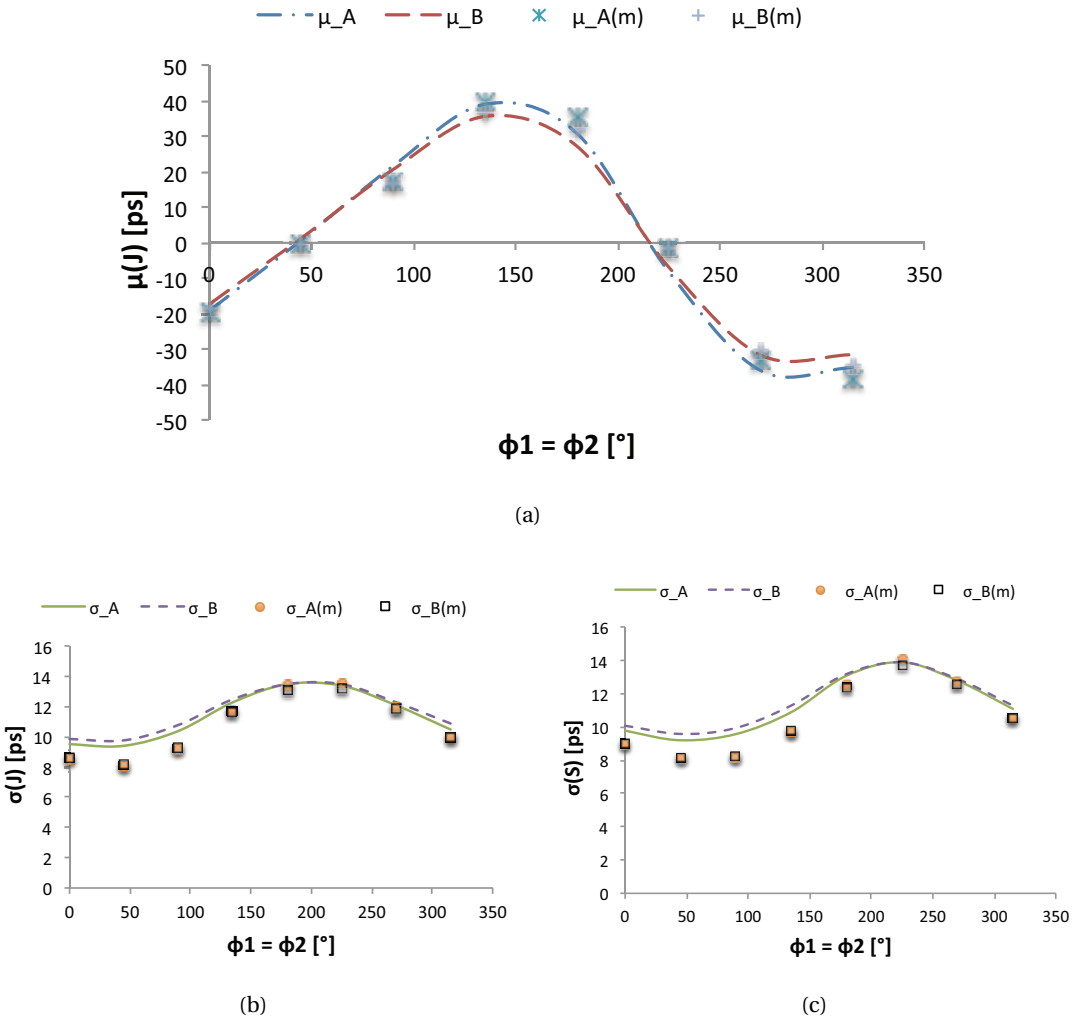


Figure 5.20: Skitter *vs.* different ϕ ($\phi_1 = \phi_2$), where (a) is the change of $\mu_{J_{1,2}}$. (b) and (c) are the change of $\sigma_{J_{1,2}}$ and $\sigma_{S_{1,2}}$, respectively.

Setup skitter $J_{1,2}$ *vs.* ϕ

As shown in Figs. 5.20(a) and 5.20(b), the difference in ϕ results in significant change not only in $\mu_{J_{1,2}}$, but also in $\sigma_{J_{1,2}}$. For instance, the highest $\sigma_{J_{1,2}}$ is 41% higher than the lowest one for Distribution (A) in Fig. 5.20(b). The worst $\mu_{J_{1,2}}$ occurs when ϕ_1 and ϕ_2 are both around 270°, similar to the conclusion for 2-D ICs in [94]. The worst $\sigma_{J_{1,2}}$, however, occurs when $\phi \approx 225°$.

Therefore, if the initial phase is not 270°, the skitter can be still high due to the high $\sigma_{J_{1,2}}$. The difference in $\sigma_{J_{1,2}}$ is low between distributions (A) and (B) since in both cases, the two clock paths are symmetrically distributed among tiers.

Hold skitter $S_{1,2}$ *vs.* $\phi_{\mathbf{n}}$

The effect of ϕ_1 and ϕ_2 on $S_{1,2}$ is shown in Fig. 5.20(c). Due to the similarity between the two clock paths, the resulting $\mu_{S_{1,2}}$ is relatively low. The standard deviation, however, is significantly affected by ϕ. As illustrated in Figs. 5.20(b) and 5.20(c), the change of $\sigma_{S_{1,2}}$ is similar to $\sigma_{J_{1,2}}$.

Observation 5.4. *For the setup and hold skitter, σ changes considerably with the phase of the power supply noise. The highest σ and μ of skitter do not happen at the same initial phase of the supply noise.*

Considering the clock paths and waveforms shown in Fig. 5.15, ϕ is determined by the time when the first clock edge arrives at the input of clock paths. The worst σ can be obtained by traversing all the possible ϕ. Due to the excessive time required by Monte-Carlo simulations, the proposed model is highly efficient to determine the worst skitter and the corresponding ϕ for multi-tier circuits, as compared with Monte-Carlo simulations.

The effect of phase-shifting of the supply noise on skitter

Several techniques, such as *RC* filtered buffers and "stacked" phase-shifted buffers [29], have been proposed to shift the ϕ seen by the clock paths. In 3-D clock distribution networks, these techniques can be applied to a part of the clock paths in a different tier to adapt $\Delta\phi$ among tiers. The change of $\sigma_{J_{1,2}}$ versus the shifted (ϕ_1, ϕ_2) for distribution (A) is shown in Figs. 5.21(a) and 5.21(b). As shown in Fig. 5.21(b), the dashed line depicts the $\sigma_{J_{1,2}}$ for $\phi_1 = \phi_2$, which denotes the skitter without phase-shifting. As shown by the arrow, the highest $\sigma_{J_{1,2}}$ decreases with $\Delta\phi = \phi_2 - \phi_1$. In this case, since ϕ_2 and ϕ_1 are not simultaneously equal to 270°, the worst $\mu_{J_{1,2}}$ is also decreased.

In Fig. 5.21(c), however, $\sigma_{J_{1,2}}$ of distribution (B) highly depends on ϕ_2. This behavior is due to that $\sigma_{J_{1,2}}$ is dominated by the supply noise in the second tier. In this case, shifting ϕ among tiers provides less than 1.5 ps decrease in $\sigma_{J_{1,2}}$, as shown by the dashed line with arrows.

Observation 5.5. *For equally distributed clock paths across 3-D ICs, the worst skitter can be decreased by properly shifting ϕ among tiers with phase-shifted clock distribution.*

Note that the proper $\Delta\phi$ should be determined by traversing all the combinations of ϕ in different tiers. The number of combinations increases exponentially with the number of tiers, which implies a large number of simulations. Again, the proposed model provides a highly efficient way to determine a valid shift in ϕ for multi-tier circuits to decrease skitter.

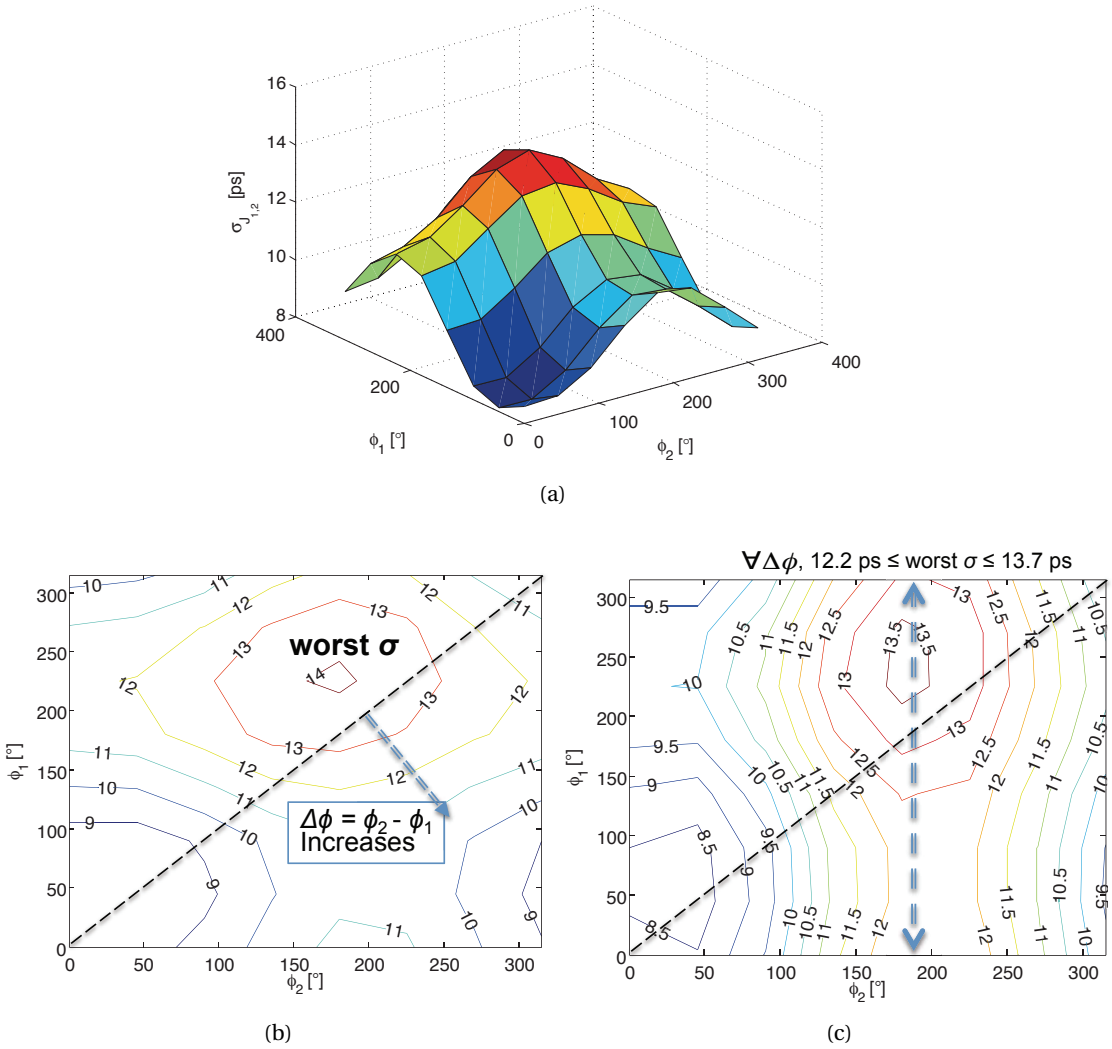
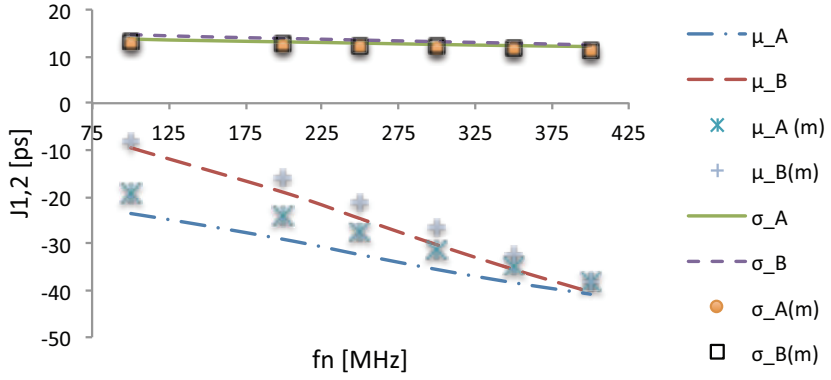


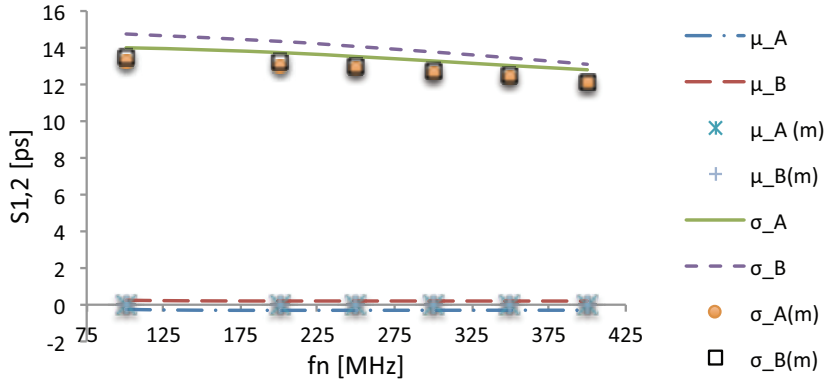
Figure 5.21: Skitter $J_{1,2}$ *vs.* shifted ϕ_1 and ϕ_2, where (a) and (b) are the 3-D plot and contour map of $\sigma_{J_{1,2}}$ *vs.* (ϕ_2, ϕ_1) for distribution (A), respectively. (c) is the contour map of $\sigma_{J_{1,2}}$ for distribution (B).

5.3.6 The effect of f_n on skitter

The effect of the frequency of power supply noise on skitter is investigated in this subsection. This frequency is usually considered similar among tiers [34], as shown in Figs. 4.8(b) and 4.9(c). Different f_n are investigated, herein, to demonstrate the change of skitter with the frequency of supply noise. The amplitude V_n and phase ϕ are assumed to be the same among tiers, where $V_{n1} = V_{n2} = 90$ mV and $\phi_1 = \phi_2 = 270°$. The simulation results are illustrated in Fig. 5.22.



(a)

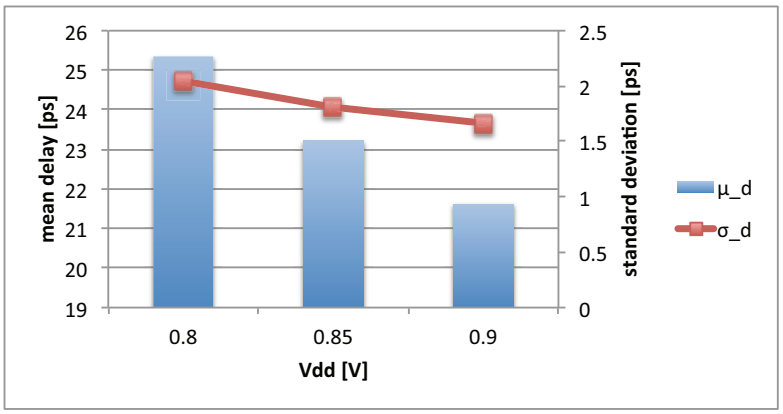


(b)

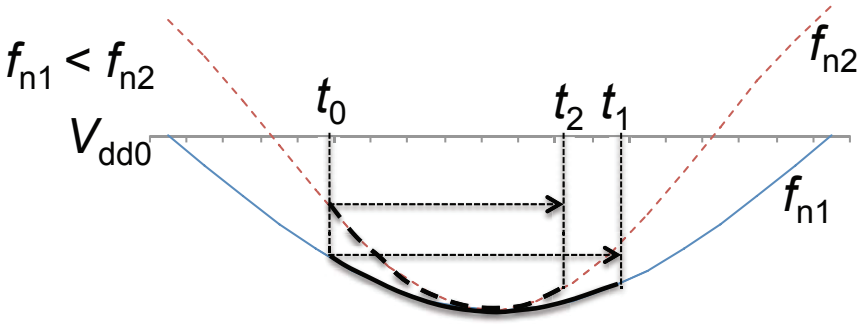
Figure 5.22: Skitter *vs.* f_n. The change of $J_{1,2}$ and $S_{1,2}$ are illustrated in (a) and (b), respectively.

Similar to the effect of V_n, f_n greatly affects $\mu_{J_{1,2}}$. For instance, $\mu_{J_{1,2}}$ increases with f_n up to 70% for distribution (B). The variation of skitter, however, decreases with f_n. The resulting $\Delta\sigma_{J_{1,2}}$ and $\Delta\sigma_{S_{1,2}}$ are up to 15% for both distributions (A) and (B). This behavior is due to the decreased voltage seen by the clock buffers during the clock propagation. The change of μ_d and σ_d for the delay of two inverters (a clock buffer) in series is illustrated in Fig. 5.23(a). Both μ_d and σ_d decrease with V_{dd}. As shown in Fig. 5.23(b), assume that the clock edge seeing the

worst σ_J arrives at the input of the clock path at t_0. When f_n increases from f_{n1} to f_{n2}, the propagation time of this edge decreases from t_1 to t_2 and the supply voltage within this time interval increases. This higher supply voltage introduces lower σ in the buffer delay, which causes lower $\sigma_{J(1,2)}$ and $\sigma_{S(1,2)}$ according to (5.30) and (5.35).



(a)



(b)

Figure 5.23: The effect of the change of f_n on delay variation, where (a) is the mean and standard deviation of buffer delay *vs.* V_{dd}. (b) is the supply voltage to a clock path during the propagation of a clock edge.

Observation 5.6. *The mean setup skitter increases significantly with the frequency of power supply noise, while both $\sigma_{J_{1,2}}$ and $\sigma_{S_{1,2}}$ decrease with this frequency.*

As shown in Figs. 5.16 - 5.22, the proposed statistical model for skitter exhibits reasonably high accuracy as compared with SPICE-based simulations. For the worst-case $\mu_{J_{1,2}}$ ($\sigma_{J_{1,2}}$) in Figs. 5.16 - 5.22, the error is -11% (-12%), -7% (-10%), -8% (-4%), and -10% (-9%), respectively. The behavior of skitter under different scenarios of supply noise can correctly be described by the proposed model. Since $\sigma_{J_{1,2}}$ varies with power supply noise, process variations and power supply noise need to be simultaneously modeled to correctly describe clock uncertainty.

The difference in mean skitter varies up to 60% due to the different V_{n} among tiers. $\sigma_{J_{1,2}}$ can vary up to 41% due to different ϕ (see Figs. 5.20(b) and 5.20(c)). Decreasing the variation as well as the mean skitter helps to improve the robustness of 3-D clock distribution networks.

5.4 Methodologies for Skitter Mitigation in 3-D ICs

Potential methods for skitter mitigation in 3-D ICs are discussed in this section. The effect of the number and size of clock buffers on skitter is investigated in the following subsection. The tradeoff between skitter and power consumption is discussed in Section 5.4.2. Based on the simulation results and the previous propositions, a set of design guidelines for skitter mitigation is proposed in Section 5.4.3.

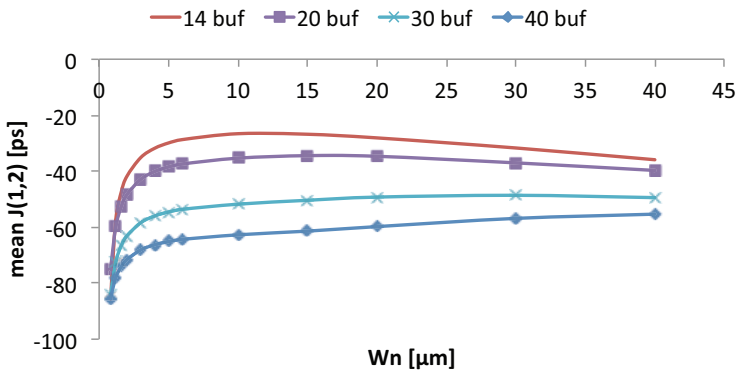
5.4.1 Skitter for different buffer insertion

The change of skitter with the size and numbers of clock buffers inserted along the clock paths is discussed in this subsection. A pair of clock paths with a length of 5 mm are simulated. These paths are both equally distributed across two tiers, where $V_{\mathrm{n1}} = 0.09$ V and $V_{\mathrm{n2}} = 0.08$ V. The skitter is determined by Monte-Carlo simulations in the remainder of this section.

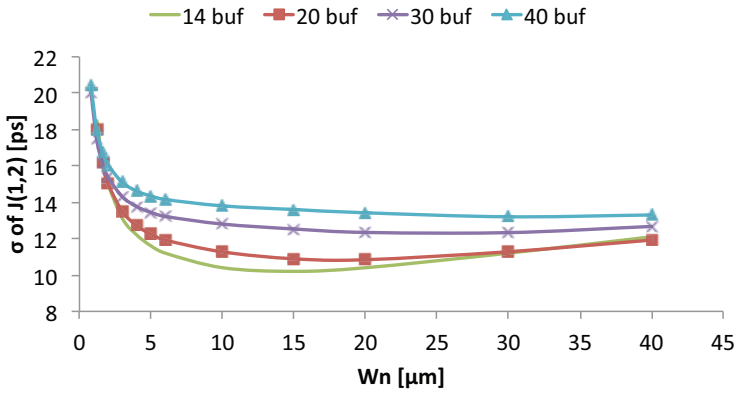
The worst $\mu_{J_{1,2}}$ ($\phi = 270°$) for different numbers and size of buffers is shown in Fig. 5.24(a). The $\sigma_{J_{1,2}}$ from Monte-Carlo simulations for different buffer solutions is illustrated in Fig. 5.24(b). As shown in this figure, both the mean and standard deviation of $J_{1,2}$ increase with the number of buffers. For hold skitter, the change of $\sigma_{S_{1,2}}$ is highly similar to $\sigma_{J_{1,2}}$.

Considering the Gaussian distribution of $J_{1,2}$ in (5.29), $J_{1,2}$ falls in the range $[\mu_{J_{1,2}} - 3\sigma_{J_{1,2}}, \mu_{J_{1,2}} + 3\sigma_{J_{1,2}}]$ with a probability of 99.7%. Within this range, $\max(J_{1,2})$ is used to indicate the worst (maximum) skitter. For improved readability, the absolute value of $\max(J_{1,2})$ is shown, where $\max(J_{1,2}) = |\mu_{J_{1,2}}| + 3\sigma_{J_{1,2}}$. The relation between $\max(J_{1,2})$ and the transition time at the clock sinks for different buffer insertion is illustrated in Fig. 5.25.

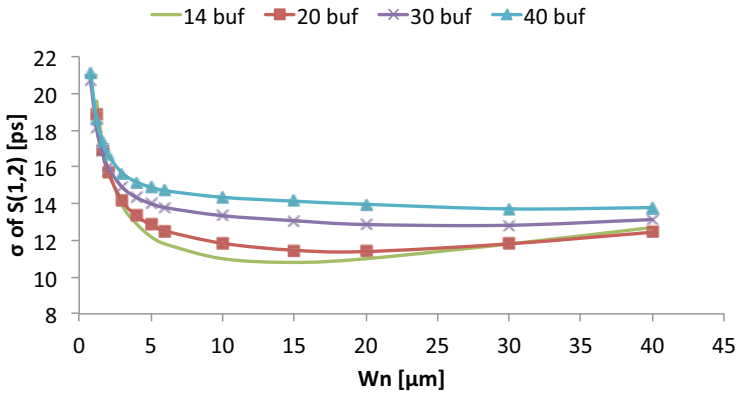
As shown in Fig. 5.25, for the same $\max(J_{1,2})$, the clock paths with fewer buffers produce a considerably longer transition time. For the clock paths with 14 buffers, the shaded area contains the inferior buffer solutions, which can be replaced by smaller buffers with the same skitter but lower transition time. For instance, the buffer solution B_1 can be replaced by the iso-skitter point B_2 with a lower rise time. In the unshaded area, both the transition time and skitter decrease with the buffer size W_{n} in region R_1. In R_2, however, larger W_{n} decreases skitter with an increase in transition time.



(a)



(b)



(c)

Figure 5.24: Skitter for different buffer insertion, where the mean of $J_{1,2}$ is illustrated in (a) and $\sigma_{J_{1,2}}$ and $\sigma_{S_{1,2}}$ are shown in (b) and (c), respectively.

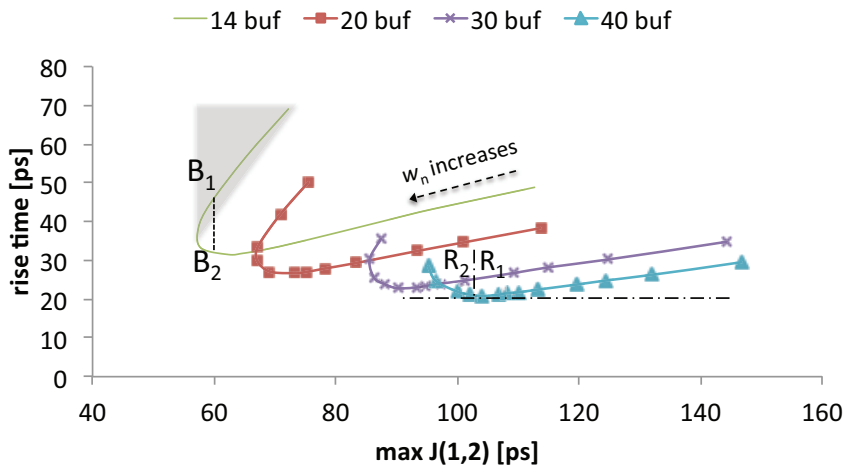


Figure 5.25: Transition time *vs.* max($J_{1,2}$) for different buffer insertion.

5.4.2 Tradeoffs between skitter and power consumption

The power consumed by the clock distribution networks constitutes a significant portion of the total power consumed by a circuit [18, 169]. The power consumption of the clock network under different constraints on skitter is investigated in this subsection.

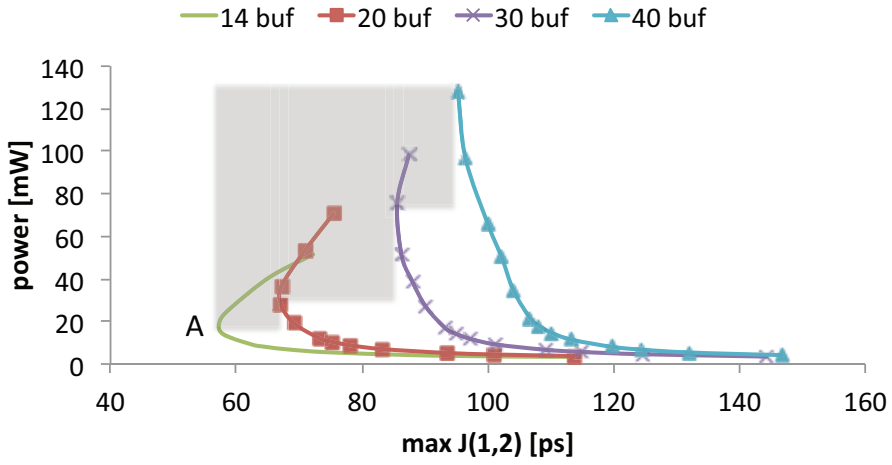
For the investigated clock paths, the total power consumption under different constraints on max($J_{1,2}$) and max($S_{1,2}$) is illustrated in Fig. 5.26. The shaded area depicts the inferior buffer solutions. Point A denotes the lowest skitter that can be obtained. In the unshaded area, skitter decreases as the buffer size and power increase. For the same constraint in skitter, the clock paths with fewer buffers are more power-efficient.

For the clock paths with 14 buffers, as the constraint becomes lower than 68 ps, significant power overhead is shown. For example, to decrease the max($J_{1,2}$) from 68 ps to 58 ps (15% improvement), the buffers are sized up from 4 μm to 10 μm. The resulting power consumption increases from 6.9 mW to 14.4 mW (109% increase). In conclusion, pursuing extreme constraints on clock skitter results in high overhead in power.

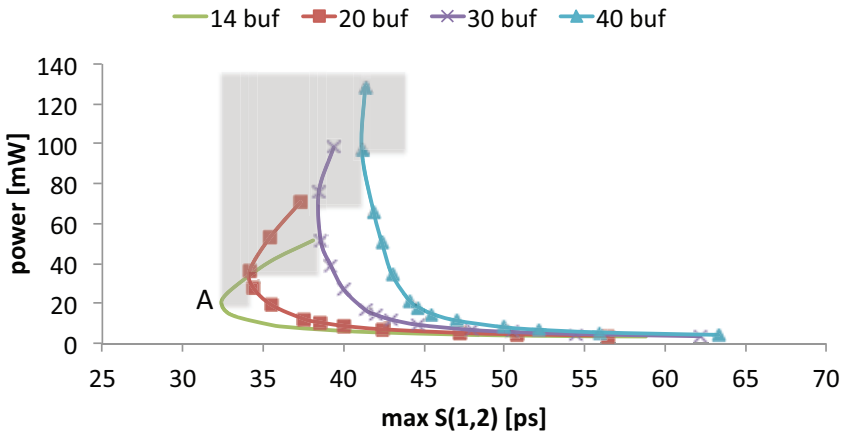
Observation 5.7. *Skitter can be decreased by sizing up clock buffers at the expense of power consumption.*

5.4.3 Guidelines to mitigate skitter

Based on Observations 5.1 to 5.7 presented in the previous sections, a set of guidelines is provided to support the design of robust 3-D clock distribution networks. The objective of these guidelines is to decrease skitter in 3-D ICs.



(a)



(b)

Figure 5.26: Tradeoff between power and timing. Power *vs.* $\max(J_{1,2})$ and $\max(S_{1,2})$ are illustrated in (a) and (b), respectively.

Guideline 5.1. *Given the freedom to choose among tiers for the clock paths in a 3-D circuit, the mean skitter can be decreased by placing most of the clock path length in those tiers that exhibit the lowest supply noise.*

Guideline 5.2. *For 3-D clock paths equally distributed among tiers, the worst-case* $\mu_{J_{1,2}}$ *and* $\sigma_{J_{1,2}}$ *can be decreased by shifting* ϕ *among different tiers.*

Guideline 5.3. *By decreasing the frequency of resonant supply noise, mean skitter can be decreased by trading off the standard deviation of skitter.*

Guideline 5.4. *Reducing the number of buffers along the clock path can decrease the skitter at the expense of input slew.*

Guideline 5.5. *By properly sizing up the clock buffers, a tradeoff between skitter and power consumption can be exploited.*

5.5 Case Study of 3-D Clock Trees

To illustrate the role of these guidelines, several examples of synthesized 3-D clock trees are simulated and analyzed in this section. The 3-D circuits are generated from IBM clock network benchmarks [175] by randomly distributing the clock sinks to different tiers [173]. The 3-D clock trees are synthesized with a 3D MMM+DME algorithm based on [20]. This algorithm is briefly introduced in the following subsection. The skitter in the synthesized clock trees is discussed in Section 5.5.2. Clock buffers are inserted under a specified constraint in the capacitive load to limit the input transition time. Meanwhile, the buffer insertion should be optimized to produce a short clock delay. A fast buffer insertion algorithm for 3-D clock trees is presented in Section 5.6.

5.5.1 3-D clock tree synthesis

The objective of clock tree synthesis is to determine the topology and routing of a clock tree to propagate the clock signal from the source to all the sinks of this tree. The buffers are inserted to satisfy specifications on clock skew and slew rate under constraints on the area and power consumed by these buffers [18, 176, 177]. In 3-D ICs, since the stacked tiers can separately be fabricated, pre-bond testing issues can arise, where different parts of a clock tree need to be separately tested. In addition, the number of TSVs used in a 3-D clock tree needs to be constrained both to save area and to increase yield [20]. The clock tree should also be designed to tolerate the failure of a certain number of TSVs such that the yield and robustness are improved.

Several clock tree synthesis algorithms have been proposed for 3-D clock trees. These algorithms focus on different optimization objectives. Low power clock tree synthesis algorithms considering pre-bond testing problems are proposed in [83, 84]. A TSV-fault-tolerant algorithm is proposed in [85]. A low power synthesis algorithm minimizing the number of TSVs is proposed in [20]. A synthesis algorithm considering temperature variations is proposed in [82]. The 3D MMM+DME algorithm proposed in [20] is implemented in this chapter to generate 3-D clock trees. Note that the proposed model of skitter is applicable to different types of synthesized 3-D clock trees.

The 3D MMM+DME algorithm consists of two phases. In the first phase, the topology of the clock tree is determined by the 3D-MMM algorithm under a specified limit on the number of TSVs. In the second phase, based on the topology from the first phase, the abstract clock tree is traversed twice to determine the location of each internal node, the routing path, and the location of TSVs and clock buffers.

1. In the first phase, the 3D-MMM algorithm based on the classical MMM algorithm [177] is used. An abstract binary clock tree is generated top-down. The connection among clock sinks and the topology of the clock tree are determined. The exact locations of the internal nodes, TSVs, and clock paths, however, are not determined. The basic idea is to recursively bipartite a given set of clock sinks. When the limit on the number of TSVs for the current set of sinks is higher than one, this set is bipartited according to the horizontal distance among sinks. Otherwise, this set is bipartited based on the tiers where these sinks are located.

2. In the second phase, the 3D-DME algorithm based on the traditional DME algorithm [78] is used. According to the tree topology from the first phase, the location of the internal nodes, TSVs, and clock buffers is determined. The routing of clock paths is also determined. 3D-DME also consists of two stages. In the first stage, the clock tree is traversed from sinks towards the tree root (clock source). The parent node p of two nodes (children) i and j should be placed such that the skew between the two subtrees (rooted from i and j, respectively) is zero. The possible positions for p are denoted by a merging segment, as illustrated by ms_p in Fig. 5.27. Clock buffers are inserted to satisfy the constraint on the capacitive load and zero-skew. In the second stage of 3D-DME, the clock tree is traversed top-down again. The exact location of each internal node along the corresponding merging segment is determined to minimize the length of clock paths, as denoted by p along ms_p in Fig. 5.27. The clock tree synthesis is accomplished after all the clock sinks are reached.

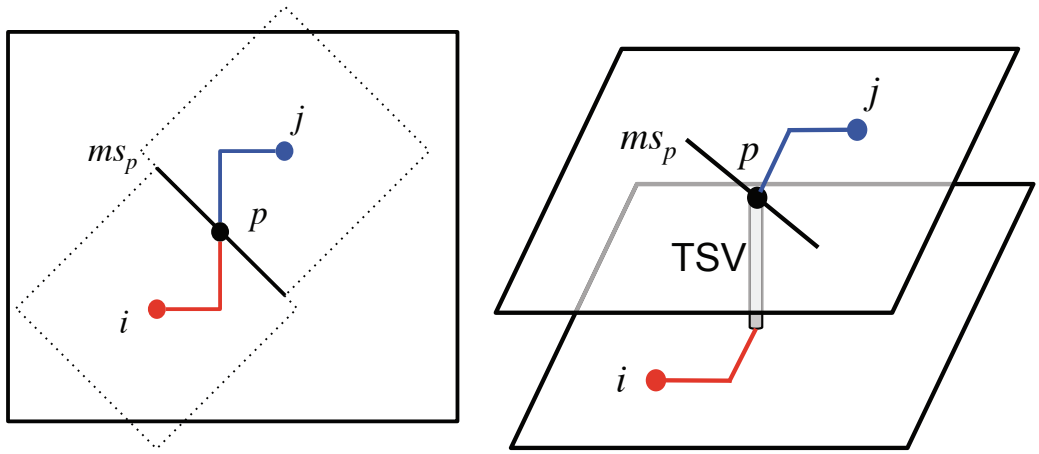


Figure 5.27: An example of merging two nodes in 3D-DME algorithm, where (a) and (b) are the top and 3-D views of a 3-D circuit, respectively.

The 3-D clock trees generated from the 3D MMM+DME algorithm achieve near-zero clock skew and satisfy the constraint on the number of TSVs. In the following subsection, clock trees

Table 5.6: 3-D ICs Based on IBM Clock Benchmarks

	# sinks	# buffers	area [mm^2]	t_s [h]	t_m [s]	speedup
r3	862	2128	9.8×9.6	1.8	45	142×
r4	1903	4695	12.7×12.7	1.9	53	129×
r5	3101	7496	14.5×14.3	2.4	56	154×

for different benchmarks and different numbers of tiers are simulated. The skitter of these 3-D clock trees is investigated.

5.5.2 Skitter in synthesized 3-D clock trees

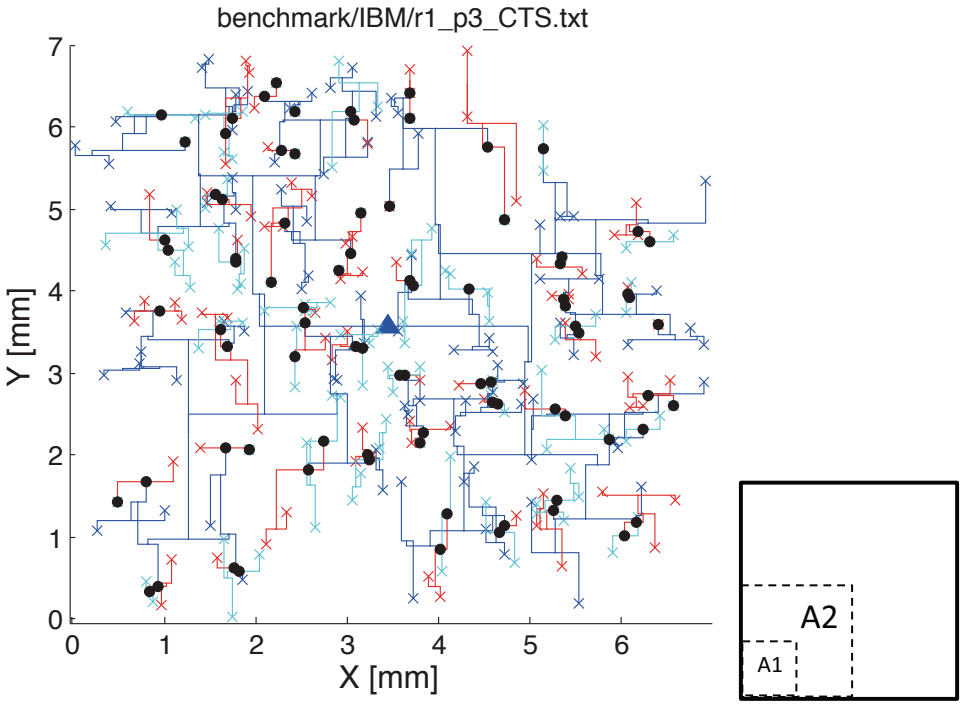
The effect of the design guidelines proposed in Section 5.4.3 is illustrated in this subsection by taking Guideline 5.1 as an example. Several clock benchmarks have been simulated. The buffers are inserted with a constraint of 50 fF in the capacitive load. Each clock buffer is formed by an inverter ($W_n = 4.83\mu$m and $W_p = 2.1W_n$). An example of the resulting 3-tier clock trees for "r1" benchmark (267 sinks) is illustrated in Fig. 5.28(a). The clock source, clock sinks, and TSVs are denoted by ▲, ×, and •, respectively. The clock networks in tiers 1, 2, and 3 are denoted in blue, red, and green, respectively.

The skitter is measured within two different regions, as illustrated in Fig. 5.28(b). For both regions A_1 and A_2, the skitter is reported between the pair of the farthest sinks. The three largest IBM benchmarks r3, r4, and r5 are simulated. SPICE simulations are performed for the paths of interest with 2000 Monte-Carlo simulations. The features of these benchmarks are shown in Table 5.6, where the CPU time is also listed. Note that the simulation time is only for the selected clock paths, not for the entire clock tree. The initial phase and the frequency of the supply noise are assumed to be the same among the three tiers ($f_{n1} = f_{n2} = f_{n3} = 400$ MHz). The amplitudes V_n are assumed to differ among tiers ($V_{n1} = 0.09$ V, $V_{n2} = 0.08$ V, $V_{n3} = 0.065$ V).

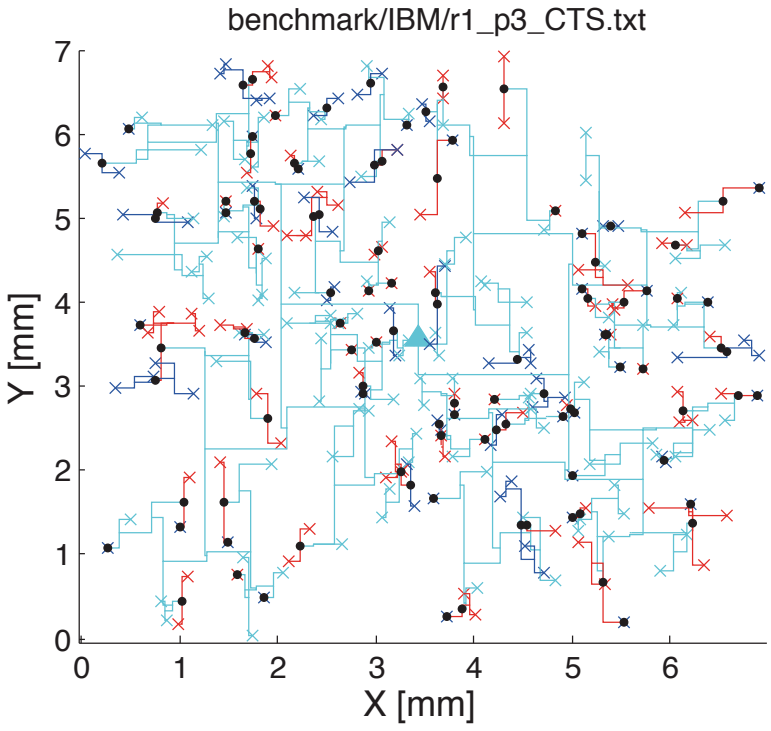
The skitter is reported in Table 5.7. As mentioned before, the mean hold skitter is close to zero. Consequently, only the mean setup skitter is reported. The highest mean skitter is obtained when $\phi_1 = \phi_2 = \phi_3 = 270°$ and the highest σ is reported for $\phi_1 = \phi_2 = \phi_3 = 200°$. Four design practices are compared with each other.

- Case 1 (C1), the majority of the clock tree is located in Tier 1, which is adjacent to the heat sink. Most of the clock buffers are placed to this tier to constrain the increase in the temperature of the circuit. The power supply noise and process variations are separately considered for $\mu_{J_{1,2}}$ and $\sigma_{J_{1,2}}$, respectively.

- Case 2 (C2), the majority of the clock tree is also located in Tier 1, but the power supply noise and process variations are simultaneously modeled.



(a) (b)



(c)

Figure 5.28: A synthesized 3-D clock tree with the majority of clock buffers in the first (a) and third tier (c). The regions where the skitter is measured are illustrated in (b).

- Case 3 (C3), the majority of the clock tree is placed in the middle tier (Tier 2) to decrease the number of TSVs and power consumption, as proposed in [20].
- Case 4 (C4), based on Guideline 5.1, the majority of the clock tree is located in Tier 3 (with the lowest V_{n}) as illustrated in Fig. 5.28(c).

Table 5.7: Skitter in 3-D ICs Generated from the IBM Clock Distribution Network Benchmarks

Benchmark		Region A_1							Region A_2						
		C1	C2	C3	C4	Impr1[1]	Impr2	Error[2]	C1	C2	C3	C4	Impr1	Impr2	Error
Setup μ [ps]	r3	-52.6	-52.1	-44.0	-35.9	31%	18%	-5%	-53.7	-53.1	-44.8	-36.4	31%	19%	-7%
	r4	-66.3	-65.0	-58.6	-48.8	25%	17%	-3%	-69.3	-68.6	-62.1	-52.0	24%	16%	-7%
	r5	-64.8	-62.9	-56.8	-47.6	24%	16%	3%	-67.3	-66.5	-59.9	-50.2	25%	16%	-1%
Setup σ [ps]	r3	8.5	11.2	9.6	10.5	7%	-9%	-10%	11.5	15.2	13.9	13.1	14%	6%	-6%
	r4	10.7	16.6	12.0	11.3	32%	6%	-8%	10.8	15.4	16.0	15.6	-2%	2%	-7%
	r5	8.5	12.9	11.6	12.5	2%	-8%	-9%	11.8	16.0	13.9	18.5	-16%	-33%	-8%
Hold σ [ps]	r3	8.5	11.4	10.1	10.3	10%	-1%	-7%	11.5	15.6	14.4	13.2	15%	9%	-7%
	r4	10.7	14.5	13.6	11.5	21%	16%	-7%	10.8	15.1	15.6	15.6	-3%	0%	-9%
	r5	8.5	11.5	11.1	11.5	0%	-4%	-5%	11.8	15.9	15.6	17.6	-10%	-13%	-6%

[1] Impr1 and Impr2 are the improvements of C4 over C2 and C3, respectively.

[2] Error is the maximum error of the proposed model as compared with SPICE-based Monte-Carlo simulations.

As shown in Table 5.7, $\mu_{J_{1,2}}$ in Case 1 is similar to Case 2. Nevertheless, $\sigma_{J_{1,2}}$ and $\sigma_{S_{1,2}}$ are significantly underestimated in Case 1, for both regions A_1 and A_2. As compared to Case 2, the difference in $\sigma_{J_{1,2}}$ and $\sigma_{S_{1,2}}$ is up to 36%. This difference shows the necessity of simultaneously modeling process variations and power supply noise.

Observation 5.8. *Separately modeling process variations and power supply noise significantly underestimates the variation of skitter.*

The difference between the proposed model and SPICE-based Monte-Carlo simulations is listed in the Error column of Table 5.7. For all $\sigma_{J_{1,2}}$ and $\sigma_{S_{1,2}}$, the error of the proposed model is below 10% as compared to Monte-Carlo simulations. The error in μ is below 7% for $J_{1,2}$. Considering the greater than 129× speedup in CPU time as reported in Table 5.6, the proposed model provides an efficient way to accurately model skitter.

In Case 2, the majority of the CDN is placed in the tier adjacent to the heat sink. In Case 3, the majority of the CDN is placed in the middle tier to reduce the number of TSVs and power consumption [20]. The number of TSVs and the power consumption of the entire tree for Cases 2 to 4 are illustrated in Fig. 5.29. The results are normalized over Case 4. As proposed in [20], Case 3 produces the fewest TSVs (see "#TSV(C2/C4)" and "#TSV(C3/C4)"). The total power is similar among the three cases due to the similar number of clock buffers, as shown by "Power (C2/C4)" and "Power (C3/C4)". The distribution of this power, however, differs due to the different distribution of buffers among tiers.

Case 4 mitigates the mean skitter trading off the number of TSVs and the power distribution. As illustrated in Figs. 4.8(a) and 4.9(a), the tier next to the package has the lowest V_n. Consequently, $\mu_{J_{1,2}}$ of Case 4 is significantly improved over Cases 2 and 3, as shown by the first three rows of Impr1 and Impr2, respectively. This improvement ranges from 16% up to 31%. This comparison shows the efficiency of Guideline 1 in decreasing mean skitter. For several paths, however, $\sigma_{J_{1,2}}$ and $\sigma_{S_{1,2}}$ in Case 4 increase over Cases 2 and 3. This situation is due to the change of the topology of the clock trees. For instance, for the pair of paths in A2 and circuit r5, the number of buffers after the merging point of these paths increases as compared to Case 2. These buffers are located in different tiers. Consequently, $\sigma_{J_{1,2}}$ and $\sigma_{S_{1,2}}$ both increase.

5.6 Fast Buffer Insertion for 3-D Trees

Although 3-D ICs are expected to greatly reduce the wire length as compared to planar circuits, methods to further improve the interconnect delay are required. This situation is due to the length of the global interconnects, the high fanout of the interconnect trees, and possibly the high capacitance of the vertical interconnects [70, 71] that determine the overall performance of a 3-D circuit.

Many buffer insertion algorithms have been proposed for 2-D interconnects. The optimal number and size of the buffers to achieve the minimum interconnect delay for a distributed *RC*

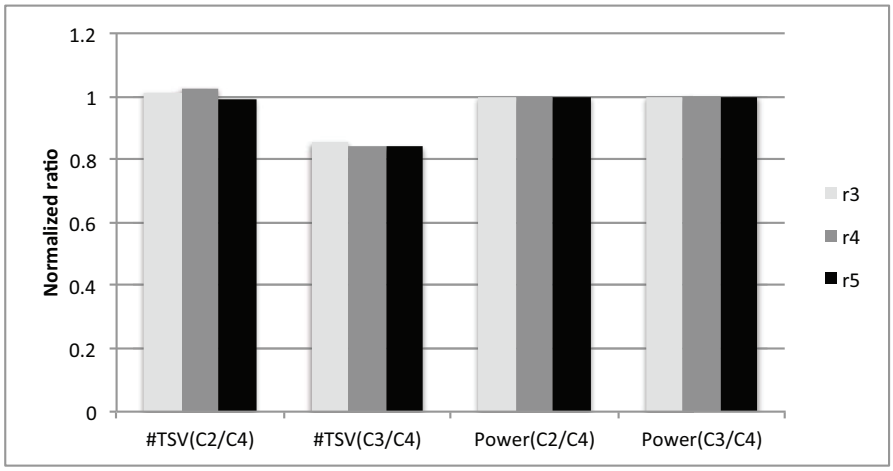


Figure 5.29: Normalized number of TSVs and power for Cases 2 to 4.

interconnect are described in [39], [178]. A uniform buffer design methodology for efficiently driving *RC* tree structures is presented in [40]. All of these methods are based on uniform buffer insertion, which utilizes the same size of buffers with the same interval within one segment (branch). Nevertheless, applying the uniform buffer insertion techniques for 2-D interconnects to 3-D trees traversing multiple tiers does not result in the minimum interconnect delay. In a 3-D system, each physical tier can be fabricated with a different process or technology node. This situation results in diverse interconnect impedance characteristics and buffer libraries among the tiers of a 3-D circuit. In addition, the various manufacturing technologies for the vertical interconnects (*e.g.*, through silicon via (TSV)) affect the delay of the intertier interconnects [7, 179]. All of these factors complicate the buffer insertion task for 3-D interconnects. An optimal non-uniform buffer insertion algorithm is proposed in [180–182]. The required CPU time, however, is at least $O(n^2)$, where n is the number of nodes within a clock tree.

Few buffer planning algorithms for 3-D circuits have been presented in [183], [184], where the size and number of the buffers required by 3-D interconnects are considered known. Nevertheless, there is no method to provide the size and number of the buffers, which are required by multi-tier interconnect trees considering the disparate impedance characteristics of interconnects and dissimilar buffer libraries. The contribution of the proposed algorithm is twofold.

- The number, size, and location of the buffers required for 3-D interconnect trees are determined. The proposed method extends over traditional uniform buffer insertion techniques to consider the inherent heterogeneity of 3-D circuits.

- The proposed technique achieves improvement up to 25% in the timing performance of 3-D trees over conventional uniform buffer insertions used for 2-D circuits. The time complexity of the approach is linear to the number of nodes of a 3-D tree.

5.6.1 Uniform buffer insertion

The uniform buffer insertion problem for a 3-D interconnect tree is formulated as follows. A 3-D interconnect tree with buffers is illustrated in Fig. 5.30. R_{so} is the resistance of the driver. C_{sinki} is the capacitive load of sink i of the tree.

The nodes of the tree in Fig. 5.30 are labeled from N_1 to N_9 by pre-order depth-first traversal ($i.e.$, the tree is traversed depth-first and the parent node is labeled before all of the child nodes). The nodes of a 3-D tree include the source, roots of the subtrees, and the sinks of the tree. For this specific example, the roots of the subtrees are the TSVs at nodes N_2, N_3, and N_7 and the root of the subtree at node N_6. A branch B_i ($e.g.$, B_1) of the tree is a two-terminal 2-D wire segment that ends at N_{i+1} ($e.g.$, N_2), as illustrated in Fig. 5.30. There are $n-1$ branches in a 3-D tree, where n is the number of nodes of this tree.

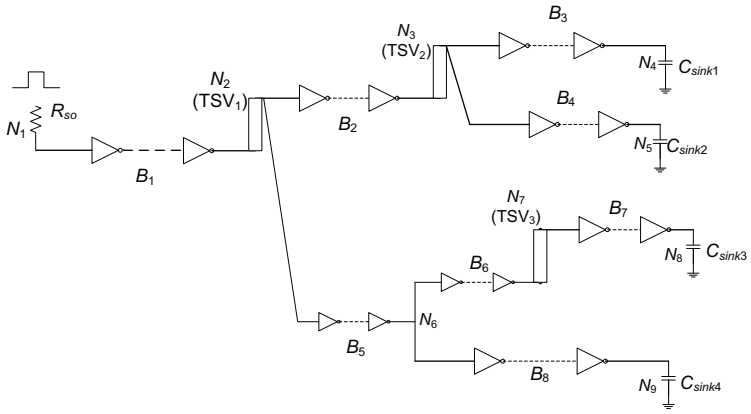


Figure 5.30: A 3-D interconnect tree with buffers.

The maximum wire delay of a 3-D tree, t_{max} is the maximum delay of each path from the source (root) of the tree to any sink of the tree (path delay). The total wire delay of a 3-D tree, t_{total} is the sum of all the path delays. The delay of a path consists of the delay of the branches and TSVs. The delay of a TSV is modeled as a fixed wire segment delay where buffers cannot be inserted. The buffer insertion problem for a 3-D tree is to determine the number, size, and location of the buffers required along B_i ($1 \le i \le n-1$), which minimizes t_{max} and t_{total}.

5.6.2 Delay model of 3-D interconnects for buffer insertion

The delay model of a two-terminal 2-D net with uniform buffers and the method to determine the number, size, and location of the buffers for a point-to-point 2-D net is first presented. Following this method, the delay model of 3-D trees is presented. The vertical interconnect technology considered in this section is based on the MIT Lincoln Labs 3-D integration process [70], [185]. The diameter and length of TSVs are 1.75 μm and 11 μm, respectively. The resistance and capacitance of TSVs are 170 mΩ and 2 fF, respectively.

Delay model of a two-terminal 2-D net

The delay model of a point-to-point 2-D net i is illustrated in Fig. 5.31. x_i is the distance between the source and the first buffer of net i. y_i is the distance between the last buffer and the sink of net i. k_i is the number of buffers inserted in net i. Uniform buffer insertion achieves minimum delay for a two-terminal net in a single tier [178]. The size of the buffers is denoted by h_i, which is the multiple of the minimum buffer size that can be used in the tier that includes net i.

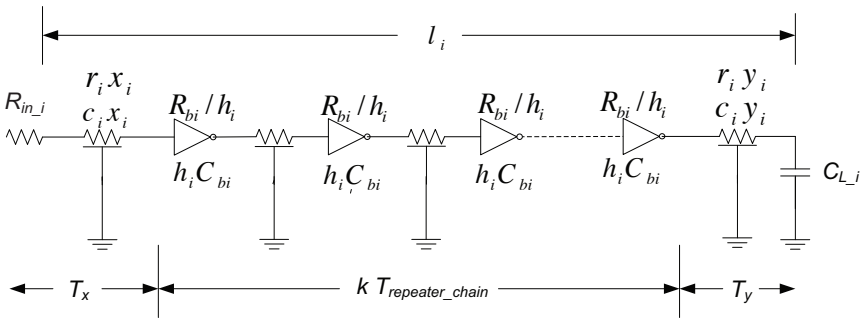


Figure 5.31: The electrical model of a 2-D net with buffers.

As shown in this figure, R_{in_i} is the output resistance of the driver of net i. C_{L_i} is the capacitance of the load of net i. R_{bi} and C_{bi} are the output resistance and input capacitance, respectively, of the smallest buffer used in the tier that contains net i. r_i and c_i are the interconnect resistance and capacitance per length, respectively, in this tier. For net i where k_i ($k_i \geq 2$) buffers with size h_i are inserted, the resulting delay of net i using the distributed Elmore delay model [111] is

$$\begin{aligned} t_{net\_i} &= t_{x_i} + t_{buffer\_chain} + t_{y_i} \\ &= R_{bi}C_{bi}(k_i - 1) + \frac{(l_i - x_i - y_i)^2 r_i c_i}{2(k_i - 1)} + \frac{R_{bi}(C_{L\_i} + (l_i - x_i)c_i)}{h_i} + \\ &\quad C_{bi}(R_{in\_i} + (l_i - y_i)r_i)h_i + R_{in\_i}c_i x_i + \frac{x_i^2 r_i c_i}{2} + \frac{y_i^2 r_i c_i}{2} + y_i r_i C_{L\_i}. \end{aligned} \tag{5.36}$$

The physical constraints for the variables h_i, k_i, x_i, and y_i, respectively, are

$$h_i \geq 1;\ k_i \geq 2;\ 0 \leq x_i \leq l_i;\ 0 \leq y_i \leq l_i;\ 0 \leq x_i + y_i \leq l_i. \tag{5.37}$$

To provide a closed form expression that minimizes (5.36) is a rather formidable task [178]. Expressions of k_i, x_i, and y_i are provided by [182] where the size of the inserted buffers is considered fixed. Alternatively, an efficient method to determine (k_i, h_i, x_i, y_i) is proposed in this section, where h_i is not considered fixed. Let $\frac{\partial t_{net\_i}}{\partial h_i} = 0$ and $\frac{\partial t_{net\_i}}{\partial k_i} = 0$, (k_i, h_i) can be

written as a function of (x_i, y_i),

$$k_i = (l_i - x_i - y_i)\sqrt{\frac{r_i c_i}{2R_{b\_i}C_{b_i}}} + 1, \tag{5.38}$$

$$h_i = \sqrt{\frac{R_{bi}(C_{L\_i} + (l_i - x_i)c_i)}{C_{bi}(R_{in\_i} + (l_i - y_i)r_i)}}. \tag{5.39}$$

By replacing (5.38) and (5.39) in (5.36) and since x_i and y_i are constrained according to (5.37), the minimum wire delay t_{net_i} and a feasible solution (x_i, y_i) can be determined with numerical methods. If there is only one buffer inserted along the net, applies that $k_i = 1$, $y_i = l - x_i$, and $t_{net\_i} = t_{x_i} + t_{l-x_i}$. The location and size of the single buffer is also provided through (5.36) and (5.39).

Delay model of a 3-D tree

For a 3-D tree denoted as Tr with n nodes, let rt denote the root of Tr. The delay of branch B_i connecting N_j and N_{i+1} is modeled as a 2-D net. Expression (5.36) is properly adapted to describe the delay of B_i. Consequently, the expression for R_{in_i} includes the resistance of the TSV, the section y_{j-1} and the output resistance of the last buffer inserted in B_{j-1}. C_{L_i} includes the capacitance of the TSV and the sum of the input capacitance of the child branches of B_i,

$$R_{in\_i} = \begin{cases} R_{so}, & \text{if } N_j = rt, \\ \frac{R_{b(j-1)}}{h_{j-1}} + r_{j-1}y_{j-1} + R_{tsv}, & \text{if } N_j \text{ is a TSV}, \\ \frac{R_{b(j-1)}}{h_{j-1}} + r_{j-1}y_{j-1}, & \text{otherwise.} \end{cases} \tag{5.40}$$

$$C_{L\_i} = \begin{cases} C_{sinkz}, & \text{if } N_{i+1} \text{ is the sink } z \text{ of } Tr, \\ C_{eff\_(i+1)} + C_{tsv}, & \text{if } N_j \text{ is a TSV}, \\ C_{eff\_(i+1)}, & \text{otherwise.} \end{cases} \tag{5.41}$$

$$C_{eff\_(i+1)} = \sum_{B_e \in DS_{i+1}} C_{in\_e}, \tag{5.42}$$

$$C_{in\_e} = c_e x_e + C_{be} h_e. \tag{5.43}$$

$C_{eff\_(i+1)}$ is the effective capacitive load at N_{i+1}. C_{in_e} is the input capacitance of branch B_e, which is the sum of the capacitance of wire section x_e and the input capacitance of the first buffer in branch B_e. DS_{i+1} is the set of all of the downstream branches starting from N_{i+1}. In Fig. 5.30, $C_{eff\_2} = C_{in\_2} + C_{in\_5}$. Due to (5.40)-(5.43), the buffers inserted in one branch considerably affect the number and size of buffers required in other branches. For a 3-D tree as in Fig. 5.30, the total delay of the tree and the maximum sink delay of the tree are, respectively,

$$t_{total} = \sum_{p \in PT} \Big(\sum_{B_i \in p} (t_{net\_i} - R_{in\_i}C_{in\_i}) + R_{so}C_{in\_1} \Big), \tag{5.44}$$

$$t_{max} = \max_{p \in PT} \sum_{B_i \in p} (t_{net\_i} - R_{in\_i}C_{in\_i}) + R_{so}C_{in\_1}, \tag{5.45}$$

where PT is the set of the paths of the tree, and R_{in_i} depends on $1/h$ as shown in (5.40). Choosing variables x, y, h, and k to globally minimize (5.44) requires computationally expensive optimization techniques, since (5.44) is essentially a multi-variable non-polynomial function. A global optimization technique for 2-D trees is proposed in [40]. In 3-D circuits, due to the TSVs and the disparate characteristics of buffers and interconnects in different physical tiers, this optimization method is not applicable. Alternatively, a heuristic approach is proposed to minimize the delay of each branch iteratively resulting in a near-optimal buffer insertion in 3-D interconnect trees. This novel approach completes the buffer insertion in $O(n)$ time, where n is the number of nodes of the tree. Note that the effect of inserting buffers in adjacent branches on the delay of the investigated branch is considered in (5.36) through (5.38)-(5.43).

5.6.3 Iterative buffer insertion algorithm

The proposed algorithm determines a near-optimal solution $S = \{s_i | 1 \leq i \leq n-1\}$ for minimizing t_{total} and t_{max} by iteratively optimizing the delay of each branch based on (5.36). The solution for each branch B_i is denoted by $s_i = (k_i, h_i, x_i, y_i)$. The pseudo-code of the Iterative Repeater Insertion Algorithm for 3-D ICs (IRI-3D) is illustrated in Algorithms 5.1 and 5.2. The proposed algorithm comprising two phases is described in the following subsections.

Algorithm 5.1 Pseudo-code of IRI-3D.

Input: A 3-D tree Tr with n nodes.
Output: $t_{total}, t_{max}, \{(h_i, k_i, x_i, y_i) | 1 \leq i \leq n\}$.

▷ first phase

1: $buildTree(Tr)$;
2: $rt \leftarrow$ root of Tr;
3: $uptRC(rt)$;
4: $optBranch(rt)$;
5: $[t_{total}, t_{max}] = uptDelay(rt)$;

▷ second phase

6: **while** $\Delta t_{total} > target\_ratio$ **do**
7: $uptRC(rt)$;
8: $optBranch(rt)$;
9: $[t'_{total}, t'_{max}] = uptDelay(rt)$;
10: $\Delta t_{total} \leftarrow (t_{total} - t'_{total})/(t_{total})$;
11: $t_{total} \leftarrow t'_{total}$;
12: **end while**

Initial allocation of buffers

In the first phase, an initial placement of buffers for a 3-D tree with n nodes is obtained. The minimum delay of each branch is successively determined from B_{n-1} to B_1, assuming that a minimum size buffer is inserted exactly before the start node of each branch (except for the

Algorithm 5.2 Pseudo-code of $optBranch(nd)$.

Input: A node N_i in tree Tr.

▷ B_{i-1} is the branch before N_i

```
1: for all child nodes chd of N_i do
2:     optBranch(chd)
3: end for
4: uptRC(N_i);
5: if N_i ≠ rt then
6:     if number of iterations ≤ 1 then
7:         B_{i-1}.(k,h,x,y) ← calc1(B_{i-1});
8:     else
9:         B_{i-1}.(k,h,x,y) ← calc2(B_{i-1});
10:    end if
11: end if
```

root of the tree), as illustrated in Fig. 5.32. By considering a buffer in the previous branch, the effect of the TSVs and the different libraries of buffers used in different tiers are considered.

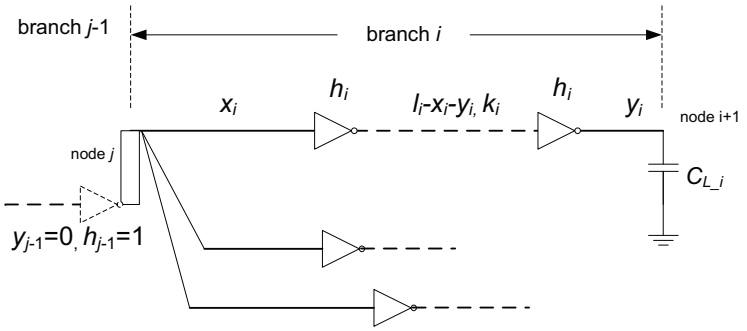


Figure 5.32: A minimum size buffer exactly before the starting node is assumed. C_{L_i} is determined by (5.41).

The algorithm commences from the root of the tree and the tree is traversed depth-first by recursively invoking the procedure *optBranch*(). The nodes and branches in the tree are labeled as described in Section 5.6.1. B_i is processed only after every child branch of B_i has been processed. The corresponding s_i is determined by the procedure $calc1()$ based on (5.36)-(5.41).

In the procedure $uptRC(nd)$, the capacitive load seen at node nd and the resistance of the driver of the child nodes are updated by (5.40)-(5.43). If nd is the root of the tree, R_{in} and C_L of all the nodes in the tree are updated. The delay t^0_{total} of the tree is determined after all the branches have been processed, where the superscript indicates the number of iterations. At the end of the first phase where S^0 is obtained, the resistance of the driver and capacitive load $\{(R_{in\_i}, C_{L\_i}) | 1 \le i \le n-1\}$ for each branch are updated.

Refinement of the buffer allocation

The second phase starts with the updated set $\{(R_{in\_i}, C_{L\_i}) | 1 \leq i \leq n-1\}$ obtained during the first phase. The tree is traversed similar to the first phase, where a new S is obtained. The R_{in_i} used for each branch is updated by considering the effect of the location and size of the last buffer inserted at the preceding branch (determined during the first phase) as compared to the initial assumption that a minimum size buffer is placed right before the starting node.

The buffers inserted in B_i described by $s_i = (k_i, h_i, x_i, y_i)$ affect the total delay of the tree. To capture the dependency between any s_i and t_{total}, the terms of t_{total} that include the variables of s_i are described by

$$t_{term\_i} = q_i t_{net\_i} + (q_j - q_i) R_{in\_i} C_{in\_i}. \tag{5.46}$$

q_i and q_j denote the number of the paths containing B_i and B_j, respectively, where B_j is the preceding branch of B_i. The expressions for h_i and k_i are accordingly modified. This expression is used in $calc2()$ to evaluate the effect of s_i on t_{total}.

During the first phase the assumption of a minimum size buffer at the start node of each investigated branch results in particularly high R_{in_i}. Consequently, the contribution of the $R_{in\_i} C_{in\_i}$ to t_{total} as described by the second term of (5.46) is excessive and rather artificial. Therefore, (5.36) is used instead since the assumption of a minimum size buffer is adopted to provide a rough initial solution for the first phase.

The t_{total} determined at each iteration of the second phase is smaller or at least no greater than the previously determined delay. The following proposition supports this statement.

Theorem 5.1. *Given the initial delay of a 3-D tree t^0_{total} resulting from s^0 obtained during the first phase and the delay t^1_{total} resulting from S^1 determined by the first iteration of the second phase, $t^1_{total} \leq t^0_{total}$.*

Proof. Theorem 5.1 is proved by induction.

1. After the first phase a solution s^0_i for each branch $B_i (1 \leq i \leq n-1)$ has been determined where the superscript indicates the number of iteration. For B_i, s^0_i is determined based on $C^0_{L\_i}$ and the assumption of placing a minimum size buffer in B_{j-1}, as depicted by the buffer drawn with the dashed line in Fig. 5.33(a). At the first iteration, denote the total delay of the tree in Fig. 5.33(a) as $t^1_{i'}$, where i' indicates that s^1_i for B_i has not been determined. The solutions at iteration one for all the branches in the subtree rooted at N_{i+1} have been determined since the child nodes of B_i are processed first. The solution for the branch B_{j-1} preceding B_i, however, is that of the previous iteration s^0_{j-1}, as illustrated in Fig. 5.33(a). Consequently, when (5.46) is evaluated during iteration one, s^0_i does not provide the minimum delay from the last buffer (depicted by the solid line) in B_{j-1} to N_{i+1}. This behavior is due to the $R^0_{in\_i}$ updated at the end of the first phase

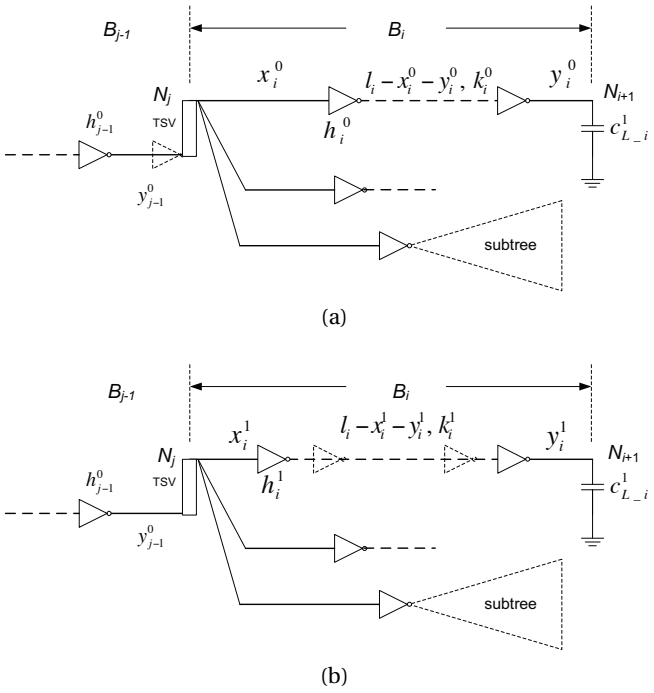


Figure 5.33: Iterative procedure to insert buffers along a branch B_i. (a) An initial solution for branch B_i and (b) refinement of the solution.

(according to s_{j-1}^0) and $C_{L\_i}^1$ (according to the new solution for the child branches of B_i).

2. Branch B_i ($1 \leq i \leq n-1$) is now processed. The allocation of the buffers in B_i after iteration one is depicted in Fig. 5.33(b). t_{total} of the 3-D tree in Fig. 5.33(b) is now t_i^1 ($= t_{(j-1)'}^1$). The contribution of B_i to t_i^1 and $t_{i'}^1$ is described by (5.46). Since s_i^1 is determined by minimizing (5.46) for $R_{in\_i}^0$ and $C_{L\_i}^1$ through (5.38)-(5.43), applies that $t_i^1 \leq t_{i'}^1$. Since B_{i+1} is processed before B_i and $t_{i'}^1 = t_{i+1}^1$, implies that $t_i^1 \leq t_{i+1}^1$.

3. For branch B_{n-1}, which is the first branch processed in iteration one, $C_{L\_n-1}^1 = C_{sink1}$. Similar to the above, $t_{(n-1)}^1 \leq t_{(n-1)'}^1 = t_{total}^0$.

Consequently, applies that $t_{total}^1 = t_1^1 \leq t_{n-1}^1 \leq t_{total}^0$. □

After the first iteration, S^1 and t_{total}^1 are obtained, along with a new set $\{(R_{in\_i}^1, C_{L\_i}^1) | 1 \leq i \leq n-1\}$. Since h_{j-1}^1 and y_{j-1}^1 can be different from h_{j-1}^0 and y_{j-1}^0, $R_{in\_i}^1$ also differs from $R_{in\_i}^0$. The solution s_i^1 for B_i, however, is determined based on $R_{in\_i}^0$. Consequently, the total wire delay is further decreased by refining the solution for B_i based on $R_{in\_i}^1$. Based on S^1 and $\{(R_{in\_i}^1, C_{L\_i}^1) | 1 \leq i \leq n\}$, the second iteration commences. According to Theorem 5.1, $t_{total}^2 \leq t_{total}^1$. As illustrated in line 6 of Algorithm 5.1, when Δt is smaller than the $target\_ratio$, the algorithm terminates. The $target\_ratio$ is considered to be user-specified.

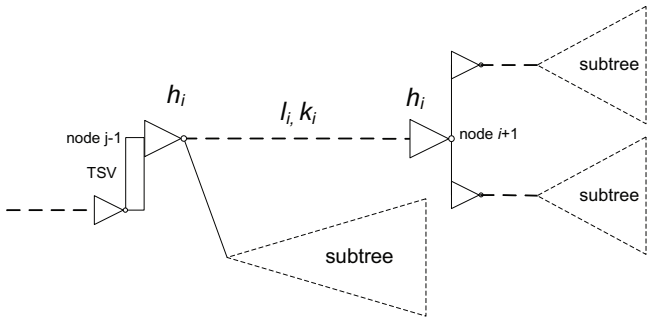


Figure 5.34: Application of a conventional buffer insertion method [39], [40] in a 3-D tree.

Considering a 3-D tree consisting of n nodes, the time used to minimize (5.36) and (5.46) is $O(1)$. The time used by $uptRC()$ is $O(n)$ for each iteration. The time consumed in lines 3 and 4 in Algorithm 5.1 is $O(n)$. The time consumed in lines 6-12 in Algorithm 5.1 is $O(mn)$, where m is the number of iterations. The simulation results show that for a typical 3-D tree, m is quite small. Consequently, the complexity of the proposed algorithm is practically $O(n)$.

5.6.4 Simulation results

IRI-3D is applied to several randomly generated 3-D interconnect trees and related results are presented in this section. To investigate the effectiveness of the proposed algorithm, IRI-3D is compared with a widely used technique for 2-D interconnects [39], [40].

This approach assumes that the buffers are equally spaced in each branch of the tree [39]. If more than one buffer is required in branch B_i, there is a buffer inserted right after the start node and before the end node of this branch, respectively, as illustrated in Fig. 5.34. With this assumption, each segment is effectively treated as a 2-D interconnect. The delay of the segments is decoupled and buffers are individually inserted in each segment based on the methods described in [39], [171]. In [39], applies that $\{x_i = 0, y_i = 0 | 1 \leq i \leq n-1\}$. The optimum number k_i and size h_i of the buffers are directly determined by (5.38) and (5.39).

Both the IRI-3D and the conventional approach are applied to 3-D trees with different length spanning several physical tiers varying from three to six tiers. The ASU predictive technology model (PTM) [41] is used to extract the parameters of the interconnect and the buffers for the 130 nm, 90 nm, and 65 nm technology nodes. The length of the longest path ranges from 0.8 mm to 17.16 mm. The output resistance of the source is $R_{so} = 800\ \Omega$. C_{sink} ranges from 1 fF to 2.5 fF. The resistance per length r ranges from 170 Ω/mm to 450 Ω/mm for intermediate interconnects and from 30 Ω/mm to 40 Ω/mm for global interconnects. The capacitance per length c ranges from 190 fF/mm to 250 fF/mm. The output resistance of the minimum size buffer on different tiers ranges from 800 Ω to 1100 Ω. The input capacitance of the minimum size buffer varies from 0.9 fF to 1.5 fF.

Table 5.8: Delay of 3-D interconnect trees after buffers are inserted.

# tiers	# sinks	t_{total}		t_{max}		Impr(k)
		Impr1[1]	Impr2	Impr1	Impr2	
2	3	86.52%	10.32%	85.87%	9.95%	34.31%
3	4	91.20%	10.54%	90.10%	10.63%	28.59%
3	7	97.47%	15.57%	97.22%	15.99%	24.92%
4	4	97.01%	17.94%	96.42%	19.46%	22.49%
4	9	98.13%	16.71%	97.77%	17.65%	23.52%
4	13	98.55%	16.69%	98.31%	16.95%	23.93%
5	4	97.34%	19.05%	96.71%	20.52%	20.91%
5	8	98.27%	17.86%	97.83%	19.33%	21.59%
5	16	98.85%	17.35%	98.59%	18.18%	22.29%
6	4	97.35%	22.89%	96.71%	24.75%	24.46%
6	8	98.28%	21.48%	97.79%	23.67%	25.38%
6	16	98.87%	20.56%	98.57%	22.46%	25.98%
6	32	99.27%	20.38%	99.10%	21.42%	26.51%
Average		96.70%	17.49%	96.23%	18.54%	24.99%

[1] For each case, 500 trees are simulated. Impr1 is the average improvement of IRI-3D over the trees without buffers. Impr2 is the average improvement of IRI-3D over the conventional method. Impr(k) is the reduction in k of IRI-3D as compared to the conventional method.

3-D trees consisting of intermediate and global interconnects are investigated. The average improvement in t_{total} and t_{max} before and after inserting buffers by employing the two methods are listed in Table 5.8. As listed in Table 5.8, after buffer insertion, both t_{total} and t_{max} are significantly decreased, demonstrating a more than 90% improvement. As reported by $Impr1$, the improvement in delay increases with the number of sinks of the trees, the number of tiers, and the length of the longest path. This situation is due to the increase in the capacitive load of 3-D trees without buffers. Although 3-D ICs reduce the interconnect length, the total capacitance of the tree remains considerable.

IRI-3D achieves significant improvement over the approach for 2-D trees. The additional decrease in delay ranges from 10% to 23% in t_{total} and from 10% to 25% in t_{max}. The improvement in t_{total} increases with the number of tiers that the tree spans and the path length.

The variation of improvement in t_{total} with the number of iterations is reported in Table 5.9. The improvement decreases fast with the number of iterations. Beyond the second iteration, the improvement in delay is negligible. The number of iterations is fewer than four in all of the simulation results.

The number of buffers inserted in 3-D trees by IRI-3D is smaller than that of the conventional method as reported in the last column of Table 5.8. According to (5.38), fewer buffers are required by IRI-3D by properly adjusting x_i and y_i for each branch. The size of these buffers is typically large wasting silicon area. This area overhead is mitigated by permitting a small

Table 5.9: The improvement in total delay vs. number of iterations.

# Iteration	2 tiers	3 tiers	4 tiers	5 tiers	6 tiers
1	89.76%	95.03%	95.56%	97.24%	98.29%
2	7.66%	16.30%	19.35%	22.51%	24.96%
3	0.01%	1.07%	1.91%	2.07%	3.76%
4	0.00%	0.01%	0.18%	0.43%	0.55%

Table 5.10: The improvement in delay and area under diverse area constraints.

# tiers	# sinks	Impr ($A_{con} \leq A_{2-D}$)[1]			Impr ($A_{con} \leq 1.2A_{2-D}$)			Impr (no constraint)		
		t_{total}	t_{max}	area[2]	t_{total}	t_{max}	area	t_{total}	t_{max}	area
2	3	7.9%	7.6%	27.3%	9.2%	9.0%	18.3%	10.3%	10.0%	-2.5%
3	4	6.8%	6.2%	22.7%	8.5%	8.1%	11.0%	10.5%	10.6%	-20.2%
3	7	7.8%	8.3%	20.3%	10.4%	11.1%	5.9%	15.6%	16.0%	-91.7%
4	4	9.5%	10.1%	19.8%	12.3%	13.4%	5.3%	17.9%	19.5%	-100.5%
4	9	8.3%	8.7%	20.1%	11.0%	11.7%	5.1%	16.7%	17.7%	-97.0%
4	13	8.1%	8.2%	20.1%	10.9%	11.2%	4.9%	16.7%	17.0%	-95.3%
5	4	9.8%	10.3%	18.5%	12.9%	13.8%	3.4%	19.1%	20.5%	-109.9%
5	8	8.9%	9.4%	18.6%	11.9%	12.8%	3.2%	17.9%	19.3%	-105.9%
5	16	8.3%	8.4%	18.8%	11.3%	11.8%	3.3%	17.4%	18.2%	-101.3%
6	4	12.0%	13.0%	21.8%	15.4%	16.8%	7.2%	22.9%	24.8%	-122.4%
6	8	11.1%	12.2%	22.1%	14.3%	15.9%	7.2%	21.5%	23.7%	-116.1%
6	16	10.5%	11.3%	22.2%	13.6%	14.9%	7.3%	20.6%	22.5%	-114.1%
6	32	10.3%	10.6%	22.2%	13.4%	14.1%	7.2%	20.4%	21.4%	-111.1%
Average		9.2%	9.6%	21.2%	11.9%	12.7%	6.9%	17.5%	18.5%	-91.4%

[1] For each case, 500 trees are simulated. Impr is the improvement of IRI-3D over the conventional method.

[2] Area = $\sum_{i=1}^{n-1} k_i h_i$, where $n-1$ is the number of branches of a tree.

penalty in the delay improvement achieved by IRI-3D. Consequently, an area constraint is applied to IRI-3D, causing a deviation from the optimal h_i but with important benefits in area. This area constraint, denoted as A_{con}, is added to IRI-3D such that while determining s_i for each branch B_i through (5.46), the maximum $\sum h_i k_i$ does not exceed A_{con}. The average improvement in delay and area under different area constraints is reported in Table 5.10.

In Table 5.10, setting $A_{con} \leq A_{2-D}$ means that the resulting buffer area A_{con} cannot be larger than the area of the buffers resulting by the conventional method for 2-D interconnects denoted as A_{2-D}. As listed in columns three to five, both delay and area improvements are achieved. The area constraint can be employed in IRI-3D trading off area for greater delay improvement. This improvement, however, comes with a significant area overhead. By relaxing the area constraint by 20% ($i.e.$, $A_{con} \leq 1.2A_{2-D}$) an additional delay decrease of about 3% is produced while the required area is about 7% smaller than A_{2-D}. Removing the area constraint yields an additional 6% yet the area penalty exacerbates.

Limiting the area occupied by the inserted buffers is particularly important, since silicon area is also used for the TSVs that go through the substrate. In addition, the power dissipated by

the buffers is implicitly limited; an important issue for 3-D circuits where thermal effects are expected to be more pronounced [7].

5.7 Summary

The combined effect of process variations and power supply noise on the timing uncertainty of clock distribution networks is investigated in this chapter. Skitter consisting of clock skew and jitter is used to describe the clock uncertainty. The contributions of this chapter include:

- Statistical models of skitter are developed for both 2-D and 3-D clock trees. Simulation results show that separately modeling process variations and power supply noise will significantly underestimate the variation of clock uncertainty. This behavior demonstrates the necessity to simultaneously model process variations and power supply noise.
- The effect of the number and size of buffers on skitter is investigated. For the same paths, using fewer buffers produces lower skitter.
- Skitter in different scenarios of power supply noise in 3-D ICs are discussed. Skitter is shown to be significantly affected by the different amplitudes, frequencies, and initial phases of supply noise among tiers.
- A 3-D clock tree synthesis algorithm is implemented to analyze skitter in different clock trees based on industrial benchmarks.
- A fast buffer insertion algorithm for 3-D trees is proposed to decrease the total and maximum delay of interconnect trees.

Based on the analysis and simulation results, a set of design guidelines have been proposed to facilitate the design of robust clock trees:

- Using fewer buffers decreases skitter at the expense of input slew. Properly sizing up buffers help to decrease skitter by trading off power consumption.
- Recombining clock paths and increasing supply voltage both help to decrease skitter at the expense of power.
- Given the freedom to choose among tiers, for the clock paths in a 3-D circuit, the mean skitter can be decreased by placing most of the clock path length in those tiers that exhibit the lowest supply noise.
- For 3-D clock paths equally distributed among tiers, the worst-case skitter can be decreased by shifting the phase of supply noise among different tiers.
- By decreasing the frequency of resonant supply noise, the mean skitter can be decreased by trading off the standard deviation of skitter.

6 Heat Transfer Model of Thermal TSVs

Thermal issues are one of the primary challenges in 3-D integrated circuits. *Thermal through-silicon vias* (TTSVs) are considered as an effective means to reduce the temperature of 3-D ICs. The effect of the physical and technological parameters of TTSVs on the heat transfer process within 3-D ICs is investigated in this chapter.

Thermal issues in 3-D ICs are discussed in Section 6.1. The structure of TTSVs is introduced in Section 6.2. Two resistive networks are utilized to model the thermal behavior of TTSVs in Section 6.3. Based on these models, closed-form expressions are provided describing the flow of heat through TTSVs within a 3-D IC. The accuracy of these models is compared with the results from a commercial FEM tool. The effect of the physical parameters of TTSVs on the resulting temperature is described through the proposed models in Section 6.4. For example, the temperature changes non-monotonically with the thickness of the silicon substrate. This behavior is not described by traditional single thermal resistance models. In Section 6.4.5, the new models are used for the thermal analysis of a 3-D DRAM-μP system where the conventional model is shown to considerably overestimate the temperature of the system.

6.1 Thermal Issues in 3-D ICs

As introduced in Section 2.3, thermal issues become increasingly important as technology scales and the density of circuits increases. The resulting increase in temperature leads to non-negligible increase and variations in the timing and power of circuits. In 3-D integrated circuits, thermal issues are forecast to be a major challenge. This situation is due to the high power density, the low thermal conductivity along the primary heat transfer path, and the smaller footprint area of the circuit attached to the heat sink [186–188].

For instance, the different layers of materials included in a typical volumetric circuit are illustrated in Fig. 6.1. The substrates and device layers are typically made of silicon. The *inter layer dielectric* (ILD) is assumed being SiO_2 and metal interconnects (*i.e.*, BEOL) are assumed made of copper. The bonding layer adhering adjacent tiers is made of polyimide. The thermal

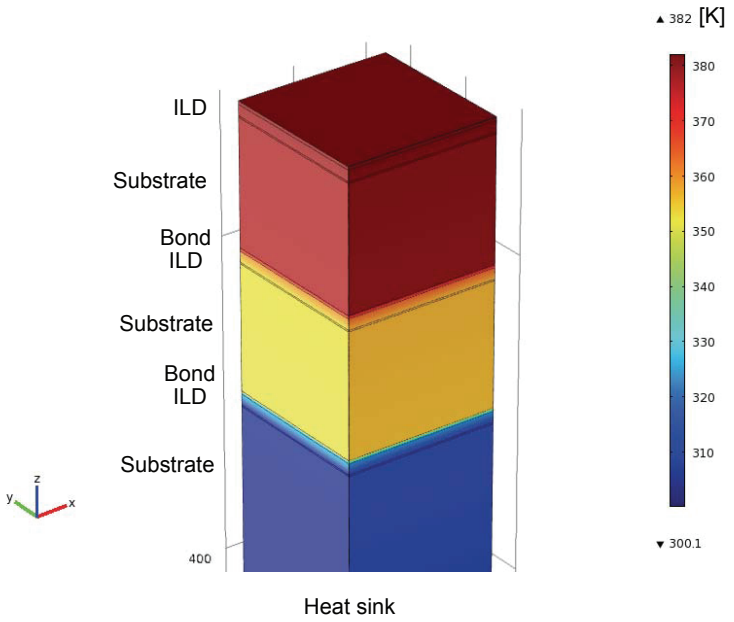


Figure 6.1: A typical 3-D circuit with different layers.

Table 6.1: The thermal conductivity of different materials used in 3-D ICs

	silicon	SiO_2	copper	polyimide
Thermal conductivity [W/(m*K)]	163	1.4	400	0.15

conductivity of the different materials is listed in Table 6.1. As listed in this table, the thermal conductivity of the ILD and bonding layer is relatively low, which hinders heat transfer and introduces large difference in temperature among tiers, as illustrated in Fig. 6.1. Assuming the temperature of the heat sink can be maintained at 80°C, the resulting maximum temperature in a 3-D circuit with different numbers of tiers is illustrated in Fig. 6.2. As shown in this figure, even if the heat sink can be maintained at a temperature of 80°C, the maximum temperature increases dramatically with the number of tiers. For a circuit with two tiers, the maximum temperature is already over 100°C. Consequently, improving the transfer of heat to decrease the temperature is necessary for 3-D ICs.

Several techniques have been developed to facilitate the heat transfer within 3-D circuits to reduce the temperature, such as *thermal through-silicon via* (TTSV) planning [127, 189], thermal wire insertion, liquid cooling, and thermal-driven floorplanning [7, 188]. TTSVs are vertical vias used only to convey heat. Using thermal vias to facilitate the transfer of heat has been traditionally utilized in the design of packages and printed circuit boards [123]. Several papers have demonstrated that TTSVs can alleviate the thermal problem in 3-D circuits [127, 189]. Analyzing how the TTSVs affect the developed temperature in 3-D ICs

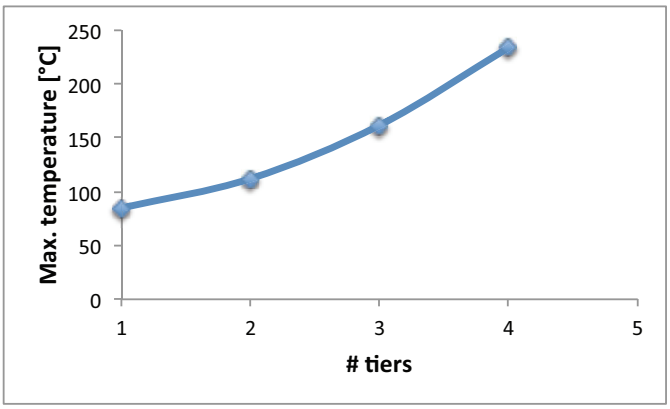


Figure 6.2: The maximum temperature of a 3-D circuit *vs.* the number of tiers.

is important for efficient TTSVs insertion. The thermal properties of TTSVs, in turn, are determined by several physical and technological parameters.

6.2 Application and Structure of Thermal TSVs

3-D ICs utilizing wafer bonding technology are considered in this chapter [7]. A segment of a typical three-plane 3-D circuit with a single TTSV is illustrated in Fig. 6.3. The physical structure is illustrated in Fig. 6.3(a). The cross section of the circuit and the temperature distribution from COMSOL Multiphysics is illustrated in Fig. 6.3(b). Although for different fabrication technologies the materials and geometries of the circuit can vary, the underlying structure remains the same.

As labeled in Fig. 6.3(a), each plane of the circuit consists of three layers describing the silicon substrate (Si), the *inter layer dielectric* (ILD) and metal interconnects (*i.e.*, BEOL), and the bonding layer, respectively. The heat sources include the power generated by the active devices on the top surface of the Si substrates and the Joule heat generated by the interconnects surrounded by the ILD. The cross section of Fig. 6.3(a) and the temperature distribution is illustrated in Fig. 6.3(b). Different paths of the flow of heat are depicted with the dashed lines in Fig. 6.3(b).

The traditional analytic approach is to thermally model a TTSV as a vertical lumped thermal resistor in each physical plane, which is proportional to the length and inversely proportional to the diameter of the TTSV [123, 186, 190, 191]. A TTSV is considered as a one-dimensional (1-D) network implying a flow of heat only in the vertical direction towards the heat sink of the system. This method is shown to be insufficient in capturing the thermal behavior of the TTSVs since the lateral heat transfer through these structures is neglected. Compact thermal models can capture the heat transfer in all directions by representing a circuit with a set of nodes connected with thermal resistors [192, 193]. Alternatively, a highly accurate and

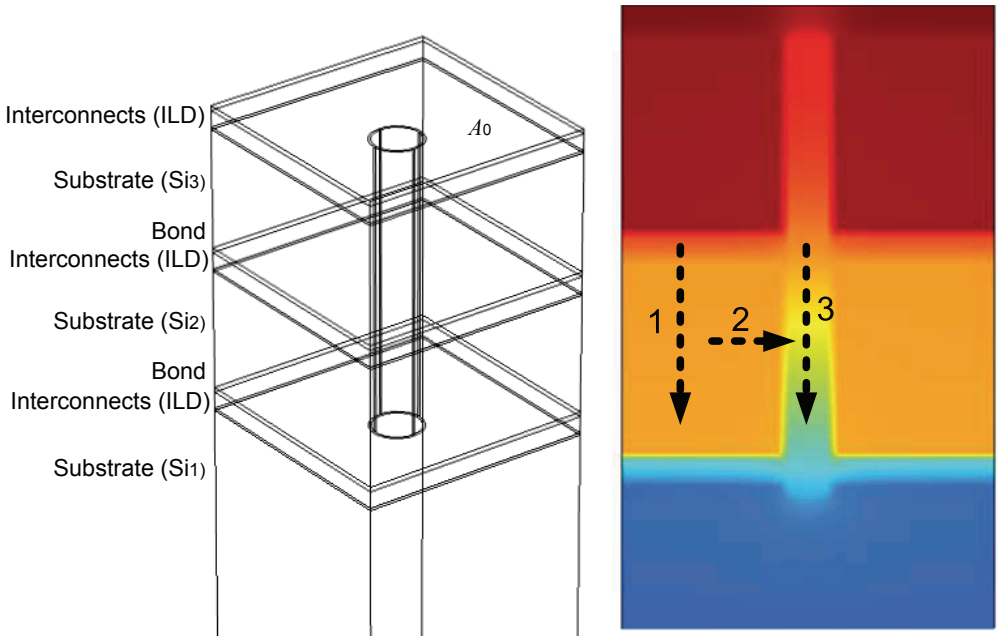


Figure 6.3: A segment of a three-plane 3-D IC with a TTSV, where (a) is the geometric structure and (b) is the cross section. The footprint area of the circuit is denoted by A_0. Three paths of heat transfer are depicted with the dashed lines.

common mesh-based method to analyze the thermal behavior of a system including TTSVs is the *finite element method* (FEM). Nevertheless, neither the compact models nor the FEM approaches offer a useful link between the heat transfer process and the physical parameters of the TTSVs.

6.3 Analytical Heat Transfer Model for Thermal TSVs

To accurately describe and offer insight about the thermal properties of these structures, novel analytical thermal models for the TTSVs are presented in this section. These models include the most important physical and technological parameters related to TTSVs, such as the thickness of the insulator liner and the thickness of silicon layers.

A thermal TSV is modeled as a compact resistive network rather than a single resistor. The accuracy of the resulting models is verified by the COMSOL Multiphysics tool [194]. The effect of the parameters related to TTSVs on temperature reduction is discussed through simulations in [195, 196]. There is, however, no analytical method to describe and quantify these effects.

The traditional thermal model of a TTSV only considers the vertical transfer of heat through TTSVs. Consequently, a single thermal resistance is assumed to suffice. Alternatively, the compact models consider heat conveyed in all directions for full-circuit analysis [192]. Im-

proved steady-state analytical models integrating the advantages of these two approaches are presented in the following subsections.

6.3.1 Lumped heat transfer model for TTSVs

The proposed lumped thermal resistance network (Model A) describing the thermal conductivity of TTSVs is illustrated in Fig. 6.4. Due to the similarity between heat transfer and electrical current flow [192], the heat sources are modeled as current sources (q_1 - q_3 in Fig. 6.4) and the temperature is analogous to the voltage at a node. In Fig. 6.4, T_0 - T_5 are used to denote the difference between the temperature in different planes and the temperature at the bottom of the first plane, which is adjacent to the heat sink and is considered as a reference temperature. A voltage source and/or another resistor can be included to describe the ambient temperature and/or the thermal resistance of the package. These elements, however, are not required for modeling the thermal behavior of the TTSVs (but rather for the temperature rise within a 3-D IC). In the proposed model, a TTSV is considered as a stack of TSVs through all the planes, as depicted in Fig. 6.4. Based on Kirchhoff's Current Law (KCL),

$$q_3 = \frac{T_5 - T_3}{R_7} + \frac{T_5 - T_4}{R_8 + R_9}, \tag{6.1}$$

$$q_2 + \frac{T_5 - T_3}{R_7} = \frac{T_3 - T_4}{R_6} + \frac{T_3 - T_1}{R_4}, \tag{6.2}$$

$$\frac{T_3 - T_4}{R_6} + \frac{T_5 - T_4}{R_8 + R_9} = \frac{T_4 - T_2}{R_5}, \tag{6.3}$$

$$q_1 + \frac{T_3 - T_1}{R_4} = \frac{T_1 - T_2}{R_3} + \frac{T_1 - T_0}{R_1}, \tag{6.4}$$

$$\frac{T_1 - T_2}{R_3} + \frac{T_4 - T_2}{R_5} = \frac{T_2 - T_0}{R_2}, \tag{6.5}$$

$$T_0 = R_s \sum_{i=1}^{3} q_i. \tag{6.6}$$

The resistances R_2, R_5, and R_8 are the thermal resistances of the filling material (*e.g.*, copper) of the TTSV. The resistances R_3, R_6, and R_9 denote the lateral thermal resistances of the insulator liner (*e.g.*, SiO_2) of the TTSV. The resistances R_1, R_4, and R_7 denote the thermal resistances of the surroundings of the TTSV (see Fig. 6.3(a)) for each of the three physical planes. The thermal resistance of the silicon substrate of the first plane is denoted by R_s due to the considerably different thickness of the substrate. These thermal resistances based on the physical parameters of the TTSVs are determined by

$$R_1 = \frac{1}{k_1 A}\left(\frac{t_D}{k_D} + \frac{l_{ext}}{k_{Si}}\right), \tag{6.7}$$

$$R_2 = \frac{t_D + l_{ext}}{k_1 k_f \pi r^2}, \tag{6.8}$$

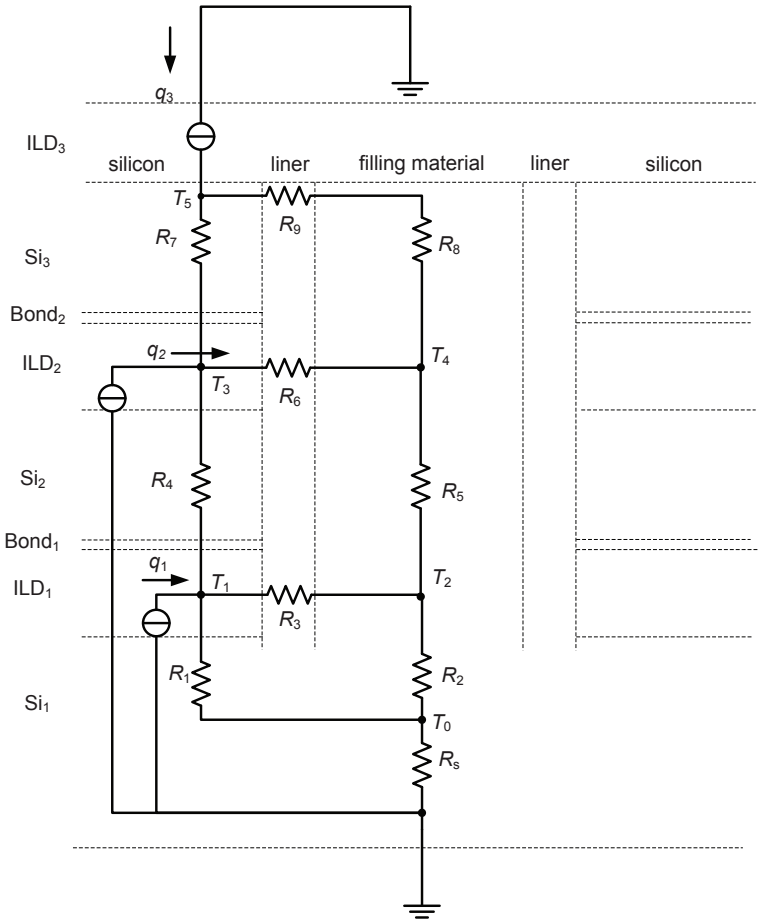


Figure 6.4: Thermal model of a TTSV in a three-plane circuit (Model A).

$$R_3 = \int_0^{t_L} \frac{1}{2\pi k_L (t_D + l_{ext})(r+x)}\, dx = \frac{\ln(r+t_L) - \ln r}{2\pi k_2 k_L (t_D + l_{ext})}, \tag{6.9}$$

$$R_4 = \frac{1}{k_1 A}\left(\frac{t_D}{k_D} + \frac{t_{Si_2}}{k_{Si}} + \frac{t_b}{k_b}\right), \tag{6.10}$$

$$R_5 = \frac{t_D + t_{Si_2} + t_b}{k_1 k_f \pi r^2}, \tag{6.11}$$

$$R_6 = \frac{\ln(r+t_L) - \ln r}{2\pi k_2 k_L (t_D + t_{Si_2} + t_b)}, \tag{6.12}$$

$$R_7 = \frac{1}{k_1 A}\left(\frac{t_D}{k_D} + \frac{t_{Si_3}}{k_{Si}} + \frac{t_b}{k_b}\right), \tag{6.13}$$

$$R_8 = \frac{t_{Si_3} + t_b}{k_1 k_f \pi r^2}, \tag{6.14}$$

$$R_9 = \frac{\ln(r+t_L) - \ln r}{2\pi k_2 k_L (t_{Si_3} + t_b)}, \tag{6.15}$$

$$R_s = \frac{t_{Si_1} - l_{ext}}{k_1 k_{Si} A_0}, \tag{6.16}$$

$$A = A_0 - \pi(r + t_L)^2. \tag{6.17}$$

The footprint area of the investigated structure is denoted by A_0, as shown in Fig. 6.3(a). The geometric parameters related to TTSVs are illustrated in Fig. 6.5. The radius of the TTSV and

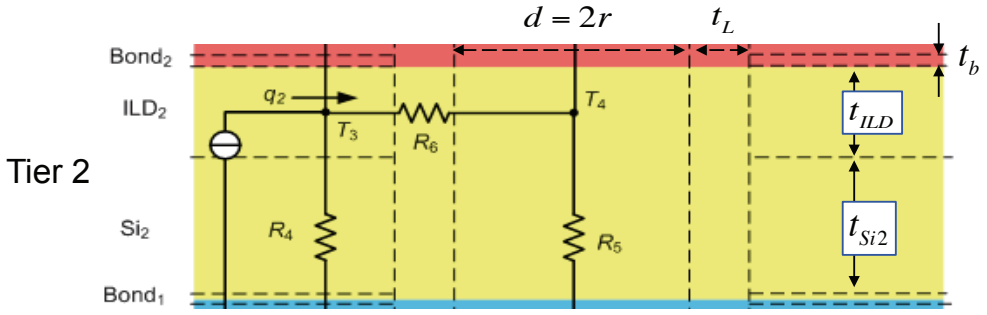


Figure 6.5: Geometric parameters related to TTSVs in the second tier.

the thickness of the insulator liner are denoted by r and t_{L}, respectively. The thickness of the silicon substrate, the BEOL layer, and the bonding layer are denoted by t_{Si}, t_{D}, and t_{b}, respectively. The thermal conductivity of these layers, the liner and filler materials of the TTSV are denoted by k_{Si}, k_{D}, k_{b}, k_{L}, and k_{f}, respectively. Since the horizontal heat transfer is more complex than the specific paths described by R_3, R_6, and R_9, the fitting coefficients k_1 and k_2 are used to decrease the discrepancy of the model from FEM simulations. If the TTSV extends into the silicon substrate in the first plane, this extended segment is denoted by l_{ext}.

Substituting (6.7)-(6.16) into (6.1)-(6.6), the resulting temperature in the three planes can be determined. Note that Model A can be extended to any number of planes. For a 3-D IC consisting of N planes, the TTSVs in the first plane are modeled by $R_1 - R_3$. The TTSVs in the N^{th} plane are modeled by $R_7 - R_9$. The TTSVs in other planes are modeled similar to $R_4 - R_6$.

6.3.2 Distributed heat transfer model for TTSVs

As mentioned before, the fitting coefficients k_1 and k_2 are required in Model A. This situation is due to the transfer of heat within one plane, modeled by three primary paths 1, 2, and 3, as shown in Fig. 6.3(b). To eliminate the need of fitting coefficients, Model A is extended to a distributed TTSV model (Model B). The lumped thermal resistors of each plane in Model A are replaced by distributed segments of thermal resistors. The resulting structures in the second plane are illustrated in Fig. 6.6. As demonstrated in this section, this model can be used to capture the thermal behavior of TTSVs with reasonable accuracy without curve fitting.

As illustrated in Fig. 6.6, the second plane is modeled by n_2 π-segments. There are n_{D2} segments for the ILD layer and n_{S2} segments for the silicon layer, where $n_2 = n_{\mathrm{D2}} + n_{\mathrm{S2}}$. For an N-plane circuit, assuming there are n_A π-segments in total, $n_A = \Sigma_{i=1}^{N} n_i$. Consequently, there

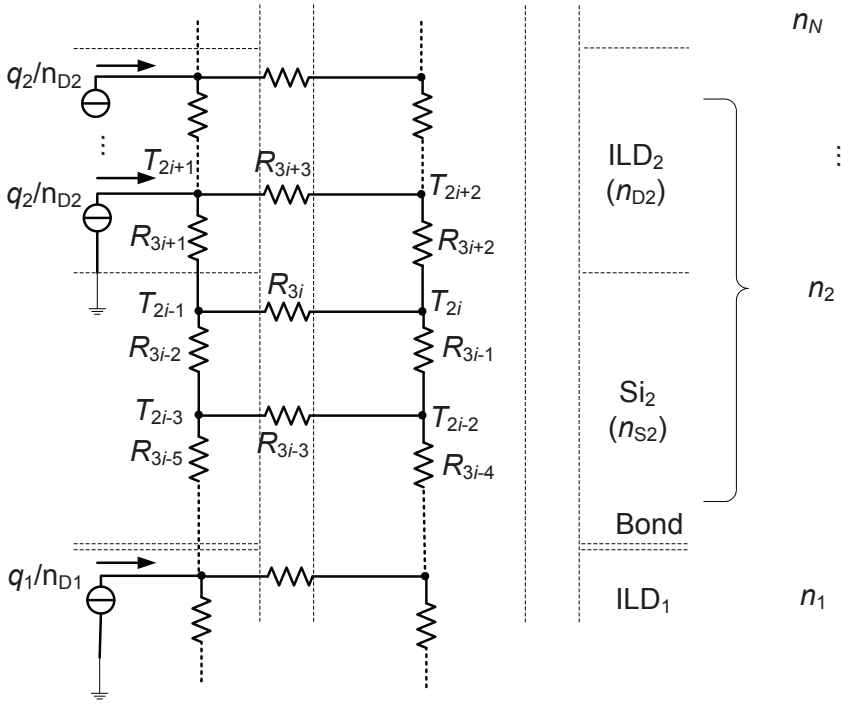


Figure 6.6: Distributed thermal model of a TTSV in the second plane (Model B).

are $2n_A$ temperature nodes ($T_1, ..., T_{2n_A}$) and $3n_A$ resistances ($R_1, ..., R_{3n_A}$) for the entire circuit, as exemplified in Fig. 6.6. The heat generated in each plane is denoted by q_i ($1 \le i \le N$).

A set of expressions similar to (6.1) - (6.5) can be obtained at each temperature node using KCL. For example, for T_{2i-1} and T_{2i} in Fig. 6.6,

$$\frac{T_{2i+1} - T_{2i-1}}{R_{3i+1}} - \frac{T_{2i-1} - T_{2i}}{R_{3i}} - \frac{T_{2i-1} - T_{2i-3}}{R_{3i-2}} = 0, \tag{6.18}$$

$$\frac{T_{2i+2} - T_{2i}}{R_{3i+2}} + \frac{T_{2i-1} - T_{2i}}{R_{3i}} - \frac{T_{2i} - T_{2i-2}}{R_{3i-1}} = 0. \tag{6.19}$$

Consequently, a linear equation system can be obtained for a 3-D circuit,

$$\mathbf{A} \cdot \mathbf{T} = \mathbf{b}. \tag{6.20}$$

In (6.20), $\mathbf{T}$ is a $2n_A \times 1$ vector corresponding to the temperatures $[T_1, T_2, ..., T_{2n_A}]'$. $\mathbf{A}$ is a $2n_A \times 2n_A$ matrix generated from the KCL expressions similar to (6.18) and (6.19). $\mathbf{b}$ is a $2n_A \times 1$ vector corresponding to the heat input at each π-segment,

$$\forall b_i \in \mathbf{b} \, (1 \le i \le 2n_A), \, b_i = \begin{cases} q_j / n_{Dj}, & \text{if } b_i \in j^{\text{th}} \text{ ILD}, \\ 0, & \text{otherwise.} \end{cases} \tag{6.21}$$

The notation "$b_i \in j^{\text{th}}$ ILD" implies that the node to which the i^{th} expression corresponds is in the j^{th} ILD layer.

The resistances $(R_1, ..., R_{3n_A})$ are the distributed resistances within each plane. For the i^{th} segment within the j^{th} plane, the related resistances are determined as follows,

$$\forall 1 \leq i \leq n_A, R_{3i-1} = R_{M_j}/n_j,\ R_{3i} = n_j R_{L_j}, \tag{6.22}$$

$$R_{3i-2} = \begin{cases} R_{ILD_j}/n_{D_j}, & \text{if } R_{3i-2} \in j^{\text{th}} \text{ ILD}, \\ R_{S_j}/n_{S_j} + R_{B_j}, & \text{for the } 1^{\text{st}} \text{ segment in } S_j, \\ R_{S_j}/n_{S_j}, & \text{otherwise.} \end{cases}$$

The horizontal resistance of the liner, the vertical resistances of the metal, the ILD, the silicon, and the bonding layer in the j^{th} plane are denoted by $R_{L_j}, R_{M_j}, R_{ILD_j}, R_{S_j}$, and R_{B_j}, respectively. These resistances are obtained similar to (6.7) - (6.15) without k_1 and k_2.

Equation (6.20) can be solved by linear system solvers. The complexity of the linear system and the required solving time are determined by the number of resistor networks. By increasing n_A, the heat transfer process related to the TTSV is more precisely described, while the time required to solve (6.20) also increases. A comparison of Model A and Model B is provided in the following section.

6.4 Effect of the Physical Parameters of TTSVs on 3-D ICs

The proposed models are compared with the results of a FEM tool [194]. The effect of the parameters related to TTSVs in the heat transfer process is also discussed in this section. Furthermore, a three-plane 3-D IC is thermally analyzed applying the proposed models.

The materials used for the ILD and bonding layers are assumed to be SiO_2 (k_D = 1.4 W/(m·K)) and polyimide (k_b = 0.15 W/(m·K)), respectively [7]. Since metal interconnects are embedded in the ILD, the k_D can be adapted to include the effect of the metal within the ILD layer. The liner of the TTSV is SiO_2 (k_L = 1.4 W/(m·K)). The footprint area, A_0, of the investigated 3-D circuit block is 100 μm × 100 μm. The thickness of the silicon substrate of the first plane is 500 μm and l_{ext} = 1 μm. Without loss of generality, the temperature of the bottom surface of the circuit adjacent to the heat sink is assumed to be 27°C. The device heat sources are assumed to be uniformly distributed on the top surface of each silicon substrate and the power density is 700 W/mm^3 [195]. The heat generated by the interconnects is assumed to be uniformly distributed in each ILD layer with a power density of 70 W/mm^3. The filling material of the TTSV is copper (k_f = 400W/(m·K)). The reduction in temperature due to the TTSV is discussed in the following subsections where different parameters are varied. To emphasize the importance of considering the lateral heat transfer in the analytical thermal model of TTSV, the proposed models are also compared with a traditional 1-D heat transfer model [186, 187, 191].

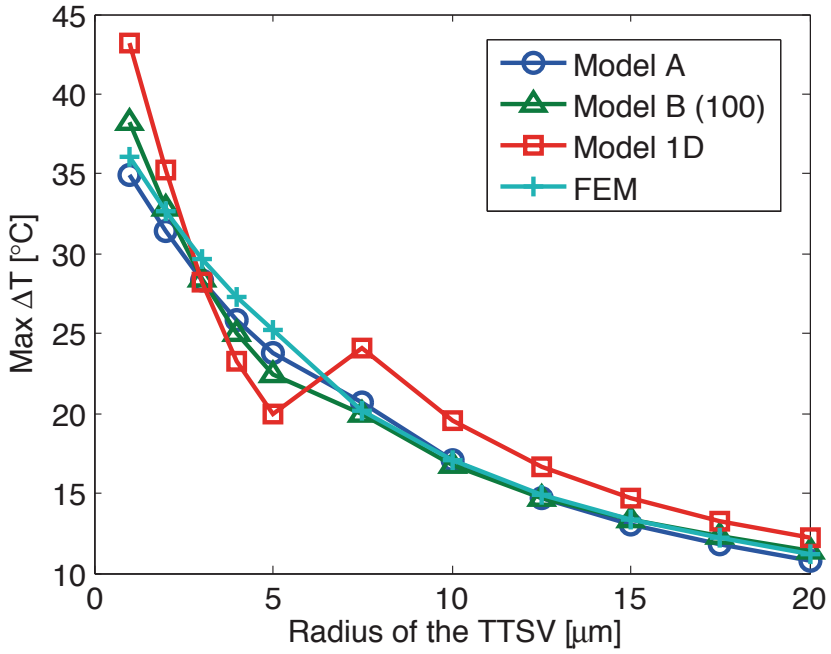


Figure 6.7: Maximum temperature rise in a three-plane 3-D IC due to different TTSV radius. $t_L = 0.5\,\mu$m, $t_D = 4\,\mu$m, $t_b = 1\,\mu$m. For $1\,\mu\text{m} \le r \le 5\,\mu$m, $t_{Si_2} = t_{Si_3} = 5\,\mu$m; for $5\,\mu\text{m} < r \le 20\,\mu$m, $t_{Si_2} = t_{Si_3} = 45\,\mu$m. $k_1 = 1.3$, and $k_2 = 0.55$.

6.4.1 The effect of the diameter of TTSVs

The effect of the diameter of the TTSV on the temperature reduction is discussed in this section. In the simulations, the radius of the TTSV, r, ranges from 1 μm to 20 μm. The other parameters are fixed except for t_{Si_2} and t_{Si_3}. Due to fabrication limitations, the aspect ratio of the TTSV (typically lower than ten [7,68]) has to be adapted according to the variation of r. The plots "Model A", "Model B (100)", "1-D", and "FEM" denote the results of Model A, Model B with 100 segments in planes 2 and 3, the traditional 1-D model, and the FEM tool, respectively.

As illustrated in Fig. 6.7, the maximum temperature rise ΔT decreases as the diameter (or radius) of the TTSV increases. When r increases, as shown in (6.7)-(6.16), the resistances R_2, R_3, R_5, R_6, R_8, and R_9 significantly decrease. Consequently, the resulting temperature T_5 (see Fig. 6.4) decreases.

In Fig. 6.7, compared with the FEM, the maximum difference (absolute value) in the steady-state temperature of Model A, Model B (100), and 1-D model is 6%, 11%, and 21%, respectively. The average difference is 3%, 3%, and 13%, respectively. Model A is more accurate than Model B, since in the first model fitting coefficients are adopted to decrease the discrepancy. Model B, however, also achieves reasonably high accuracy without the need of fitting coefficients. The 1-D model also captures the relation between r and ΔT, but the error is higher when the aspect ratio is high. This situation is because as the aspect ratio increases, the lateral

heat transfer becomes nontrivial as compared with the vertical flow of heat. Consequently, neglecting path 2 (see Fig. 6.3(b)) introduces a higher error.

6.4.2 The effect of the thickness of the dielectric liner

The effect of the thickness of the dielectric liner surrounding the TTSV on the temperature reduction is discussed in this section. The dielectric liner ranges from 0.5 μm to 3 μm. The other parameters are provided in Fig. 6.8.

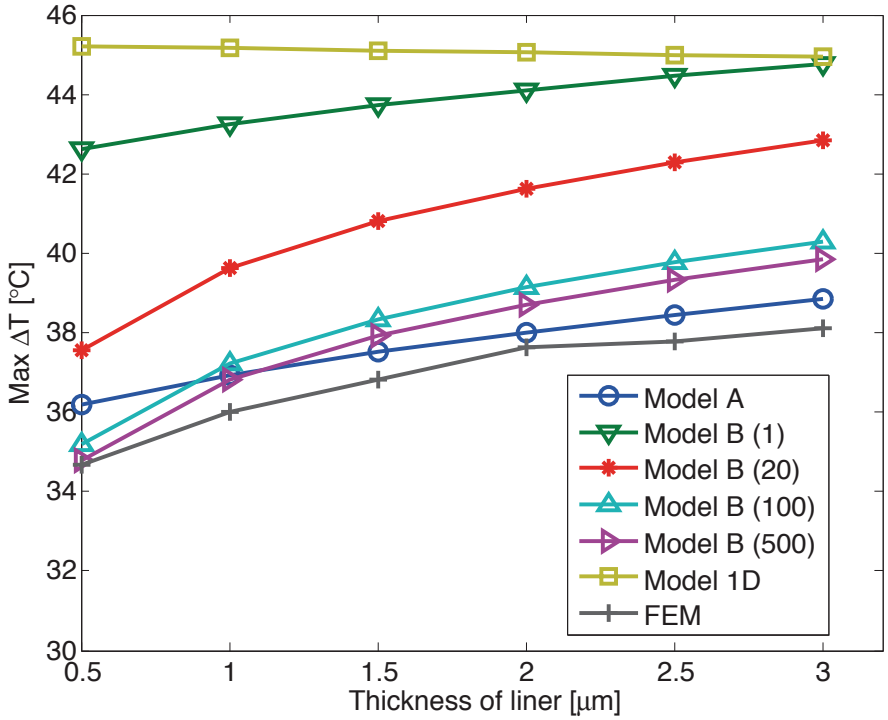


Figure 6.8: Maximum temperature rise in a three-plane 3-D IC for different thickness of the dielectric liner, where $r = 5\,\mu$m. The other parameters are $t_D = 7\,\mu$m. $t_b = 1\,\mu$m, $t_{Si_2} = t_{Si_3} = 45\,\mu$m. $k_1 = 1.3$ and $k_2 = 0.55$.

As shown in Fig. 6.8, the thickness of the TTSV dielectric liner considerably affects the resulting temperature, a behavior not captured by the conventional 1-D thermal TTSV model. For different t_L, the ΔT from FEM differs up to 11%, which is approximately 4°C for this specific setup.

As expressed by (6.9), (6.12), and (6.15), the resistances R_3, R_6, and R_9 increase as t_L increases. The resulting temperatures, consequently, increase significantly. These resistances R_3, R_6, and R_9 increase linearly with $\ln t_L$. Consequently, the change of ΔT with t_L is smaller than the change of ΔT with r. Since the traditional TTSV model only considers vertical 1-D heat transfer through the liner [123, 186, 190], the lateral or horizontal heat transfer is ignored.

Table 6.2: The Error and Run Time vs. # of Segments in Model B.

Model	B (1)	B (20)	B (100)	B (500)	A	1-D
Max. Error	23%	12%	6%	5%	4%	30%
Av. Error	19%	11%	4%	3%	2%	23%
Time [ms]	1	3	32	2475	-	-

For various t_{L}, different number of segments are investigated for Model B. The numbers of segments within the first plane and the other planes are (1, 1), (2, 20), (10, 100), and (50, 500), respectively, as denoted by Model B (1) - Model B (500). The maximum and average difference in the temperature between the four cases and FEM are reported in Table 6.2. The run time for these four cases is also reported. The accuracy of Model B increases with the number of segments within each plane, while the run time also increases significantly.

6.4.3 The effect of the thickness of the silicon substrate

The effect of the thickness of the silicon substrate (t_{Si_2} and t_{Si_3}) on the temperature reduction is discussed in this section. For 1-D heat transfer models, the temperature increases as t_{Si_2} and t_{Si_3} increase. The results of FEM simulations, however, exhibit a different behavior.

The change of ΔT according to different t_{Si_2} and t_{Si_3} is illustrated in Fig. 6.9. The thickness of the silicon substrate ranges from 5 μm to 80 μm. The other parameters are listed in Fig. 6.9. ΔT changes non-monotonically with the thickness of the silicon substrates, another behavior that cannot be described by the 1-D heat transfer model. As illustrated in Fig. 6.9, within the range $5\,\mu\mathrm{m} \leq t_{\mathrm{Si}_2} \leq 20\,\mu\mathrm{m}$, ΔT decreases as t_{Si_2} increases. For $t_{\mathrm{Si}_2} > 20\,\mu\mathrm{m}$, ΔT increases with t_{Si_2}.

Both Model A and Model B capture this behavior. As described by (6.10)-(6.16), the vertical thermal resistances (R_4, R_5, R_7, and R_8) in Fig. 6.4 increase as t_{Si_2} and t_{Si_3} increase, which implies that the thermal resistance along the vertical path of the heat transfer increases. Nevertheless, the horizontal thermal resistances described by (6.12) and (6.15) decrease as t_{Si_2} and t_{Si_3} increase, which indicates that the thermal resistance along the horizontal path of the heat transfer through the liner of the TTSV decreases. The combination of these two effects leads to the non-monotonic change in temperature. Consequently, thinning the silicon substrate cannot always improve the heat transfer within a 3-D IC with TTSVs, since wafer thinning limits the lateral spreading of heat within the substrate [196].

As shown in Fig. 6.9, the average error of Model A, Model B (100), and 1-D model is 4%, 6%, and 17%, respectively. The maximum error is 7%, 18%, and 32%, respectively. When the silicon layer is thin, all the models introduce a high error.

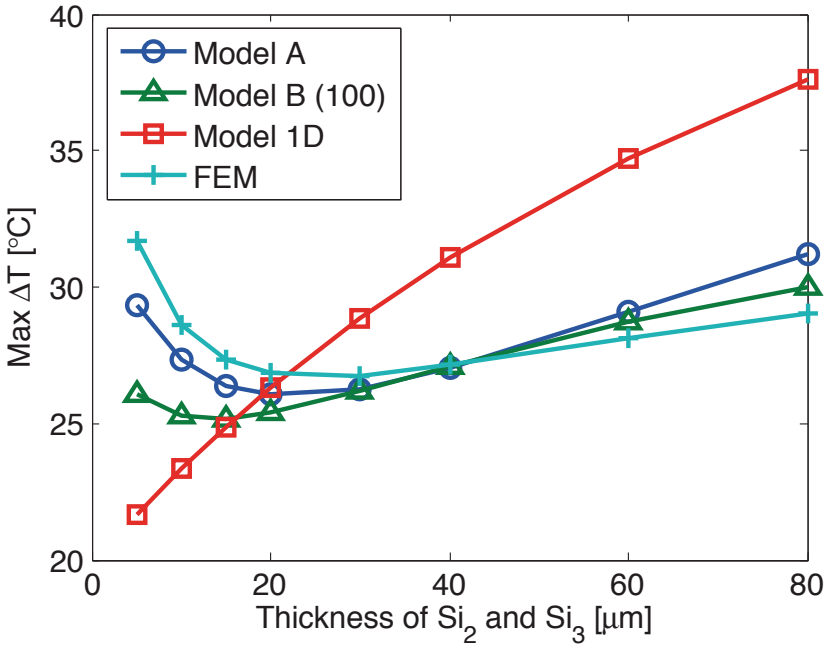


Figure 6.9: Maximum temperature rise in a three-plane 3-D IC due to different thickness of the silicon substrate. The other parameters are $t_L = 1\,\mu$m, $t_D = 7\,\mu$m, $t_b = 1\,\mu$m, $r = 8\,\mu$m, $k_1 = 1.3$ and $k_2 = 0.55$.

6.4.4 The effect of TTSV density

The effect of dividing a large TTSV into a cluster of thinner TTSVs on the temperature reduction is discussed in this section. Several works have shown that by replacing a large-diameter TTSV with a cluster of small-diameter TTSVs, the temperature of a 3-D IC can be further reduced [195].

While dividing a TTSV into a cluster of thin TTSVs, the total area of the metal forming the TTSVs is assumed to be the same, as illustrated in Fig. 6.10. As a result, if a TTSV with radius

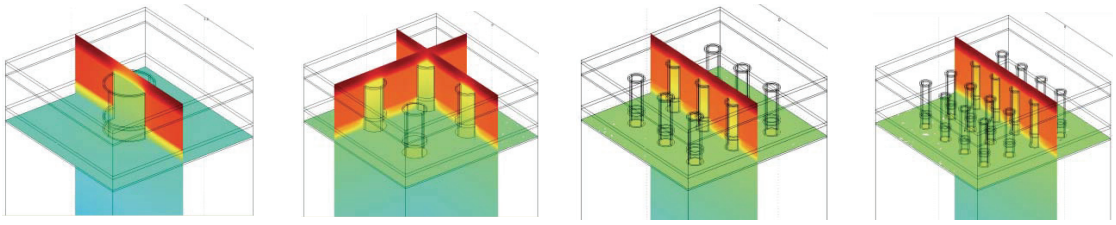


Figure 6.10: Dividing a large TSV into four, nine, and 16 smaller TSVs.

r_0 is divided into n TTSVs, the radius of the new TTSVs is $r_n = \frac{r_0}{\sqrt{n}}$. The other parameters remain the same. In the proposed model, the new cluster of TTSVs is modeled as an equivalent thermal resistance network $R'_i\,(1 \le i \le 9)$. Since the total metal area within the TTSVs remains the same, the vertical thermal resistances remain the same, $R'_i = R_i\,(i \ne 3,6,9)$. The horizontal

resistances are updated from (6.9) as the total lateral surface of the TTSVs increases,

$$R'_3 = \frac{\ln(t_L\sqrt{n}+r_0)-\ln r_0}{2n\pi k_2 k_L(t_D+l_{ext})}. \tag{6.23}$$

R'_6 and R'_9 are updated similar to (6.23). In the simulations, a TTSV is divided into 2, 4, 9, and 16 TTSVs.

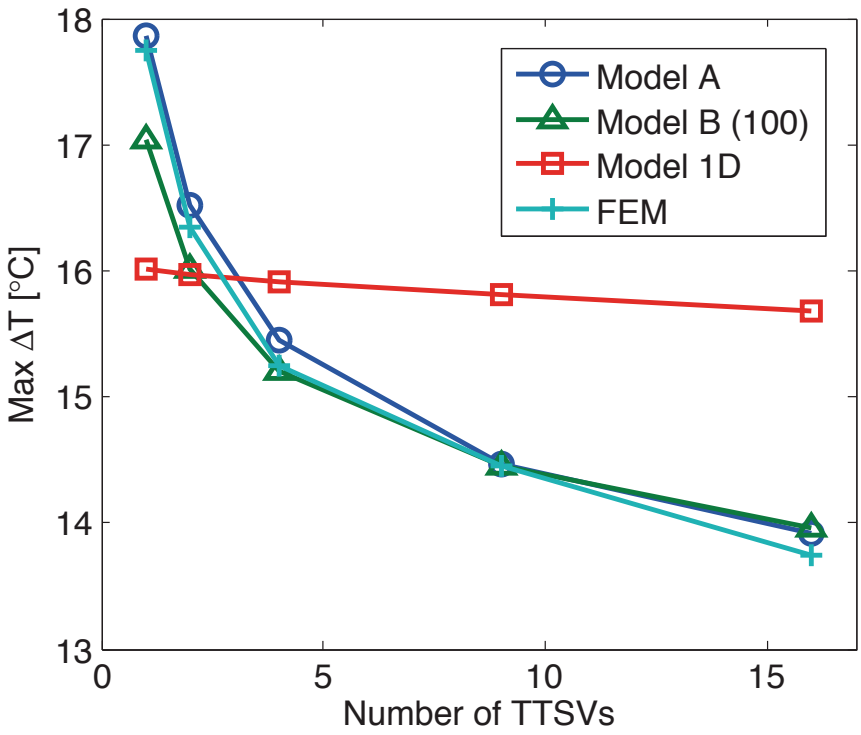


Figure 6.11: Maximum temperature rise in a three-plane 3-D IC due to different thickness of the silicon substrate. $t_L = 1\,\mu$m, $t_D = 4\,\mu$m, $t_b = 1\,\mu$m, $t_{Si_2} = t_{Si_3} = 20\,\mu$m, $r_0 = 10\,\mu$m, $k_1 = 1.3$, and $k_2 = 0.55$.

As shown in Fig. 6.11, ΔT decreases as a single TTSV is divided into more TTSVs. This behavior is because as a TTSV is divided into more TTSVs, the total lateral surface increases and more heat is conducted through the TTSVs. According to (6.23), as n increases, R'_3, R'_6, and R'_9 decrease, which causes the temperature to decrease. As depicted by the three plots, the decrease in temperature with the number of TTSVs saturates as n increases. Consequently, dividing a TTSV into more and thinner TTSVs exhibits a diminishing improvement after a specific n.

The average error of Model A, Model B (100), and 1-D model is 1%, 2%, and 8%, respectively. The maximum error is 1%, 4%, and 14%, respectively. As illustrated in Fig. 6.11, both of Model A and Model B correctly describe the expected behavior of dividing a TTSV. Since the area of

the metal forming the TTSV remains the same for any n, the 1-D model cannot describe the temperature reduction with n.

6.4.5 3-D DRAM-μP Case Study

The proposed models are used to evaluate the temperature rise in a 3-D circuit. The physical parameters of the circuit are based on [191, 195], where a 1-D heat transfer model is used for the system. The circuit consists of three physical planes with face-to-back bonding. The footprint area is 10 mm × 10 mm. The thickness of the silicon substrate (t_{Si}) in each plane is 300 μm. The power dissipated by the μP and DRAM planes is 70 W and 7 W, respectively. The TTSVs are uniformly distributed with a density of 0.5% of the total circuit area. The proposed models are embedded in the analytic thermal analysis model of the system. The FEM simulation and 1-D TTSV model are also implemented for comparison. The structure of the circuit and the other parameters are illustrated in Fig. 6.12.

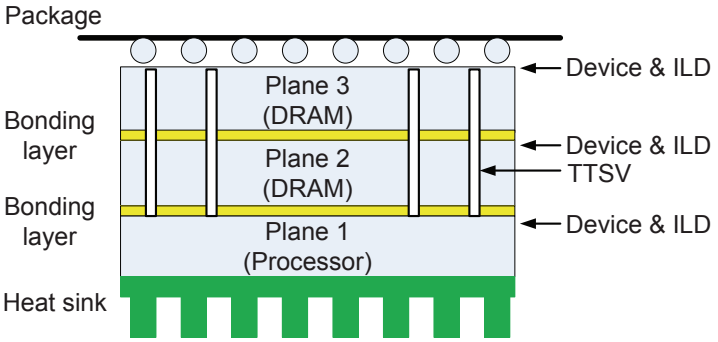


Figure 6.12: A three-plane 3-D circuit with TTSVs. $t_L = 1\,\mu$m, $t_D = 20\,\mu$m, $t_b = 10\,\mu$m, $t_{Si_1} = t_{Si_2} = t_{Si_3} = 300\,\mu$m, $r = 30\,\mu$m, $k_1 = 1.6$, $k_2 = 0.8$, and $c_{1,2} = 3.5$.

Since $t_{Si} = 300\,\mu$m, the TTSVs in the second and third planes are divided into 1000 segments for Model B. For the investigated 3-D circuit, the resulting maximum temperature rise from the heat sink for Model A, Model B (1000), FEM, and 1-D model are 12.8°C, 13.9°C, 12°C, and 20°C, respectively. The runtime of FEM is 59 minutes. The fitting coefficients of Model A are determined by the simulation of a block of the investigated circuit as shown in Fig. 6.3, the runtime of which is 1.9 minutes. The runtime for Model B (1000) is 8.5 seconds. As demonstrated by this example, the proposed models are efficient and reasonably accurate while the 1-D model is highly inaccurate even for this first-order analysis.

6.5 Summary

Heat transfer models of thermal TSVs are investigated in this chapter. Two analytic steady-state models for TTSVs are proposed. Compared with the traditional 1-D analytic model, the proposed models produce significantly higher accuracy. Compared with FEM methods,

the proposed models are significantly faster with reasonably high accuracy. Based on these thermal models, the thermal behavior of 3-D ICs under different physical parameters of TTSVs is investigated:

- The maximum temperature of 3-D ICs decreases with the diameter of TTSVs.
- The maximum temperature of 3-D ICs increases with the thickness of the dielectric liner of TTSVs.
- The maximum temperature of 3-D ICs changes non-monotonically with the thickness of silicon substrates. When this thickness is relatively small, the temperature decreases with the thickness of substrates. After a specific thickness, the temperature increases with the thickness of substrates.
- Dividing a large TTSV into a cluster of thinner TTSVs helps to further decrease the temperature of 3-D circuits.

Ignoring these effects can result in significant overestimate of the temperature increase in 3-D ICs where TTSVs are utilized, as demonstrated by a first-order thermal analysis of a 3-D DRAM-μP system. Adapting a 1-D model, therefore, in a TTSV insertion/planning methodology can result in excessive usage of TTSVs (a critical resource in 3-D ICs), with an immediate increase in the total cost of the system.

7 Conclusions and Future Directions

The conclusions of this dissertation are drawn in the following section, where the contributions of this thesis are also summarized. The potential future directions in this area are discussed in Section 7.2.

7.1 Conclusions

Process variations, power supply noise, and the thermal behavior of 3-D ICs are all investigated in this dissertation. Novel models for 3-D ICs are proposed to correctly describe process, voltage, and temperature variations. Design techniques and guidelines are presented to mitigate the negative clock uncertainty caused by different sources of variations.

Process variations in 3-D ICs are investigated in Chapter 3. The effect of process variations in 3-D clock distribution networks is modeled and analyzed. A novel model to describe the distribution of process-induced skew in 3-D clock trees, which exhibits reasonably high accuracy, is proposed. Typical 3-D clock distribution networks are compared among each other in terms of clock skew variation. 3-D clock grids exhibit the lowest skew variation but with a significant cost in power consumption. For 3-D clock trees, the multi-via topology outperforms the single-via topology in terms of the maximum skew variation and power consumption, since the single-via topology requires a larger number of buffers. For clock sinks within the same tier, however, single-via 3-D clock trees usually produce a lower skew variation due to the smaller number of buffers per tier.

A new 3-D clock tree topology is proposed to combine the advantages of both the multi- and single-via topologies, which produces a low skew variation for the clock sinks within the same group. The skew variation in multi-domain clock trees is also investigated. It is shown that placing different clock domains in different tiers does not necessarily produce the lowest skew. Skew variation can be decreased by locating different clock domains within the same tier and vertically extending these domains. For spatially correlated WID process variations, increasing the number of tiers a clock domain spans increases the skew variation between the sinks

located within a short distance. The maximum skew variation, however, is determined by the sinks with the farthest distance. The change of the maximum skew depends on the relation between D2D and WID variations.

The power supply noise in 3-D PDNs is investigated in Chapter 4. A fast steady-state *IR*-drop analysis method is developed for 3-D power grids. In this method, the row-based algorithm for 2-D power grids is extended to consider the influence of P/G TSVs and the interaction among tiers. Compared to SPICE-based simulations, the proposed method achieves reasonably high accuracy and savings in the computing resources.

The resonant noise in 3-D PDNs is investigated based on the one-dimensional model. Under different scenarios of 3-D PDNs, the resonant noise exhibits different characteristics among tiers. For various schemes of switching current and turn-on time, the tier adjacent to the package and the heat sink experience the lowest and highest amplitude of resonant noise, respectively. The difference in the amplitude of resonant noise increases with the resistance of P/G TSVs and the number of tiers. The frequency of resonant noise slightly differs among tiers, with a difference lower than 10% in the simulations.

The combined effect of process variations and power supply noise on the timing uncertainty of clock distribution networks is investigated in Chapter 5. Skitter consisting of clock skew and jitter is used to describe the clock uncertainty. Statistical models of skitter are developed for both 2-D and 3-D clock trees. Simulation results show that separately modeling process variations and power supply noise will significantly underestimate the variation of clock uncertainty. The effect of the number and size of buffers on skitter is investigated. For the same paths, using fewer buffers produces lower skitter. Skitter in different scenarios of power supply noise in 3-D ICs are discussed. Skitter is shown to be significantly affected by the different amplitudes, frequencies, and initial phases of supply noise among tiers. A 3-D clock tree synthesis algorithm is implemented to analyze skitter in different clock trees based on industrial benchmarks. A fast buffer insertion algorithm for 3-D trees is proposed to decrease the total and maximum delay of interconnect trees.

Based on the analysis and simulation results, a set of design guidelines have been proposed to facilitate the design of robust clock trees:

- Using fewer buffers decreases skitter at the expense of input slew. Properly sizing up buffers helps to decrease skitter by trading off power consumption.

- Recombining clock paths and/or increasing supply voltage both help to decrease skitter at the expense of power.

- Given the freedom to choose among tiers, for the clock paths in a 3-D circuit, the mean skitter can be decreased by placing most of the clock path length in those tiers that exhibit the lowest supply noise.

- For 3-D clock paths equally distributed among tiers, the worst-case skitter can be decreased by shifting the phase of supply noise among different tiers.

- By decreasing the frequency of resonant supply noise, the mean skitter can be decreased by trading off the standard deviation of skitter.

The thermal behavior of 3-D ICs is investigated in Chapter 6. Two analytic steady-state models for thermal TSVs are proposed. Compared with the traditional 1-D analytic model, the proposed models produce significantly higher accuracy. Compared with FEM methods, the proposed models are significantly faster with reasonably high accuracy.

Based on these thermal models, the thermal behavior of 3-D ICs under different physical parameters of TTSVs is investigated. The maximum temperature of 3-D ICs decreases with the diameter of TTSVs and increases with the thickness of the dielectric liner of TTSVs. When the substrate thickness is relatively small, the temperature decreases with this thickness. Nevertheless, after a specific thickness, the temperature increases with the thickness of substrates. In addition, dividing a large TTSV into a cluster of thinner TTSVs helps to further decrease the temperature in 3-D circuits. Ignoring these effects can result in significant overestimate of the temperature increase in 3-D ICs where TTSVs are utilized, as demonstrated by a first-order thermal analysis of a 3-D DRAM-μP system.

The proposed models for skew, skitter, and thermal TSVs are used to fast and correctly estimate the timing and thermal behavior of 3-D ICs under different sources of variations. The design guidelines provided with these models facilitate the design of robust 3-D circuits. Since PVT variations increase significantly as technology advances, efficiently estimating and mitigating the negative effect of PVT variations is critical to the design of 3-D ICs.

7.2 Future Directions

In addition to the topics investigated in this dissertation, there are other important issues relating to the variation-aware design of 3-D ICs. Potential future research directions in the design of 3-D ICs under PVT variations are presented next.

Process-induced skew in other topologies of 3-D clock distribution networks

The skew variation of clock trees is carefully modeled in this thesis. Nevertheless, the skew in other topologies, such as clock meshes, recombinant clock trees, and clock spines, have not yet been analytically modeled. Since hybrid clock distribution networks are widely used in modern high-density ICs [38, 169, 197], it is necessary to model the skew in different clock distribution networks. The deterministic skew in non-tree based topologies has been modeled for 2-D ICs. For instance, the skew in 2-D clock meshes is investigated in [198]. The skew in 3-D non-tree based topologies under PVT variations has not been modeled. Fast and accurate

methods to measure this skew are required to facilitate the design of robust and cost-efficient hybrid 3-D clock distribution networks.

Accurate models for power supply noise in 3-D power distribution networks

Steady-state *IR*-drop analysis and resonant noise are discussed separately in this thesis. The DC *IR*-drop differs among devices. When analyzing the temporal change of resonant supply noise the devices within the same tier are assumed to experience similar voltage variations. An accurate model to fast describe the different transient response of devices to the power supply noise is required. This type of models has been proposed for 2-D power grids [118]. For 3-D power distribution networks, the correlation among tiers through P/G TSVs has to be considered.

Clock-data compensation in 3-D ICs

Clock uncertainty is modeled independently from the delay variation of the data signals in this dissertation. Nevertheless, the highest clock frequency of a circuit is determined by the setup and hold time slacks, which are determined by the combination of clock and data delays. Research on 2-D ICs [29, 30] has shown that, when clock and data variations are considered together, the effect of power supply noise on the speed of circuits significantly varies. The negative effect of power supply noise can be mitigated utilizing this clock-data compensation. In 3-D ICs, the effect of clock-data compensation on the speed of circuits needs to be investigated to provide more efficient design guidelines and methods to distribute the clock and data signals.

Timing uncertainty under spatial thermal variations

Timing uncertainty of 3-D ICs is modeled in this thesis considering process and power variations. The timing uncertainty due to the spatial differences in temperature within a 3-D circuit has not yet been investigated. In 3-D ICs, temperature differs both within a tier and significantly across tiers [34]. Since the delay of transistors and wires is sensitive to temperature, the resulting timing uncertainty needs to be investigated.

As 3-D integration provides significantly higher device density, faster interconnection, and easier heterogeneous integration, complicated PVT variations also result in more challenges in the design of robust 3-D ICs. Coping with these challenges is critical to fully exploit the advantages of 3-D circuits. More research efforts are required to solve the above challenges and, thereby, improve the performance of 3-D circuits. This dissertation has provided the necessary means including both models and methods to significantly improve the robustness of 3-D ICs under these variations.

Bibliography

[1] Intel Corporation, "Microprocessor Quick Reference Guide," 2012. [Online]. Available: http://ark.intel.com

[2] N. Weste and D. M. Harris, *CMOS VLSI Design: A Circuits and Systems Perspective*, 4th ed. Pearson/Addison-Wesley, 2011.

[3] YOLE Developpement, "3D IC & TSV Report," November 2007.

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

[5] J. Gibbons and K. Lee, "One-Gate-Wide CMOS Inverter on Laser-Recrystallized Polysilicon," *IEEE Electron Device Letters*, Vol. 1, No. 6, pp. 117–118, June 1980.

[6] J. Joyner *et al.*, "Impact of Three-Dimensional Architectures on Interconnects in Gigascale Integration," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, Vol. 9, No. 6, pp. 922–928, December 2001.

[7] V. Pavlidis and E. Friedman, *Three-Dimensional Integrated Circuit Design*. Morgan Kaufmann Pub, 2009.

[8] M. Karnezos, "3D Packaging: Where All Technologies Come Together," in *Proceedings of IEEE/CPMT/SEMI International Electronics Manufacturing Technology Symposium*, July 2004, pp. 64–67.

[9] S. Al-Sarawi, D. Abbott, and P. Franzon, "A Review of 3-D Packaging Technology," *IEEE Transactions on Components, Packaging, and Manufacturing Technology: Part B*, Vol. 21, No. 1, pp. 2–14, February 1998.

[10] K. Tanida *et al.*, "Ultra-High-Density 3D Chip Stacking Technology," in *Proceedings of Electronic Components and Technology Conference*, May 2003, pp. 1084–1089.

[11] K. Saban, "Xilinx Stacked Silicon Interconnect Technology Delivers Breakthrough FPGA Capacity, Bandwidth, and Power Efficiency." Xilinx, October 2011.

[12] Intel Corporation, "Intel 22nm 3-D Tri-Gate Transistor Technology," April 2011. [Online]. Available: http://newsroom.intel.com/docs/DOC-2032

[13] X. Wu *et al.*, "A Three-Dimensional Stacked Fin-CMOS Technology for High-Density ULSI Circuits," *IEEE Transactions on Electron Devices*, Vol. 52, No. 9, pp. 1998–2003, September 2005.

[14] P. S. Andry *et al.*, "Fabrication and Characterization of Robust Through-Silicon Vias for Silicon-Carrier Applications," *IBM Journal of Research and Development*, Vol. 52, No. 6, pp. 571–581, November 2008.

[15] K. W. Lee *et al.*, "Development of Three-Dimensional Integration Technology for Highly Parallel Image-Processing Chip," *Japanese Journal of Applied Physics*, Vol. 39, No. 4B, pp. 2473–2477, April 2000.

[16] P. Andry *et al.*, "A CMOS-Compatible Process for Fabricating Electrical Through-Vias in Silicon," in *Proceedings of Electronic Components and Technology Conference*, 2006, pp. 831–837.

[17] D. M. Jang *et al.*, "Development and Evaluation of 3-D SiP with Vertically Interconnected Through Silicon Vias (TSV)," in *Proceedings of Electronic Components and Technology Conference*, July 2007, pp. 847–852.

[18] T. Xanthopoulos, *Clocking in Modern VLSI Systems*. Springer, 2009.

[19] V. F. Pavlidis, I. Savidis, and E. G. Friedman, "Clock Distribution Networks in 3-D Integrated Systems," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, Vol. 19, No. 12, pp. 2256–2266, December 2011.

[20] X. Zhao, J. Minz, and S. Lim, "Low-Power and Reliable Clock Network Design for Through-Silicon via (TSV) Based 3D ICs," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, Vol. 1, No. 2, pp. 247–259, February 2011.

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

[22] M. Orshansky, S. R. Nassif, and D. Boning, *Design for Manufacturability and Statistical Design: A Constructive Approach*. Springer, 2008.

[23] K. A. Bowman *et al.*, "Impact of Extrinsic and Intrinsic Parameter Fluctuations on CMOS Circuit Performance," *IEEE Journal of Solid-State Circuits*, Vol. 35, No. 8, pp. 1186–1193, August 2000.

[24] S. Borkar *et al.*, "Parameter Variations and Impact on Circuits and Microarchitecture," in *Proceedings of IEEE/ACM Design Automation Conference*, June 2003, pp. 338–342.

[25] R. Jakushokas *et al.*, *Power Distribution Networks with On-Chip Decoupling Capacitors*, 2nd ed. Springer, 2011.

[26] D. Blaauw, R. Panda, and R. Chaudhry, "Design and Analysis of Power Distribution Networks," in *Design of High-Performance Microprocessor Circuits*, A. P. Chandrakasan, W. J. Bowhill, and F. Fox, Eds. Wiley-IEEE Press, 2001, pp. 499–522.

[27] S. R. Nassif, "Power Grid Analysis Benchmarks," in *Proceedings of the Asia and South Pacific Design Automation Conference*, January 2008, pp. 376–381.

[28] L. Cao and J. P. Krusius, "Bonded ICs and Packages of Future Deep Sub-Micron ULSI," in *Proceedings of Electronic Components and Technology Conference*, May 1997, pp. 1138–1145.

[29] D. Jiao, J. Gu, and C. Kim, "Circuit Design and Modeling Techniques for Enhancing the Clock-Data Compensation Effect Under Resonant Supply Noise," *IEEE Journal of Solid-State Circuits*, Vol. 45, No. 10, pp. 2130–2141, October 2010.

[30] K. Wong *et al.*, "Enhancing Microprocessor Immunity to Power Supply Noise With Clock-Data Compensation," *IEEE Journal of Solid-State Circuits*, Vol. 41, No. 4, pp. 749–758, April 2006.

[31] M. Pant, "Microprocessor Power Impacts," *Tutorials of Great Lakes Symposium on VLSI Systems*, May 2010.

[32] G. Taylor, "Energy Efficient Circuit Design and the Future of Power Delivery," in *Tutorial of IEEE Conference on Electrical Performance of Electronic Packaging and Systems*, October 2009.

[33] A. Vassighi and M. Sachdev, *Thermal and Power Management of Integrated Circuits*. Springer, 2006.

[34] P. Jain *et al.*, "Thermal and Power Delivery Challenges in 3D ICs," in *Three Dimensional Integrated Circuit Design*, ser. Integrated Circuits and Systems, Y. Xie, J. Cong, and S. Sapatnekar, Eds. Boston, MA: Springer US, 2010.

[35] K. Skadron *et al.*, "Temperature-Aware Microarchitecture," in *Proceedings of the ACM International Symposium on Computer Architecture*, June 2003, pp. 2–13.

[36] ——, "Temperature-Aware Microarchitecture: Modeling and Implementation," *ACM Transactions on Architecture and Code Optimization*, Vol. 1, No. 1, pp. 94–125, March 2004.

[37] A. Agarwal, D. Blaauw, and V. Zolotov, "Statistical Timing Analysis for Intra-Die Process Variations with Spatial Correlations," in *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, 2003, pp. 900–907.

[38] P. Restle *et al.*, "The Clock Distribution of the Power4 Microprocessor," in *Digest of Technical Papers of IEEE International Solid-State Circuits Conference*, Vol. 88, February 2002, pp. 144–145.

[39] H. Bakoglu, *Circuits, Interconnections, and Packaging for VLSI*. Addison-Wesley Pub, 1990.

[40] V. Adler and E. G. Friedman, "Uniform Repeater Insertion in *RC* Trees," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, Vol. 47, No. 10, pp. 51–61, October 2000.

[41] Nanoscale Integration and Modeling (NIMO) Group, "Predictive Technology Model (PTM)." Arizona State University, USA, 2008. [Online]. Available: http://ptm.asu.edu

[42] C. Weiner, "How the Transistor Emerged," *IEEE Spectrum*, Vol. 10, No. 1, pp. 24–33, January 1973.

[43] "International Technology Roadmap for Semiconductors (ITRS)," 2011. [Online]. Available: http://www.itrs.net

[44] T. Ghani, "Challenges and Innovations in Nano-CMOS Transistor Scaling." Intel Corporation, October 2009.

[45] Samsung Electronics Co. Ltd., "Samsung Readies Green Memory with Advanced Chip Stacking Technology after Extensive System-level Testing," 2010. [Online]. Available: http://www.samsung.com/global/business/semiconductor/minisite/Greenmemory/main.html

[46] U. Kang *et al.*, "8 Gb 3-D DDR3 DRAM Using Through-Silicon-Via Technology," *IEEE Journal of Solid-State Circuits*, Vol. 45, No. 1, pp. 111–119, January 2010.

[47] V. Jain *et al.*, "A Highly Reconfigurable Computing Array: DSP Plane of a 3-D Heterogeneous SoC," in *Proceedings of IEEE International SOC Conference*, September 2005, pp. 243–246.

[48] M. Koyanagi *et al.*, "Future System-on-Silicon LSI Chips," *IEEE Micro*, Vol. 18, No. 4, pp. 17–22, July 1998.

[49] Y. Xie, J. Cong, and S. Sapatnekar, Eds., *Three-Dimensional Integrated Circuit Design: EDA, Design and Microarchitectures*. Springer, 2010.

[50] R. Tummala, "SOP: What Is It and Why? A New Microsystem-Integration Technology Paradigm-Moore's Law for System Integration of Miniaturized Convergent Systems of the Next Decade," *IEEE Transactions on Advanced Packaging*, Vol. 27, No. 2, pp. 241–249, May 2004.

[51] Y. Liu *et al.*, "Fine Grain 3D Integration for Microarchitecture Design through Cube Packing Exploration," in *Proceedings of International Conference on Computer Design*, October 2007, pp. 259–266.

[52] P. Batude *et al.*, "Advances, Challenges and Opportunities in 3D CMOS Sequential Integration," in *Proceedings of IEEE International Electron Devices Meeting*, December 2011, pp. 7.3.1–7.3.4.

[53] 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, December 2000.

[54] P. Chan and M. Chan, "Stacked 3-D Fin-CMOS Technology," *IEEE Electron Device Letters*, Vol. 26, No. 6, pp. 416–418, June 2005.

[55] R. Gutmann *et al.*, "Three-Dimensional (3D) ICs: a Technology Platform for Integrated Systems and Opportunities for New Polymeric Adhesives," in *Proceedings of IEEE Conference on Polymers and Adhesives in Microelectronics and Photonics*, October 2001, pp. 173–180.

[56] W. Davis *et al.*, "Demystifying 3D ICs: The Pros and Cons of Going Vertical," *IEEE Design and Test of Computers*, Vol. 22, No. 6, pp. 498–510, June 2005.

[57] E. Culurciello and A. G. Andreou, "Capacitive Inter-Chip Data and Power Transfer for 3-D VLSI," *IEEE Transactions on Circuits and Systems II: Express Briefs*, Vol. 53, No. 12, pp. 1348–1352, December 2006.

[58] K. Sakuma *et al.*, "3D Chip-Stacking Technology with Through-Silicon Vias and Low-Volume Lead-Free Interconnections," *IBM Journal of Research and Development*, Vol. 52, No. 6, pp. 611–622, November 2008.

[59] R. Reif, A. Fan, and S. Das, "Fabrication Technologies for Three-Dimensional Integrated Circuits," in *Proceedings of International Symposium on Quality Electronic Design*, March 2002, pp. 33–37.

[60] E.-P. Li, *Electrical Modeling and Design for 3D System Integration: 3D Integrated Circuits and Packaging, Signal Integrity, Power Integrity and EMC*. Wiley-IEEE Press, 2012.

[61] J. U. Knickerbocker *et al.*, "Three-Dimensional Silicon Integration," *IBM Journal of Research and Development*, Vol. 52, No. 6, pp. 553–569, November 2008.

[62] Semiconductor Equipment and Materials International (SEMI), "3D Integration: A Progress Report," 2009. [Online]. Available: http://www.semi.org

[63] K. Sakuma, "Development of 3D Chip Integration Technology," in *Nano-Semiconductors: Devices and Technology*, 1st ed., K. Iniewski, Ed. CRC Press, 2011, ch. 7, p. 599.

[64] C. Bower *et al.*, "High Density Vertical Interconnects for 3-D Integration of Silicon Integrated Circuits," in *Proceedings of Electronic Components and Technology Conference*, May 2006, pp. 399–403.

[65] R. Patti, "Three-Dimensional Integrated Circuits and the Future of System-on-Chip Designs," *Proceedings of the IEEE*, Vol. 94, No. 6, pp. 1214–1224, June 2006.

[66] D. Henry *et al.*, "Low Electrical Resistance Silicon Through Vias: Technology and Characterization," in *Proceedings of Electronic Components and Technology Conference*, July 2006, pp. 1360–1366.

[67] L. Yu *et al.*, "Methodology for Analysis of TSV Stress Induced Transistor Variation and Circuit Performance," in *Proceedings of the International Symposium on Quality Electronic Design*, March 2012, pp. 216–222.

[68] G. Katti *et al.*, "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.

[69] ——, "Temperature Dependent Electrical Characteristics of Through-Si-Via (TSV) Interconnections," in *Proceedings of the Internationial Interconnect Technology Conference*, June 2010, pp. 26–28.

[70] 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.

[71] ——, "Electrical Modeling and Characterization of 3-D Vias," in *Proceedings of IEEE International Symposium on Circuits and Systems*, May 2008, pp. 784–787.

[72] X.-P. Wang, W.-Y. Yin, and S. He, "Multiphysics Characterization of Transient Electrothermomechanical Responses of Through-Silicon Vias Applied With a Periodic Voltage Pulse," *IEEE Transactions on Electron Devices*, Vol. 57, No. 6, pp. 1382–1389, June 2010.

[73] L. Cadix *et al.*, "Modelling of Through Silicon Via RF Performance and Impact on Signal Transmission in 3D Integrated Circuits," in *Proceedings of IEEE International Conference on 3D Systems Integration*, September 2009, pp. 28–30.

[74] D. Messerschmitt, "Synchronization in Digital System Design," *IEEE Journal on Selected Areas in Communications*, Vol. 8, No. 8, pp. 1404–1419, October 1990.

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

[76] J. L. Neves and E. G. Friedman, "Optimal Clock Skew Scheduling Tolerant to Process Variations," in *Proceedings of IEEE/ACM Design Automation Conference*, June 1996, pp. 623–628.

[77] G. E. Téllez and M. Sarrafzadeh, "Minimal Buffer Insertion in Clock Trees with Skew and Slew Rate Constraints," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 16, No. 4, pp. 333–342, April 1997.

[78] K. Boese and A. Kahng, "Zero-Skew Clock Routing Trees with Minimum Wirelength," in *Proceedings of IEEE International ASIC Conference and Exhibit*, September 1992, pp. 17–21.

[79] J. Cong *et al.*, "Bounded-Skew Clock and Steiner Routing," *ACM Transactions on Design Automation of Electronic Systems*, Vol. 3, No. 3, pp. 341–388, July 1998.

[80] G. Gerosa *et al.*, "A Sub-1W to 2W Low-Power IA Processor for Mobile Internet Devices and Ultra-Mobile PCs in 45nm Hi-K Metal Gate CMOS," in *Digest of Technical Papers IEEE International Solid-State Circuits Conference*, February 2008, pp. 256–611.

[81] V. F. Pavlidis, I. Savidis, and E. G. Friedman, "Clock Distribution Networks for 3-D Integrated Circuits," in *Proceedings of IEEE Custom Integrated Circuits Conference*, September 2008, pp. 651–654.

[82] J. Minz, X. Zhao, and S. K. Lim, "Buffered Clock Tree Synthesis for 3D ICs Under Thermal Variations," in *Proceedings of the Asia and South Pacific Design Automation Conference*, March 2008, pp. 504–509.

[83] T.-Y. Kim and T. Kim, "Clock Tree Embedding for 3D ICs," in *Proceedings of the Asia South Pacific Design Automation Conference*, January 2010, pp. 486–491.

[84] X. Zhao *et al.*, "Low-Power Clock Tree Design for Pre-Bond Testing of 3-D Stacked ICs," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 30, No. 5, pp. 732–745, May 2011.

[85] C. Lung *et al.*, "Fault-Tolerant 3D Clock Network," in *Proceedings of IEEE/ACM Design Automation Conference*, June 2011, pp. 645–651.

[86] H. Chang and S. Sapatnekar, "Statistical Timing Analysis Under Spatial Correlations," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 24, No. 9, pp. 1467–1482, September 2005.

[87] ——, "Full-Chip Analysis of Leakage Power under Process Variations, Including Spatial Correlations," in *Proceedings of IEEE/ACM Design Automation Conference*, June 2005, pp. 523–528.

[88] S. Garg and D. Marculescu, "System-Level Process Variability Analysis and Mitigation for 3D MPSoCs," in *Proceedings of Design, Automation and Test in Europe Conference*, March 2009, pp. 604–609.

[89] ——, "3D-GCP: An Analytical Model for the Impact of Process Variations on the Critical Path Delay Distribution of 3D ICs," in *Proceedings of the International Symposium on Quality of Electronic Design*, March 2009, pp. 147–155.

[90] H. Chang *et al.*, "Parameterized Block-Based Statistical Timing Analysis with Non-Gaussian Parameters, Nonlinear Delay Functions," in *Proceedings of IEEE/ACM Design Automation Conference*, June 2005, pp. 71–76.

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

[92] K. A. Bowman *et al.*, "Impact of Die-to-Die and Within-Die Parameter Variations on the Clock Frequency and Throughput of Multi-Core Processors," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, Vol. 17, No. 12, pp. 1679–1690, December 2009.

[93] F. Akopyan *et al.*, "Variability in 3-D Integrated Circuits," in *Proceedings of IEEE Custom Integrated Circuits Conference*, September 2008, pp. 659–662.

[94] J. Jang, O. Franza, and W. Burleson, "Compact Expressions for Supply Noise Induced Period Jitter of Global Binary Clock Trees," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, Vol. 20, No. 1, pp. 66–79, January 2012.

[95] M. Ietrich and J. Haase, Eds., *Process Variations and Probabilistic Integrated Circuit Design*. Springer, 2011.

[96] D. Blaauw, K. Chopra, and A. Srivastava, "Statistical Timing Analysis: From Basic Principles to State of the Art," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 27, No. 4, pp. 589–607, April 2008.

[97] M. Orshansky, L. Milor, and K. Keutzer, "Impact of Spatial Intrachip Gate Length Variability on the Performance of High-Speed Digital Circuits," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 21, No. 5, pp. 544–553, May 2002.

[98] D. Boning and S. R. Nassif, "Models of Process Variations in Device and Interconnect," in *Design of High Performance Microprocessor*, A. Chandrakasan, W. J. Bowhill, and F. Fox, Eds. Wiley-IEEE Press, 2000, ch. 6.

[99] A. Narasimhan and R. Sridhar, "Variability Aware Low-Power Delay Optimal Buffer Insertion for Global Interconnects," *IEEE Transactions on Circuits and Systems I: Regular Papers*, Vol. 57, No. 12, pp. 3055–3063, December 2010.

[100] R. Rao, A. Devgan, and D. Blaauw, "Parametric Yield Estimation Considering Leakage Variability," in *Proceedings of IEEE/ACM Design Automation Conference*, June 2004, pp. 442–447.

[101] A. Nardi *et al.*, "Impact of Unrealistic Worst Case Modeling on the Performance of VLSI Circuits in Deep Submicron CMOS Technologies," *IEEE Transactions on Semiconductor Manufacturing*, Vol. 12, No. 4, pp. 396–402, November 1999.

[102] K. A. Bowman *et al.*, "A Physical Alpha-Power Law MOSFET Model," *IEEE Journal of Solid-State Circuits*, Vol. 34, No. 10, pp. 1410–1414, October 1999.

[103] K. Shinkai *et al.*, "A Gate Delay Model Focusing on Current Fluctuation over Wide-Range of Process and Environmental Variability," in *Proceedings of the IEEE/ACM International Conference on Computer Aided Design*, November 2006, pp. 47–53.

[104] T. Sakurai and A. Newton, "A Simple MOSFET Model for Circuit Analysis," *IEEE Transactions on Electron Devices*, Vol. 38, No. 4, pp. 887–894, April 1991.

[105] ——, "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.

[106] J. Croix and D. Wong, "Blade and Razor: Cell and Interconnect Delay Analysis Using Current-Based Models," in *Proceedings of IEEE/ACM Design Automation Conference*, June 2003, pp. 386–389.

[107] H. Fatemi, S. Nazarian, and M. Pedram, "Statistical Logic Cell Delay Analysis Using a Current-Based Model," in *Proceedings of the IEEE/ACM Design Automation Conference*, June 2006, pp. 253–256.

[108] V. Zolotov *et al.*, "Compact Modeling of Variational Waveforms," in *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, November 2007, pp. 705–712.

[109] A. Goel and S. Vrudhula, "Statistical Waveform and Current Source Based Standard Cell Models for Accurate Timing Analysis," in *Proceedings of IEEE/ACM Design Automation Conference*, June 2008, pp. 227–230.

[110] T. Sakurai, "Closed-Form Expressions for Interconnection Delay, Coupling, and Crosstalk in VLSIs," *IEEE Transactions on Electron Devices*, Vol. 40, No. 1, pp. 118–124, January 1993.

[111] W. Elmore, "The Transient Response of Damped Linear Networks with Particular Regard to Wideband Amplifiers," *Journal of Applied Physics*, Vol. 19, No. 1, p. 55, January 1948.

[112] C. V. Kashyap *et al.*, "Closed Form Expressions for Extending Step Delay and Slew Metrics to Ramp Inputs," in *Proceedings of International Symposium on Physical Design*, April 2003, pp. 24–31.

[113] Y. Ismail, E. Friedman, and J. Neves, "Equivalent Elmore Delay for *RLC* Trees," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 19, No. 1, pp. 83–97, January 2000.

[114] R. Venkatesan and J. Davis, "Compact Distributed *RLC* Interconnect Models-Part IV: Unified Models for Time Delay, Crosstalk, and Repeater Insertion," *IEEE Transactions on Electron Devices*, Vol. 50, No. 4, pp. 1094–1102, April 2003.

[115] M. Gowan, L. Biro, and D. Jackson, "Power Considerations in the Design of the Alpha 21264 Microprocessor," in *Proceedings of IEEE/ACM Design Automation Conference*, June 1998, pp. 726–731.

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

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

[118] S. Pant and E. Chiprout, "Power Grid Physics and Implications for CAD," in *Proceedings of IEEE/ACM Design Automation Conference*, July 2006, pp. 199–204.

[119] S. Nassif and S. Sapatnekar, "Random Walks in a Supply Network," in *Proceedings of IEEE/ACM Design Automation Conference*, June 2003, pp. 93–98.

[120] Y. Zhong and M. Wong, "Fast Algorithms for *IR* Drop Analysis in Large Power Grid," in *Proceedings of the IEEE/ACM International Conference on Computer-Aided Designided design*, November 2005, pp. 351–357.

[121] T. Enami, S. Ninomiya, and M. Hashimoto, "Statistical Timing Analysis Considering Spatially and Temporally Correlated Dynamic Power Supply Noise," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 28, No. 4, pp. 541–553, April 2009.

[122] R. Andrews, "Solving Conductive Heat Transfer Problems with Electrical-Analogue Shape Factors," *Chemical Engineering Progress*, Vol. 51, No. 2, pp. 67–71, February 1955.

[123] R. Li, "Optimization of Thermal Via Design Parameters Based on an Analytical Thermal Resistance Model," in *Proceedings of InterSociety Conference on Thermal Phenomena*, May 1998, pp. 475–480.

[124] M. N. Özisik, *Heat Transfer: A Basic Approach.* New York: McGraw-Hill, 1985.

[125] C.-h. Tsai and S.-M. Kang, "Cell-Level Placement for Improving Substrate," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 19, No. 2, pp. 253–266, February 2000.

[126] Y. Zhan, S. V. Kumar, and S. S. Sapatnekar, "Thermally Aware Design," *Foundations and Trends in Electronic Design Automation*, Vol. 2, No. 3, pp. 255–370, October 2007.

[127] B. Goplen, "Advanced Placement Techniques for Future VLSI Circuits," PhD Thesis, University of Minnesota, 2006.

[128] "HotSpot." [Online]. Available: http://lava.cs.virginia.edu/HotSpot/index.htm

[129] T. Wang and C. C.-P. Chen, "3-D Thermal-ADI : A Linear-Time Chip Level Transient Thermal Simulator," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 21, No. 12, pp. 1434–1445, December 2002.

[130] N. Meteopolis and S. Ulam, "The Monte Carlo Method," *Journal of the American Statistical Association*, Vol. 44, No. 247, pp. 335–341, September 1949.

[131] Y. Okayama *et al.*, "Methodology of MOSFET Chracteristics Fluctuation Description Using BSIM3v3 SPICE Model for Statistical Circuit Simulations," in *Proceedings of International Workshop on Statistical Metrology*, June 1998, pp. 14–17.

[132] A. Gattiker *et al.*, "Timing Yield Estimation from Static Timing Analysis," in *Proceedings of the IEEE International Symposium on Quality Electronic Design*, March 2001, pp. 437–442.

[133] J. J. Liou *et al.*, "Fast Statistical Timing Analysis by Probabilistic Event Propagation," in *Proceedings of IEEE/ACM Design Automation Conference*, June 2001, pp. 661–666.

[134] D. J. Pilling and H. B. Sun, "Computer-Aided Prediction of Delays in LSI Logic Systems," in *Proceedings of IEEE/ACM Design Automation Conference*, June 1973, pp. 182–186.

[135] R. Kamikawai, M. Yamada, and T. Chiba, "A Critical Path Delay Check System," in *Proceedings of IEEE/ACM Design Automation Conference*, June 1981, pp. 118–123.

[136] R. B. Hitchcock, G. L. Smith, and D. D. Cheng, "Timing Analysis of Computer Hardware," *IBM Journal Research Development*, Vol. 26, No. 1, pp. 100–105, January 1982.

[137] R. B. Hitchcock, "Timing Verification and the Timing Analysis Program," in *Proceedings of IEEE/ACM Design Automation Conference*, 1982, pp. 594–604.

[138] N. Maheshwari and S. S. Sapatnekar, *Timing Analysis and Optimization of Sequential Circuits*. Norwell, MA, USA: Kluwer Academic Publishers, 1998.

[139] A. Agarwal, V. Zolotov, and D. Blaauw, "Statistical Clock Skew Analysis Considering Intradie-Process Variations," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 23, No. 8, pp. 1231–1242, August 2004.

[140] S. Reda, A. Si, and R. I. Bahar, "Reducing the Leakage and Timing Variability of 2D ICs Using 3D ICs," in *Proceedings of IEEE/ACM International Symposium on Low Power Electronics and Design*, August 2009, pp. 283–286.

[141] M. Hashimoto, T. Yamamoto, and H. Onodera, "Statistical Analysis of Clock Skew Variation in H-tree Structure," in *Proceedings of the International Symposium on Quality of Electronic Design*, Vol. 88, No. 12, December 2005, pp. 402 – 407.

[142] X. Jiang and S. Horiguchi, "Statistical Skew Modeling for General Clock Distribution Networks in Presence of Process Variations," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, Vol. 9, No. 5, pp. 704–717, October 2001.

[143] E. Malavasi *et al.*, "Impact Analysis of Process Variability on Clock Skew," in *Proceedings of the International Symposium on Quality Electronic Design*, March 2002, pp. 129–132.

[144] D. Harris and S. Naffziger, "Statistical Clock Skew Modeling with Data Delay Variations," *IEEE Transactions on Very Large Scale Integration(VLSI) Systems*, Vol. 9, No. 6, pp. 888–898, December 2001.

[145] S. Sundareswaran *et al.*, "A Timing Methodology Considering Within-Die Clock Skew Variations," in *Proceedings of IEEE International SOC Conference*, September 2008, pp. 351–356.

[146] "UMC Foundry Design Kit (FDK) User Guide." United Microelectronics Corporation, August 2007.

[147] A. Devgan and C. Kashyap, "Block-Based Static Timing Analysis with Uncertainty," in *Proceedings of IEEE/ACM International Conference on Computer-Aided Design*, November 2003, pp. 607–614.

[148] "Virtuoso Spectre Circuit Simulator User Guide." Cadence Design Systems, Inc., 2008.

[149] M. Mondal *et al.*, "Thermally Robust Clocking Schemes for 3D Integrated Circuits," in *Proceedings of Design, Automation and Test in Europe Conference*, April 2007, pp. 1206–1211.

[150] X. Zhao and S. K. Lim, "Power and Slew-Aware Clock Network Design for Through-Silicon-Via (TSV) Based 3D ICs," in *Proceedings of Asia and South Pacific Design Automation Conference*, January 2010, pp. 175–180.

[151] J. Joyner, P. Zarkesh-Ha, and J. Meindl, "Global Interconnect Design in a Three-Dimensional System-on-a-Chip," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, Vol. 12, No. 4, pp. 367–372, April 2004.

[152] N. MohammadZadeh *et al.*, "Multi-Domain Clock Skew Scheduling-Aware Register Placement to Optimize Clock Distribution Network," in *Proceedings of Design, Automation and Test in Europe Conference*, March 2009, pp. 833–838.

[153] G. H. Loh, Y. Xie, and B. Black, "Processor Design in 3D Die-Stacking Technologies," *IEEE Micro*, Vol. 27, No. 3, pp. 31–48, May 2007.

[154] J. Kozhaya, S. Nassif, and F. Najm, "A Multigrid-Like Technique for Power Grid Analysis," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 21, No. 10, pp. 1148–1160, October 2002.

[155] M. Zhao *et al.*, "Hierarchical Analysis of Power Distribution Networks," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 21, No. 2, pp. 159–168, February 2002.

[156] S. Nassif and S. Sapatnekar, "Power Grid Analysis Using Random Walks," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 24, No. 8, pp. 1204–1224, August 2005.

[157] T.-H. Chen and C. C.-P. Chen, "Efficient Large-Scale Power Grid Analysis Based on Preconditioned Krylov-Subspace Iterative Methods," in *Proceedings of IEEE/ACM Design Automation Conference*, 2001, pp. 559–562.

[158] Z. Feng, X. Zhao, and Z. Zeng, "Robust Parallel Preconditioned Power Grid Simulation on GPU With Adaptive Runtime Performance Modeling and Optimization," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 30, No. 4, pp. 562–573, April 2011.

[159] C. Zhang, V. Pavlidis, and G. De Micheli, "Voltage Propagation Method for 3-D Power Grid Analysis," in *Proceedings of Design, Automation and Test in Europe Conference*, March 2012, pp. 844–847.

[160] N. H. Khan *et al.*, "Power Delivery Design for 3-D ICs Using Different Through-Silicon Via (TSV) Technologies," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, Vol. 19, No. 4, pp. 647–658, April 2011.

[161] "HSPICE." Synopsys, Inc., 2010. [Online]. Available: http://www.hspice.com

[162] B. Razavi, *Phase-Locking in High-Performance Systems: From Devices to Architectures*. New York: John Wiley & Sons, Inc., 2003.

[163] R. Franch *et al.*, "On-chip Timing Uncertainty Measurements on IBM Microprocessors," in *Proceedings of the IEEE International Test Conference*, October 2007, pp. 1–7.

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

[165] T. Enami *et al.*, "Statistical Timing Analysis Considering Clock Jitter and Skew due to Power Supply Noise and Process Variation," in *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, Vol. E93-A, No. 12, December 2010, pp. 2399–2408.

[166] M. Gupta *et al.*, "Tribeca: Design for PVT Variations with Local Recovery and Fine-Grained Adaptation," in *Proceedings of the IEEE/ACM International Symposium on Microarchitecture*, December 2009, pp. 435–446.

[167] R. Chen and H. Zhou, "Fast Buffer Insertion for Yield Optimization Under Process Variations," in *Proceedings of the IEEE/ACM Design Automation Conference*, June 2007, pp. 338–343.

[168] J. Xiong and L. He, "Fast Buffer Insertion Considering Process Variations," in *Proceedings of International Symposium on Physical Design*, April 2006, pp. 128–135.

[169] S. Tam, J. Leung, and R. Limaye, "Clock Generation & Distribution for a 45nm, 8-Core Xeon® Processor with 24MB Cache," in *Proceedings of Symposium on VLSI Circuits*, August 2009, pp. 154–155.

[170] X. Liang, G.-Y. Wei, and D. Brooks, "ReVIVaL: A Variation-Tolerant Architecture Using Voltage Interpolation and Variable Latency," in *Proceedings of International Symposium on Computer Architecture*, June 2008, pp. 191–202.

[171] G. Chen and E. Friedman, "Low-Power Repeaters Driving *RC* and *RLC* Interconnects with Delay and Bandwidth Constraints," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, Vol. 14, No. 2, pp. 161–172, February 2006.

[172] J. Yang *et al.*, "Robust Clock Tree Synthesis with Timing Yield Optimization for 3D-ICs," in *Proceedings of Asia and South Pacific Design Automation Conference*, January 2011, pp. 621–626.

[173] X. Zhao and S. Mukhopadhyay, "Variation-Tolerant and Low-Power Clock Network Design for 3D ICs," in *Proceedings of Electronic Components and Technology Conference*, June 2011, pp. 2007–2014.

[174] P. Friedberg, J. Cain, and C. Spanos, "Modeling Within-Die Spatial Correlation Effects for Process-Design Co-Optimization," in *Proceedings of the International Symposium on Quality of Electronic Design*, No. 510, March 2005, pp. 516–521.

[175] R. S. Tsay, "IBM Clock Benchmarks." [Online]. Available: http://vlsicad.ucsd.edu/GSRC/bookshelf/Slots/BST/#III

[176] L. Wang, Y. Chang, and K.-T. Cheng, Eds., *Electronic Design Automation: Synthesis, Verification, and Test*. Morgan Kaufmann Publishers, 2009.

[177] M. Jackson, A. Srinivasan, and E. Kuh, "Clock Routing for High-Performance ICs," in *Proceedings of IEEE/ACM Design Automation Conference*, March 1990, pp. 573–579.

[178] S. Dhar and M. A. Franklin, "Optimum Buffer Circuits for Driving Long Uniform Lines," *IEEE Journal of Solid-State Circuits*, Vol. 26, No. I, pp. 32–40, January 1991.

[179] V. Pavlidis and E. Friedman, "Timing-Driven Via Placement Heuristics for Three-Dimensional ICs," *Integration, the VLSI Journal*, Vol. 41, No. 4, pp. 489–508, July 2008.

[180] L. P.P.P. van Ginneken, "Buffer Placement in Distributed RC-tree Networks for Minimal Elmore Delay," in *Proceedings of IEEE International Symposium on Circuits and Systems*, May 1990, pp. 865–868.

[181] J. Lillis, C. Cheng, and Ting-Ting Y. Lin, "Optimal Wire Sizing and Buffer Insertion for Low Power and a Generalized Delay Model," *IEEE Journal of Solid-State Circuits*, Vol. 31, No. 3, pp. 437–447, March 1996.

[182] C. J. Alpert and A. Devgan, "Wire Segmenting for Improved Buffer Insertion," in *Proceedings of the IEEE/ACM Design Automation Conference*, June 1997, pp. 588–593.

[183] X. He *et al.*, "Simultaneous Buffer and Interlayer Via Planning for 3D Floorplanning," in *Proceedings of International Symposium on Quality Electronic Design*, March 2009, pp. 740–745.

[184] S. Dong *et al.*, "Buffer Planning for 3D ICs," in *Proceedings of IEEE International Symposium on Circuits and Systems*, May 2009, pp. 1735–1738.

[185] MIT Lincoln Laboratory, "MITLL Low-Power FDSOI CMOS Process Application Notes," June 2006.

[186] S. Im and K. Banerjee, "Full Chip Thermal Analysis of Planar (2-D) and Vertically Integrated (3-D) High Performance ICs," in *Technical Digest of International Electron Devices Meeting*, Vol. 94305, December 2000, pp. 727–730.

[187] A. Rahman and R. Reif, "Thermal Analysis of Three-Dimensional (3-D) Integrated Circuits (ICs)," in *Proceedings of the IEEE International Interconnect Technology Conference*, Vol. 15, No. 4, June 2001, pp. 157–159.

[188] C. Bachmann, "Thermal Modeling and Analysis of Three Dimensional (3D) Chip Stacks," Thesis, University of Maryland (College Park, Md.), 2007. [Online]. Available: http://www.lib.umd.edu/drum/handle/1903/7419

[189] J. Cong and Y. Zhang, "Thermal Via Planning for 3-D ICs," in *Proceedings of IEEE/ACM International Conference on Computer-Aided Design*, November 2005, pp. 745–752.

[190] T.-Y. Chiang *et al.*, "Thermal Analysis of Heterogeneous 3D ICs with Various Integration Scenarios," in *Technical Digest of International Electron Devices Meeting*, December 2001, pp. 31.2.1–31.2.4.

[191] A. Jain *et al.*, "Analytical and Numerical Modeling of the Thermal Performance of Three-Dimensional Integrated Circuits," *IEEE Transactions on Components and Packaging Technologies*, Vol. 33, No. 1, pp. 56–63, 2010.

[192] P. Wilkerson, A. Raman, and M. Turowski, "Fast, Automated Thermal Simulation of Three-Dimensional Integrated Circuits," in *Proceedings of Intersociety Conference on Thermal and Thermomechanical Phenomena In Electronic Systems*, June 2004, pp. 706–713.

[193] M. Sabry and H. Saleh, "Compact Thermal Models: A Global Approach," in *Proceedings of International Conference on Thermal Issues in Emerging Technologies: Theory and Application*, January 2007, pp. 33–39.

[194] COMSOL Inc., "COMSOL Multiphysics," 2011.

[195] S. G. Singh and C. S. Tan, "Impact of Thermal Through Silicon Via (TTSV) on the Temperature Profile of Multi-Layer 3-D Device Stack," in *Proceedings of IEEE International Conference on 3D System Integration*, September 2009, pp. 1–4.

[196] J. H. Lau and T. G. Yue, "Thermal Management of 3D IC Integration with TSV (Through Silicon Via)," in *Proceedings of Electronic Components and Technology Conference*, May 2009, pp. 635–640.

[197] N. Kurd *et al.*, "Next Generation Intel Core™ Micro-Architecture (Nehalem) Clocking," *IEEE Journal of Solid-State Circuits*, Vol. 44, No. 4, pp. 1121–1129, April 2009.

[198] H. Chen *et al.*, "A Sliding Window Scheme for Accurate Clock Mesh Analysis," in *Proceedings of IEEE/ACM International Conference on Computer-Aided Design*, November 2005, pp. 939–946.

List of Abbreviations

3-D	Three-dimensional
ACLV	across-chip linewidth variation
ADE	alternating-direction-implicit method
BEOL	back-end of line
CDF	cumulative distribution function
CDN	clock distribution networks
CMP	chemical-mechanical polishing
CTS	clock tree synthesis
D2D	die-to-die or inter-die process variations
DVS	dynamic voltage scaling
FDA	finite differential analysis
FEA	finite element analysis
FEOL	front-end of line
Fin-FET	fin field effect transistors
I/O	input/output
IC	integrated circuits
ILD	interlayer dielectric
KCL	Kirchholff's Current Law
NLDM	non-linear delay models
P/G	power and ground
PCB	printed circuit board

PDE partial differential equation

PDN power distribution networks

PLL phase-locked loop

POD point of divergence

PSN power supply noise

PV process variations

PVT process, voltage, and temperature

SiP System-in-Package

SOI Silicon-on-Insulator

SoP System-on-Package

SSTA statistical static timing analysis

STA static timing analysis

TSV through silicon via

TTSV thermal through silicon via

WID within-die or intra-die process variations

WJ worst case period jitter

Curriculum Vitae

Name:	Hu XU
Address:	Ch. du Bochet 18, CH-1024, Ecublens(VD), Switzerland
Phone:	+41 (0)78 6815624
Email:	hughesxuh@gmail.com
Nationality:	Chinese
Gender:	Male

PROFILE

- Three years of academic/industry collaboration on the design of SoCs and CPUs
- Four years of experience on the clock and power design of Through Silicon Via-based 3-D ICs
- CAD/EDA development for VLSI Systems

EDUCATIONAL BACKGROUND

2008 – 2012 **Ph.D. in Computer Science**, Swiss Federal Institute of Technology at Lausanne (EPFL), Switzerland.
- Advisor: Prof. Giovanni De Micheli

2005 – 2008 **M.S. in Computer Architecture**, Peking University, Beijing, China
- Advisors: Prof. Dong Tong and Prof. Jason Cong (UCLA)
- Thesis: *Setting Constraints and Logic Synthesis for Novel PKUnity© SoC*
- China Reconstruct Scholarship for Academic Excellence

2001 – 2005 **B.E. in Automation**, Tsinghua University, Beijing, China

TECHNICAL SKILLS

- **VLSI design**
  - Logic synthesis and verification, backend/physical design (floorplanning, placement, clock tree synthesis, routing, and Static Timing Analysis) with Synopsys tools and scripts.
  - RTL design in Verilog and VHDL and simulations with ModelSim.
  - Transistor-level simulations with Cadence Virtuoso and related scripts.
  - 3-D IC design and modeling down to 32 nm technology.
  - Thermal simulations with COMSOL Multiphysics.
- **CAD/EDA development and programming**
  - Clock tree synthesis, buffer insertion, and statistical timing analysis in C++, Java, and Matlab.
  - Tcl and Linux Shell programming.

PROFESSIONAL EXPERIENCE

2008 – 2012 **Integrated Systems Laboratory, EPFL: Research assistant**

- *Timing Uncertainty Analysis and Mitigation for 3-D ICs*, 2010 – 2012

Developed the first statistical model including both skew and jitter for 3-D ICs. Modeled the combined effect of process variations and power supply noise on 3-D clock distribution networks.

- *Power Grid Analysis for 3-D ICs*, 2011 – 2012
  Developed an efficient *IR*-drop analysis tool for 3-D power grids of industrial benchmarks. Collaborated with and funded by Intel Braunschweig Labs, Germany.
- *Modeling Thermal Through Silicon Vias in 3-D ICs*, 2010 – 2011
  Proposed novel thermal resistor networks to analytically model the heat transfer through TTSVs. Investigated the effect of TTSVs on the temperature of 3-D ICs.

2005 – 2008 Microprocessor R&D Center (MPRC), Peking Univ.: IC designer

- *SK-M© SoC Project*: **Group project manager**, 2007 – 2008
  Organized the physical design of SK-M© System-on-Chip. Implemented logic synthesis with multi-voltage technique. Optimized the clock distribution networks and the timing performance of the Floating-Point Unit.
- *PKUnityX86© SoC Project*, 2007
  Developed the first semi-custom physical design flow for an X86-based CPU authorized by AMD in MPRC. Implemented logic synthesis and verification for the SoC.
- *PKUnity863II© C8 CPU Project*, 2005 – 2006
  Implemented logic synthesis and verification for MPRC UniCore II© CPU. Developed the physical design flow with multi-threshold technique for this CPU.
- *Clock tree synthesis and clock skew scheduling*, 2007 – 2008
  Developed a JAVA platform for prescribed-skew clock tree synthesis and skew scheduling.

2004 Shenzhen Yoky Filters Co., Ltd.: two-month intern

Designed an automatic control system for Can Sealing Machines with PLCs.

COMPETITION AWARDS

- **The First Winner**, STMicroelectronics InnovationCup, 2012.
  Designed an electronic system for badminton training with MEMS-based inertia sensors.
- **Finalist**, the Asia-Pacific Robot Contest China, 2005.
  Designed embedded control systems for automatic vehicle robots for Tsinghua Univ. Team.
- **The First Prize**, the Fifth Electronic Design Competition of Tsinghua University, 2003.
  Designed the wireless control system of an automatic vehicle.

EXTRA-CURRICULAR ACTIVITIES

- **Vice President** of Chinese Student & Scholars Association Lausanne, 2009 – 2011.
- **Captain** of the football team of Dept. Automation in Tsinghua Univ., 2004 – 2005.
- **Leading player** of Lausanne University Badminton Club, 2009 – 2012.
- Assistant Mentor for undergraduate students, Peking University, 2007-2008.
- Member of Institute of Electrical and Electronics Engineers (IEEE), 2006 – present.

LANGUAGES

English	**Chinese (Mandarin)**	**Cantonese**	**French**
Fluent	Native language	Advanced	Beginner (A1-A2)

ACADEMIA SERVICES

- **Technical Program Committee Member**

Asia Symposium on Quality Electronic Design (ASQED), 2012
IEEE Asia Pacific Conference on Circuits and Systems (APCCAS), 2010

- **Journal Reviewer**

IEEE Transactions on Circuits and Systems-Part II (TCAS), 2012
Integration, the VLSI Journal, 2011 – 2012
Journal of Circuits, Systems, and Computers (JCSC), 2010 – 2012
ACM Journal on Emerging Technologies in Computing Systems (JETC), 2011

REFERENCES

- **Prof. Giovanni De Micheli**

Affiliation: Director of the Institute of Electrical Engineering, Swiss Federal Institute of Technology Lausanne (EPFL)
Telephone: +41 (0)21 693 0911
Email: giovanni.demicheli@epfl.ch
Address: EPFL-ISIM-IC-LSI1, Building INF 341, Station 14, Lausanne 1015, Switzerland.

- **Prof. Wayne Burleson**

Affiliation: Professor of Electrical and Computer Engineering, University of Massachusetts Amherst, USA
Email: burleson@ecs.umass.edu
Address: Dept. of Electrical and Computer Engineering, University of Massachusetts, Amherst, MA 01003-5110, USA

- **Prof. Vasilis F. Pavlidis**

Affiliation: Lecturer in the Department of Computer Science, University of Manchester, UK
Telephone: +44 161 275 6191
Email: vpavlidis@cs.man.ac.uk
Address: The APT Group, School of Computer Science, The University of Manchester, Oxford Road, Manchester. M13 9PL United Kingdom

SELECTED PUBLICATIONS

[1] "Effect of Process Variations in 3-D Global Clock Distribution Networks," **H. Xu**, V. Pavlidis, and G. De Micheli, *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, Vol. 8, No. 3, August 2012.

[2] "An Accurate Dynamic Power Model on FPGA Routing Resources," X. Tang, L. Wang, and **H. Xu**, *Proceedings of IEEE International Conference on Solid-State and Integrated Circuit Technology*, October 2012.

[3] "Enhanced Wafer Matching Heuristics for 3-D ICs," V. Pavlidis, **H. Xu**, and G. De Micheli, *Proceedings of IEEE European Test Symposium*, May 2012.

[4] "The Combined Effect of Process Variations and Power Supply Noise on Clock Skew and Jitter," **H. Xu**, V. Pavlidis, W. Burleson, and G. De Micheli, *Proceedings of International Symposium on Quality Electronic Design (ISQED)*, March 2012. (Extended version under review for *IEEE Transaction on Very Large Scale Integration (VLSI) Systems*)

[5] "Skew Variability in 3-D ICs with Multiple Clock Domains," **H. Xu**, V. Pavlidis, and G. De Micheli, *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS)*, pp.2221-2224, May 2011.

[6] "Analytical Heat Transfer Model for Thermal Through-Silicon Vias," **H. Xu**, V. Pavlidis, and G. De Micheli, *Proceedings of Design, Automation and Test in Europe Conference (DATE)*, March, 2011.

[7] "Synchronization and Power Integrity Issues in 3-D ICs," V. Pavlidis, **H. Xu**, I. Tsioutsios, and G. De Micheli, *Proceedings of IEEE Asia Pacific Conference on Circuits and Systems (APCCAS)*, pp. 536-539, 2010.

[8] "Process-Induced Skew Variation for Scaled 2-D and 3-D ICs," **H. Xu**, V. Pavlidis, and G. De Micheli, *Proceedings of the ACM/IEEE International Workshop on System Level Interconnect Prediction (SLIP)*, pp. 17-24, June 2010.

[9] "Repeater Insertion Techniques for 3-D Interconnects," **H. Xu**, V. Pavlidis, and G. De Micheli, *Electronic Workshop Digest of DATE Workshop on 3D Integration*, pp. 41-44, March 2010.

[10] "Repeater Insertion for Two-Terminal Nets in 3-D ICs," **H. Xu**, V. Pavlidis, and G. De Micheli, *Proceedings of the International ICST Conference on Nano-Networks*, pp. 141-150, October 2009.

[11] "Prescribed Skew Clock Routing Algorithm with Local Topology Optimization," L. Duan, **H. Xu**, K. Wang, and X. Cheng, *Chinese Journal of Computer-Aided Design & Computer Graphics*, Vol.20, No.4, pp. 452-458, April 2008.

[12] "A Fast Incremental Clock Skew Scheduling Algorithm for Slack Optimization," K. Wang, H. Fang, **H. Xu**, and X. Cheng, *Proceedings of Asia and South Pacific Design Automation Conference (ASPDAC)*, pp. 492-497, March 2008.

ADDITIONAL PRESENTATIONS

[1] "Voltage Propagation Method for 3-D Power Grid Analysis", Interactive Presentation in *Design, Automation and Test in Europe (DATE)*, March 2012, Dresden, Germany.

[2] "Modeling Issues for Thermal TSVs," Invited Talk in *Design for 3D Silicon Integration Workshop (D43D)*, June 2011, MINATEC, Grenoble, France.

[3] "Combined Effect of Process Variations and Power Supply Noise on Clock Skew and Jitter," Poster in *Intel European Research&Innovation Conference*, October 2011, Leixlip, Ireland.