Test and Design-for-Testability Solutions for Monolithic 3D Integrated Circuits

by

Abhishek Koneru

Department of Electrical and Computer Engineering
Duke University

Date: ______________________
Approved:

____
Krishnendu Chakrabarty, Supervisor

____
Yiran Chen

____
Sukeshwar Kannan

____
Xin Li

____
Aida Todri-Sanial

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Electrical and Computer Engineering
in the Graduate School of Duke University
2019

Abstract

Test and Design-for-Testability Solutions for Monolithic 3D Integrated Circuits

by

Abhishek Koneru

Department of Electrical and Computer Engineering
Duke University

Date: ____________________________
Approved:

____
Krishnendu Chakrabarty, Supervisor

____
Yiran Chen

____
Sukeshwar Kannan

____
Xin Li

____
Aida Todri-Sanial

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Electrical and Computer Engineering
in the Graduate School of Duke University
2019

Abstract

Three-dimensional (3D) integration is a promising way to sustain Moore's Law beyond device- and interconnect-scaling limits. 3D technologies enable the integration of heterogeneous fabrication processes, and provide high-speed interconnects, high device-integration density, and low power consumption. Today's 3D technologies can broadly be classified into two categories based on the integration process: (i) 3D die/wafer stacking, in which separately manufactured dies/wafers are integrated onto the same package, and (ii) monolithic 3D (M3D) integration, in which transistor layers are processed sequentially on the same wafer. Through-silicon-vias (TSVs) are used to connect dies to each other in a 3D stacked integrated circuit (IC). In contrast, M3D ICs use inter-layer-vias (ILVs) of much smaller dimensions to connect a metal line in one transistor layer to a metal line in another transistor layer.

TSV-based 3D stacked ICs can be manufactured without requiring substantial changes to the existing fabrication flow. Considerable research efforts have therefore been directed towards the development of TSV-based 3D stacking technology, and products based on this technology have been successfully introduced into the marketplace, e.g., the AMD Fiji chip. However, the keep-out-zone (KOZ) required for TSVs and limitations on the die alignment precision impose limits on the device integration density that can be achieved using TSV-based 3D stacking. A minimum KOZ of 3 μm is required for ICs fabricated at the 20 nm technology node, and the die alignment precision is currently limited to 0.5 μm.

The above limitations on integration density can be overcome by adopting M3D integration. High-density integration in M3D is enabled by the alignment precision of ILVs, which is determined by the lithography stepper accuracy and has been reported to be 10 nm for the 22 nm technology node. In addition, the size and pitch of an ILV are typically one to two orders of magnitude smaller than those of a TSV. Therefore, M3D integration can result in reduced area and higher performance when compared to 3D die stacking.

Due to the above benefits of M3D integration, there is growing interest in industry towards the adoption of this technology. However, test challenges for M3D integration have remained largely unexplored. This thesis is focussed on four key test challenges for M3D integration: (i) performance variations due to high-density integration, (ii) defect analysis and modeling, (iii) defect isolation and yield enhancement, and (iv) yield loss due to voltage droop. For each test challenge, we motivate the need to study its impact on an M3D IC, analyze the effectiveness of existing test solutions, and develop new solutions.

This dissertation first addresses challenges (i) and (ii). We quantify the impact of electrostatic coupling and wafer-bonding defects on the threshold voltage of a top-layer transistor in an M3D IC. In addition, we show that wafer-bonding defects can lead to a change in the resistance of ILVs, and in some cases, lead to an open in an ILV or a short between two ILVs. We also study the impact of these defects on path delays and on the effectiveness of delay-test patterns for large benchmarks. Our results show that the timing characteristics of an M3D IC can be significantly altered due to coupling and wafer-bonding defects if the thickness of its inter-layer dielectric is less than 100 nm.

Next, this dissertation presents a new DfT solution for M3D ICs based on dedicated test layers, which are inserted between functional layers. We evaluate the cost associated with the proposed DfT solution and compare it with that for a potential

DfT solution based on the proposed IEEE Std. P1838. Our results show that the proposed DfT solution is more cost-efficient than the P1838-based DfT solution for a wide range of ILV density, ILV yield, and defect density. We also present a test scheduling and optimization technique for wafer-level testing of M3D ICs.

This dissertation then presents an ILV BIST solution for M3D ICs to address isolation of ILV defects and yield enhancement. In the proposed ILV BIST solution, interface-register cells in a test layer are stitched into a TRC using their functional outputs and the ILVs. We show that the proposed solution detects all hard opens and shorts in the ILVs. We validate the detection of all hard opens and shorts using HSpice simulations. We also implement an artificial neural network-based diagnosis framework to estimate the size of ILV defect and show that the prediction accuracy of the proposed framework is extremely high.

Finally, this dissertation describes an optimization approach for reliable power delivery in M3D ICs to address challenge (iv). We analyze the voltage droop during testing and compare it with that observed during functional operation. We also quantify the impact of voltage droop during testing on yield loss. Our results show that the proposed power delivery optimization approach significantly reduces the worst-case voltage droop and yield loss due to voltage droop compared to a baseline.

In summary, the dissertation targets important design and optimization problems related to testing of M3D ICs. This research has led to theoretical insights, significant academic and industrial collaborations, simulations results using advanced process design kits, and a set of test and DfT solutions.

To my parents, Kumudini and Kishore, for their unconditional love and encouragement. And my fiancé, Shruti, for her endless love and support.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to acknowledge the guidance and support I received over the course of my journey at Duke that made it academically, intellectually, and personally rewarding.

First, I would like thank my advisor, Professor Krishnendu Chakrabarty, for his excellent mentorship and guidance. Krish has been a mentor to me since my first exposure to research as an undergrad. He has motivated me to work on difficult problems and helped me cultivate the habit of paying attention to detail. Despite his busy schedule, I always found him to be extremely approachable and open to discuss any problem that concerned me. This dissertation would not have been possible without his guidance.

I would like to thank Professor Hisham Z. Massoud, who has greatly shaped my research at Duke. Professor Massoud has always been open for a discussion, asked insightful questions, and provided valuable feedback. I am grateful to him for shaping me as a researcher and helping me breakdown complex problems into smaller pieces.

I also wish to thank Professor Aida Todri-Sanial for our research collaboration and stimulating discussions on various topics during my stay in France. I am grateful to her for recommending me to the prestigious Eiffel Excellance Scholarship and supporting me throughout my stay in France.

I would like to thank my collaborator and mentor, Dr. Sukeshwar Kannan, for his insightful comments and feedback throughout the course of my PhD. I also thank my mentors from my internships at Intel, TSMC, and Apple: Dr. Maheshwar

Chandrasekar, Dr. Sandeep Goel, Dr. Anuja Banerjee, and Dr. Samy Makar, for shaping my problem-solving skills.

I would also like to thank the rest of my thesis committee: Prof. Yiran Chen and Prof. Xin Li, for their encouragement, insightful comments, and questions.

I thank people in my research group and colleagues from Duke University: Dr. Mukesh Agrawal, Dr. Sergej Deutsch, Dr. Fangming Ye, Dr. Ran Wang, Dr. Zipeng Li, Dr. Mohamed Ibrahim, Dr. Shi Jin, Rana Elnaggar, Mengyun Liu, and Zhanwei Zhong, for their useful advice on a range of issues and fun times. My other friends from Duke: Aditi Dighe, Dr. Ugonna Ohiri, Callie Woods, and Shwetha Mangalesh, made my stay in Durham a memorable one.

Finally, I would like to thank my family: my parents, my sister, my grandparents, and my fiancé. It is their love, encouragement, critique, and support that helped me reach this important milestone in my career.

1

Introduction

The semiconductor industry has been able to meet the demand for high-performance integrated circuits (ICs) with added functionality by relentlessly scaling device sizes. However, it is becoming increasingly difficult to sustain device scaling in an economically viable manner [7]. Such escalating costs are mainly due to the challenges associated with: (i) lithography of small features, (ii) interconnect scaling, and (iii) reducing and mitigating process variations.

Three-dimensional (3D) integration is a promising way to achieve high-performance ICs with more functionality and reduced die footprint. The improvement in performance and device-integration density is mainly due to the decrease in interconnect length compared to conventional ICs. 3D technologies also enable the integration of heterogenous fabrication processes, thus paving the way for complex systems such as memory-on-logic [8], 3D optical sensors with a coprocessor [9], and 3D CMOS-MEMS ICs [10]. Today's 3D integration process is primarily based on die/wafer stacking since it does not require substantial changes to the existing fabrication flow. In this process, separately manufactured dies/wafers are integrated onto the same package, and through-silicon-vias (TSVs) are used to connect dies to each other. Considerable research efforts have therefore been directed towards the development of TSV-based 3D stacking technology [8, 9, 11], and products based on this technology have been successfully introduced into the market, e.g., the AMD Fiji chip [12]. However, the keep-out-zone (KOZ) required for TSVs and limitations on the die alignment precision impose limits on the device integration density that can be achieved using TSV-based 3D stacking. A minimum KOZ of 3 μm is required for ICs fabricated at the 20 nm technology node [13], and the die alignment precision is currently limited to 0.5 μm [14].

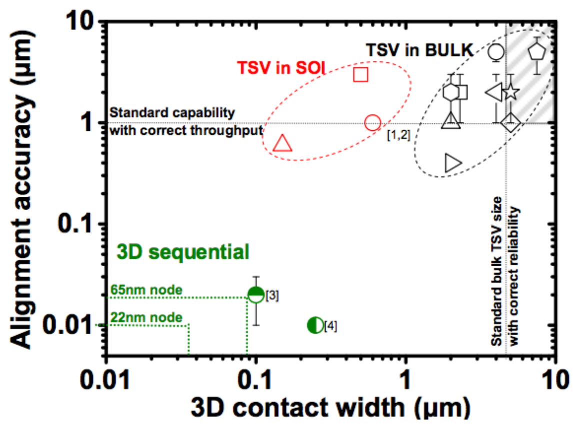


FIGURE 1.1: Alignment accuracy versus 3D contact width in TSV-based 3D and M3D integration approaches [3].

Monolithic 3D (M3D) integration is an emerging technology in which transistors are processed layer-by-layer on the same wafer. Sequential integration of transistor layers enables high-density vertical interconnects, known as inter-layer vias (ILVs). Fig. 1.1 presents a comparison of the inter-layer interconnect width between TSV-based 3D and M3D integration approaches. The size and pitch of an ILV is typically one to two orders of magnitude smaller than those of a through-silicon via (TSV) [3]. Therefore, M3D integration can result in significantly reduced area and higher performance when compared to 3D die stacking. Another key benefit of M3D integration is that it allows transistor-level partitioning of a design, i.e., placement of P-channel transistors in one layer and N-channel transistors in another layer. Such a placement is not feasible in 3D die stacking technology due to the large keep-out zones for TSVs, necessitated by TSV-induced stress and the limitations in alignment precision.

Due to the above benefits of M3D integration, there is a growing interest in industry towards the adoption of this technology [4, 7, 10, 15]. In order to ensure that M3D integration is primed for use when industry is ready to adopt it, researchers have started investigating solutions to various problems related to the design and fabrication of M3D ICs. For instance, a low-temperature process to fabricate high-performance top transistor layers without degrading the performance of bottom transistor layers was proposed in [1]. A process to manufacture field-programmable gate-arrays using gate-level M3D integration was demonstrated [3]. Design techniques to reduce the interconnect length, critical path delay, and die area were presented in [16, 17]. Despite these advances, test challenges for M3D integration have remained largely unexplored, as highlighted at the 2015 IEEE 3D Test Workshop [18]. A recent patent on M3D testing [19] is insightful, but it does not consider any of the important technology-specific details of M3D such as the sequential processing of device layers and the high density of inputs and outputs in each layer.

In this chapter, we introduce the M3D technology and provide motivation for the thesis research. Section 1.1 presents an overview of the M3D fabrication flow and computer-aided design of M3D ICs. Section 1.2 describes the challenges associated with testing an M3D IC and motivation for the thesis research. An outline of this dissertation is provided in Section 1.3.

1.1 M3D Technology Overview

In this section, we present an overview of the M3D fabrication flow and design-partitioning strategies.

1.1.1 *Fabrication Process*

M3D integration involves sequential processing of transistor layers on the same wafer. Fig. 1.2 illustrates the steps involved in the fabrication of an M3D IC. First, the bottom-layer transistors and their associated interconnects are processed using a standard high-temperature process, e.g., silicon-on-insulator (SOI), FinFET, or bulk CMOS technology. Next, a thin silicon layer is created over the bottom layer. The top-layer transistors are then processed under a strict thermal budget. Finally, ILVs are processed to connect the two layers. The last three steps are repeated to fabricate additional layers.

Sequential integration of device layers would not have been possible without the development of: (i) a low-temperature process to create a thin silicon film over the bottom layer, and (ii) a process to realize top-layer transistors without damaging the underlying interconnects and degrading the properties of the bottom-layer transistors. Several approaches have been proposed in the literature to create a thin silicon film at low temperatures: (i) thin-film polysilicon deposition [20] (which is based on the method used to create thin-film transistors and is currently for 3D-NAND flash memories), (ii) pulsed-laser crystallization of amorphous silicon [21], (iii) excimer laser crystallization [22], (iv) epitaxial layer transfer (ELTRAN) [23], (v) wafer bonding followed by grinding and etchback [5], and (vi) modified Smart-Cut process [6] (originally used to manufacture SOI wafers). The M3D technology developed by CEA Leti, which is know as CoolCube$^{\text{TM}}$ and has attracted the attention of companies such as IBM, ST Microelectronics, and Qualcomm, uses the last two techniques for creating a thin silicon film over the bottom layer [4].

Dopant activation, which is usually performed at temperatures higher than 800 ^oC , is one of the key steps in forming transistors. However, dopant activation at such high temperatures to realize top-layer transistors can damage the underlying

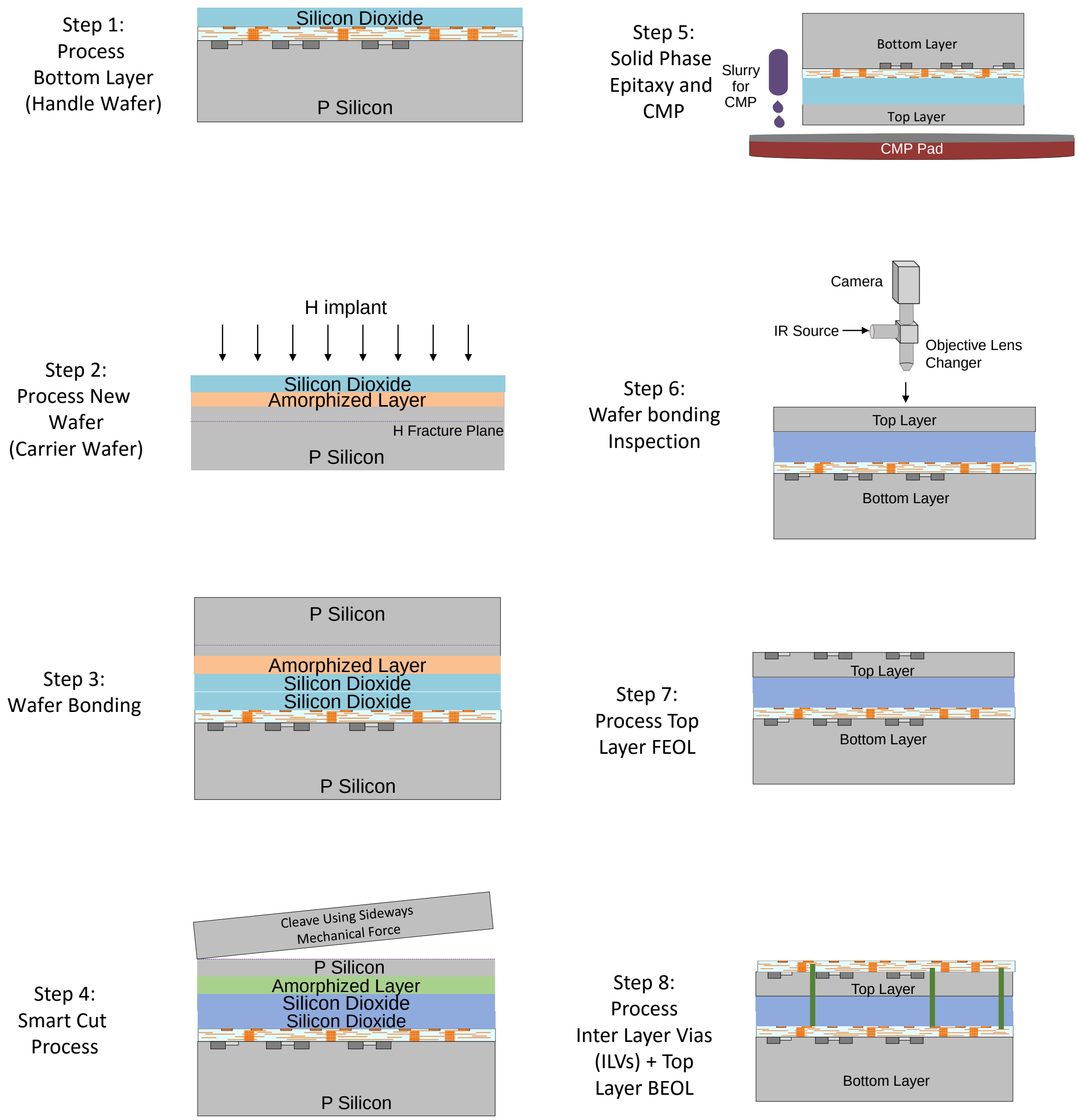


FIGURE 1.2: Steps involved in the fabrication of an M3D IC [3–6].

interconnects and degrade the properties of the bottom-layer transistors. Several process flows have been proposed in the literature to perform dopant activation at low temperature: (i) solid-phase epitaxy at 600 ^oC along with the use of tungsten back-end-of-the-line (BEOL) in the bottom layer [5], (ii) recessed channel array transistor (RCAT) process flow [24], and (iii) pulsed laser annealing along with a shielding layer between the top- and bottom-layer [25]. The last two processes achieve dopant activation at temperatures below 400 ^oC; therefore, these processes are compatible with Cu/low-k BEOL in the bottom layer.

1.1.2 Low-Temperature Wafer Bonding

A key enabler for M3D integration is the ability to create a defect-free thin silicon film over the bottom layer. Low-temperature wafer bonding is a key step in most of the techniques to realize such a thin silicon film. It is therefore important to carefully understand the bonding process. To illustrate the wafer-bonding process, we describe the method to sequentially stack device layers based on the Smart-Cut process. Fig. 1.3 illustrates the key steps in this process. A silicon-on-insulator (SOI) wafer is used as the carrier wafer during the bonding process, and the handle wafer consists of the bottom-layer transistors and their associated interconnects.

The first step in the bonding process is to deposit the bonding surfaces. Oxide layers are the prime candidates for these layers due to the presence of hydroxyl (OH) groups that lead to high bond strengths. An oxide layer can be grown or deposited over the carrier wafer, which is the wafer on which the top-layer transistors will be processed. However, thermal oxidation at 1000 ^oC cannot be used to grow an oxide layer over the handle wafer (on which bottom-layer transistors have been processed) due to the strict thermal budget constraints. Therefore, techniques such as plasma-enhanced chemical vapor deposition (PECVD) and atomic layer deposition (ALD) are used to deposit an oxide layer over the handle wafer at temperatures below 400 ^oC [26].

Silicon atoms are implanted into the carrier wafer to form an amorphous layer. Hydrogen ions are then implanted to create a fracture plane. The energy of hydrogen implantation is chosen such that the fracture plane is formed below the amorphous layer. The condition of the bonding surfaces plays a crucial role in achieving a defect-

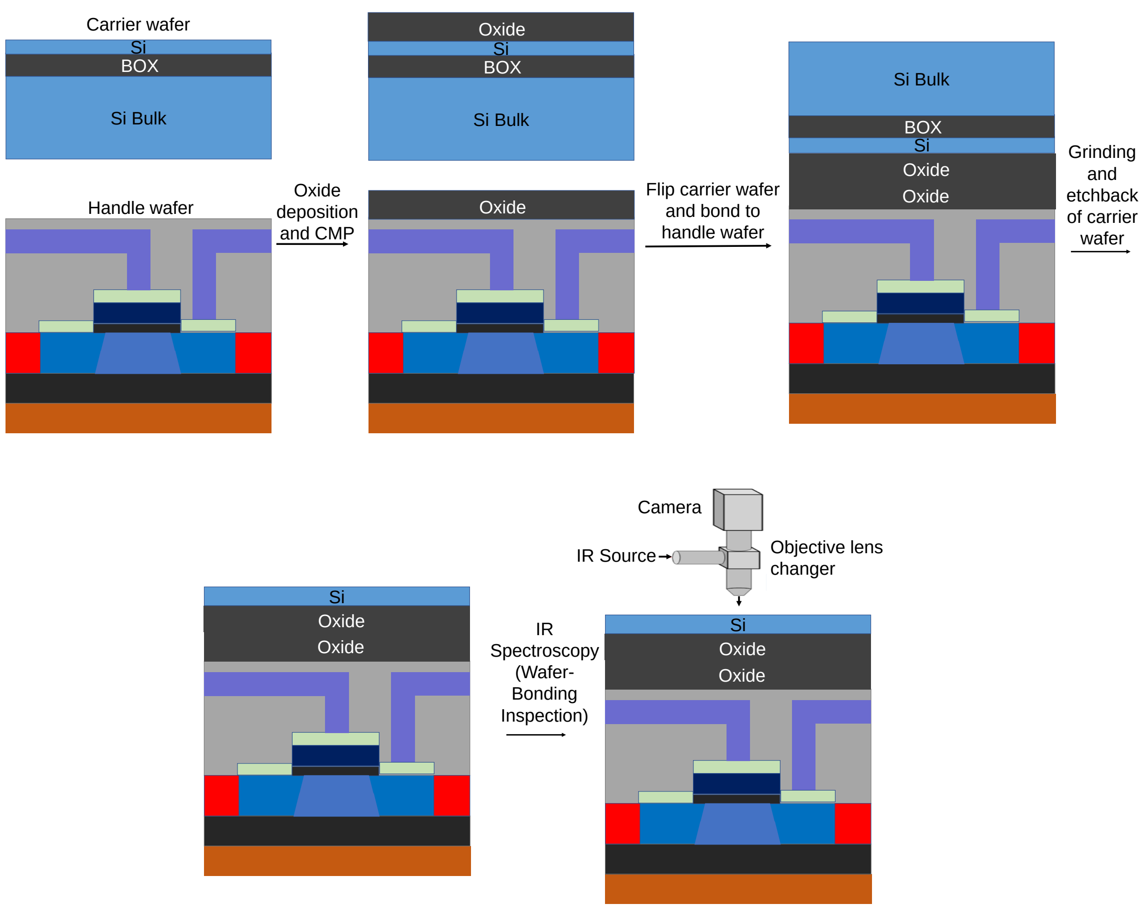


FIGURE 1.3: Key steps in the Smart-Cut process.

free bond. Chemical mechanical planarization (CMP) is therefore performed after the oxide layers have been deposited to ensure that the flatness and smoothness requirements of the bonding surfaces are met. The oxide layers are then treated with nitrogen, oxygen, or argon plasma to activate their surfaces. Plasma treatment enables us to achieve high bond strength at low temperature, without which we would require a high-temperature anneal (incompatible with M3D integration) after bonding to achieve sufficient bond strength.

Next, the carrier wafer is flipped and bonded to the handle wafer in ambient air at room temperature. Since plasma activation leads to an increase in the number of hydroxyl groups at the bonding surfaces, the wafers bond spontaneously when brought in contact. The bonded wafers are then annealed at 200 ^oC for one to two hours. This post-bond annealing step is required to achieve sufficient bond strength. An appropriate force is then applied to break the wafer at the fracture plane. Finally, solid phase epitaxy is performed to re-crystallize the amorphous layer, and CMP is then carried out to remove end-of-range defects.

1.1.3 Computer-Aided Design of M3D ICs

An M3D IC can be partitioned at the core-level, at the block-level, at the gate-level, and at the transistor-level. Core-level and block-level design-partitioning approaches do not fully utilize the high integration density offered by M3D. The transistor-level design partitioning approach is illustrated in Fig. 1.4(a). In this approach, the P-channel and N-channel transistors are placed on different layers, thereby allowing the fabrication process to be optimized separately for each layer. The gate-level design partitioning approach is illustrated in Fig. 1.4(b). In this approach, all the transistors in a standard cell are fabricated in one layer. Note that 3D standard cells have to be designed for transistor-level M3D integration. However, existing design automation tools can be reused to place and route these 3D standard cells [17]. On the other hand, 3D-aware physical layout tools are required for gate-level design partitioning [27].

The floor-planning for M3D ICs is performed in two steps: (i) 3D placement, and (ii) MIV insertion [28]. 3D placement algorithms for M3D ICs that re-use existing 2D placement tools have been proposed in [27] and [29]. 3D routing and clock-tree synthesis algorithms for TSV-based 3D ICs can be re-used for M3D ICs with minimal changes [30]. The 3D power-network design algorithms for M3D ICs focus on reducing routing congestion while satisfying IR drop constraints [31].

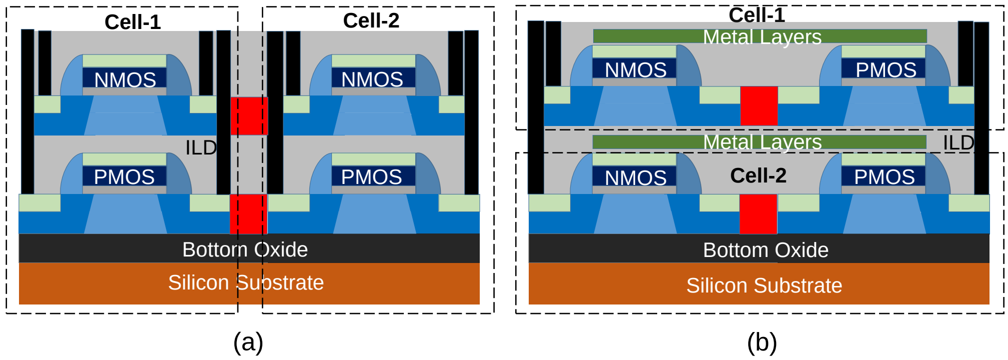


FIGURE 1.4: Design partitioning: (a) Transistor-level; (b) Gate-level.

1.2 Test Challenges for M3D ICs

In this section, we first present four test challenges for M3D integration: (i) timing variations due to high-density integration, (ii) defect analysis and modeling, (iii) defect isolation and yield enhancement, and (iv) yield loss due to voltage droop. For each test challenge, we motivate the need to study its impact on an M3D IC, analyze the effectiveness of existing test solutions, and develop new solutions.

1.2.1 Timing Variations due to High-Density Integration

In order to realize high-density vertical interconnects , the inter-layer dielectric (ILD) thickness in M3D ICs is being aggressively scaled [10, 17], which leads to electrostatic coupling between device layers. Electrostatic coupling was first observed in SOI transistors in which the front- and back-gates compete to induce charge in the silicon film (body); therefore, the electrical state of the silicon film at any point in time depends on the bias at both these gates. As a result, transistor parameters such as the threshold voltage and drain current depend on the bias at both these gates. Transistors in non-bottom layers in an M3D IC are SOI transistors since those layers are separated from their underneath layer by a dielectric. Therefore, transistors in non-bottom layers can be influenced by transistors underneath them.

The structure of an M3D IC partitioned at the transistor level is shown in

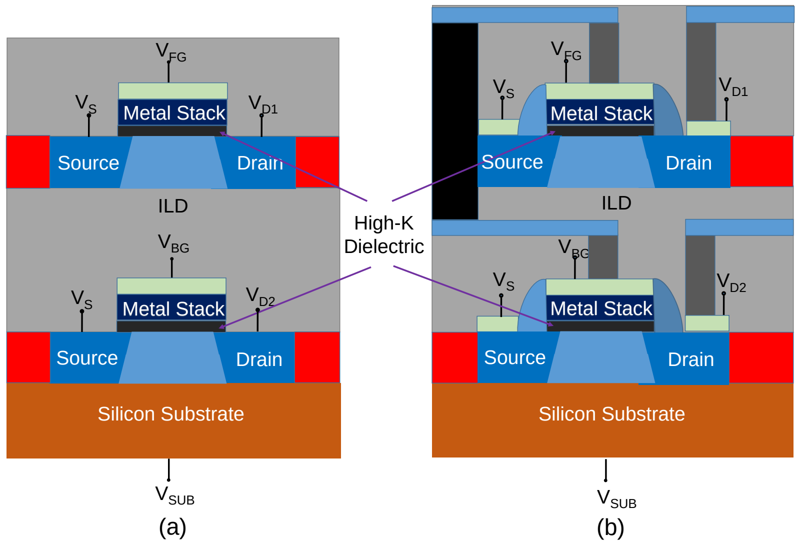


FIGURE 1.5: Structure of a partitioned M3D IC: (a) transistor level; (b) gate level.

Fig. 1.5(a). In such an M3D IC, the front-gate of a bottom-layer transistor acts as the back-gate for a top-layer transistor, and the ILD acts as the back-gate dielectric. Moreover, the front- and back-gates of a top-layer transistor will be connected to each other if all the standard cells are designed using static CMOS logic. Therefore, a top-layer transistor in an M3D IC partitioned at the transistor level can be regarded as a double-gate SOI transistor. Note that the double-gate transistor is asymmetric, i.e., the front- and back-gate work functions and dielectric thicknesses are different. Since the electrical state of the silicon film in a top-layer transistor depends on the voltage applied at the gate of the transistor below it, these transistors are coupled to each other.

Fig. 1.5(b) shows the structure of an M3D IC partitioned at the gate level. In such an M3D IC, a metal line (typically, the uppermost metal line) from the BEOL of the bottom layer acts as the back-gate for a top-layer transistor, and the ILD acts as the back-gate dielectric. The back gate will vary from transistor to transistor, and the actual 3D layout of the IC has to be examined to determine it. However, we can broadly classify the types of back gates we will encounter as follows: (i) a metal line

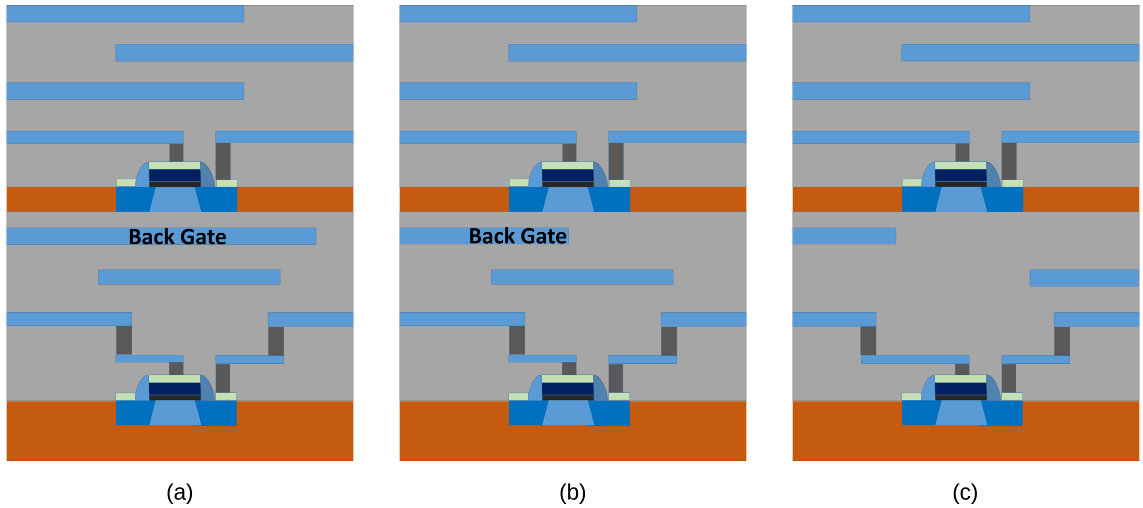
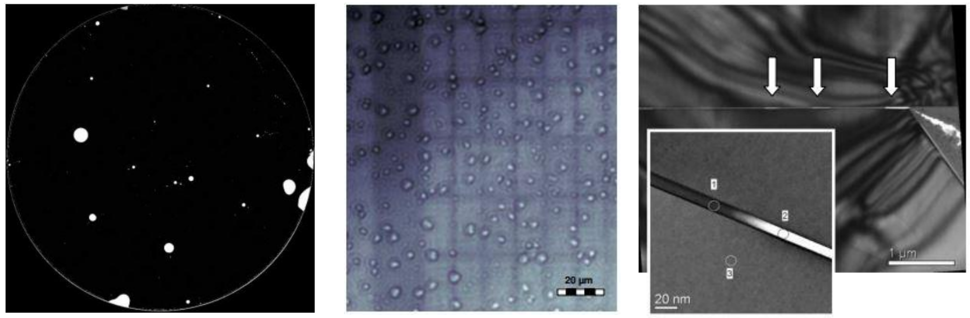


FIGURE 1.6: The three types of back gates encountered in gate-level integration: (a) complete overlap with the channel, (b) partial overlap with the channel, and (c) no overlap.

from the BEOL of the bottom layer which completely covers the channel, (ii) a metal line from the BEOL of the bottom layer which partially covers the channel, and (iii) a metal line from the BEOL of the bottom layer that overlaps with the channel is far away. Fig. 1.6 illustrates the three types of back gates. In the first two cases, the top-layer transistor can be regarded as an asymmetric double-gate SOI transistor in which the front- and back-gates are independent of each other. In the third case, the top-layer transistor can be viewed as conventional SOI transistor with a very thick bottom oxide, where the thickness of the bottom oxide depends on thickness of the BEOL of the bottom layer, and the transistor is not coupled to the bottom-layer devices. However, in the first two cases, the voltage on the metal line impacts the electrical state of the channel in a top-layer transistor, thereby coupling the top and bottom device layers.

1.2.2 Defect Analysis and Modeling

The condition of the bonding surfaces plays a crucial role in achieving a defect-free bond during the wafer-bonding step in the fabrication of an M3D IC. Oxide layers



(a) (b) (c)

FIGURE 1.7: Bond defects: (a) SAM image of bonded wafers showing mm size voids [32], (b) SAM image of bonded wafers showing µm size voids [33], (c) TEM image of bonded wafers showing sub-µm size voids [33].

are the prime candidates for bonding surfaces due to the presence of hydroxyl (OH) groups that lead to high bond strengths. These oxide layers also act as the ILD. Therefore, defects that arise during the wafer-bonding step can impact the top-layer transistors, as well as the ILVs. It is important to understand and analyze wafer-bonding defects, and develop methods to test for these defects.

The defects during the wafer-bonding process occur at the bond interface. Defects can often be attributed to voids, delaminations, and foreign particles [32–37]. Fig. 1.7 shows scanning acoustic microscopic (SAM) and transmission electron microscopic (TEM) images of bonded wafers with defects due to voids [32, 33]. The size of these defects can be in the mm to nm range. Considerable research has been reported on the root-cause analysis and methods to reduce the occurence of these defects [32–37]. The major factors that affect the quality of the bond can be summarized as follows:

- The root-mean-square value of surface roughness of the wafers to be bonded must be less than 0.5 nm.

- The total thickness variation of the bonding surfaces needs to be less than 5 µm.

- The warpage of the wafers prior to bonding has to be less than 50 µm.

- Accidental particle contamination or process-related residuals, e.g., from the CMP step, can degrade the bond strength. In addition, the presence of particle contaminants can result in large unbonded areas. Therefore, particle contamination levels are tightly controlled during the bonding process, and the wafers are cleaned thoroughly prior to the bonding step.

- The condition of the equipment used to handle the wafers has to be checked continuously. Equipment wear-out can lead to a variation in the initiation of contact between two wafers and the bond initiation force. Initiation of contact at more than one point will lead to the propagation of multiple bonds fronts, thereby resulting in voids. A very high or a very low bond initiation force can lead to defects in the bonded wafer or misalignment between the wafers, respectively.

1.2.3 Defect Isolation and Yield Enhancement

Industry presentations have highlighted concerns about defects that may arise in the bottom layer in an M3D IC when additional layers are processed [15]. In addition, non-bottom layers are susceptible to process variations [38] and electrostatic coupling, and the ILVs are prone to defects that arise during wafer bonding. Therefore, there is a need for low-cost design-for-test (DfT) solutions to enable defect isolation and yield enhancement.

Recently proposed DfT solutions for M3D, such as [39] and [19], suffer from several drawbacks. For example, [19] does not consider technology-specific details of M3D such as the sequential processing of device layers and the high density of I/Os in each layer. In [39], Cu-fuse structures are proposed to isolate the bottom layer. Cu-fuses not only impose considerable area overhead but also introduce feedback loops in the bottom layer. Such loops introduce sequential behavior to otherwise combinational logic, thereby necessitating the use of sequential ATPG. Feedback loops may also

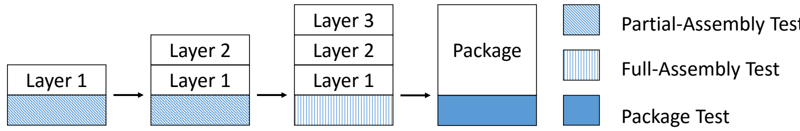


FIGURE 1.8: Test insertions associated with M3D integration.

lead to races and hazards, resulting in unpredictable behavior.

In the manufacturing of an M3D IC, tests can be applied at multiple stages, e.g., on a partial assembly of layers and on the full assembly with all the layers. Tests can also be applied after the full assembly is diced and packaged. During partial-assembly or full-assembly test, individual layers can be tested, or their testing can be omitted. Fig. 1.8 shows the typical integration flow for a three-layer M3D IC; it involves the incremental assembly of one layer at a time, and test insertions can be associated with these steps. These tests can cover the intra-layer logic, as well as inter-layer interconnections.

The IEEE Std P1838 can potentially be extended for M3D ICs. P1838 mandates the insertion of a die wrapper register for TSV-based 3D ICs that provides controllability and observability [40]. A similar register for M3D can enable modular testing by supporting inward-facing and outward-facing test modes. Such a register can enable modular testing by supporting inward-facing and outward-facing test modes. However, the overhead due to this register at the boundary of every layer can be significant since the number of ILVs in M3D ICs is expected to be an order of magnitude higher compared to TSV-based 3D ICs [3].

Although the extension of P1838 to M3D enables reuse of methods developed for TSV-based 3D ICs, new test solutions are needed due to the significant differences between M3D and TSV-based 3D in terms of design, fabrication, failure modes, and test constraints [3]. Moreover, a die (including the TSVs) can be tested pre-bond and a known-good die can be used for a TSV-based 3D stack [40]. On the other hand, ILVs are absent in the uppermost layer during partial-assembly test. For layers other than the top layer, there is no functional interface through which tests can be applied. The ILVs are not fabricated until the next silicon layer is assembled and the front-end is processed. Therefore, even if we insert probe pads on the uppermost layer, the logic connected to the ILVs will still remain untested. Wrapper cells are required for each ILV in order to test the connected logic. However, since the number of ILVs in an M3D IC can be very large, the area overhead due to wrapper cells for each ILV will be high.

1.2.4 Yield Loss Due to Voltage Droop

Fig. 1.9 presents an overview of a typical power-delivery network (PDN) for M3D ICs. Significant resources from the uppermost metal layers in each tier are dedicated for power delivery depending on the number of supply voltages used in the design. In addition, the uppermost metal layer in the bottom tier is used to design landing pads for power vias from metal-1 (M1) in the top tier. Significant routing blockages can occur in the uppermost metal layer of the bottom tier if it is heavily used for power delivery. Because of these blockages, the signal MIV count and wire length get impacted, thereby adversely affecting performance.

Voltage droop during testing is a major concern for M3D ICs since it can cause defect-free chips to fail on the tester, i.e., yield loss. Voltage droop is typically higher in magnitude during testing when compared to functional operation due to high switching activity during scan shift or capture [41]. Several power-aware test solutions have been proposed to carry out testing without violating specified power budgets [41, 42]. However, there is no previous work that analyzes voltage droop during test in M3D ICs.

Reliable power delivery is challenging in M3D ICs due to high power density and current demand per unit area. In addition, the high susceptibility of interconnects in

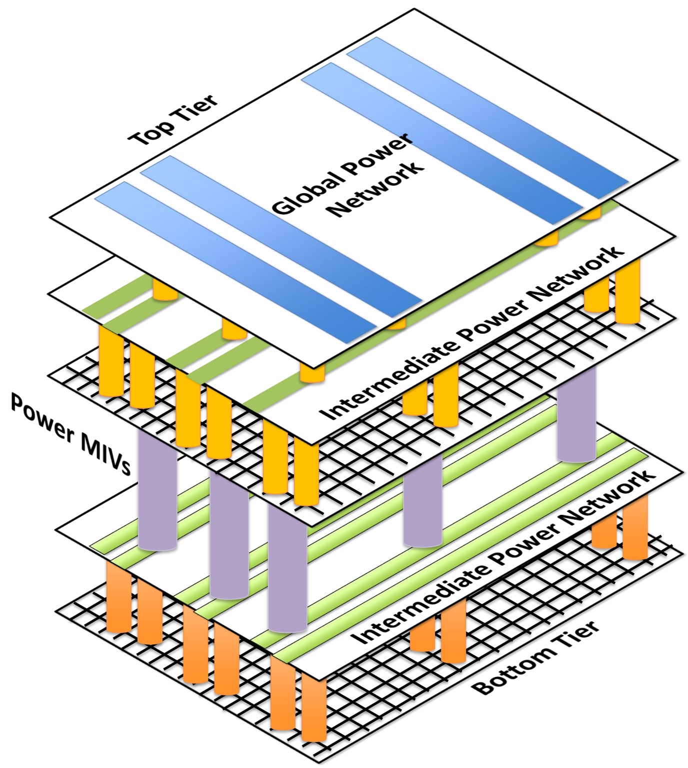


FIGURE 1.9: Overview of a typical PDN for M3D ICs.

M3D ICs to electromigration (EM) and stress migration (SM) increase the complexity of PDN design. Although several recent studies on power delivery for M3D ICs have been reported [31, 43, 44], the reliability of PDN in M3D and its impact on testing has not been explored.

1.3 Outline of Dissertation

The adoption of M3D technology by the industry can be impeded due to an insufficient understanding of test issues and the lack of DfT solutions. These concerns provide a strong motivation for research in testing of M3D ICs. In this dissertation, we carefully study the test challenges identified in Section 1.2 and develop solutions to tackle them. We first present the analysis of the test challenges, then we move to the test solutions. The remainder of this dissertation is organized as follows.

In Chapter 2, we first analyze electrostatic coupling in M3D ICs. We then analyze defects that arise due to voids created during wafer bonding, a key step in most M3D fabrication processes. We quantify the impact of these defects on the threshold voltage of a top-layer transistor in an M3D integrated circuit. We also show that wafer-bonding defects can lead to a change in the resistance of ILVs, and in some cases, lead to an open in an ILV or a short between two ILVs. We then analyze the impact of these defects on path delays using HSpice simulations. We study their impact on the effectiveness of delay-test patterns for multiple instances of IWLS'05 benchmarks in which these defects were randomly injected. Our results show that the timing characteristics of an M3D IC can be significantly altered due to coupling and wafer-bonding defects if the thickness of its inter-layer dielectric is less than 100 nm. Therefore, for such M3D ICs, test-generation methods must be enhanced in order to take M3D fabrication defects into account.

Chapter 3 presents a new DfT solution for M3D ICs based on dedicated test layers. The proposed DfT solution consists of a set of cooperating test layers, one for each interface between two functional layers. We describe the operating modes and configurations supported by the test layers. We also derive a cost model accounting for the additional steps involved in the fabrication of an M3D IC. We then compare the cost of the proposed DfT solution with the cost of a potential test solution based on P1838. We also present a test scheduling and optimization technique for wafer-level testing of M3D ICs.

Chapter 4 presents an ILV BIST solution for M3D ICs since conventional scan-based test pattern generation do not provide the diagnostic resolution to diagnose and characterize ILV faults. In the proposed ILV BIST solution, interface-register cells in a test layer are stitched into a TRC using their functional outputs and the ILVs. We show that the proposed solution detects all hard opens and shorts in the ILVs. We validate the detection of all hard opens and shorts using HSpice simulations. We also propose an artificial neural network-based framework to estimate the size of ILV defect and show that the prediction accuracy of the proposed framework is extremely high.

Chapter 5 describes an optimization approach for reliable power delivery in M3D ICs. We leverage genetic programming to explore the design space to optimize power delivery for M3D. We also analyze voltage droop during scan-based testing and compare it with that observed during functional operation. We then quantify the impact of voltage droop during scan-based testing on yield loss. Our results show that the proposed PDN design significantly reduces the worst-case voltage droop and yield loss due to voltage droop compared to a baseline.

Finally, Chapter 6 summarizes the contributions of the dissertation and identifies directions for future work.

2

Analysis of Coupling and Wafer-Bonding Defects

In this chapter, we analyze electrostatic coupling between device layers in M3D ICs and quantify its impact on circuit timing. We carry out detailed device simulations to understand the impact of coupling on the threshold voltage of a top-layer transistor for both transistor- and gate-level integration. We also study the impact of wafer-bonding defects on the threshold voltage of a top-layer transistor and on the ILVs [45–47]. We use HSpice simulations to quantify the change in path delays due to electrostatic coupling between device layers and the presence of defects at the bond interface. We also analyze the impact of coupling and wafer-bonding defects on the effectiveness of delay-test patterns using the statistical delay quality level (SDQL) metric [48]. This metric is correlated to the test escape rate due to small-delay defects (SDDs).

The remainder of the chapter is organized as follows. A study of the impact of coupling on the threshold voltage of devices is presented in Section 2.1. Section 2.2 presents a study of the impact of wafer-bonding defects on the threshold voltage of top-layer transistors. Section 2.3 discusses the impact of wafer-bonding defects on

ILVs. Section 2.4 and Section 2.5 analyze the impact of coupling and wafer-bonding defects on path delays and test quality, respectively. Finally, Section 2.6 concludes the chapter.

2.1 Impact of Coupling on Threshold Voltage

Due to the coupling between the top and bottom transistor layers, the threshold voltage of a transistor in the top layer can differ considerably from that of a conventional transistor (SOI or bulk CMOS), and depends on the bias and properties of the back gate. We carried out device simulations to accurately evaluate the impact of coupling on the threshold voltage of a top-layer transistor. Simulations were performed using Silvaco Atlas, a physics-based numerical device simulator.

We simulated two device structures using Atlas: (i) an asymmetric double-gate SOI transistor to study the impact of coupling in transistor-level M3D integration, and (ii) an asymmetric double-gate SOI transistor in which the front- and back-gates are independent of each other to study the impact of coupling in gate-level M3D integration. In transistor-level M3D integration, a top-layer transistor cannot be regarded as a double-gate transistor in all cases, especially if the design consists of pass-transistor and dynamic CMOS logic. However, in most practical circuits, pass-transistor and dynamic CMOS logic form a small part of the total chip area; therefore, neglecting the impact of coupling in such transistors can be justified.

The values of various parameters for the devices in our simulations were obtained from [1], and they are listed in Table 2.1. In our simulations, we have included physical models such as Shockley-Read-Hall recombination, Lombardi's mobility,

Table 2.1: Device parameters obtained from [1].

Gate Length	HfO_2 Thickness	Poly-Si Layer Thickness	TiN Layer Thickness	Silicon Film Thickness	Buried-Oxide Thickness
50 nm	2.5 nm	50 nm	5 nm	10 nm	23 nm

and Selberherr's impact ionization. To get realistic results, we calibrated the model parameters using experimental values of on- and off-currents reported in [1]. The values of the work functions of the front and back gate in P- and N-channel transistors were adjusted during the calibration process.

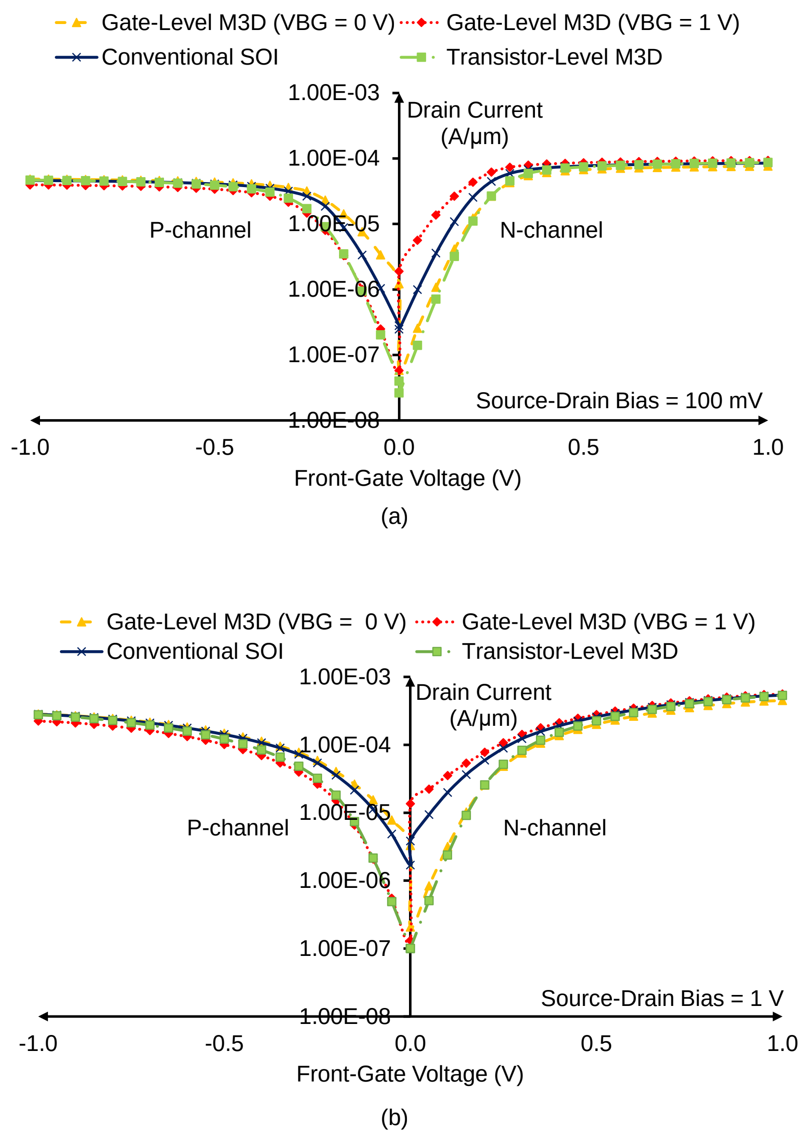


FIGURE 2.1: Plot of drain current as a function of front-gate voltage showing the shift in the threshold voltage in: (a) linear bias range, and (b) saturation bias range. Voltage notations used in this figure are illustrated in Fig. 1.5.

Fig. 2.1 presents several plots of the drain current (I_D) as a function of the front-gate voltage (V_{FG}) for double-gate transistors used to analyze the impact of coupling on transistor- and gate-level M3D. Each of these plots also include the $I_D - V_{FG}$ plot for a conventional SOI transistor. The thickness of buried oxide in the SOI transistor was chosen to be 500 nm. We chose this value since the back-gate does not impact the state of the channel for this oxide thickness. We observe that the $I_D - V_{FG}$ plots for a top-layer transistor in transistor- and gate-level M3D integration are shifted from those obtained for a conventional SOI device, in both linear and saturation bias ranges. Therefore, we expect a shift in the threshold voltage of these devices from a planar SOI device.

In the following paragraphs, we present the threshold-voltage values obtained from our simulations. The threshold-voltage values are calculated from $I_D - V_{FG}$ plots using the constant-current threshold voltage extraction method [49] for a drain-source bias of 100 mV. In this method, threshold voltage of a device is the gate-source voltage at which the drain current is equal to a constant current (I_{D0}) multiplied by the ratio of gate width to gate length. The value of I_{D0} is process dependent, and it typically ranges between 50 and 500 nA.The value of I_{D0} was chosen to be 300 nA for N-channel and 70 nA for P-channel transistors. These values for obtained from fab measurements for 32-nm partially-depleted SOI technology [49]. Note that the absolute value of the threshold voltage obtained using this method may not be accurate since the I_{D0} values used in this work have not been optimized for M3D technology. Nevertheless, the shift in threshold-voltage values that we report will still be accurate.

2.1.1 Impact of Coupling on a Top-Layer Transistor (Transistor-Level Partitioning)

Table 2.2 shows the threshold-voltage values of a double-gated device created to simulate a top-layer transistor in the case of transistor-level M3D integration. For

Table 2.2: Threshold-voltage values of a double-gated device for transistor-level M3D integration.

Channel Type	Conventional SOI	Perfect Alignment	Misalignment			
			5 nm	10 nm	15 nm	20 nm
N	0.117 V	0.168 V	0.167 V	0.166 V	0.164 V	0.161 V
P	-0.058 V	-0.109 V	-0.109 V	-0.108 V	-0.106 V	-0.104 V

the N-channel double-gate transistor, the threshold voltage is 0.168 V in the ideal case in which the front gate and the back gate are perfectly aligned with each other, which is 51 mV greater than the threshold voltage of a conventional-SOI transistor (obtained to be 0.117 V). For the P-channel double-gate transistor, the threshold voltage is -0.109 V in the ideal case in which the front gate and the back gate are perfectly aligned with each other, which is 51 mV less than the threshold voltage of a conventional-SOI transistor (obtained to be -0.058 V).

In addition, there can be some changes in the threshold voltage of a top-layer transistor due to the misalignment between the front and back gates. The misalignment between these gates results from the limitations of the lithography stepper accuracy, and the extent of this misalignment can be as high as 40% of the minimum feature size in current technology nodes [1, 50]. We observe that the change in threshold voltage for the case in which the back-gate is shifted towards the drain is similar to the case in which the back-gate is shifted towards the source. Nevertheless, our results show that misalignment between the front and back gates does not have a significant impact on the threshold voltage (only a 7 mV change in the worst case).

2.1.2 Impact of Coupling on a Top-Layer Transistor (Gate-Level Partitioning)

Table 2.3 shows the threshold-voltage values of a double-gated device created to simulate a top-layer transistor for gate-level M3D integration. For this transistor, the threshold voltage depends on the layout of the metal line, which in this case acts

Table 2.3: Threshold-voltage values of a double-gated device for gate-level M3D integration.

Channel Type	Conventional SOI	Back-gate Voltage	Perfect Alignment	Misalignment		
				10 nm	30 nm	50 nm
N	0.117 V	0.0 V	0.161 V	0.158 V	0.147 V	0.126 V
		1.0 V	0.052 V	0.057 V	0.082 V	0.107 V
P	-0.058 V	0.0 V	-0.004 V	-0.008 V	-0.030 V	-0.054 V
		1.0 V	-0.109 V	-0.106 V	-0.090 V	-0.070 V

as the back gate, as well as the voltage on the metal. We observe that the change in threshold voltage for the case in which the metal line is shifted towards the drain is similar to the case in which the metal line is shifted towards the source.

For the N-channel double-gated device, we observe a 44 mV increase from the threshold voltage of a conventional-SOI transistor for the case in which the metal line overlaps completely with the top-layer transistor. In addition, the threshold-voltage value decreases by 65 mV when the voltage on the metal line changes from 0 V to 1 V (assuming that the power supply voltage is 1 V). However, the change in threshold voltage due to a change in the voltage on the metal line decreases when the metal line does not completely overlap with the transistor.

For the P-channel double-gated device, we observe a 54 mV increase from the threshold voltage of a conventional-SOI transistor for the case in which the metal line overlaps completely with the top-layer transistor. In addition, the threshold-voltage value decreases by 51 mV when the voltage on the metal line changes from 0 V to 1 V. However, the change in threshold voltage due to a change in the voltage on the metal line decreases when the metal line does not completely overlap with the transistor.

2.1.3 Impact of ILD Thickness on Threshold Voltage

Thus far, we have evaluated the impact of coupling on the threshold voltage of a top-layer transistor by considering a 3D benchmark with an ILD thickness of 23 nm.

Although the usage of a thin ILD enables high-density ILVs and high-quality wafer bonding, we showed that it can lead to a significant shift in the threshold voltage (ΔV_{th}) from that of a conventional SOI transistor for both transistor- and gate-level M3D integration. We next vary the ILD thickness to evaluate the impact of ILD thickness on the ΔV_{th} due to coupling.

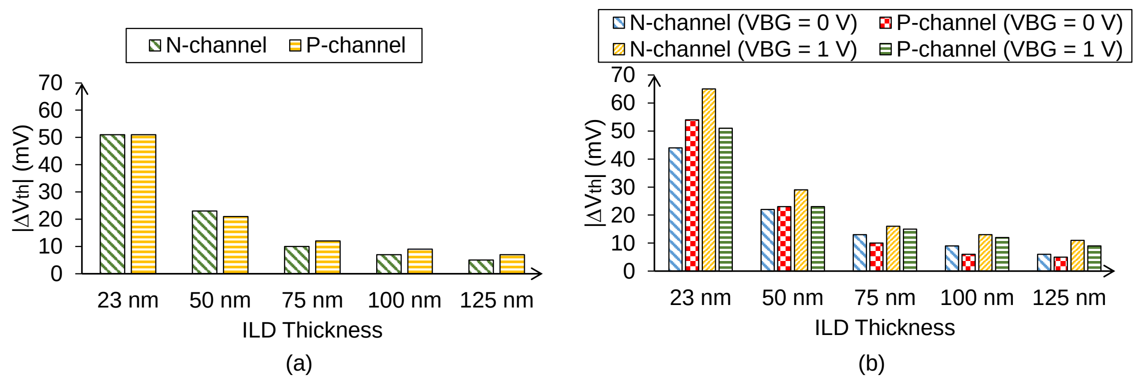


FIGURE 2.2: Change in the threshold voltage from that of a conventional SOI transistor for: (a) transistor-level M3D, and (b) gate-level M3D.

Fig. 2.2(a) shows the $|\Delta V_{th}|$ of a double-gated device created to simulate a top-layer transistor in the case of transistor-level M3D integration. For both N-channel and P-channel double-gate transistors, the $|\Delta V_{th}|$ was obtained to be 51 mV for an ILD thickness of 23 nm. However, for an ILD thickness of 100 nm, the $|\Delta V_{th}|$ reduces to 7 mV and 9 mV for N-channel and P-channel transistors, respectively. Therefore, we conclude that for ILD thicknesses greater than 100 nm, the impact of bottom layer on threshold voltage of a top-layer transistor in transistor-level M3D is minimal. For ILD thickness less than 50 nm, the impact of coupling on threshold voltage of a top-layer transistor is significant, and cannot be ignored.

Fig. 2.2(b) shows the $|\Delta V_{th}|$ of a double-gated device created to simulate a top-layer transistor in the case of gate-level M3D integration. For the N-channel double-gate transistor, the $|\Delta V_{th}|$ was obtained to be 44 mV and 65 mV for a back-gate (metal line from bottom layer) voltage of 0 V and 1 V, respectively. However, for an ILD thickness of 100 nm, the $|\Delta V_{th}|$ reduces to 9 mV and 13 mV for a back-gate voltage of 0 V and 1V, respectively. Similarly, for the P-channel double-gate transistor, we observe that the impact of a metal line from the bottom layer on the threshold voltage of a top-layer transistor is minimal for ILD thickness greater than 100 nm. Therefore, we conclude that for ILD thickness greater than 100 nm, coupling between the top- and bottom-layers is minimal. For ILD thickness less than 50 nm, the impact of coupling on threshold voltage of a top-layer transistor is significant, and cannot be ignored.

2.2 Impact of Wafer-Bonding Defects on Device Threshold Voltage

Although the occurence of bond defects can be reduced by tightly controlling the factors described in Section 1.2.2, it is not possible to completely prevent them from occuring due to increasing process variations at nanoscale technology nodes. It is especially difficult to control the occurence of nm size voids since they can easily form with a small variation in the surface roughness. It is therefore important to understand these defects, and analyze their impact on the functionality and performance of an M3D IC. Such an analysis will provide insights into the test methods required to screen defective and marginal M3D ICs.

The presence of a void at the bond interface impacts the threshold voltage of top-layer transistors. The oxide layers, which act as the bond interface, form the inter-layer dielectric (ILD). If a void is present in the back-gate dielectric of a double-gate transistor, the back-gate capacitance will be lower compared to the defect-free case, since the dielectric constant of air is lower than that of a dielectric material such as SiO_2 or Al_2O_3. Let us consider a void that is perfectly aligned with the channel of a top-layer transistor as shown in Fig. 2.3(a). We assume a void in the shape of a ellipsoid of revolution (oblate spheroid) since most voids reported in the literature were observed of that shape [51]. We also assume the diameter of the void to be greater than the channel length of the transistor under consideration but not large enough to affect transistors in other gates. This is a valid assumption since voids that impact transistors in multiple gates lead to catastrophic faults, and they can be easily detected using existing test methods. We are interested in analyzing the impact of voids that cause parametric faults as they are more difficult to test. With these assumptions, the effective back-gate dielectric capacitance can be expressed as the capacitance of the void in series with the capacitances of the dielectric layers above and below the void:

$$\frac{1}{C_{effective}} = \frac{t'_{box}}{\epsilon_{ox}} + \frac{t_{void}}{\epsilon_{void}} + \frac{t''_{box}}{\epsilon_{ox}} \tag{2.1}$$

where t_{box} is the thickness of the back-gate dielectric in the absence of a void, t'_{box} and t''_{box} are the thicknesses of the dielectric layers above and below the void, respectively, t'_{void} is the thickness of the void, and ϵ_{ox} and ϵ_{air} are the dielectric constants of the ILD and void, respectively. From (2.1), we can conclude that the effective back-gate dielectric capacitance ($C_{effective}$) will be less than the capacitance of the back-gate dielectric in the absence of a void (t_{box}/ϵ_{ox}) since $\epsilon_{void} < \epsilon_{ox}$. The dielectric constants corresponding to the ILD and a void are 3.9 and 1, respectively.

Fig. 2.3(b) and Fig. 2.3(c) show the case in which a void is not perfectly aligned with the channel of a top-layer transistor. In this case, computation of the effective back-gate dielectric capacitance is more complex since the potential across the back interface is not constant. However, the effective back-gate dielectric capacitance in this case will lie between that for the perfectly aligned case and the defect-free case. Therefore, we will observe maximum shift in the threshold voltage when the void is perfectly aligned with the channel of a top-layer transistor.

We carried out device simulations to evaluate the impact of a void on the threshold voltage of a top-layer transistor. In Section. 2.2.3, we concluded that the impact of

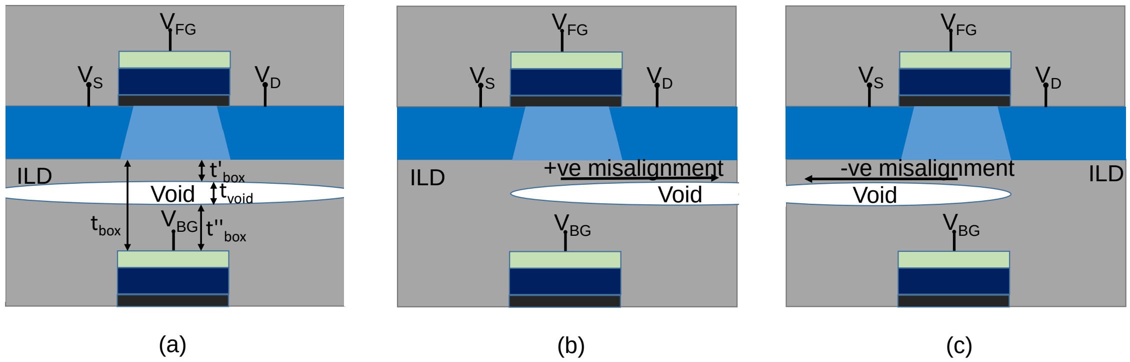


FIGURE 2.3: Alignment of a void with the channel of a top-layer transistor: (a) perfect alignment, (b) positive misalignment, and (c) negative misalignment.

coupling on threshold voltage of a top-layer transistor is minimal for ILD thickness greater than 100 nm. Since the impact of a void on the threshold voltage of a top-layer transistor depends on the extent of coupling between the top layer and the bottom layer, we carry out simulations for an ILD thickness of: (i) 23 nm, and (ii) 100 nm.

Fig. 2.4, Fig. 2.5, Fig. 2.6 and Fig. 2.7 present several plots of the drain current (I_D) as a function of the front-gate voltage (V_{FG}) for double-gate transistors used to analyze transistor- and gate-level M3D, respectively, for: (i) defect-free case, and (ii) the case in which a 12.5 nm thick void is present in the back-gate dielectric. Each of these plots also include the $I_D - V_{FG}$ plot for a conventional SOI transistor. We observe that the $I_D - V_{FG}$ plots for a top-layer transistor in transistor-level M3D integration with a void in the back-gate dielectric have shifted to the left and to the right from those obtained for a defect-free transistor for N- and P-channel devices, respectively. This shift occurs for both linear and saturation bias ranges. For a top-layer transistor in gate-level M3D integration, we observe that the direction of shift of the $I_D - V_{FG}$ plot depends on the back-gate voltage of the transistor in addition to the channel type.

Therefore, we expect a shift in the threshold voltage of these devices from the

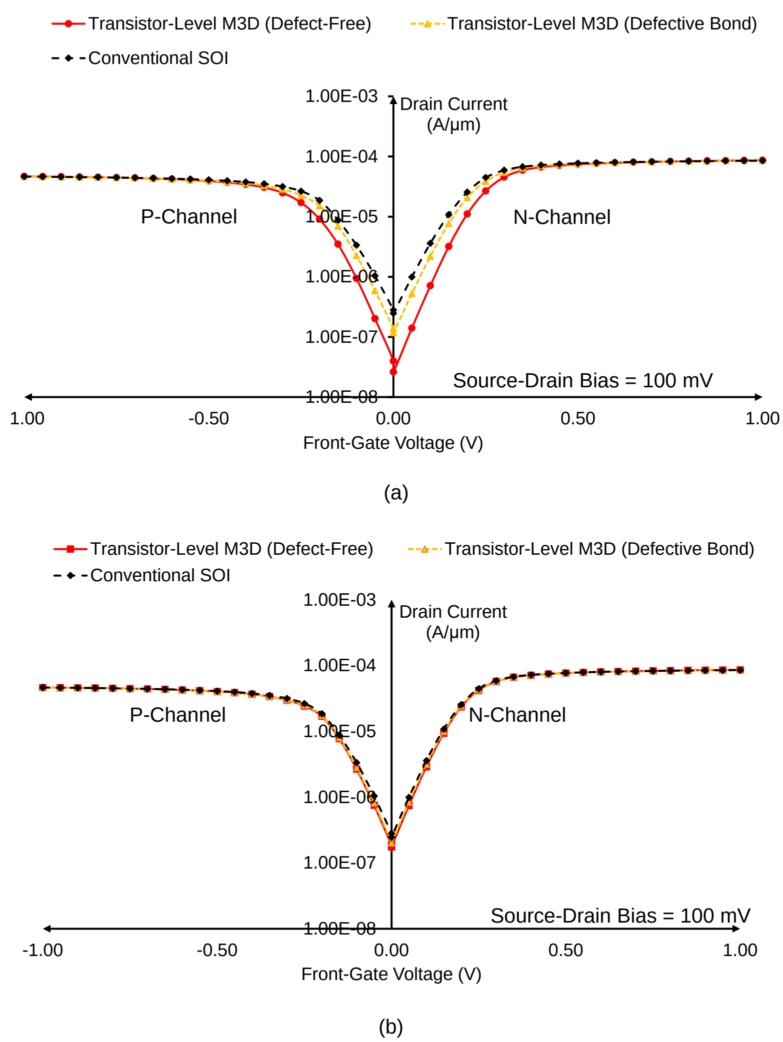
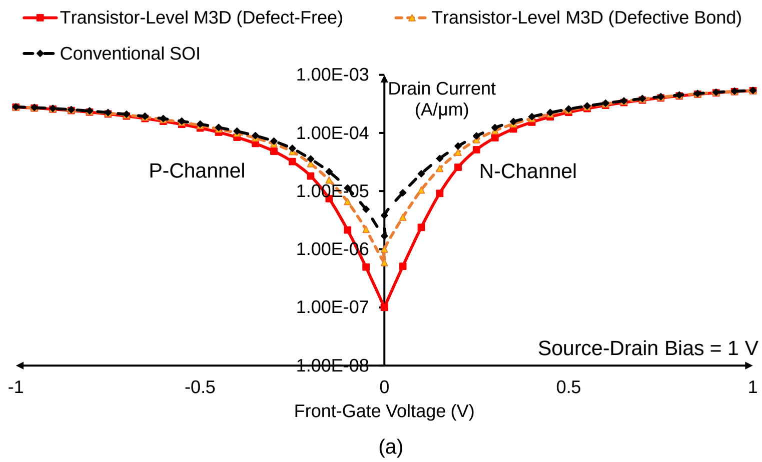


FIGURE 2.4: Plots of drain current as a function of front-gate voltage showing the shift in the threshold voltage for a top-layer transistor in transistor-level M3D in linear bias range for an ILD thickness of: (a) 23 nm; (b) 100 nm.

defect-free device for transistor- and gate-level M3D. In the following paragraphs, we present a detailed analysis of the threshold-voltage values obtained from our simulations. The threshold-voltage values are calculated from the $I_D - V_{FG}$ plot using the constant-current threshold voltage extraction method [49] for a drain-source bias of 100 mV.



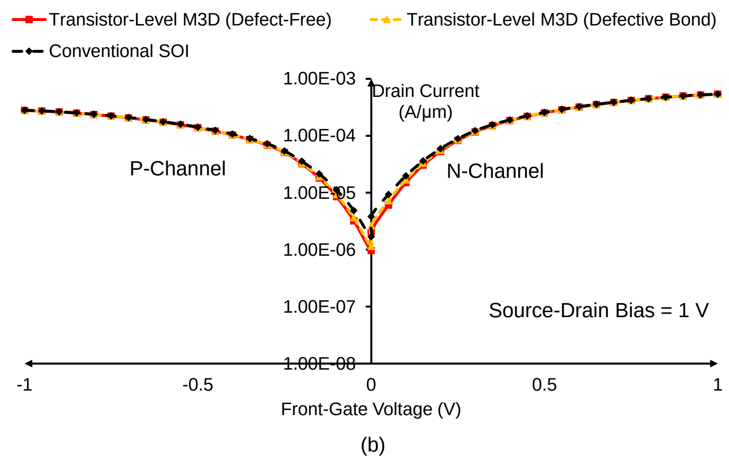


FIGURE 2.5: Plots of drain current as a function of front-gate voltage showing the shift in the threshold voltage for a top-layer transistor in transistor-level M3D in saturation bias range for an ILD thickness of: (a) 23 nm; (b) 100 nm.

2.2.1 Impact of Voids on a Top-Layer Transistor (Transistor-Level Partitioning)

Table 2.4 shows the threshold-voltage values of a double-gate transistor used to study the impact of a void on the threshold voltage of a top-layer transistor in transistor-level M3D integration. In this design partitioning approach, we use the threshold-voltage values of a defect-free top-layer transistor as the baseline since the $I_D - V_{FG}$ curves obtained for defect-free transistors match those obtained for conventional SOI transistors.

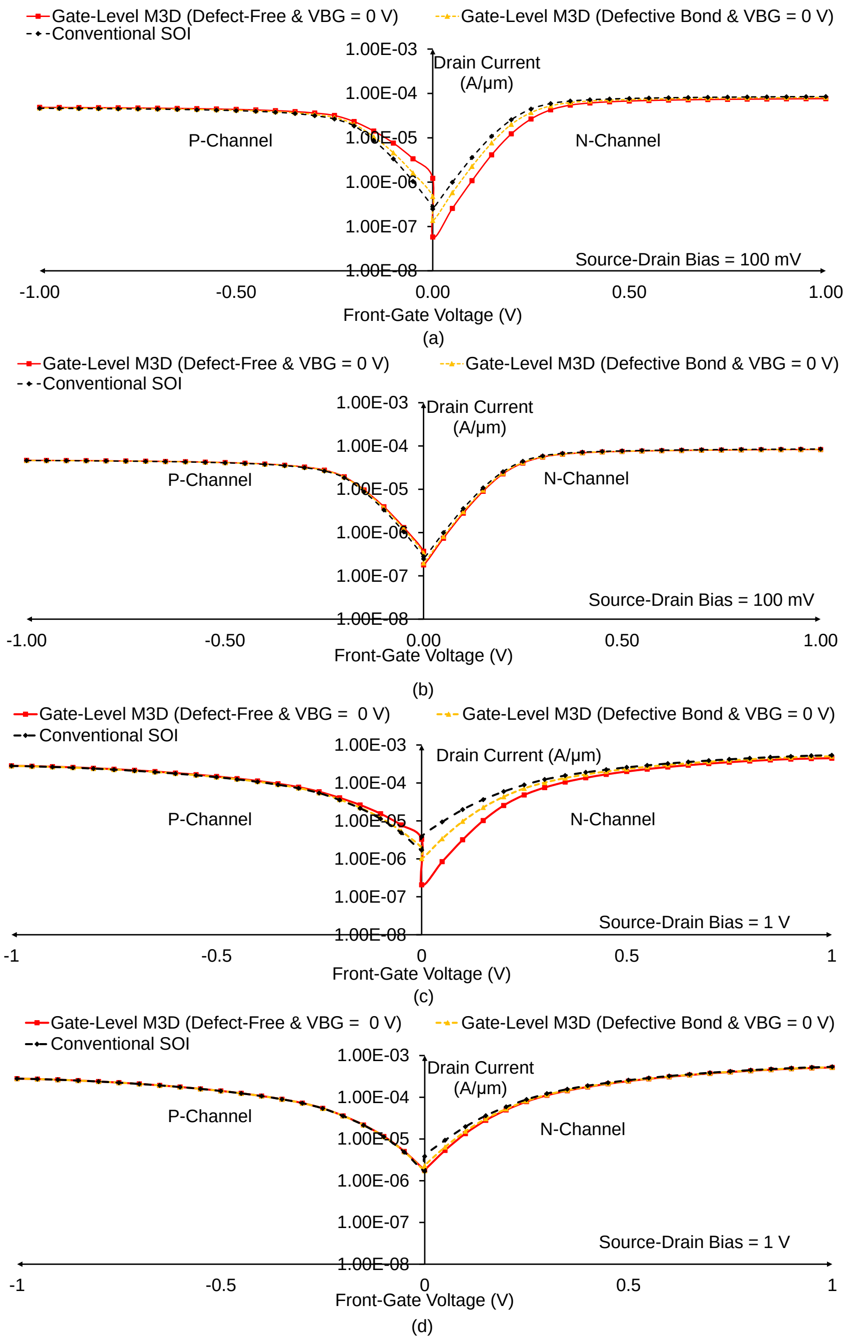


FIGURE 2.6: Plot of drain current as a function of front-gate voltage showing the shift in the threshold voltage for a top-layer transistor in gate-level M3D in: (a) linear bias range for a back-gate voltage of 0 V and ILD thickness of 23 nm, (b) linear bias range for a back-gate voltage of 0 V and ILD thickness of 100 nm, (c) saturation bias range for a back-gate voltage of 0 V and ILD thickness of 23 nm, and (d) saturation bias range using for a back-gate voltage of 0 V and ILD thickness of 100 nm.

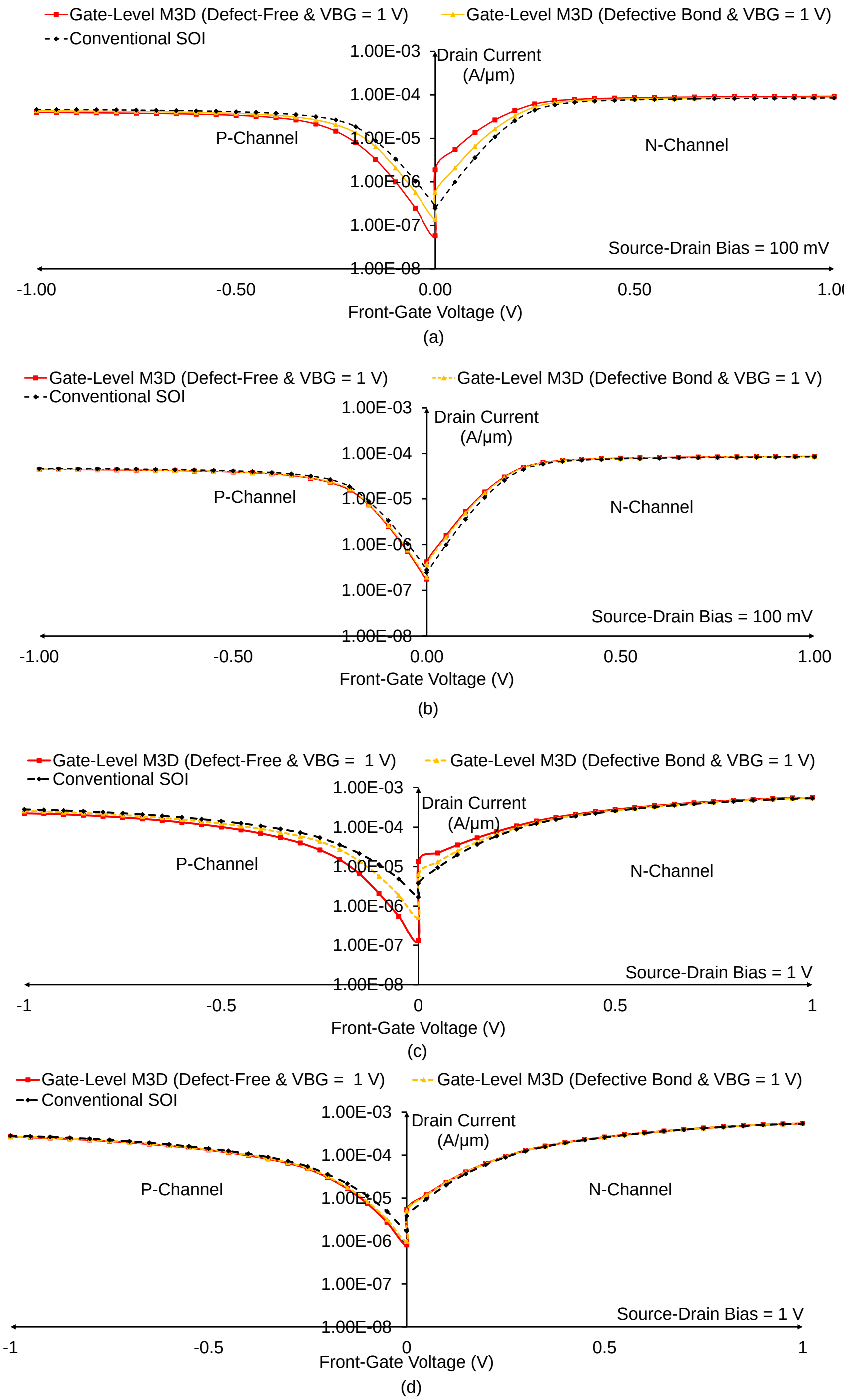


FIGURE 2.7: Plot of drain current as a function of front-gate voltage showing the shift in the threshold voltage for a top-layer transistor in gate-level M3D in: (a) linear bias range for a back-gate voltage of 1 V and ILD thickness of 23 nm, (b) linear bias range for a back-gate voltage of 1 V and ILD thickness of 100 nm, (c) saturation bias range for a back-gate voltage of 1 V and ILD thickness of 23 nm, and (d) saturation bias range using for a back-gate voltage of 1 V and ILD thickness of 100 nm.

For the defect-free P-channel double-gate transistor, the threshold voltage was obtained to be -0.109 V and -0.067 V for an ILD thickness of 23 nm and 100 nm, respectively. We observe a 8 mV increase from the threshold voltage of a defect-free double-gate transistor when the simulation was carried out for an ILD thickness of 23 nm and the back-gate dielectric capacitance was modified to account for a 2.5 nm thick void. The change in threshold voltage increases to 35 mV when the simulation was carried out for a 12.5 nm thick void. On the other hand, the change in threshold voltage from the defect-free case was obtained to be only 2 mV when the simulation was carried out for an ILD thickness of 100 nm for a 12.5 nm thick void.

For the defect-free N-channel double-gate transistor, the threshold voltage was obtained to be 0.168 V and 0.124 V for an ILD thickness of 23 nm and 100 nm, respectively. We observe a 10 mV decrease from the threshold voltage of a defect-free double-gate transistor when the simulation was carried out for an ILD thickness of 23 nm and the back-gate dielectric capacitance was modified to account for a 2.5 nm thick void. The change in threshold voltage increases to 32 mV when the simulation was carried out for a 12.5 nm thick void. On the other hand, the change in threshold voltage from the defect-free case was obtained to be only 3 mV when the simulation was carried out for an ILD thickness of 100 nm for a 12.5 nm thick void. Therefore, our results show that the presence of a void in the back-gate dielectric can have a notable impact on the threshold voltage of a top-layer transistor in an M3D

Table 2.4: Threshold-voltage values of a double-gated device for transistor-level M3D integration.

ILD Thickness	Channel Type	Defect-Free Case	Bubble thickness				
			2.5 nm	5 nm	7.5 nm	10 nm	12.5 nm
23 nm	P	-0.109 V	-0.101 V	-0.089 V	-0.083 V	-0.077 V	-0.074 V
	N	0.168 V	0.158 V	0.151 V	0.146 V	0.139 V	0.136 V
100 nm	P	-0.067 V	-0.066 V	-0.066 V	-0.065 V	-0.065 V	-0.065 V
	N	0.124 V	0.123 V	0.123 V	0.122 V	0.121 V	0.121 V

Table 2.5: Threshold-voltage values of a double-gated device for gate-level M3D integration.

ILD Thickness	Channel Type	Back-Gate Voltage	Defect-Free Case	Bubble thickness		
				2.5 nm	7.5 nm	12.5 nm
23 nm	P	0 V	-0.004 V	-0.016 V	-0.031 V	-0.041 V
	N	0 V	0.161 V	0.154 V	0.143 V	0.136 V
	P	1 V	-0.109 V	-0.101 V	-0.085 V	-0.077 V
	N	1 V	0.052 V	0.066 V	0.082 V	0.094 V
100 nm	P	0 V	-0.052 V	-0.052 V	-0.053 V	-0.054 V
	N	0 V	0.126 V	0.125 V	0.123 V	0.123 V
	P	1 V	-0.070 V	-0.069 V	-0.068 V	-0.067 V
	N	1 V	0.104 V	0.105 V	0.106 V	0.107 V

IC partitioned at the transistor-level (35 mV and 32 mV change in the worst case for P- and N-channel devices, respectively) when the ILD thickness is as small as 23 nm. However, their impact on threshold voltage is minimal when the ILD thickness is greater than 100 nm.

2.2.2 Impact of Voids on a Top-Layer Transistor (Gate-Level Partitioning)

Table 2.5 shows the threshold-voltage values of a double-gate transistor used to study the impact of a void on the threshold voltage of a top-layer transistor in gate-level M3D integration. In this design partitioning approach, we use the threshold-voltage values of a conventional SOI transistor as the baseline since the $I_D - V_{FG}$ curves obtained for defect-free transistors do not match those obtained for conventional SOI transistors. This mismatch is due to the electrostatic coupling between the top and bottom layers.

For the defect-free P-channel double-gate transistor, the $|\Delta V_{th}|$ from that of a conventional SOI transistor was obtained to be 54 mV and 6 mV for a back-gate voltage of 0 V for an ILD thickness of 23 nm and 100 nm, respectively. The $|\Delta V_{th}|$ is 51 mV and 12 mV, respectively, when the back-gate voltage is increased to 1 V. We observe that the $|\Delta V_{th}|$ decreases to 43 mV when the simulation was carried out for an ILD thickness of 23 nm for a back-gate voltage of 1 V and when the back-gate dielectric capacitance was modified to account for a 2.5 nm thick void. The $|\Delta V_{th}|$ decreases to 17 mV when the simulation was carried out for an ILD thickness of 23 nm for a 12.5 nm thick void. On the other hand, the $|\Delta V_{th}|$ is not impacted when the simulation was carried out for an ILD thickness of 100 nm. A similar trend in the threshold voltage values was obtained when we carried out simulations for the N-channel double-gate transistor. Therefore, we conclude that the changes in threshold-voltage values due to wafer-bonding defects are significant for gate-level M3D integration when the ILD thickness is as small as 23 nm. However, their impact on threshold voltage is minimal when the ILD thickness is greater than 100 nm.

2.3 Impact of Voids on Inter-Layer Vias

We now analyze the impact of wafer-bonding defects on ILVs. In the M3D fabrication flow, ILVs are etched after the top-layer transistors have been processed. The CoolCube$^{\text{TM}}$ technology uses tungsten for the ILVs and in the BEOL of the bottom layer due to its ability to withstand higher processing temperatures [5, 27]. Copper is used for the interconnects in the top layer. In order to make the M3D fabrication flow compatible with copper interconnects in the bottom layer, low-temperature steps to process the top-layer transistors are currently being developed [24, 25]. We therefore analyze the impact of wafer-bonding defects on copper ILVs.

Copper interconnects in today's ICs are processed using the dual-damascene process. In this process, both the metal line and the via are filled with copper in a single step. If the dual-damascene approach is used in the fabrication of an M3D IC, the ILVs are processed along with the first metal layer of the top layer. The key steps in the dual-damascene process are: (i) resist deposition and patterning, (ii) reactive-ion etching, (iii) deposition of barrier and seed layers using atomic-layer deposition (ALD), (iv) electroplating to fill the etched pattern with copper, and (v) chemical-mechanical planarization to remove excess copper.

The presence of a void in the dielectric at the position where an ILV is being etched impacts the electroplating process. To fill trenches and vias completely, without any voids or seams, a superconformal deposition of the barrier and seed layers is required [52]. However, if a void is present in the dielectric, superconformal deposition may or may not happen depending on the size of the void; thus, leading to the formation of voids or seams, and resulting in an increase in the resistance of the ILV. In some cases, the void can be large enough to cause an open defect in the ILV. On the other hand, if the size of the void is large enough to impact two neighboring ILVs, then it leads to a resistive short between the two ILVs. The resistance of the defect depends on the thickness of the void. ILVs are affected by voids since they are processed after the wafer-bonding step.

2.4 Impact on Path Delays

We carry out HSpice simulations on a small design to analyze the impact of coupling and wafer-bonding defects on path delays. We use the Leti-UTSOI2.1 model in these simulations [53]. This model has been developed for ultra-thin body and buried-oxide SOI transistors, and they offer the flexibility of applying different front- and back-gate voltages, work functions, dielectric thicknesses, and dielectric constants. We calibrated the model parameters using experimental values of on- and off-currents reported in [1]. The purpose of this analysis is to show that path delays in an M3D IC depend on the extent of coupling between device layers, and on the number, location, and size of defects that arise during wafer bonding.

Fig. 4.6 shows the design on which we performed our analysis. We partition this design into two layers as shown in Fig. 4.6 in order to simulate the case of gate-level design partitioning. In the case of transistor-level design partitioning, all the N-channel transistors are placed on the top layer and all the P-channel transistors

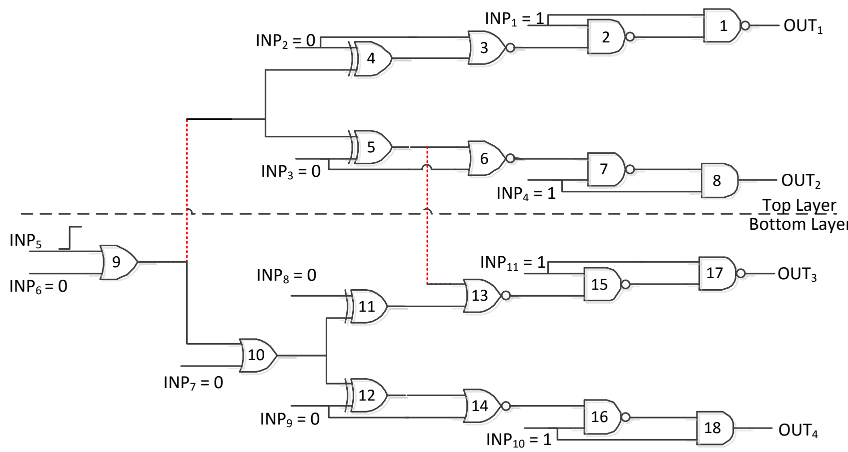


FIGURE 2.8: Design used to study the impact of bond defects on path delays.

are placed on the bottom layer. As shown in Fig. 4.6, constant DC voltages were applied to all the inputs, except on one input, on which a rising transition with a rise time of 5 ps was applied. The delay in propagating this transition to the output nodes was calculated.

We carry out simulations on 100 instances of the above design, which are created by randomly injecting defects on the gates in that design. In the case of transistor-level design partitioning, we inject four defects that include voids, which impact top-layer transistors (N-channel), and at the most one ILV defect. We only consider resistive opens and intra-cell shorts for the ILV defects. We do not consider inter-cell shorts as they are less likely to occur. The defect sizes considered for resistive opens (4 K-Ohm and 40 K-Ohm) and shorts (80 K-Ohm and 400 K-Ohm) in our simulations were obtained from those considered for regular vias [54]. For gate-level design partitioning of our design, there are only two ILVs (highlighted in the figure), if we neglect the ILVs for routing the inputs and the power supplies from the top layer to the bottom layer. We inject four defects even in this case that include voids, which impact top-layer transistors, and at the most one ILV defect (resistive open in one of the two ILVs or a resistive short between the two ILVs). We assume that a void impacts a complete gate in gate-level design partitioning, i.e., we consider the same defect size for both N- and P-channel transistors in a gate. We also vary the back-gate voltage of top-layer transistors to take into account the impact of coupling on path delays in a gate-level partitioned design.

Fig. 2.9(a) and Fig. 2.9(b) show the propagation delay to the output nodes for transistor-level design partitioning of the design under consideration for an ILD thickness of 23 nm and 100 nm, respectively. For transistor-level partitioning, we showed that the threshold voltage of top-layer transistors is different from that of conventional-SOI transistors due to coupling if the ILD thickness is less than 100 nm. However, this shift in the threshold voltage can be determined by carrying out careful simulations at design time since the variation due to misalignment is very small. Therefore, we select a defect-free transistor as the baseline for these simulations.

For each output, the delay in propagating the input transition to that node for the case in which no wafer-bonding defects were considered, and the case in which defects were considered, are shown in Fig. 2.9(a) and Fig. 2.9(b). Note that the propagation delay to all the four outputs are impacted since a defect can arise in any gate in the design. We observe that the path delays do not change significantly when the simulation was carried out for an ILD thickness of 100 nm, except when ILV defects are considered. This is expected as the impact of voids on the threshold-voltage of a top-layer transistor is not very significant. In addition, all the defects are not necessarily injected on the same path. However, we observe a notable change in the path delays when the simulation is carried out for an ILD thickness of 23 nm. For a few circuit instances, the ranking of the path delays is different from the case in which no wafer-bonding defects were considered. Such results were obtained when ILV defects were considered.

Fig. 2.10(a) and Fig. 2.10(b) show the propagation delay to the output nodes for gate-level partitioning of the design under consideration for an ILD thickness of 23

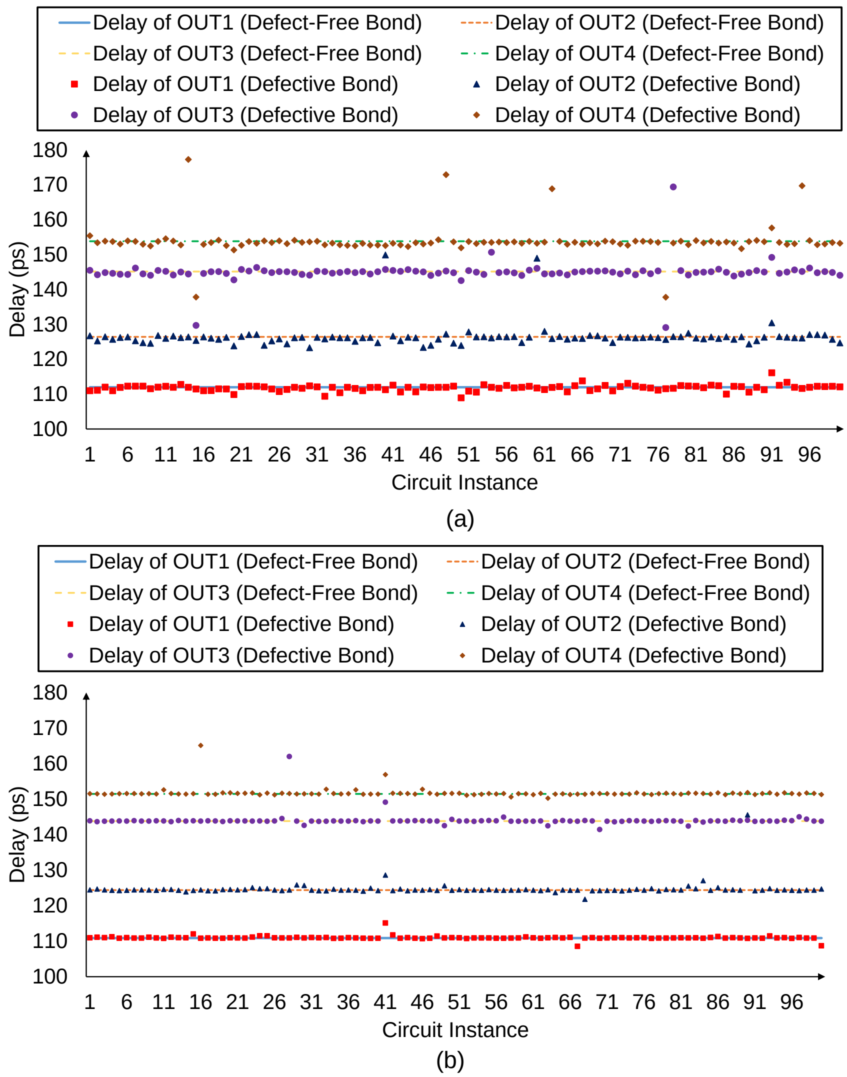
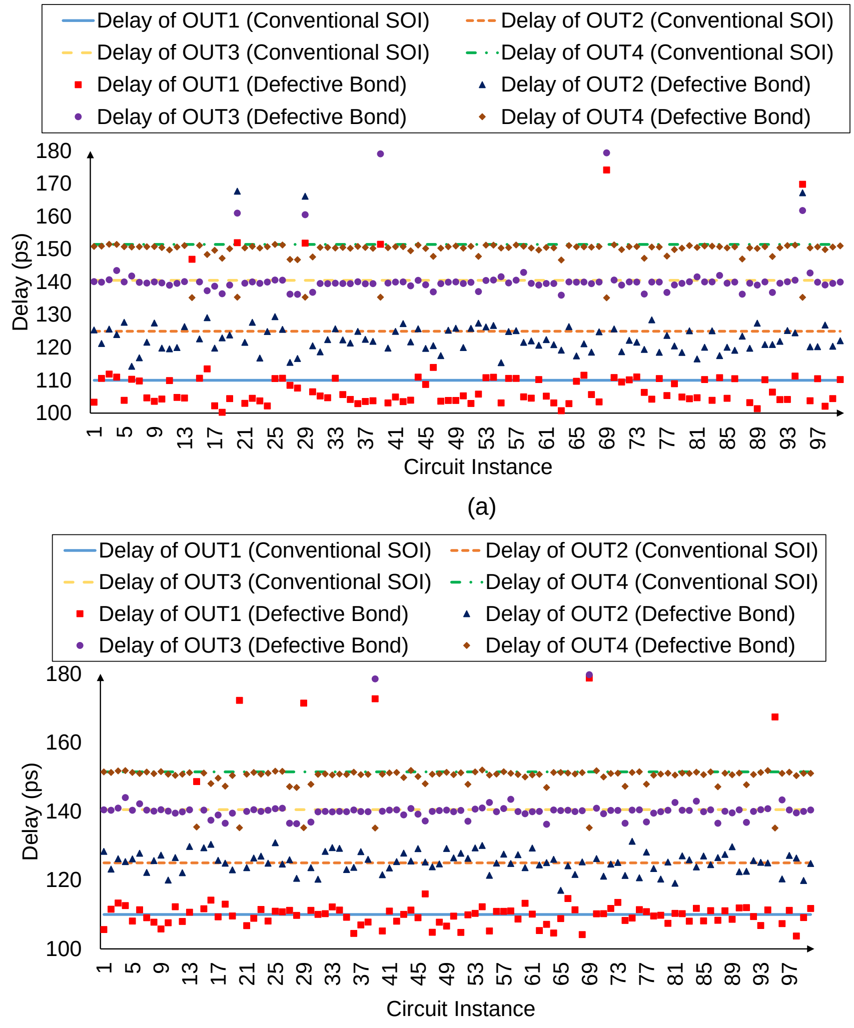


FIGURE 2.9: Propagation delay of the transition to the output nodes for the design shown in Fig. 4.6 for: (a) transistor-level partitioning for an ILD thickness of 23 nm; (b) transistor-level partitioning for an ILD thickness of 100 nm.

nm and 100 nm, respectively. For each output, the delay in propagating the input transition to that node for the case in which wafer-bonding defects were considered, and the conventional SOI case, are shown in this figure. For gate-level partitioning, we use a conventional-SOI transistor as the baseline case since the path delays can vary even in the defect-free gate-level design partitioning case due to variations in the back-gate voltage.



(b)

FIGURE 2.10: Propagation delay of the transition to the output nodes for the design shown in Fig. 4.6 for: (a) gate-level partitioning for an ILD thickness of 23 nm; (b) gate-level partitioning for an ILD thickness of 100 nm.

We observe that the change in path delays is more significant when the simulation was carried out for an ILD thickness of 23 nm compared to when the simulation was carried for an ILD thickness of 100 nm. In the former case, the change in path delays is notable even when ILV defects are not considered. For example, the average change in path delays over 10 circuit instances was obtained to be 4.57 ps and 2.64 ps for an ILD thickness of 23 nm and 100 nm, respectively. This is expected as the impact of voids, in combination with variations in the back-gate voltage, on the threshold-voltage of a top-layer transistor is significant for gate-level partitioning. Moreover, for the circuit instances in which all the defects were injected on the same path, or the instances in which ILV defects were considered, the impact on the path delays is significant. For these circuit instances, the ranking of the path delays can be different from the conventional SOI case. Therefore, coupling and wafer-bonding defects can cause a change in the path delays, thereby causing a significant change in the timing of the circuit, especially when the ILD thickness is less than 100 nm.

2.5 Impact on the Effectiveness of Delay-Test Patterns

SDDs have become a major concern at the nanoscale technology nodes due to increasing process variations. SDDs are difficult to screen because they only cause an error if the length of the path affected by them exceeds the clock period. However, effective screening of SDDs can be ensured by generating patterns specifically for those defects. Timing-aware automatic test pattern generation (ATPG) tools can be used to target the longest sensitizable path for each fault. Such ATPG tools rely on circuit timing information to generate path profiles. We showed that coupling and wafer-bonding defects significantly impact the path profiles in a gate-level-integrated M3D IC. If these changes are not considered, the ATPG tool will propagate faults through paths that are much shorter than the long paths. This will decrease the effectiveness of patterns used to screen delay defects, resulting in lower test quality.

Note that delay-test methods that consider process variations, e.g., the test-pattern selection method for screening SDDs [55], are unlikely to be effective for M3D ICs. We cannot obtain the delay distributions for the gates in an M3D IC by considering die-to-die, wafer-to-wafer, and lot-to-lot variations for different process parameters as proposed in [55] for conventional ICs. This is because the threshold voltage of top-layer transistors depends on the voltage on the uppermost metal line in the bottom layer, in addition to the number and size of the wafer-bonding defects.The

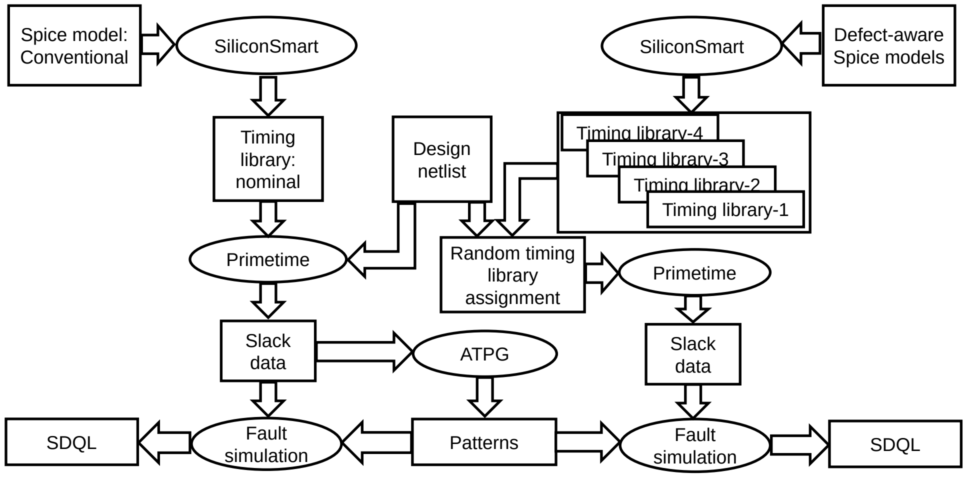


FIGURE 2.11: Tool flow used to analyze the impact of coupling on test quality.

technique presented in [55] (and other related prior work such as [56]) does not address such voltage-dependent variations.

Statistical static timing analysis (SSTA) based ATPG methods such as [57] take into account the effect of process variations on path delays. These methods rely on variability-aware delay data to generate test patterns to effectively screen SDDs. However, these methods require considerable computation time to generate test patterns. For example, in [57], each test pattern requires multiple dynamic timing analysis runs. The Gaussian pin-to-pin delay distributions are randomly sampled and provided as input to the timing analysis runs. Since the test data volumes of today's ICs are extremely high, SSTA-based ATPG methods will be inefficient.

2.5.1 Simulation Setup

We use three IWLS'05 benchmarks to analyze the impact on the effectiveness of delay-test patterns. These benchmarks were synthesized and scan-inserted using the Nangate open cell library for the 45 nm technology node. We assumed that each benchmark design constitutes the top layer in an M3D IC. In order to carry out this analysis, we developed a tool flow based on conventional library characterization, timing analysis, and ATPG tools: Synopsys SiliconSmart, Synopsys PrimeTime, and

Synopsys TetraMax, respectively. An overview of the tool flow is shown in Fig. 2.11.

We start with an open-source timing library such as the Nangate open cell library. The SPICE model used to characterize the standard cells in this library, and the netlists (post-parasitic extraction) of the standard cells can also be obtained. We then modify the SPICE model based on the threshold voltage values obtained in Section 4 to take the impact of coupling into account. The timing libraries for the gates in a defect-free M3D IC can be obtained by recharacterizing the nominal timing library with the modified SPICE models using SiliconSmart. We also generate the defect-aware timing libraries by injecting wafer-bonding defects (transistor and ILV) in the SPICE models of a defect-free M3D IC and the netlists of the standard cells, and then recharacterizing the timing library. We only consider resistive opens and intra-cell shorts for the ILV defects. We do not consider inter-cell shorts as they are less likely to occur. The range of defect sizes considered in our simulations are 4 K-Ohm - 40 K-Ohm and 80 K-Ohm - 400 K-Ohm for opens and shorts, respectively.

We use the nominal timing library to generate the test patterns for a conventional design. To generate these test patterns, we extract the slack data using PrimeTime, and then carry out timing-aware transition-fault ATPG. We use statistical delay quality level (SDQL) as the metric to evaluate the effectiveness of these test patterns to screen SDDs in an M3D IC. We carry out fault simulation and compute the SDQL values using TetraMax with: (i) the nominal slack data and the test pattern set as input, and (ii) the slack data for a defective M3D instance and the test pattern set as input. This procedure is repeated for multiple defective M3D instances, which are created by randomly assigning the defect-aware timing libraries to at most 10% of gates and the defect-free timing libraries to the rest of the gates in the design. The SDQL value computed for each M3D instance is compared with that obtained in the conventional case. We used the default values in TetraMax for the defect-distribution parameters used to compute SDQL. We repeated the simulations using a different set of parameter values and obtained similar results.

2.5.2 SDQL Results

We obtained the SDQL values for 50 instances of: (i) vga_lcd, (ii) leon3mp, and (iii) ethernet. The results are presented in Fig. 2.12. We normalized these SDQL values with respect to the SDQL value obtained using the nominal timing library for the standard cells in the corresponding benchmark. For most defective M3D instances of these benchmarks, we observe that the SDQL values obtained are greater than those obtained using the nominal timing library when the simulation was carried out for an ILD thickness of 23 nm. For example, the SDQL values for defective M3D instances of vga_lcd are upto 10% greater than those obtained using the nominal timing library. Such an outcome is possible if there is an increase in the range of defect sizes for most faults in the benchmark that are not tested by test patterns generated using the nominal timing library. For each fault, the range of untested defect size is defined as the difference between the slack on the longest-sensitizable path and the slack on the path being sensitized by the test patterns. These results show that coupling and wafer-bonding defects increase the range and size of defects that are not tested by timing-aware ATPG patterns, and therefore, the likelihood of delay-defect test escapes.

For some circuit instances, we also observe a decrease in SDQL. Such an outcome is possible if for most faults in the benchmark, the range of untested defect size decreases. For a fault, the range of untested defect size decreases if the slack on the path being sensitized by the test pattern decreases. Since we showed in Section 8 that the delay of a path can increase or decrease due to coupling and wafer-bonding defects, the slack on the path being sensitized by the test patterns can decrease in some cases, resulting in a lower SDQL value.

We also note that the change in SDQL value of vga_lcd is more significant com-

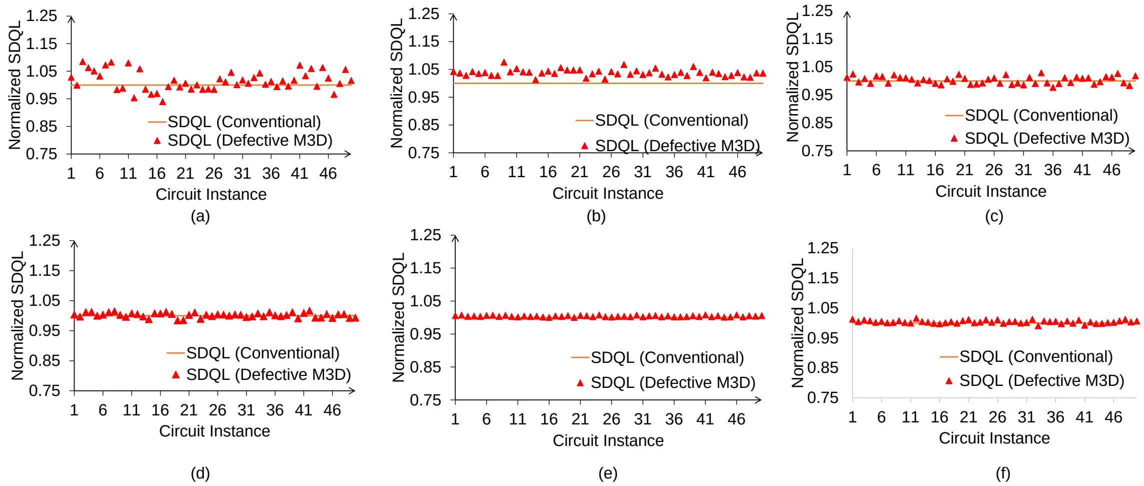


FIGURE 2.12: Normalized SDQL values for IWLS'05 benchmarks: (a) vga_lcd (23 nm thick ILD), (b) leon3 (23 nm thick ILD), (c) ethernet (23 nm thick ILD), (d) vga_lcd (100 nm thick ILD), (e) leon3 (100 nm thick ILD), (f) ethernet (100 nm thick ILD).

pared to leon3 and ethernet. SDQL is calculated as the sum of the probability of a delay fault being undetected and escaping the test over all the faults in a design. The probability of a delay fault being undetected depends on the range of untested defect size. Since the slack on a path depends on the circuit, the impact on SDQL is also circuit dependent.

When the simulations were carried out for an ILD thickness of 100 nm, we observed that the change in SDQL is minimal. For example, we observe only upto a 1% change in the SDQL values obtained for defective M3D instances of vga_lcd from that obtained using the nominal timing library. Therefore, we can conclude that coupling and wafer-bonding defects decrease the effectiveness of existing test-generation methods for screening SDDs when the ILD thickness is less than 100 nm.

2.6 Conclusions

We have analyzed the impact of electrostatic coupling and wafer-bonding defects on timing characteristics of an M3D IC. We carried out detailed device simulations to understand the impact of coupling and wafer-bonding defects on the threshold voltage of a top-layer transistor for both transistor- and gate-level integration. We then carried out HSpice simulations to understand their impact on path delays. We also showed that the presence of defects at the bond interface can lead to a change in resistance of an ILV, and in some cases, lead to an open in the ILV or a short between two ILVs. We then analyzed the impact of coupling and wafer-bonding defects on the effectiveness of delay-test patterns to screen SDDs. Our results show that test patterns generated using conventional ATPG tools are not effective for screening small-delay defects in M3D ICs that have been fabricated with an ILD less than 100 nm thick. For such M3D ICs, existing test-generation methods must be enhanced in order to take M3D fabrication defects into account.

3

DfT Solution Based on Dedicated Test Layers and Test Scheduling

In this chapter, we propose a new DfT solution for M3D ICs based on dedicated test layers, which are inserted between functional layers [58–60]. These layers provide controllability and observability to signals at the interfaces of functional layers. The main features of a test layer are: (1) a low-bandwidth serial interface, and a higher-bandwidth parallel interface, (2) dedicated probe pads on all layers, except the top layer, to enable partial-assembly testing, and (3) test structures to enable modular testing of the assembled layers.

The addition of test layers to the M3D assembly can potentially lower chip yield because of more candidate defect locations. However, the improvement in test coverage and defect-isolation capability offsets this concern. Moreover, the dedicated test layers can be manufactured using a mature technology and the number of back-endof-the-line (BEOL) layers can be minimized to reduce the impact on die yield. The impact on chip performance is also minimal since the size of an ILV that connects logic in the test layer to logic in the functional is comparable to that of a conventional via.

In the proposed solution, functional layers can have DfT structures such as scan-chains, test data CoDeCs, IEEE Std. 1500 wrappers, and test access mechanisms. However, all the layer-level test structures are located on the dedicated test layer. The placement of test structures for modular testing of M3D ICs on a separate layer is realistic since M3D integration enables significantly higher integration density compared to TSV-based 3D integration. Such high device density is achieved by using an extremely thin inter-layer dielectric (ILD) and silicon layers [3]. The dedicated test layers can be manufactured using a mature technology to reduce the impact on cost due to the additional processing steps.

The cost-per-die in M3D is impacted by the addition of DfT structures. The DfT structures can be placed on the functional layer, e.g., IEEE Std P1838 extension to M3D, or on a dedicated test layer.In order to justify the practicality of the proposed method, we carefully evaluate the DfT cost. We derive a detailed cost model by modifying the cost models for wafer-to-wafer stacked 3D ICs to account for the additional steps involved in M3D integration. We then carefully analyze and quantify the impact of the above test solution on yield and cost. Our results, for a 65 nm technology reported in the literature and an advanced foundry process, show that the proposed DfT solution is more cost-efficient than the P1838-based solution for a wide range of ILV density, ILV yield, and defect density.

We also propose an efficient test-scheduling technique for wafer-level partial-assembly or full-assembly testing of M3D ICs. The goal is to determine an optimal test schedule for wafer-level testing of an M3D IC such that the test-application time is minimized and the number of probe pads required to carry out the test does not exceed a given limit, while meeting power-budget constraints. To the best of our knowledge, this is the first test-scheduling approach reported for M3D integration.

The rest of the chapter is organized as follows. Section 3.1 describes the proposed

DfT solution. Section 3.2 analyzes the cost associated with the DfT solution. Section 3.3 presents test scheduling for wafer-level testing of M3D ICs. Section 3.4 presents the simulation results.Finally, Section 3.5 concludes the chapter.

3.1 Proposed DfT Solution

We consider M3D ICs that contain at least two functional layers. We refer to the layer that is processed first as the bottom layer, and the layer that is processed at the end as the top layer. We start with the assumption that the external I/O pins are located on the top layer. The proposed DfT architecture can also be used for an M3D IC in which the I/O pins are located on the bottom layer. The signals going from a particular layer to another layer in the direction of external I/Os is referred to as the *primary interface*, and the collection of signals going to a layer in the direction opposite to the external I/Os is referred to as a *secondary interface* corresponding to this layer. Fig. 3.1 shows the primary interface and the secondary interface for the middle layer in a three-layer M3D IC.

Fig. 3.2(a) illustrates the impact of wrapper cells on die footprint. This figure presents an example M3D IC consisting of two functional layers that implements a test solution based on P1838. The external I/Os of the IC are all located on the uppermost test layer, and are wrapped by IEEE 1149.1 boundary scan. We illustrate through this figure that the area overhead due to the insertion of wrapper cells can be prohibitively large.

The DfT solution proposed in this paper is based on dedicated test layers, which provide controllability and observability to the signals at the interfaces of functional layers. A register is added to the test layer to control and observe the signals at the interface of two functional layers. We refer to this register as the interface register. By placing the interface register on a dedicated test layer, we are not adding any test structures to the functional layer apart from conventional test structures.

An overview of the architecture is presented in Fig. 3.2(b). The figure shows an example M3D IC consisting of two functional layers and two test layers. The test layer for the uppermost functional layer is optional, and the IEEE 1149.1 boundary scan can be placed in the uppermost functional layer itself. A limited number of additional pins are required for boundary scan, of which two (TDI and TDO) are shown. The functional layers have intra-layer DfT such as scan-chains, test data CoDeCs, IEEE Std wrappers, and test-access mechanisms (TAMs). However, all the layer-level DfT is located in the test layer, thus facilitating the reuse of test structures.

We advocate the use of a dedicated test layer for all functional layers except the top layer since they can be processed using different technologies. Therefore, we may need to isolate and test each layer separately. In order to do that, we will require a test layer per functional layer except for the top layer. The test layer can be common for multiple functional layers if they are processed using the same technology and

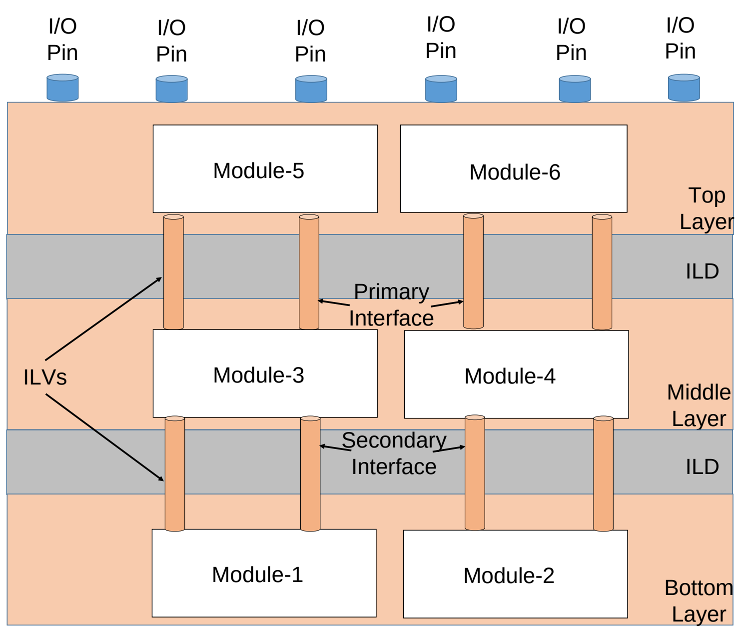


FIGURE 3.1: The primary interface and the secondary interface for the middle layer in a three-layer M3D IC.

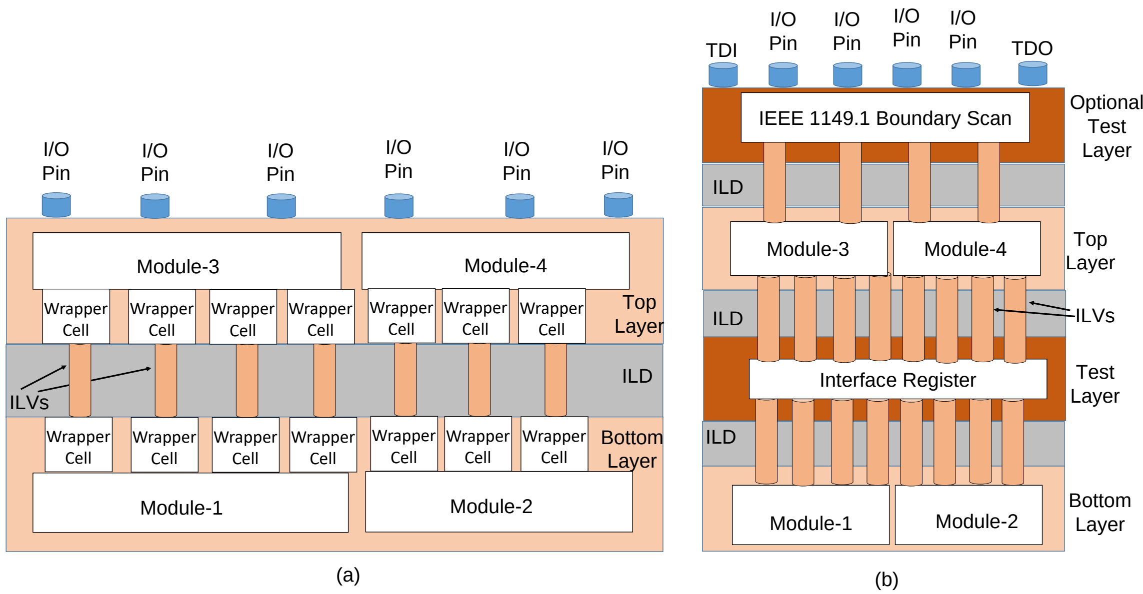


FIGURE 3.2: Overview of: (a) DfT solution based on P1838; (b) proposed DfT solution.

do not require isolation. However, we may lose test coverage in that case. In order to achieve good test coverage and die yield, we need to add a test layer for each functional layer.

The proposed DfT solution enables testing of a partial assembly of layers, and also of the full assembly with all the layers. During partial-assembly testing, the logic directly connected to the ILVs can now be tested before new functional layers are assembled since the ILVs at the primary interface of the uppermost layer are fabricated before a test layer is assembled over it. To perform partial-assembly testing, dedicated probe pads are required for each new layer.

Probe pads can be inserted on all test layers that will be used to carry out partial-assembly test and not just the top layer. In order to probe M3D ICs at a fabrication facility, low-force vertical probe cards are now available [61]. These probe cards, unlike cantilever probes, cause negligible damage to the probe pads. In addition, a standard clean-up step can be carried out as a part of the test protocol. Note that the probe pads have to be covered with a SiN passivation layer before a new layer is stacked upon the existing layers.

Tests can also be applied after the full assembly is diced and packaged into individual ICs. The typical integration flow for an M3D IC involves the incremental assembly of one layer at a time, and test insertions can be associated with each of these steps. These tests can cover the intra-layer logic, as well as inter-layer interconnections. In order to facilitate board-level testing, we assume a IEEE 1149.1 compliant package interface for the full assembly. We also assume that each layer in an M3D IC is scan-testable.

The operating modes supported by a test layer are:

1. *Partial-assembly test*: test access via dedicated probe pads;
2. *Full-assembly test*: test access via external I/O pins;
3. *Primary test*: test the functional layer for which the test layer forms the primary interface;
4. *Secondary test*: test the functional layer for which the test layer forms the secondary interface;
5. *Bypass*: only the bypass register is included in the test access chain;
6. *Interconnect test*: test the interface register;
7. *Turn*: the TAM turns upwards from the test layer;
8. *Elevate*: the TAM goes downwards towards the next layer.

3.1.1 Serial Test Mode

We designed the serial test mode based on IEEE Std. 1149.1. In order to enable the serial test mode, each test layer is equipped with an IEEE 1149.1-compliant

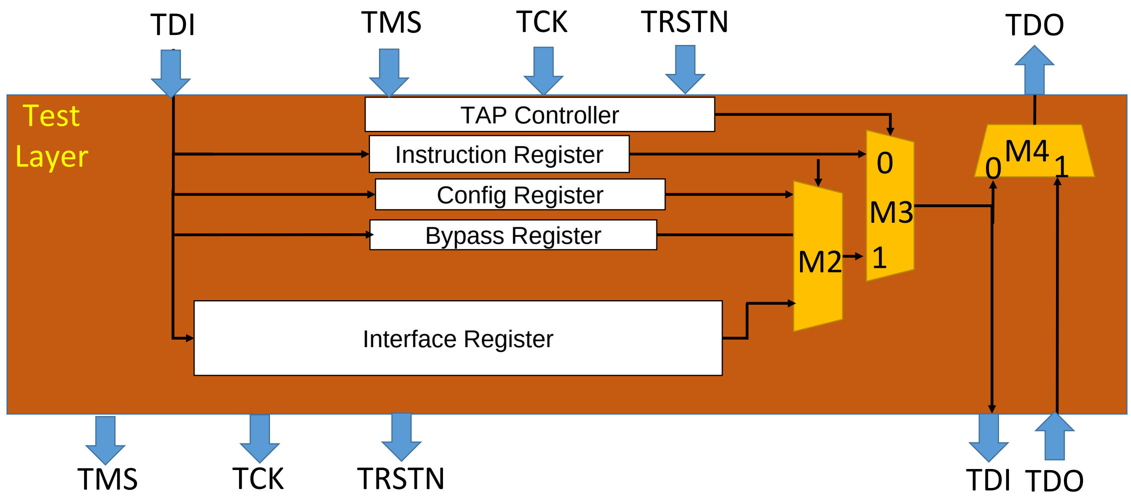


FIGURE 3.3: TAP interface for a test layer.

test access port (TAP), consisting of four input terminals TDI, TCK, TMS, and TRSTN, and an output terminal TDO. This layer also contains an IEEE 1149.1 TAP controller.

Fig. 3.3 shows the proposed TAP interface for a test layer. The registers that are included in a test layer are: (i) an instruction register to program the test mode, e.g., Intest, Extest, Bypass, etc., and to select a test data register to be active between TDI and TDO; (ii) configuration registers to determine the TAP configuration; (iii) bypass register that is used to bypass a test layer.

We propose the following instructions for the instruction register: (i) INTERFACE: to select the interface register between TDI and TDO; (ii) CONFIG: to configure the TAP; (iii) PRIMARY: to transport test data to the functional layer for which the test layer forms the primary interface; (iv) BYPASS: to bypass the test layer; (v) SECONDARY: to transport test data to the functional layer for which the test layer forms the secondary interface.

The configurations supported by the configuration register are: (i) COMPLETE-ASSEMBLY-TURN, (ii) COMPLETE-ASSEMBLY-ELEVATE, (iii) PARTIAL-ASSEMBLY-TURN, and (iv) PARTIAL-ASSEMBLY-ELEVATE. In the PARTIAL-ASSEMBLY and COMPLETE-ASSEMBLY configurations, the test access is via dedicated probe pads and ILVs, respectively. The TURN and ELEVATE configurations are used to program the serial path of an assembly. These configurations determine if a test-data register (TDR) of the functional layer for which the test layer forms the primary interface is included in the serial path. If a test layer is programmed to work in TURN mode, that register is not included in the serial path. If a test layer is programmed to work in ELEVATE mode, that register is included in the serial path. Multiplexer $M4$ determines whether that TDR is included in the serial path or not.

3.1.2 Parallel Test Mode

We have designed the parallel interface of a test layer based on the flexible parallel port (FPP) described in the proposed IEEE Std P1838. However, the test mechanism of the proposed parallel interface is significantly different from that of P1838 FPP. In P1838, die wrapper registers are inserted at the boundary of each die. On the other hand, interface registers are inserted at the primary and secondary interface of each functional layer in the proposed DfT solution. Therefore, grouping of scan chains and test control mechanism will be significantly different.

Fig. 3.4 shows the proposed parallel interface for a test layer. Since the I/O ports in an M3D IC are located on the uppermost layer, the downward parallel bus, i.e., the bus that runs from the uppermost layer to the bottom layer, transports the test stimuli to the target layer, and the upward parallel bus, i.e., the bus that runs from the bottom layer to the uppermost layer, transfers the test responses back to the I/O ports. An FPP can be configured to be a part of the downward parallel bus or the upward parallel bus. Each FPP has four bi-directional terminals: (i) FPP_Pri_Chain; (ii) FPP_Pri_Test_Layer; (iii) FPP_Sec_Chain; (iv) FPP_Sec_Test_Layer. There are two additional terminal, From_Side and To_Side, which serve as input and output, respectively. The bi-directional terminals can be configured to serve only as input

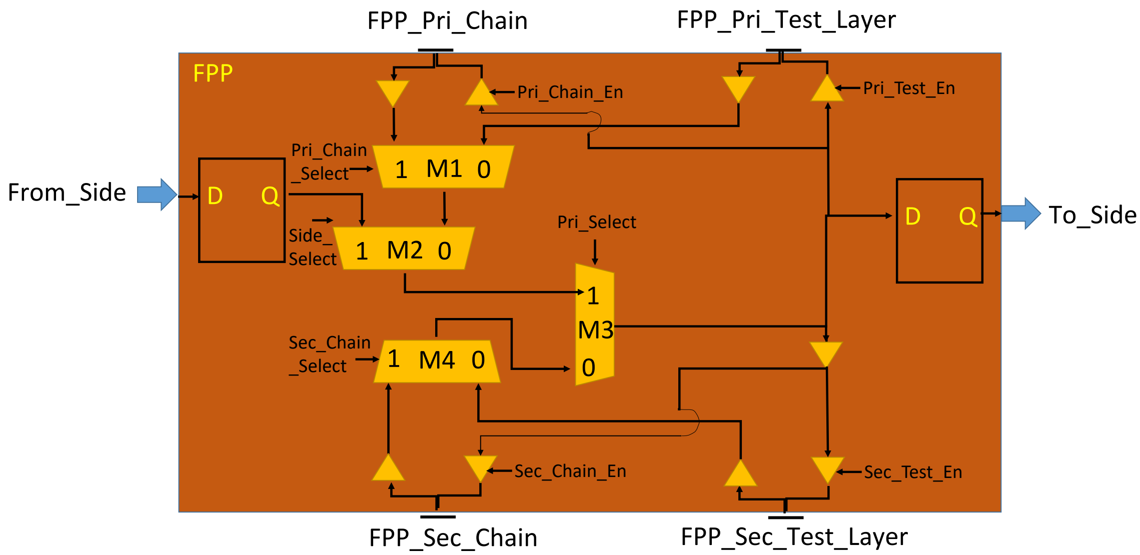


FIGURE 3.4: The parallel interface corresponding to a test layer.

or output.

The control signals needed to control the multiplexers and tri-state buffers are held in a configuration register. We refer to this register as the parallel port (PP) configuration register. The serial interface of a test layer is used to program the PP configuration register. The number of bits in the PP configuration register depends on the width of the parallel interface. A configuration hold bit is also included in the PP configuration register to block the reset signal from being propagated to the PP configuration register.

For transporting the test stimuli to a functional layer, all the FPPs, except those in the test layers at the primary and secondary interface of the target functional layer, are configured to bypass scan chains. Once the test patterns reach the test layer at the primary interface of the functional layer, they are transported from the input port (FPP_Pri_Test_Layer) in the primary test layer to the input port (FPP_Pri_Chain) in the secondary test layer through the scan chains in the functional layer. The test responses are transported back from FPP_Pri_Chain port in the secondary test layer to the output port (FPP_Pri_Test_Layer) in another FPP in the same test layer through the To_Side port.

The From_Side and To_Side ports of an FPP are used to turn the TAM upwards from a test layer. Here, we assume that the TAM originates from the uppermost layer. The TAM goes downward from a test layer and bypasses scan chains in the functional layer at the secondary interface of that test layer by using the FPP_Sec_Test_Layer as an output port. The TAM goes upward from a test layer and bypasses scan chains in the functional layer at the primary interface of that test layer by using the FPP_Pri_Test_Layer as an output port.

3.2 DfT Cost Analysis

In order to understand the cost associated with the proposed DfT solution, we develop a detailed cost model. We start with the cost models for wafer-to-wafer stacked 3D ICs and then account for the additional steps involved in the fabrication of an M3D IC. We use the cost model to compare the cost of the proposed DfT solution with the cost of a potential test solution based on P1838.

We assume that spot defects arise at the same rate on all device layers. In addition, we also assume that these defects tend to affect the device layer they are on, and not the layers above or below them. The first assumption is justified if we assume a homogenous M3D IC in which each device layer is processed using identical steps. A slight adjustment would be necessary for the first device layer since the processing steps would be slightly different between the first layer and the rest. We also realistically assume that spot defects that arise during the wafer-bonding step can only impact the ILVs, and not the devices in the layers above or below [46].

In a typical semiconductor manufacturing process, the cost-per-die of an IC (C_{die}) can be expressed as:

$$C_{die} = \frac{C_{wafer}}{N_{die} \times Y_{\text{eff}}} \tag{3.1}$$

where C_{wafer} is the total cost incurred in processing a wafer, N_{die} is the number of dies-per-wafer, and Y_{eff} is the effective wafer yield [2].

In order to process a wafer of M3D ICs that contain N layers, we need to process N device layers sequentially. In addition, we need to perform $N-1$ wafer-bonding and ILV processing steps. Therefore, the total cost incurred in processing a wafer (C_{wafer}) can be expressed as:

$$C_{wafer} = \sum_{i=1}^{N} C_{layer} + (N-1) \times C_{bonding} + (N-1) \times C_{via-layer} \tag{3.2}$$

where C_{layer} is the cost incurred in processing a device layer in an M3D IC, $C_{bonding}$ is the wafer bonding cost, and $C_{via-layer}$ is the cost incurred in processing one ILV layer.

Since we assume realistically that spot defects that arise while processing the device layers only affect the layer they are on and those that arise during wafer bonding only impact the ILVs, the effective yield of an M3D wafer (Y_{eff}) can be expressed as:

$$Y_{\text{eff}} = \prod_{i=1}^{N} Y_i \times (Y_{bonding+via})^{N-1} \tag{3.3}$$

where Y_i and $Y_{bonding+via}$ are the i^{th} device-layer die yield and ILV process yield, respectively. Note that the ILV process yield is the effective yield after taking into account the defects that arise during wafer bonding and ILV processing.

We now combine (3.1), (3.2), and (3.3) to obtain an expression for the cost-per-die of an M3D IC:

$$C_{die} = \frac{\sum_{i=1}^{N} C_{layer} + (N-1) \times C_{bonding} + (N-1) \times C_{via-layer}}{N_{die} \times \prod_{i=1}^{N} Y_i \times (Y_{bonding+via})^{N-1}} \tag{3.4}$$

From [2], the cost incurred in processing a device layer on a silicon wafer can be

$$C_{die} = \frac{\sum_{i=1}^{N} C_{FEOL} + N_{metal} \times C_{metal} + (N-1) \times C_{bonding} + (N-1) \times C_{via-layer}}{(\frac{(\pi \times (\phi_{wafer}/2))^2}{A_{die}} - \frac{\pi \times (\phi_{wafer})}{\sqrt{2 \times A_{die}}}) \times \prod_{i=1}^{N} (1 + \frac{A_{die} \times D_0}{\alpha})^{-\alpha} \times (Y_{bonding+via})^{N-1}} \tag{3.8}$$

---

expressed as:

$$C_{layer} = C_{FEOL} + N_{metal} \times C_{metal} \tag{3.5}$$

where C_{FEOL} is the FEOL process cost for each M3D layer, C_{metal} is the cost of processing a metal layer, and N_{metal} is the number of metal layers. We also obtain the expression for the number of dies-per-wafer from [2].

$$N_{die} = \frac{\pi \times (\phi_{wafer}/2)^2}{A_{die}} - \frac{\pi \times (\phi_{wafer})}{\sqrt{2 \times A_{die}}} \tag{3.6}$$

where N_{die} is the number of dies per wafer, ϕ_{wafer} is the diameter of the wafer, and A_{die} is the die area.

In order to derive an expression for the cost-per-die of an M3D IC, an accurate yield model is required. Recent studies based on actual silicon data have shown that the defects on a wafer tend to form clusters, and can be modeled using the gamma density distribution [62, 63]. Therefore, by assuming a gamma defect-density distribution, the relationship between the die area (A_{die}) and the die yield (Y_{die}) can be expressed as:

$$Y_{die} = (1 + \frac{A_{die} \times D_0}{\alpha})^{-\alpha} \tag{3.7}$$

where A_{die} is the die area, D_0 is the average defect density, and α is a parameter that can be used to account for defect clustering. The final expression for the cost-per-die of an M3D IC can be obtained by combining (3.4), (5.4), (3.6), and (3.7). It is shown in (8).

3.3 Test Scheduling for Wafer-Level Testing

Wafer-level testing, also referred to as wafer sort, is extremely important for M3D ICs to screen defective wafers prior to assembling additional layers or packaging, thereby improving yield and reducing package costs. Wafer-level testing can be carried out on a partial assembly or full assembly of layers using probe pads. Package test refers to the tests that are applied after dicing and packaging. Test planning and optimization is typically carried out for package test because the package test time is higher compared to wafer sort.

Significant research efforts have been directed towards test-access architecture and test scheduling optimization of 3D stacked ICs. In [64], ILP models for test architecture design of 3D stacked ICs are presented. However, this approach does not consider the reuse of die-level TAMs. An integer-linear programming formulation for 3D TAM and test schedule optimization considering any or all test insertions is presented in [65]. These optimization methods can be extended to M3D ICs since the post-packaging test-access architecture and test scheduling optimization problems are similar for 3D stacked ICs and M3D ICs. However, a constraint on the number of test elevators is not required since M3D integration enables one to two orders of magnitude higher interconnects between any two layers.

Although several test architecture design and optimization methods exist for pre-bond testing of 3D stacked ICs, they cannot be reused for M3D ICs because of the following reasons: (i) unlike 3D stacked ICs, wafer-level testing for M3D ICs can be carried out on an assembly consisting of one or more layers, (ii) number of probe pads that can inserted per die is significantly smaller for M3D ICs due to smaller die sizes, and (iii) test power reduction is extremely critical for M3D ICs due to limitations in heat removal caused by high density integration.

Traditional test solutions for wafer-level testing of core-based system-on-chip

(SoC) are ineffective for M3D ICs. Althought the test length and TAM optimization method proposed in [66] is effective for core-based SoC, it is not applicable for M3D IC because of: (i) the limited power budget for testing M3D ICs, and (ii) the higher complexity associated with testing M3D ICs because of their hierarchical nature, i.e., an M3D IC can have multiple layers, and each of them can consist if multiple cores. We therefore focus on TAM optimization and test scheduling for wafer-level testing of M3D ICs.

The objective here is to schedule tests such that the time required for wafer-level testing of M3D ICs is minimized. Test time can be reduced by using length of longest chain and power consumption for embedded cores to guide test scheduling. Moreover, reduced pin-count testing (RPCT) is desirable at wafer sort to reduce the number of probe pads or IC pins that need to be contacted by the tester and yield loss arising from contact problems with the wafer probe. RPCT further reduces test cost by enabling the reuse of old testers with limited channel availability. We therefore impose a constraint on the maximum number of probe pads that can be used during wafer sort.

Test scheduling refers to the problem of scheduling the embedded cores in different test sessions. Embedded cores in the same session are tested concurrently. We assume that the total TAM width is shared among all the cores in the same test session. Therefore, grouping of cores into the same test session is done based on fault coverage and power budget constraints. The number and length of test sessions determine the total test time.

We assume that each layer contains multiple IEEE 1500 wrapped cores. We choose 1500-style wrappers since they make no assumptions on the TAM and allow for greater flexibility. The test modes associated with a 1500-style wrapper, such as bypass, map naturally to the test modes associated with the proposed DfT solution. The proposed optimization technique does not make any assumptions on the specific test method being utilized. Therefore, any appropriate test method can be used with the proposed optimization method.

We show using an example the need for an optimization technique. In this example, we attempt to optimally schedule tests for an M3D IC with two functional layers and one test layer. We assume that the bottom layer consists of two cores A and B, and the top layer consists of two cores C and D. We impose a power budget of 100 arbitrary units for each layer and total power budget of 250 arbitrary units since excessive test power can significantly damage the chip, especially for M3D ICs. We also assume that the power consumed by test layers is negligible. We assume that the maximum number of probe pads, i.e., TAM width, available for wafer sort is 8.

If all the cores are tested in parallel, we can use four of the eight lanes to transport test stimuli to each core and the other four lanes to transport responses back to test pins. However, such a test schedule will violate the power budget constraint. On the other hand, if all the cores are tested in series, the test application time would be prohibitively large. Therefore, we need to find an optimal test schedule that meets the power-consumption requirements.

With the above example as motivation for better optimization methods, we now formally state the problem addressed in this work. The problem of test scheduling for wafer-level testing of M3D ICs is defined as follows.

Problem We are given an M3D IC under test consisting of a set of N layers and total number of probe pads $2W_{max}$ available for test. For each layer $n \in N$, we are given a set of cores C_n and power budget P_n. We are also given that the power consumption during each test session cannot exceed P_{max}. The goal is to determine an optimal test schedule for wafer-level testing of the given M3D IC such that the sum of test-application times of all the scheduled test sessions is minimized and the TAM width does not exceed $2W_{max}$, while meeting the power-budget constraints.

We develop an integer-linear programming (ILP) model for solving the test-scheduling problem. Although ILP models are computationally intractable, they are still feasible for small problem instances. Since practical M3D ICs consist of two or three layers, the problem instance is small enough for ILP. To facilitate our discussion of the ILP model, we use the following notations: (i) S represents a set of test sessions; (ii) s_i represents a test session belonging to S; (iii) c_n^j represents a core in layer n; (iv) x_n^{ij} is a binary variable indicating whether core c_n^j is tested in test session s_i, x_n^{ij} is 1 when c_n^j is tested in test session s_i and x_n^{ij} is 0 otherwise; (v) W represents a set of TAM widths that can be associated with all the cores; (vi) w_k is a candidate TAM width ($w_k \in W$); (vii) y_n^{jk} is a binary variable indicating whether core c_n^j is tested with TAM width $2w_k$, y_n^{jk} is 1 when c_n^j is tested with TAM width $2w_k$ and y_n^{jk} is 0 otherwise; (viii) t_n^{jk} represents a test-tuple that describes the test case when TAM width w_k is associated to core c_n^j with the associated test length and power consumption; (ix) $L_{t_n^{jk}}$ represents test length for the test case t_n^{jk}; (x) $P_{t_n^{jk}}$ represents power consumption for the test case t_n^{jk}; (xi) L_{s_i} represents the test length of s_i; (xii) $P_n^{s_i}$ represents the power consumption of layer n in test session s_i.

As described above, t_n^{jk} represents a test-tuple that describes the test case when TAM width $2w_k$ is associated to core c_n^j with the associated test length $L_{t_n^{jk}}$ and power consumption $P_{t_n^{jk}}$. Therefore, t_n^{jk} is a four-tuple:

$$t_n^{jk} = (c_n^j, w_k, L_{t_n^{jk}}, P_{t_n^{jk}}) \tag{3.9}$$

Test patterns can be generated for this test case and the power consumption $P_{t_n^{jk}}$ associated with those test patterns can be estimated. Let $N(t_n^{jk})$ denote the number of test patterns and CL_n^j denote the longest chain length associated with this test case. The test length $L_{t_n^{jk}}$ for this test case can be estimated as:

$$L_{t_n^{jk}} = ((CL_n^j + 1) \times N(t_n^{jk})) + CL_n^j \tag{3.10}$$

Since all the cores in a test session s_i are tested concurrently, the length of that session is:

$$L_{s_i} = \max\{L_{t_n^{jk}} \times y_n^{jk} \times x_n^{ij}\}, \tag{3.11}$$

where $n \in \{1, \ldots, |N|\}$, $j \in \{1, \ldots, |C_n|\}$, and $k \in \{1, \ldots, W_{max}\}$. Equation (3.11) implies that the test length of a test session equals the maximum test length of all the cores in that test session.

Since all the cores in a test session share the TAM width, we apply the following constraint:

$$\forall s_i \in S : \sum_{n=1}^{|N|} \sum_{j=1}^{|C_n|} ((\sum_{k=1}^{W_{max}} 2w_k \times y_n^{jk}) \times x_n^{ij}) \leq 2W_{max} \tag{3.12}$$

Equation (3.12) implies that for each test session $s_i \in S$, the TAM width cannot $2W_{max}$.

Moreover, for each core, y_n^{jk} should satisfy the following constraint:

$$\forall c_n^j : \sum_{k=1}^{W_{max}} y_n^{jk}) = 1, \tag{3.13}$$

where $n \in \{1, \ldots, |N|\}$ and $j \in \{1, \ldots, |C_n|\}$.

Since each layer in an M3D IC has to satisfy a power budget for every test session, we apply the following constraint:

$$\forall s_i \in S, n \in \{1, \ldots, |N|\} : \sum_{j=1}^{|C_n|} ((\sum_{k=1}^{W_{max}} P_{t_n^{jk}} \times y_n^{jk}) \times x_n^{ij}) \leq P_n \tag{3.14}$$

Constraint (3.14) implies that the power consumption of each layer cannot exceed the maximum corresponding to that layer for all test sessions.

Since we also impose a constraint on the overall power consumption for every test session, the following constraint must be satisfied:

$$\forall s_i \in S : \sum_{n=1}^{|N|} \sum_{j=1}^{|C_n|} ((\sum_{k=1}^{W_{max}} P_{t_n^{jk}} \times y_n^{jk}) \times x_n^{ij}) \leq P_{max} \tag{3.15}$$

Constraint (3.15) implies that the power consumption of each test session cannot exceed the maximum power consumption P_{max}.

To ensure that each die is tested only once, the following constraint must be satisfied:

$$\forall n \in \{1, \ldots, |N|\}, j \in \{1, \ldots, |C_n|\} : \sum_{i=1}^{|S|} x_n^{ij} = 1 \tag{3.16}$$

Now, the total test length L over all test sessions is:

$$L = \sum_{i=1}^{|S|} \max \{L_{t_n^{jk}} \times y_n^{jk} \times x_n^{ij}\}, \tag{3.17}$$

where $n \in \{1, \ldots, |N|\}$, $j \in \{1, \ldots, |C_n|\}$, and $k \in \{1, \ldots, W_{max}\}$.

We obtain the following objective function from equation (3.17):

$$\text{Minimize:} L, \tag{3.18}$$

The minmax objective in (3.18) can be linearized as follows: Minimize: $\sum_{i=1}^{|S|} z_i$, where $z_i \geq (L_{t_n^{jk}} \times y_n^{jk} \times x_n^{ij})$, $n \in \{1, \ldots, |N|\}$, $j \in \{1, \ldots, |C_n|\}$, $k \in \{1, \ldots, W_{max}\}$, and $i \in \{1, \ldots, |S|\}$.

Therefore, we obtained an ILP formulation to minimize the test time of all test sessions. This model can be solved using an ILP solver.

3.4 Simulation Results

In this section, we compare the cost of the proposed solution with that of P1838-based solution, discuss the timing overhead introduced by the proposed solution, and present test-scheduling results.

3.4.1 Cost Comparison

We compare the cost of the proposed DfT solution with the cost of a potential solution based on P1838. In [2], the relationship between the die area and the gate count has been described as: $A_{die} = N_{gate} \cdot A_{gate}$, where N_{gate} is the gate count, and A_{gate} is an empirical parameter that shows the proportional relationship between area and gate counts. Therefore, the die area of an M3D IC with DfT based on P1838 is larger compared to that without any DfT. Since the number of ILVs in an M3D IC is an order of magnitude higher compared to TSV-based 3D ICs, the impact on the die area can be significant. This increase in area impacts the die yield and the number of dies-per-wafer leading to an increase in the cost-per-die.

For P1838, we assume that a wrapper cell, which requires at least three gates [40], is added at the two ends of each ILV. We also assume a TAP interface for the external I/Os, located on the uppermost layer. A wrapper cell based on 1149.1 requires at least four gates [67]. We express the cost per die with DfT based on P1838 as shown in (3.19), where $N_{die}^{'}$ and $Y_i^{'}$ are the number of dies-per-wafer and the die yield after taking into account the area overhead due to the wrapper cells.

$$C_{die}^{1838} = \frac{\sum_{i=1}^{N} C_{layer} + (N-1) \times C_{bond} + (N-1) \times C_{via-layer}}{N_{die}^{'} \times \prod_{i=1}^{N} Y_i^{'} \times (Y_{bond+via})^{N-1}} \tag{3.19}$$

For the proposed solution, we assume that a test layer is added at each interface between two functional layers. We also assume realistically that a test layer is not added for the external I/Os, and the wrapper cells required for the TAP interface are added to the uppermost layer. Therefore, $(N-1)$ test layers are added to an M3D IC with N layers. We update the cost-per-die expression to account for the cost of performing $(N-1)$ additional wafer-bonding and via-etch steps. We also consider the impact of these steps on die yield. The updated cost-per-die expression of an M3D IC with the proposed DfT solution is shown in (3.20).

$$C_{die}^{P} = \frac{\sum_{i=1}^{N} C_{layer} + 2 \times (N-1) \times C_{bond} + 2 \times (N-1) \times C_{via-layer}}{N_{die}^{`} \times \prod_{i=1}^{2N-1} Y_i^{`} \times (Y_{bond+via})^{2N-2}} \quad (3.20)$$

We perform the cost comparison analysis for three benchmarks (two IWLS and one custom 128-bit multiplier). In [38], block-level partitioning of these benchmarks into two layers was performed, and the details are presented in Table 3.1. The partitioning tool assumes that both the layers have identical transistors and interconnects. The floor-planning algorithm considers inter-layer variation in M3D ICs and minimizes the weighted sum of wirelength, area, and longest path delay. The floor-planning algorithm does not minimize the number of ILVs since the ILVs are smaller in size compared to TSVs.

In order to carry out the cost comparison analysis, we obtain the values for the parameters in the cost model from [68]. These values were obtained for a 65 nm CMOS technology with 11 metals layers and wafer diameter of 300 mm (Table 5.1). We estimate the cost of smart-cut process by comparing the cost of a raw SOI wafer with the cost of a conventional wafer. Smart-cut, which has been used to manufacture SOI wafers for the past 15 years, is the most popular method to form crystalline-Si layers for monolithic 3D-ICs [4]. To manufacture an SOI wafer using smart-cut, a thin silicon layer is created over a blank silicon wafer instead of a processed one with transistors and wires as in the case of M3D integration. Therefore, we can get an estimate for the cost of smart-cut process by comparing the cost of an SOI wafer with that of a conventional wafer. Since the data in Table 5.1 was obtained for

Table 3.1: Block-level partitioned M3D benchmarks.

Benchmark	Number of I/Os	Number of ILVs		Number of Gates	Footprint (mm^2)
b19	76	13460	2D	82300	0.398
			M3D	80620	0.204
Des3	303	3750	2D	68900	0.328
			M3D	66200	0.156
Mul128	512	7261	2D	251000	1.096
			M3D	245000	0.550

Table 3.2: Parameters in the cost model [2].

Parameter	Value
Cost of raw CMOS wafer (A)	\$300
Cost of raw SOI wafer (B)	\$300
Cost of smart-cut (B-A)	\$1700
FEOL + BEOL process cost (C)	\$12600
Wafer diameter	300 mm
Process cost for via layer	\$160

an older process technology, the cost estimated for smart-cut is higher compared to the current cost of smart-cut, estimated to be less than \$60 for a wafer [69]. We also carried out the cost comparison analysis using data obtained for an advanced foundry process[1].

Fig. 3.5 and Fig. 3.6 present the cost-per-die of the three M3D benchmarks described in Table 3.1 as function of the ILV yield. For b19, we observe that the savings in cost-per-die is as high as 40% for the proposed DfT solution compared to the P1838-based DfT solution. On the other hand, for Mul128, we observe up to 11% increase in the cost for ILV yield less than 0.99. For ILV yield greater than 0.99, we observe that the cost per die for the proposed DfT solution is lower. The slight increase in cost for Mul128 for ILV yield less than 0.99 is due to lower ILV density. We explain the impact of ILV density on cost-per-die of an M3D IC towards the end of this section.

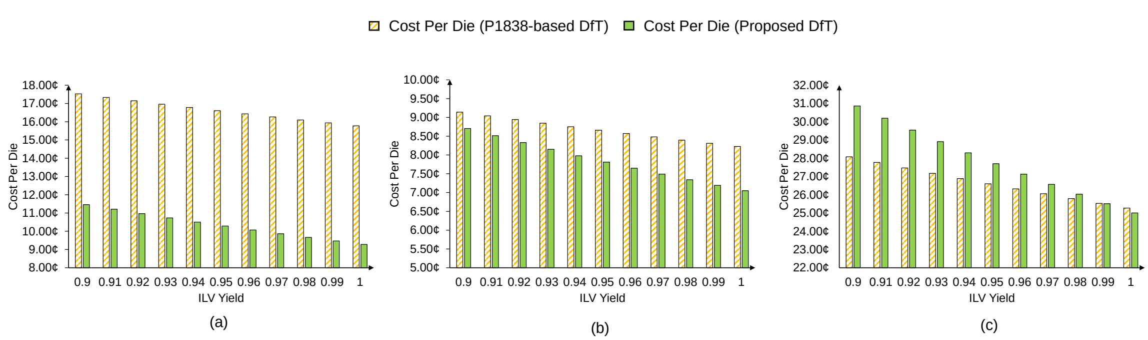


FIGURE 3.5: Cost per die of an M3D IC obtained using data from a 65 nm CMOS process for several ILV process yield values for: (a) b19, (b) Des3, and (c) Mul128.

[1] Process data withheld due to confidentiality reasons.

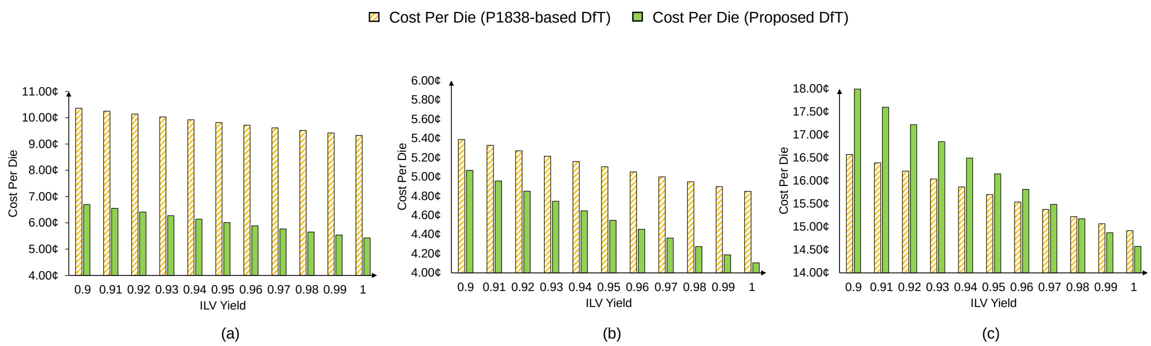


FIGURE 3.6: Cost per die of an M3D IC using data for an advanced foundry process for several ILV process yield values for: (a) b19, (b) Des3, and (c) Mul128.

Since the footprint of a typical die ranges between 500 mm^2 and 1000 mm^2 [14], we also analyze the test cost of three handcrafted benchmarks. Each of these three benchmarks are created by placing thousand dies of b19, Des3, and Mul128, respectively (Table 3.3). We assume that those thousand dies do not interact. Therefore, the footprint of each of the three handcrafted benchmarks is thousand times the footprint of the benchmark that was used to create them.

Fig. 3.7 and Fig. 3.8 present the cost-per-die of the three handcrafted benchmarks as function of the ILV yield. For b19_H, we still observe that the savings in cost-per-die is as high as 40% for the proposed DfT solution compared to the P1838-based DfT solution. On the other hand, for Mul128_H, we observe up to 11.11% increase in the cost for ILV yield less than 0.99. For ILV yield greater than 0.99, we observe that the cost per die for the proposed DfT solution is lower.

Table 3.3: Handcrafted M3D benchmarks.

Benchmark	Number of I/Os	Number of ILVs ($\times 10^3$)		Number of Gates ($\times 10^4$)	Footprint (mm^2)
b19_H	4094	13460	2D	8230	398
			M3D	8062	204
Des3_H	4094	3750	2D	6890	328
			M3D	6620	156
Mul128_H	4094	7261	2D	25100	1096
			M3D	24500	550

We next evaluate the ratio of the cost for the P1838-based solution to that for the proposed solution for a two-layer M3D IC in order to understand the impact of ILV density, ILV process yield, and defect density on the cost-per-die. The cost ratio can be obtained by dividing (3.19) by (3.20):

$$CostRatio = \frac{1}{1 + \frac{C_{bonding} + N^{i}_{vias} \times C_{via}}{2 \times (C_{layer} + C_{bonding} + N^{i}_{vias} \cdot C_{via})}} \times \frac{N_{die}}{N'_{die}} \times \frac{(Y_{die})^2}{(Y'_{die})^2} \times \frac{1}{Y_{bonding+via}} \tag{3.21}$$

In (3.21), the terms $\frac{N_{die}}{N'_{die}}$ and $\frac{(Y_{die})^2}{(Y'_{die})^2}$ depend on the ratio of the die area for the P1838-based DfT solution to that for the proposed solution. The ratio of the die area for the P1838-based DfT solution to that for the proposed solution is:

$$\frac{A_{die}(P1838)}{A_{die}(Proposed)} = \frac{G + N_{vias} \times 3 + N_{ios} \times 4}{G + N_{ios} \times 4} \tag{3.22}$$

where G and N_{vias} are the gate count and via count, respectively, for the second layer, and N_{ios} is the I/O count. Since the number of gates per layer is proportional to the area of the die, the ratio of $A_{die}(P1838)$ to $A_{die}(Proposed)$ is proportional to number of ILVs per unit area. We refer to the number of ILVs per unit area as the ILV density.

From (3.21) and (3.22), we can conclude that the savings in cost obtained by implementing the proposed DfT solution depends on: (i) cost of wafer bonding and

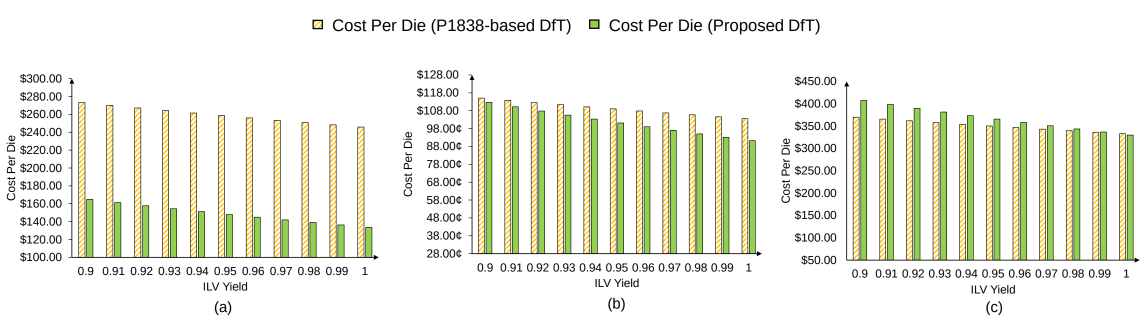


FIGURE 3.7: Cost per die of an M3D IC using data from a 65 nm CMOS process for several ILV process yield values for: (a) b19_H, (b) Des3_H, and (c) Mul128_H.

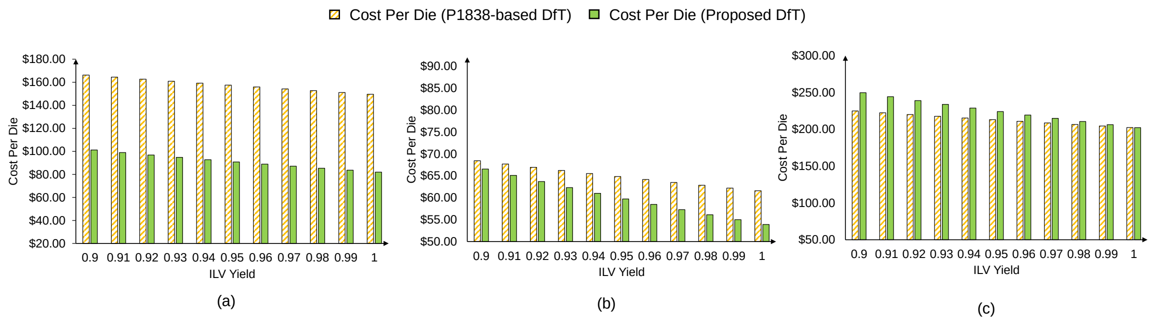


FIGURE 3.8: Cost per die of an M3D IC using data for an advanced foundry process for several ILV process yield values for: (a) b19_H, (b) Des3_H, and (c) Mul128_H.

via etching compared to total cost of processing a layer, (ii) ILV density, (iii) defect density, and (iv) ILV process yield. Since defect density and ILV yield are process dependent, and the cost of wafer bonding and via etching is expected to be low compared to the total cost of processing a single M3D layer [69], we evaluate the cost ratio for Mul128_H by varying ILV density, which is the only design parameter.

Fig. 3.9(a) presents the cost ratio for Mul128_H obtained by varying the ILV

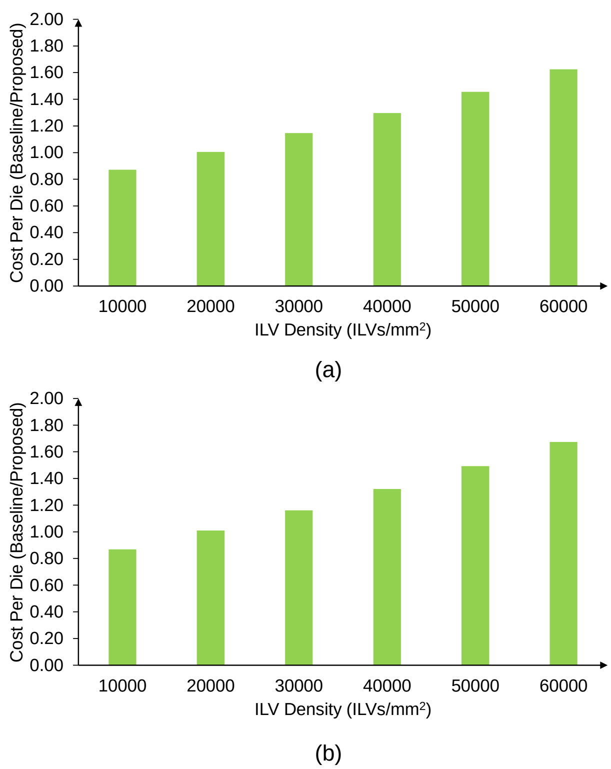


FIGURE 3.9: Cost ratio obtained for Mul128_H as function of ILV density for: (a) an advanced foundry process and (b) 65 nm CMOS process.

density using data for the foundry process. To obtain these results, we assume a defect density of 250 defects/m^2 and an ILV process yield of 0.99 [68]. We observe that the cost per die of the proposed method is lower compared to that of the P1838-based method for all values of ILV density greater than 20,000 ILVs/mm^2. Fig. 3.9(b) present the cost ratio for Mul128_H obtained using data for the 65 nm CMOS process by varying the ILV density. We make the same assumptions as we made to carry out the analysis using the 65 nm CMOS process, and the results are similar to that obtained for the foundry process.

3.4.2 Overhead Analysis

We now analyze the timing overhead introduced by the proposed DfT solution. First, we analyze the overhead introduced by a test layer when it is programmed to work in the serial mode. The serial mode of a test layer requires two IR-DR cycles: (i) to set the TAP in the appropriate configuration, and (ii) load the test data into the appropriate TDR. Since control data is loaded in the first IR-DR cycle, and test data is loaded in the second IR-DR cycle, the timing overhead introduced in the serial mode is one IR-DR cycle. An IR cycle consists of the following instruction: (i) Select˙IR (1 cycle), (ii) Capture˙IR (1 cycle), (iii) Shift˙IR (16 cycles), (iv) Exit˙IR (1 cycle), and (v) Update˙IR (1 cycle). The controller stays in the Shift˙IR state for 16 cycles since we assume a 16-bit instruction, which is sufficient to cover all possible modes. Therefore, total number of clock cycles for one IR cycle is 20. We can similarly compute the total number of clock cycles for one DR cycle to load the configuration register in a test layer. This number depends on the size of the configuration register. If we assume a 16-bit configuration register, which is sufficient to cover all the configurations of a test layer, the total number of clock cycles for one DR cycle is also 20. Therefore, we need 40 clock cycles to load the control data in the serial mode.

In the serial mode, the TDR is typically one long scan chain, which is obtained by daisy-chaining all the internal scan chains of embedded cores. Therefore, such a TDR can contain thousands of scan cells. For example, consider an M3D IC with two layers crafted using IWLS'05 benchmarks. Let us assume that one layer consists of leon3mp and the other layer consists of netcard, which have 80,238 and 97,797 scan cells, respectively. Therefore, the total number of clock cycles required to load a scan chain that connects both the layers is 178,035. Therefore, the timing overhead to apply the control data is only 0.022%, which is negligible.

Next, we analyze the timing overhead introduced by a test layer when it is programmed to work in the parallel mode. Before the parallel interface of a test layer is used, the serial interface is used to program the PP configuration register. Therefore, we use one IR-DR cycle to program the PP configuration register. Therefore, we need 40 cycles to configure the parallel interface. Let us again consider the above example M3D IC. The length of the longest internal scan chain in leon3mp and netcard are 5,027 and 4,823, respectively. The number of test patterns for leon3mp and netcard are 29,017 and 47,495, respectively. If we assume that the benchmarks are tested concurrently, the time for testing both the benchmarks is determined by the benchmark with the largest test cube, which is product of number of test patterns and longest internal scan chain. The total number of clock cycles to apply the test set is 229,068,385. Therefore, the timing overhead is 0.00002%, which is negligible.

3.4.3 Test Scheduling Results

In order to evaluate the ILP model proposed test scheduling, we consider two two-layer M3D designs. The first design (Design A) consists of four IWLS'05 benchmarks, vga˙lcd and wb˙conmax on the top layer, and des˙perf and ethernet on the bottom layer. The second design (Design B) consists of six IWLS'05 benchmarks, vga˙lcd, wb˙conmax, and pci˙bridge32 on the top layer, and des˙perf, Ethernet, and DMA on the bottom layer. Table 3.4 presents the details about each of the above benchmarks. We assume that each benchmark has 20 scan chains. These benchmarks were synthesized using the Nangate 45 nm technology library and Cadence RTL compiler.

For each benchmark, we generated multiple scan configurations corresponding to different numbers of TAM width using the Combine algorithm [70]. We used Synopsys TetraMAX to generate test sets for these configurations. The power consumption for each benchmark was estimated using Cadence RTL compiler. We solved the test-scheduling problem under constraints on maximum power consumption and TAM width. The ILP solver Xpress-MP was used to solve the ILP problem and obtain optimal test schedules.

There are two user-defined parameters in the ILP formulation, namely maximum power budget P_{max} and maximum TAM width $2W_{max}$. Both these parameters impact test scheduling. We therefore study the impact of each of these parameters on the generated test schedules. Fig. 3.10 presents the test time as a function of W_{max} and P_{max} for two M3D designs. For both these designs, we notice that the test time decreases as we increase the maximum TAM width. However, RPCT is desirable at wafer sort to reduce yield loss arising from contact problems. Therefore, the trade-off between test time and TAM width should be carefully explored.

We next study the relationship between test time and P_{max}. A higher power budget indicates that more cores can be test concurrently in each test session. However, a higher power budget does not directly translate to reduction in test time since we

Table 3.4: Details of IWLS'05 benchmarks.

Benchmark	Number of Patterns	Length of Longest Scan Chain	Fault Coverage
wb_conmax	739	39	99.98%
pci_bridge32	1,078	165	99.99%
vga_lcd	5,347	890	100%
ethernet	3,549	557	99.96%
des_perf	318	441	100%
DMA	21,451	86	100%

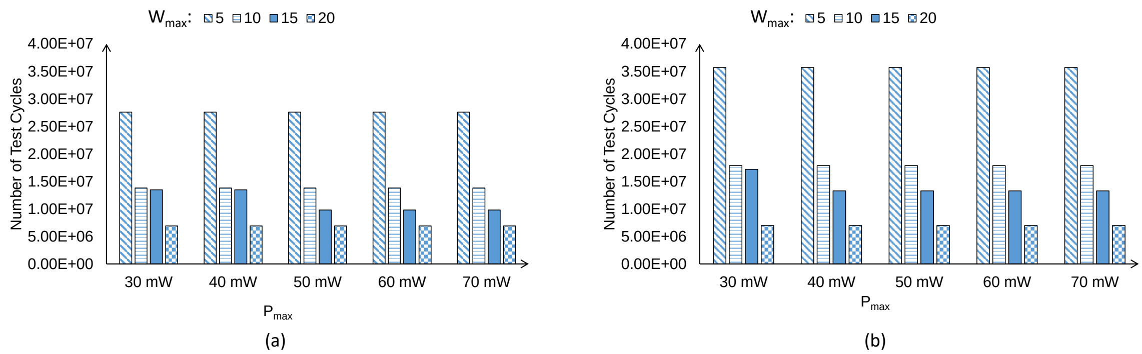


FIGURE 3.10: Test time as a function of TAM width and maximum power budget for: (a) Design A and (b) Design B.

assume that the TAM width is shared by all the cores. Test time only decreases when the TAM width is non-pareto optimal for one or more cores [71]. Note that high power dissipation during testing will increase chip temperature, which can lead to reduced reliability or yield loss. Hence, the trade-off between test time and power budget should be carefully explored. We observe that the test time decreases as we increase the power budget when W_{max} is higher than 15 since more cores can be tested concurrently as we increase the power budget. Therefore, P_{max} is an important factor for wafer-level testing of M3D ICs.

3.5 Conclusions

We have presented a DfT solution for M3D ICs that enables testing of a partial assembly as well as a full assembly with all layers. The proposed solution consists of a set of cooperating test layers, one for each interface between two functional layers. We described the operating modes and configurations supported by the test layers. We also derived a cost model accounting for the additional steps involved in the fabrication of an M3D IC. We then compared the cost of the proposed DfT solution with the cost of a potential test solution based on P1838. Our results have shown that the proposed solution is more cost-efficient than the P1838-based solution for a wide range of ILV density. We have also presented a test scheduling and optimization technique for wafer-level testing of M3D ICs.

4

Inter-Layer Interconnect BIST and Diagnosis Solution

In this chapter, we propose a post-assembly ILV built-in self-test (BIST) and diagnosis solution [72]. The proposed solution assumes a DfT architecture based on dedicated test layers. In the proposed ILV BIST solution, interface-register cells in a test layer are stitched into a twisted-ring counter (TRC) using their functional outputs and the ILVs. The interface-register cells are connected to form a long shift register, and the complement of the output of the last interface-register cell is connected to the input of the first interface-register cell.

A test is launched from one interface-register cell and captured on another interface-register cell. A metal wire is used to connect those two interface-register cells through a multiplexer. The multiplexer is inserted to switch between functional and ILV-test modes. Therefore, it is important to minimize the length of wires used to connect interface-register cells so that the impact on functional performance is minimized. The problem of minimizing the wire length can be mapped to the problem of finding a minimum-cost Hamiltonian circuit in a complete bipartite graph. The weighted

Hamiltonian circuit problem is NP-Complete on complete bipartite graphs. Therefore, we propose a heuristic technique for solving this problem efficiently.

Diagnosis of ILV BIST fails is essential to analyze yield-limiting systematic defects and design issues. If an M3D IC fails ILV BIST, we need to find the defect type (open or short), location, and size. The defect type and location can be diagnosed based on the interface cells that captured an incorrect value using an effect-cause analysis [73]. However, diagnosing the defect size is not straightforward. We therefore propose a framework to diagnose the size of a resistive open or a resistive short ILV defect. The proposed diagnosis framework is based on an artificial neural network (ANN) and developed at design time using data obtained by simulating the ILV BIST. We use features such as clock frequency during test application and the Hamming distance between the defective and defect-free states of the ILV BIST to train the ANN.

We show that the proposed solution detects all hard opens and shorts in the ILVs. We validated the detection of all hard opens and shorts using HSpice simulations. The worst-case time complexity of the proposed wire-length minimization heuristic is $O(n^3)$, where n is the number of interface-register cells. We also highlight the effectiveness of the proposed heuristic by comparing the total wire length required to form a TRC using the proposed method with a baseline that connects the interface-register cells in a greedy manner. The total wire length required to form a TRC using the baseline is 15% to 22% higher compared to the proposed method. We also evaluated the proposed diagnosis framework using simulation data for four IWLS'05 benchmarks. Our results show that the proposed framework can estimate the size of defect in an ILV with extremely high accuracy (exceeds 0.93 where 1.0 indicates perfect prediction).

The rest of the chapter is organized as follows. Section 4.1 describes the proposed BIST solution. Section 4.2 describes the proposed wirelength-minimization heuristic. Section 4.3 presents the proposed diagnosis solution. Section 4.4 presents the simulation results. Finally, Section 4.5 concludes the chapter.

4.1 Proposed ILV BIST

Although ILVs are comparable in size and pitch to conventional vias, they differ significantly in terms of fabrication steps and yield. The yield for processing ILVs is speculated to be much lower compared to a conventional via. Yield learning is especially important for a new technology such as M3D integration. Due to aggressive scaling of ILD thickness, the ILVs are prone to shorts, opens, and delay defects [46, 47]. Incomplete metal filling or voids in an ILV increase resistance and impact circuit timing [47, 54]. Large voids in the ILV can result in an open path, thereby causing a catastrophic failure. Impurities may increase resistance and interconnect delay. ILVs are also affected by voids that arise in the ILD since they are processed after the wafer-bonding step. Such voids lead to an increase in ILV resistance. If the size of a void in the ILD is large enough to impact two neighboring ILVs, then it leads to a resistive short between the two ILVs. The resistance of the defect depends on the thickness of the void.

Conventional scan-based test pattern generation cannot be re-used to test ILVs since they do not provide the diagnostic resolution to diagnose and characterize ILV faults. We may not be able to pinpoint to a specific ILV using conventional scan-based diagnosis since it may produce large number of candidates for one failure. Pre-bond TSV testing methods such as [74] is infeasible for ILVs since bare ILVs cannot be be exposed, and the state-of-the-art wafer-probe technology cannot support the pitch requirement for ILVs (100 nm to 200 nm). Techniques proposed to carry out post-bond TSV testing can be extended to M3D for post-assembly testing. For example, the proposed IEEE Std. P1838 [40], which defines a standardized architecture to test the TSVs, can be extended to M3D and used to test the ILVs. However, a wrapper cell has to be inserted at both ends of a TSV in order to test that TSV using P1838.

A similar memory element is needed at both ends of an ILV in order to test an ILV by extending P1838 to M3D ICs. However, the overhead due to the insertion of memory elements on all ILVs can be significant since the number of ILVs in M3D ICs is expected to be an order of magnitude higher compared to TSV-based 3D ICs.

At-speed transition-based delay-fault testing facilitates the detection of hard opens/shorts as well as resistive defects. A new test solution is required to enable at-speed testing of ILVs. Test solutions for TSV-based 3D ICs cannot be extended to M3D because of the following reasons. First, inserting a boundary register in every layer to test the ILVs as stand-alone objects is not efficient for M3D since the number of ILVs in M3D ICs is expected to be an order of magnitude higher. Second, testing an ILV as a part of shore logic (e.g., in [75]) by launching a transistion along an inter-layer path is not desirable since it requires significant changes to test-generation flows. The size of the Extest chain will be prohibitively large since scan flip-flops from the functional layers will be part of that chain, in addition to interface-register cells from the dedicated test layers. Moreover, test pattern generation for the ILVs may become harder since a higher number of care bits are required to justify a value to an ILV; these additional constraints will make it more difficult to obtain high fault coverage, and also result in an increase in the number of test patterns.

We propose an ILV BIST solution that enables stand-alone testing of the ILVs without incurring significant area overhead. In the proposed solution, interface-register cells are stitched into a TRC using their functional outputs and ILVs. Fig. 4.1 shows a four-bit TRC formed using interface-register cells. ILVs connected to the input and output of an interface-register cell are defined as the inbound and outbound ILVs, respectively. In a TRC, the inverse of the output of the last cell is connected to the input of the first cell. Such a counter circulates a stream of ones followed by zeros around the ring. For example, in a counter with an initial register value of 0000, the repeating pattern is: 0000, 1000, 1100, 1110, 1111, 0111, 0011, 0001, and 0000.

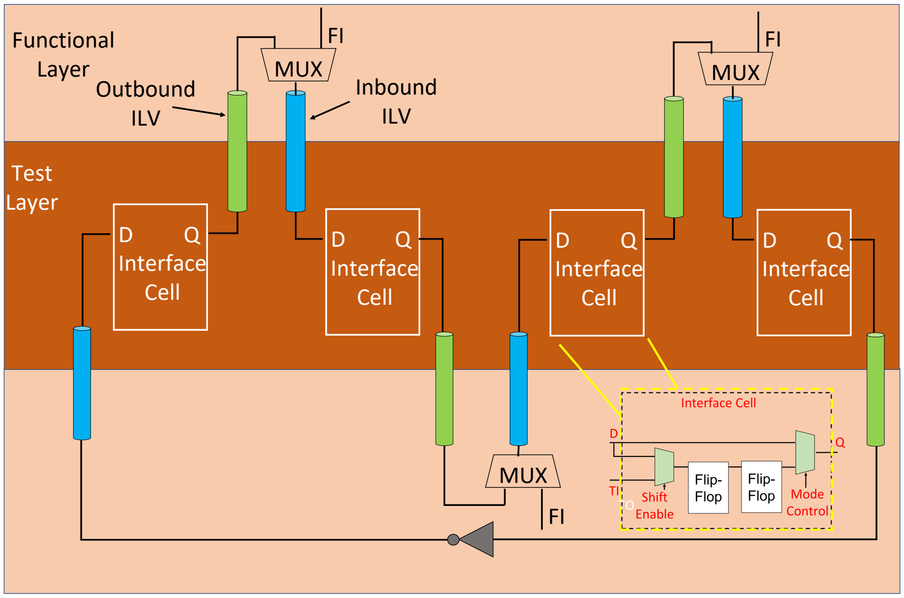


FIGURE 4.1: Illustration of a four-bit TRC using interface cells.

Once a TRC is initialized to a specific pattern, the pattern in the counter repeats every $2N$ clock cycles, where N is the number of cells in the counter. We show later that we can detect all opens and shorts in the ILVs by using these patterns. We also evaluate the size of the delay defect that can be detected by using a TRC.

In order to implement a TRC using interface-register cells, a multiplexer is inserted along with an outbound ILV to switch between functional and interconnect-test modes. The control signal for the multiplexers is generated in the dedicated test layer and fed to the functional layers. Therefore, two additional ILVs per test layer are required to enable ILV BIST. Since a test is launched from one interface-register cell and captured on another interface-register cell, interface-register cells have to equipped with at-speed launch and capture capability.

For a TRC, we define two types of operations–shift and functional. A shift operation is defined as a sequence of shifts controlled by the clock until the contents of the TRC are read out from the TDO pin. In a functional cycle, the counter is placed functional mode and clocked for one cycle. The TRC is initialized to a known value by shifting in data through the interface register. The counter is clocked at functional frequency for some clock cycles and the pattern at the end of functional cycles is shifted out through the interface register. In order to minimize the impact of test circuitry on performance, it is essential to minimize the length of wires used to connect interface-register cells into a TRC. This reduces capacitive-loading effects on the nets connected to the ILVs. Subsequently, we refer to the problem of minimizing the total length of the wires used to form the TRC as the wire-length minimization problem. We next map the wire-length minimization problem to finding a minimum cost Hamiltonian circuit in a complete bipartite graph.

In an M3D IC, the number of inbound ILVs (to a layer) is typically not equal to the number of outbound ILVs (to a layer). However, we need to match an inbound ILV with an outbound ILV to stitch the interface cells into a TRC. Therefore, we cluster the inbound (outbound) ILVs when the number of inbound (outbound) ILVs is greater than outbound (inbound) ILVs such that the total number of clusters matches the outbound (inbound) ILV count. Fig. 4.2 presents the hardware inserted to cluster three outbound ILVs. In the layer from which the signals originate, we add inverters to invert the input from a driver interface cell. We add inverters in such a way that all ILVs adjacent to a given ILV receive inverted value of that being fed into the given ILV. We feed complementary values into adjacent ILVs to detect shorts between ILVs in the same cluster. Section IV describes the method used to determine which interface cells receive inverted input from a driver interface cell. In the layer which forms the destination for the signals, we insert an N-input AND gate, where N corresponds to the number of ILVs in the cluster. We use an AND gate to aggregate the values from all the interface cells in the cluster. We feed this aggregated value to the load interface cell. Note that signals corresponding to ILVs which were fed an inverted input from a driver interface cell are inverted back to their original value before the AND gate.

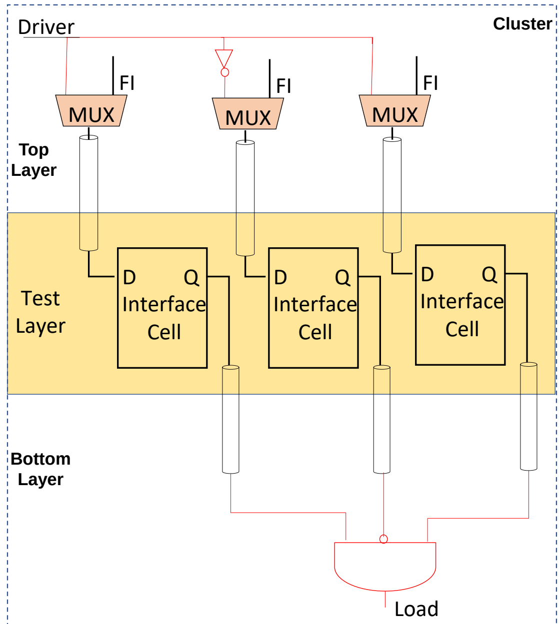


FIGURE 4.2: Hardware inserted to cluster outbound ILVs to the top layer.

4.1.1 Problem Description

In order to implement a TRC using interface-register cells, we need to connect an outbound ILV to an inbound ILV. If the ILVs are treated as nodes in a graph and an edge between two nodes is used to denote a valid connection between two ILVs (inbound ILV and outbound ILV), then the resultant graph is bipartite. A *bipartite graph* is a graph whose vertices can be divided into two disjoint and independent sets such that every edge connects a vertex in one set to one in another. Fig. 4.3 shows an example of a complete bipartite graph. The Manhattan distance between an inbound ILV and an outbound ILV is used to represent the weight of an edge between the corresponding ILVs. The wire-length minimization problem then maps to finding a minimum cost Hamiltonian circuit in the weighted bipartite graph. The wire-length minimization problem is different from scan-chain routing problem since we can only visit an input (output) ILV from an output (input) ILV. There is no such constraint for the scan-chain routing problem, which has traditionally been mapped to the traveling-salesman problem [76]. Next, we show that the weighted

Hamiltonian circuit problem is NP-Complete for complete bipartite graphs.

Claim: The weighted Hamiltonian circuit problem is NP-Complete for complete bipartite graphs.

Proof: To prove that the problem is NP-Complete, we need to prove that the problem is in NP and it is NP-Hard. To prove that the problem is in NP, we consider the decision version of this problem, which can be defined as: Given a graph $G = (V, E)$ and a number k, does there exist an ordering of vertices S such that sum of the edge weights between every edge in S is less than or equal to k.

Let S be a sequence of vertices $v_1, v_2, v_3, \cdots, v_n$. To prove that the problem is in NP, we have to check three things: (i) every edge between two adjacent vertices v_i, v_j in the sequence S is an edge in G (trivially satisfied in this case since G is complete), (ii) the sum of these edge weights is less than or equal to k, and (iii) every vertex in

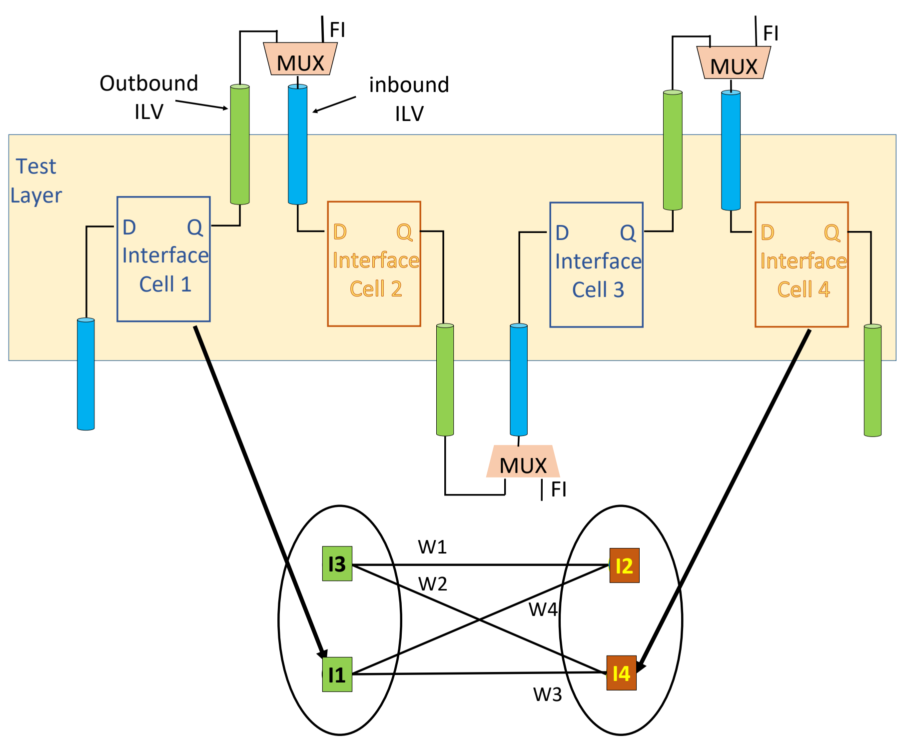


FIGURE 4.3: Illustration of mapping a set of input and outbound ILVs to a complete bipartite graph.

G is in the sequence S. Since all these three checks can be performed in polynomial time, the problem is in NP.

To prove that the problem is NP-Hard, we use reduction from a general instance of a known NP-Hard problem to a specific instance of the problem at hand. We start with a directed graph $G(V, E)$. We first reduce the problem of checking for a Hamiltonian circuit in a directed graph (known to be NP-Hard) to that of checking for a Hamiltonian circuit in a bipartite graph. We construct an undirected graph G^* from the directed graph as follows:

- Replace every vertex v_i in V by four vertices: $v_i^0, v_i^1, v_i^2, v_i^3$
- Connect the following pair of vertices with an edge: (i) (v_i^0, v_i^1), (ii) (v_i^1, v_i^2), (iii) (v_i^2, v_i^3), and (iv) (v_i^3, v_j^0) for all (i, j) such that (v_i, v_j) is an edge in G.

We can see that G^* is bipartite with the following two vertex-set partitions: $V_1 = \{\{v_i^0\}|v_i \in V \cup \{v_i^2\}|v_i \in V\}$ and $V_2 = \{\{v_i^1\}|v_i \in V \cup \{v_i^3\}|v_i \in V\}$, where $V_1 \cap V_2 = \phi$ and $V_1 \cup V_2 = V$.

It can be seen that G contains a Hamiltonian circuit iff G^* contains a Hamiltonian circuit. This statement can be proved by contradiction. Assume that G^* does not contain a Hamiltonian circuit and G contains a Hamiltonian circuit. Let the ordered sequence of vertices that form the Hamiltonian circuit in G be $\{v_1, v_2, \cdots, v_n\}$. Because of the way G^* is constructed from G, we can see that the ordered sequence of vertices in G^* $\{v_1^0, v_1^1, v_1^2, v_1^3, v_2^0, v_2^1, v_2^2, v_2^3, \cdots, v_n^0, v_n^1, v_n^2, v_n^3\}$ is a Hamiltonian circuit. We have therefore arrived at a contradiction.

We next reduce the problem of checking for a Hamiltonian circuit in a bipartite graph to finding a Hamiltonian circuit of weight n in a weighted complete bipartite graph. Take the bipartite graph $G^* = (V^*, E^*)$ and construct a weighted complete bipartite graph G^{**} by setting the weights of all edges in E^* to 1 and the weight of any edge not in E^* to 2. We use the method of contradiction to show that G^* contains a Hamiltonian circuit iff the weighted complete bipartite graph contains a Hamiltonian circuit of weight n, where n is the number of vertices in G^{**}. Assume that G^* contains a Hamiltonian circuit but G^{**} does not contain a Hamiltonian circuit of weight n. Let the ordered sequence of vertices that form the Hamiltonian circuit in G^* be $\{v_1, v_2, \cdots, v_n\}$. Because of the way G^{**} is constructed from G^*, we can see that the ordered sequence of vertices in G^{**}, $\{v_1, v_2, \cdots, v_n\}$, is a Hamiltonian circuit. Since the weight of an edge between any two vertices in the ordered sequence is 1, the total cost of the weight of the Hamiltonian circuit is n. Therefore, G^{**} contains a Hamiltonian circuit of weight n. This is a contradiction. Therefore, finding a Hamiltonian circuit of weight n in a weighted complete bipartite graph is NP-hard. Since we proved that the problem is in NP and it is NP-Hard, we conclude that it is NP-Complete. $\square$

4.1.2 Detectability of Opens and Shorts

A short between two wires is commonly referred to as a bridging fault. It can be detected by driving opposite values on the two interconnects that make up the bridge. Let us consider an n-bit TRC. We assume that the TRC is initialized to $b_1, b_2, \cdots, b_n$. We know that an n-bit TRC has a period of $2n$ functional cycles, i.e., the pattern repeats every $2n$ functional cycles. We can stop the functional clock after a predetermined number of cycles and use shift operations to read out the contents of the TRC. The shifted-out data can then be compared with the expected response in order to detect a fault.

We first show using an example that not all shorts can be detected if a shift operation is performed only once after $2n$ functional cycles. Assume that a short arises between the ILV connected to the input of the first interface cell, and the ILV connecting the n^{th} interface cell and the inverter. This fault shorts the inverter, i.e., the value from the n^{th} interface cell is directly fed to the first cell instead of being inverted. Since the inverter is shorted because of the fault, the value from the n^{th} interface is fed to the first interface in each functional cycle without being inverted. The period of the counter becomes n, i.e., the pattern repeats after every n cycles. Therefore, the pattern after $2n$ functional cycles is the same as the initial pattern. As a result, the fault is not detected if a shift operation is carried out after $2n$ cycles.

The following theorem shows that all shorts are detected if the shift operation is performed twice in a specific manner.

Theorem 1: The sequence of patterns generated by the TRC detects all shorts between ILVs if the contents of the TRC are shifted-out after n and $2n$ cycles.

Proof: The proof is by contradiction. Assume that a bridging fault arises between the i^{th} interface-register cell and j^{th} interface-register cell (assume i ¡ j without loss of generality) is not detected by the TRC. The fault can arise between the ILVs connecting the i^{th} and $(i+1)^{th}$ interface cell, and the ILVs connecting the j^{th} and the $(j+1)^{th}$ interface cell. The ILVs connecting the i^{th} and $(i+1)^{th}$ interface cell are the outbound ILV of the i^{th} interface cell and inbound ILV of the $(i+1)^{th}$ interface cell. The ILVs connecting the j^{th} and the $(j+1)^{th}$ interface cell are the outbound ILV of the j^{th} interface cell and inbound ILV of the $(j+1)^{th}$ interface cell. Therefore, a short can arise between any combination of these ILVs.

We assume that the TRC is initialized to $b_1, b_2, \cdots, b_n$. For a bridging fault to go undetected, the two shifted-out patterns should be same as their expected value. Therefore, the patterns at the end of n and $2n$ functional cycles should be $b'_1, b'_2, \cdots, b'_n$ and $b_1, b_2, \cdots, b_n$, respectively. Here, b'_i denotes the inver of b_i. However, because of a bridging fault between the i^{th} interface-register cell and j^{th} interface-register cell, the values at the i^{th} interface-register cell and j^{th} interface-register cell will change to: b_i & b_j (wired-AND) or b_i — b_j (wired-OR). The state of the counter gets corrupted if the i^{th} interface-register cell and j^{th} interface-register cell are driven by opposite values. The proof for this statement is by contradiction. Let us assume that the i^{th} interface-register cell and j^{th} interface-register cell are never driven by opposite values. Therefore, i^{th} interface-register cell and j^{th} interface-register cell have the same value for $2n$ functional cycles. This implies that there is no inversion in the TRC. This is a contradiction. So, the i^{th} interface-register cell and j^{th} interface-register cell will be driven opposite values for at least one functional cycle.

We proved that the state of the counter gets corrupted for at least one cycle. If this happens, the pattern at the end of n or $2n$ functional is different from expected. Since the pattern in the TRC repeats every $2n$ cycles, we can assume that the i^{th} interface-register cell and j^{th} interface-register cell are driven by opposite values (b_i and b_i) after the first clock pulse. So, these values will change to 0 (1) due to the wired-AND (wired-OR) bridging fault in this cycle. After n cycles, the i^{th} interface-register cell and j^{th} interface-register cell are driven by b_i and b_i, respectively. Therefore, the i^{th} interface-register cell and j^{th} interface-register cell will change to 0 (1) due to the wired-AND (wired-OR) short in this cycle. Hence, the pattern at the end of n is not as expected. This is a contradiction. Hence, the bridging fault is detected. Therefore, the sequence of patterns generated by the TRC detects all shorts between ILVs if the contents of the TRC are shifted-out after n and $2n$ cycles. □

An open tends to behave like a stuck-at fault when the ILV is used to connect gates, which is true for gate-level M3D. We next prove that the patterns generated by a TRC can detect all opens. First, we show using a counter-example that not all opens are detected if a shift operation is performed only once after $2n$ functional cycles. Assume that an open arises on the ILV connected to the input of the first interface cell. This fault disconnects the inverter from the first interface cell. Since the inverter is open because of the fault, the value from the n^{th} interface is not inverted and fed to the first interface in each functional cycle. The TRC is no longer periodic and the fault may not be detected if a shift operation is carried out after

$2n$ cycles.

The next theorem shows that all opens are detected if the shift operation is performed twice.

Theorem 2: The sequence of patterns generated by the TRC detects all opens on ILVs if the contents of the TRC are shifted-out after n and $2n$ cycles.

Proof: The proof is by contradiction. Assume that an open arises on the ILV connected to output of the i^{th} interface-register cell is not detected by the TRC. The wire connecting the i^{th} and $(i+1)^{th}$ interface cell contains the outbound ILV of the i^{th} interface cell and inbound ILV of the $(i+1)^{th}$ interface-register cell. So an open fault on these two ILVs is equivalent.

We assume that the TRC is initialized to $b_1, b_2, \cdots, b_n$. For an open to go undetected, the two shifted-out patterns should be the same as the expected state of the TRC after n and $2n$ functional cycles. Because of an open on the ILV connected to output of the i^{th} interface-register cell, the value on the $(i+1)^{th}$ interface cell will not change to: b_i after one functional cycle. We can show by contradiction that the state of the counter gets corrupted if the $(i+1)^{th}$ interface cell does not change to b_i. Let us assume that the state of the counter does not get corrupted if the $(i+1)^{th}$ interface-register cell does not change to b_i. If the $(i+1)^{th}$ interface-register cell does not change to b_i, the $(i+1)^{th}$ interface-register cell and the $i+2^{th}$ interface-register cell will have the same value after one functional cycle. This is a contradiction. So, the state of the counter indeed gets corrupted if the $(i+1)^{th}$ interface cell does not change to b_i.

After one functional cycle, the $(i+1)^{th}$ interface-register cell and the $(i+2)^{th}$ interface-register cell have the same value, b_{i+1}. Therefore, after n functional cycles, the entire TRC will have b_{i+1}. The TRC will have the same state after $2n$ cycles as well. Therefore, the shifted-out patterns will be different from their expected value unless all the cells in the TRC are initialized to the same value. In that case, only

FIGURE 4.4: Algorithm for connecting interface-register cells into a twisted-ring counter

Input: Coordinates of inbound ILVs $V = \{v_1, v_2, ..., v_n\}$ and Coordinates of outbound ILVs $U = \{u_1, u_2, ..., u_m\}$
Output: A minimum cost Hamiltonian Circuit C that traverses through all ILVs, and every pair of adjacent ILVs in C contains an inbound ILV and an outbound ILV

if $(n > m)$ **then**
  Construct a graph $G = (V, E)$: $v_i \in V$ represents inbound ILV i, and the distance between inbound ILV i and inbound ILV j denotes the weight of an edge between v_i and v_j
  Reduce $V = \{v_1, v_2, ..., v_n\}$ to $\{v'_1, v'_2, ..., v'_m\}$ using k-means clustering, where v'_i is the centroid of cluster i
else
  Construct a graph $G = (V, E)$: $v_i \in V$ represents outbound ILV i, and the distance between outbound ILV i and outbound ILV j denotes the weight of an edge between v_i and v_j
  Reduce $U = \{u_1, u_2, ..., u_m\}$ to $U = \{u'_1, u'_2, ..., u'_n\}$ using k-means clustering, where u'_i is the centroid of cluster i
end if
$D \leftarrow$ Distance matrix between V and U
Construct a complete-weighted bipartite graph $B = (V \cup U, E)$, where the edge weights are given by D
$T \leftarrow TSP(G)$, where T is a tour that traverses through all vertices in G
$M \leftarrow MWBM(B)$, where M is a perfect matching for B
$C \leftarrow T \cup M$
return P

the pattern after n functional cycles will be different. Nonetheless, the two shifted-out patterns will not be the same as the expected state of the TRC after n and $2n$ functional cycles if an open fault arises on the $(i+1)^{th}$ interface cell. This is a contradiction. Therefore, the sequence of patterns generated by the TRC detects all opens on ILVs if the contents of the TRC are shifted-out after n and $2n$ cycles. □

4.2 Wire-Length Minimization Heuristic

Since the wire-length minimization problem is NP-Complete, we propose an efficient heuristic to solve this problem. The proposed heuristic can be used to carry out wire-length optimization during the physical design phase of an M3D IC. Fig. 5.1 describes the proposed heuristic. First, we balance the number of inbound and outbound ILVs by carrying out k-means clustering on the set that has higher number of nodes. For each cluster, we identify ILVs which receive inverted value from the driver interface cell for this cluster using graph coloring. To construct a graph for a given cluster, we represent each ILV in that cluster by a node. We then insert an edge between two nodes if the distance between the corresponding ILVs is less than a given threshold. The threshold for ILV spacing is used to identify ILVs which are likely to have a short between them. If the spacing between ILVs is greater than the defined threshold, those ILVs are less likely to have a short between. In the graph-coloring step, we enforce the algorithm to use only two colors since the ILVs can only be of two types: (i) receive input directly from driver, and (ii) receive inverted value of input from driver. This constraint may lead to two adjacent ILVs having the same color and result in coverage loss. However, accepting the coverage loss is a more practical solution since re-balancing the clusters every time we obtain a graph coloring with more than two colors will make the run time of the heuristic prohibitively large.

Next, we represent each cluster by a representative node at the center of the cluster. We then construct a balanced-complete-weighted bipartite graph using the resulting set of input and outbound ILVs. Next, we find a Hamiltonian circuit in the balanced bipartite graph by using a well-known heuristic proposed in . In this heuristic, a minimum-cost matching consisting of edges that connect input and outbound ILVs is found using a heuristic for maximum-weighted bipartite matching (MWBM). A minimum-cost tour that visits all the inbound ILVs is also found using a heuristic for the traveling salesman problem (TSP). The resulting Hamiltonian circuit consists of visiting the inbound ILVs in the sequence specified by the tour and using the matching edges to make sure that an outbound ILV is visited after an inbound ILV.

The worst-case computational complexity of k-means clustering algorithm is $O(nki)$, where n is the number of vertices, k is the number of clusters, and i the number of iterations needed until convergence. The number of iterations until convergence is often small, and the results only improve slightly after the first 10 iterations. Therefore, the number of iterations until convergence is limited to 10 in most implementations of k-means clustering algorithm. As a result, we can consider i as a constant in our analysis. We then obtain the runtime of the k-means clustering algorithm as $O(n^2)$ as the number of clusters can be at most $n-1$. We use the DSATUR algorithm for graph coloring. The run time of DSATUR algorithm is $O(n^2)$ in the worst case. Therefore, the run time of the graph coloring step for all the cluster is $O(n^3)$ as the number of clusters can be at most $n-1$.

We use the Munkres algorithm [77] to find minimum cost matching for the balanced-complete-weighted bipartite graph. The run time of the Munkres algorithm is $O(n^3)$ in the worst case. We then use the Christofides algorithm to find a minimum cost tour that traverses through all the vertices in one set in the balanced bipartite graph. The worst-case run time of the Christofides algorithm is $O(n^3)$ [**?**]. Finally, we obtain the minimum-cost Hamiltonian circuit by visiting the ILVs in the sequence specified by the tour and using the matching edges to make sure that an outbound ILV is visited after an inbound ILV. The worst-case runtime of this step is $O(n)$. Therefore, the runtime of the proposed heuristic is $O(n^3)$ in the worst case.

4.3 Diagnosis of Defect Size

Conventional diagnosis methods such as effect-cause analysis [73] can be used to obtain defect type and location from the position of the interface cells in the TRC that captured an incorrect value. However, estimation of defect size cannot be carried out using conventional methods. We therefore propose an ANN-based framework to estimate the size of resistive opens and resistive shorts in ILVs. In order to estimate the actual defect size with reasonable accuracy, we need prohibitively large number of simulation data to train the ANN. Therefore, we propose to divide the range of defect sizes into multiple bins and use an ANN-based classifier to predict the defect bin in which the defect lies. We can achieve significant prediction accuracy for a

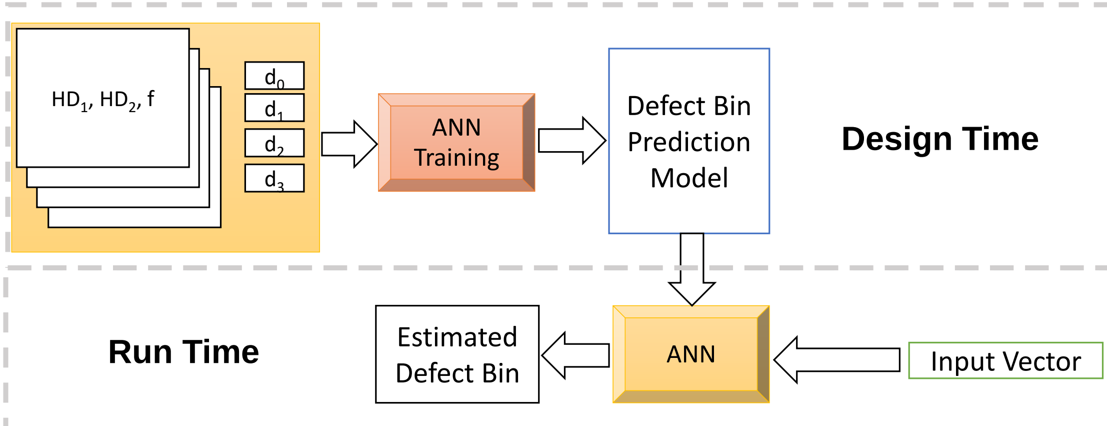


FIGURE 4.5: Training and implementation of defect-size estimation model.

defect classifier with a moderate-size data sample.

The proposed defect classifier is trained at design time using simulation data. Fig. 4.5 illustrates the training and implementation method of the proposed defect classifier. We use three features to train the defect classifier: (i) Hamming distance between the expected and captured values (HD_1) in the TRC before the first shift operation; (ii) Hamming distance between the expected and captured values (HD_2) in the TRC before the second shift operation; (iii) clock frequency at which the test was carried out. To generate training data, simulation can be carried for different clock-frequency values in range of clock frequencies specified for functional operation and defects of different sizes.

In this work, we use ANNs to estimate ILV defect size since they were recently shown to be effective for a related diagnosis problem [78]. We have utilized ANN because of its efficiency to model complex systems and high prediction accuracy for a wide range of applications. For training an ANN-based classifier, the input vectors are used to adjust the weight of the neurons as learning proceeds. Prediction for new input vectors are made using the trained neural network.

4.4 Simulation Results

In this section, we present simulation results to show the detectability of opens and shorts, evaluate size of minimum detectable resistive open and maximum detectable resistive short, discuss the effectiveness of the wirelength-minimization heuristic, and present defect size diagnosis results.

4.4.1 HSpice Simulation Results

We carry out HSpice simulations on a four-bit TRC designed using interface cells to show the detectability of opens and shorts, and evaluate size of minimum (maximum) detectable resistive open (short). The simulations were carried out using a 7 nm predictive technology model [79]. The schematic of a four-bit twisted-ring counter is shown in Fig. 4.1. The mode control signal in each interface-register cell is assigned a DC value such that the each cell operates in functional mode (FI is selected). The counter is initialized to the all-zero state. Such a TRC circulates a stream of ones followed by zeros around the ring (Fig. 4.6).

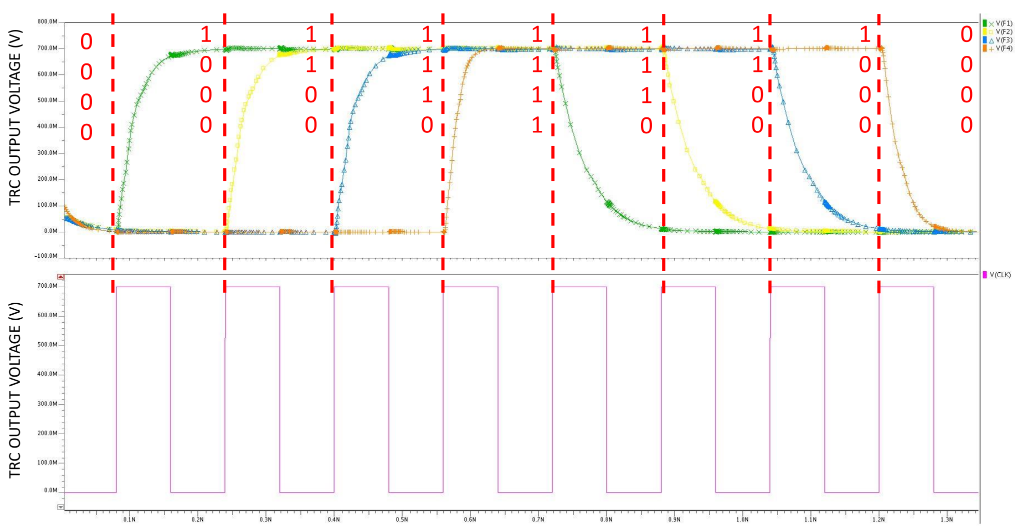


FIGURE 4.6: Simulation result for a fault-free four-bit TRC.

We next inject opens and shorts in the TRC. Fig. 4.7(a) presents simulation results for the TRC after an open is injected at the output of the first cell. We can see

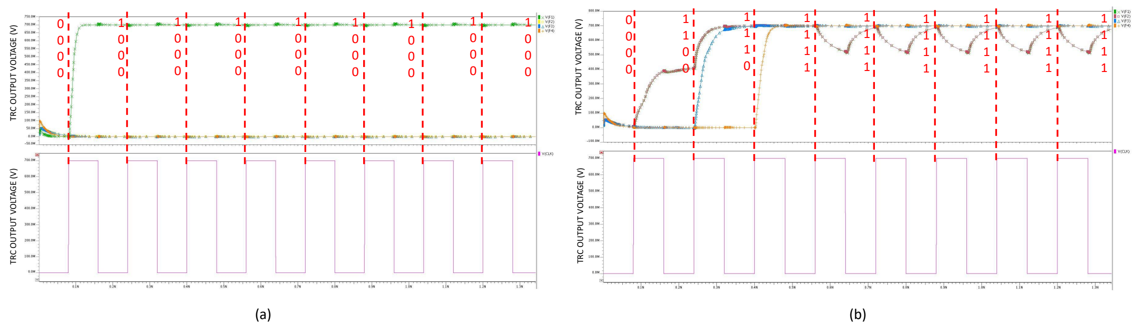


FIGURE 4.7: Simulation results after fault injection: (a) open on output of first cell, and (b) short between outputs of first and second cell.

that the pattern after four and eight functional cycles is 1000, which is different from the expected patterns (1111 and 0000). Therefore, this fault is detected. Fig. 4.7(b) presents the simulation results after a short is injected between the outputs of the first and second cell. The pattern after four and eight functional cycles is 1111, which is different from the expected patterns (1111 and 0000). Therefore, this fault is detected as well.

We also evaluate the size of a defect that can be detected using the proposed solution. We inject resistive opens and shorts at different locations in the TRC and evaluate the size of the smallest (for opens) and largest (for shorts) defect resistance that can be detected. The simulations were carried out for the four-bit TRC shown in Fig. 4.1. There are five likely fault locations in that TRC–four outputs of the interface-register cells (F1-F4) and output of the inverter (F5). An open defect can occur at any of these five locations and a short can arise between any combinations of these five locations. We carry out simulations for all these scenarios and the results are presented in Fig. 4.8. We observe that the size of the detectable defect depends on the resistance and capacitance at the fault location. For example, since the fan-out resistance and capacitance at the output of the fourth interface cell is that of an inverter, the size of the minimum detectable open at that location is much larger compared to other locations.

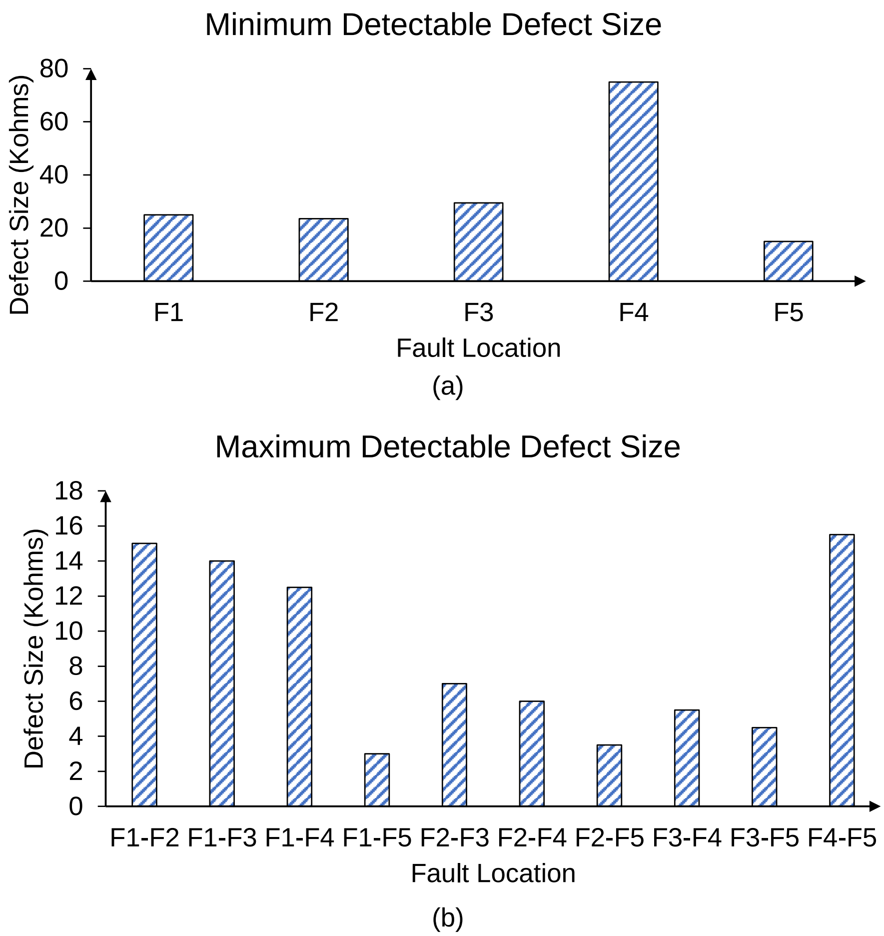


FIGURE 4.8: Detectable defect size in a four-bit TRC: (a) opens and (b) shorts.

4.4.2 Effectiveness of Wirelength-Minimization Heuristic

We utilize four IWLS'05 benchmarks to evaluate the effectiveness of the proposed algorithm. These benchmarks were implemented using the Arizona State Predictive PDK (ASAP) 7 nm technology library. We used Cadence RTL compiler to carry out synthesis and Cadence Innovus to carry out placement and routing. We used six metal layers of the metal stack from ASAP 7 nm library for routing the benchmark.

Each benchmark is first partitioned into two tiers using a modified version of Shrunk-2D [80]. We start with a placed and routed 2D design that satisfies timing constraints. From this design, only the placement information is retained. Next, we scale the placement coordinates of each standard cell in the design by $1/\sqrt{2}$ since the footprint of an M3D design is half the footprint of the corresponding 2D design. This results in the overlap of standard cells. The standard cells are then partitioned into two tiers to remove the overlap by defining regular partitioning grids, and performing an area-balanced global min-cut [80]. We scale the placement coordinates instead of standard cell dimensions, as proposed in [80], since this method does not require modification to standard cell and technology LEF, which can be time-consuming and error-prone.

In order to carry out signal ILV insertion using existing tools, we duplicate the metal stack in the technology LEF [29]. The original metal layers now constitute the metal stack for the bottom layer, and the duplicated metal layers constitute the metal stack for the top layer. In addition, the standard cell LEF is also duplicated, and the pins in the duplicated LEF are modified to the top tier metal layers. Next, we build the power-delivery network with the full metal stack before inserting signal ILVs. This avoids the placement of signal MIVs at the location of power ILVs. The dimensions of the voltage-supply rails are similar to the ones defined in the ASAP 7 nm technology library.

We then determine the location of signal ILVs by routing the design with a duplicated metal stack, and obtaining the location of vias that connect the uppermost metal layer of bottom tier with lowermost metal layer of top tier. More details about signal ILV insertion can be found in [29]. Note that duplication of the metal stack will lead to cell overlap during placement, but there will be no overlap during routing. However, we can reuse the placement of standard cells obtained from the 2D baseline. The location of signal ILVs obtained from this step is used as input to the wirelength-minimization heuristic.

We generate the distance matrices for each benchmark based on the ILV locations. Once the distance matrices are generated, they are provided as input to the

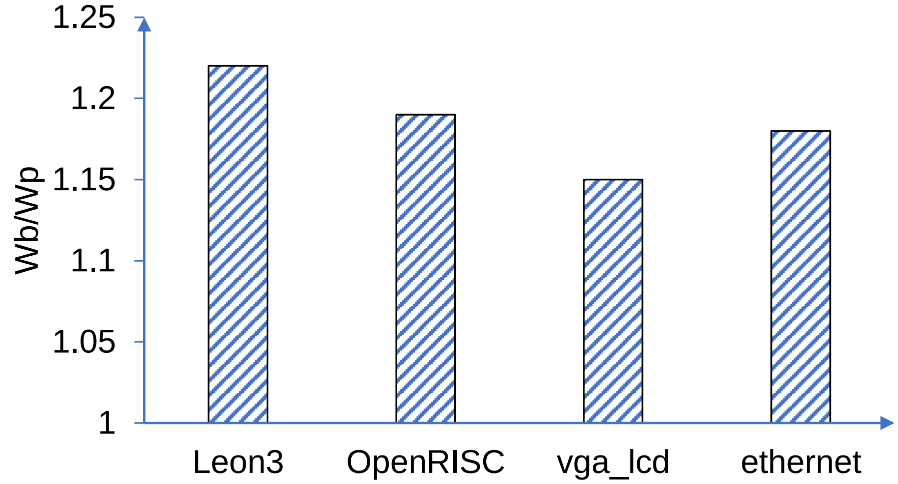


FIGURE 4.9: Ratio of total wire length obtained using the proposed heuristic and baseline.

proposed heuristic implemented in a Python script, and the Hamiltonian circuit that corresponds to a tour of ILVs with minimum wire length is generated. We define a baseline method in which we start at a random ILV and pick the ILV that is visited next in a greedy manner, i.e., an output (input) ILV nearest to an input (output) ILV. The total wirelength required to form the TRC using the proposed method and baseline method are denoted by W_p and W_b, respectively. Fig. 4.9 shows the results in terms of the ratio W_b/W_p. It can be seen that W_b is 15% to 22% larger than W_p. Therefore, the proposed algorithm reduces the total wire length required to form the TRC.

4.4.3 Diagnosis Results

We evaluate the proposed diagnosis framework using four IWLS'05 benchmarks. In order to generate training data for the proposed diagnosis engine, we need to inject defects in the TRC corresponding to a benchmark. The wirelength-minimization heuristic described above can be used to form the TRC corresponding to a benchmark. However, the run time for simulating TRCs with 10,000 interface cells at the minimum in HSpice can be prohibitively large. We therefore generate look-up tables that consist of defect size, delay introduced by the defect, and clock frequency for all possible driver-target-load combinations using Hspice simulations. The driver, target, and load clusters can consist of 1 interface cell at the minimum and 10 interface cells at the maximum. We inject a defect in the target cluster and obtain the delay introduced by the defect for several clock frequencies. We do not consider inter-cluster shorts since such a short cannot be modeled using a delay element. We therefore consider only resistive opens and intra-cluster resistive shorts. We implemented a Python-based simulator to simulate the TRC corresponding to a benchmark for $2N$ cycles, where N is the number of elements in the TRC. The look-up tables generated using HSpice simulations are provided as input to the simulator. The simulator generates the Hamming distance between the state of the defect-free and defective TRC after N and $2N$ cycles as output.

We inject defects in all clusters in the TRC corresponding to a benchmark and simulate for several clock frequencies to generate the training dataset for each benchmark. This dataset consists of 4-tuples: (i) Hamming distance between the expected and captured values (HD_1) in the TRC before the first shift operation; (ii) Hamming distance between the expected and captured values (HD_2) in the TRC before the second shift operation; (iii) clock frequency at which the test was carried out, and (iv) defect bin. The first three elements of the 4-tuples are used as features for the machine-learning model and fourth element is used as the target. We divide the range of defect sizes into 5 bins: (i) $\leq$ 100 KΩ, (ii) 100 KΩ - 200 KΩ, (iii) 200 KΩ - 300 KΩ, (iv) 300 KΩ - 400 KΩ, and (v) $\geq$ 400 KΩ.

The prediction accuracy is evaluated as the accuracy score (S), defined as follows:

$$S(d, \hat{d}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} 1(\hat{d}_i = d_i), \tag{4.1}$$

where $\hat{d}_i$ is the predicted defect-bin-label of the i-th sample, d_i is the corresponding true defect-bin-label, and $1(\hat{d}_i = d_i)$ is the indicator function, i.e., $1(\hat{d}_i = d_i) = 1$ if $\hat{d}_i = d_i$ and $1(\hat{d}_i = d_i) = 0$ if $\hat{d}_i \neq d_i$. In other words, S is the fraction of correct predictions over $\mathrm{n_{samples}}$. A value of '1' for S shows a perfect match between the actual defect-bin-labels and the predicted defect-bin-labels. Therefore, the closer the value of S is to '1', the higher is the prediction accuracy.

In order to generate simulation data for each benchmark, we carry out simulations by injecting defects of the following sizes in all ILVs: (i) 50 KΩ, (ii) 150 KΩ, (iii) 250 KΩ, (iv) 350 KΩ, and (v) 450 KΩ. For each defect size, we also vary the frequency of the clocks that drives the ILV BIST as follows: (i) 2 GHz; (ii) 1.5 GHz; (iii) 1 GHz; (iv) 0.5 GHz. For each benchmark, we use 80% of the generated data to train the diagnosis engine and 20% to test the trained model. The predictive model constructed from the training data set is used to predict the defect-bin-label from the input features in the test data set. A value of 1 for S shows a perfect match between the actual defect-bin-label and predicted defect-bin-label in the test data set across all samples.

Fig. 4.10 presents the accuracy score for four IWLS'05 benchmarks. We observe that the accuracy score exceeds 0.9 for all benchmarks. We also note that the accuracy score for prediction of resistive opens is similar to that obtained for resistive shorts for all benchmarks. For example, the accuracy score for prediction of resistive opens and resistive shorts in OpenRISC is 0.97 and 0.96, respectively. Therefore, the proposed diagnosis framework can classify resistive opens and shorts in different defect bins with high accuracy.

4.5 Conclusions

We have presented an ILV BIST solution for M3D that enables testing stand-alone testing of ILVs without incurring significant area overhead. The proposed solution

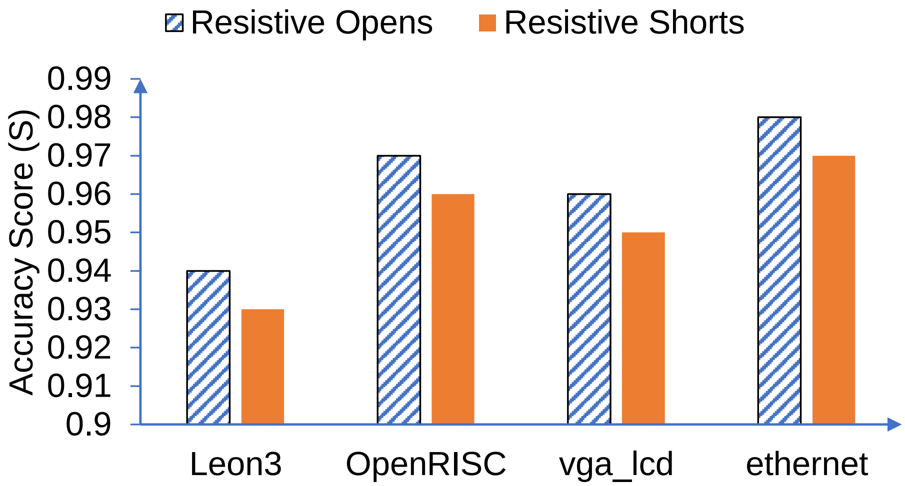


FIGURE 4.10: Defect-bin prediction accuracy score (S) for Leon3, OpenRISC, vga˙lcd, and ethernet.

can detect all opens and shorts. We proposed a heuristic to solve the wire-length minimization problem for BIST and showed that the proposed algorithm is effective in reducing the total wire length required for test application. We also proposed an artificial neural network-based framework to estimate the size of ILV defect and showed that the prediction accuracy of the proposed framework is extremely high.

5

Reliable Power Delivery and Analysis of Voltage Droop During Testing

In this chapter, we propose a framework to design a reliable PDN for M3D ICs. This framework relies on accurate electrical and reliability models to guide PDN optimization. We also analyze voltage droop during testing and propose a statistical model for quantifying the impact of voltage droop on yield loss.

We use genetic programming (GP) to explore the design space to optimize the PDN in order to achieve reliable operation. The objective is to minimize the voltage droop on power-supply rails while satisfying constraints on resilience to EM and SM. We analyze voltage droop during functional operation and compare it with that observed during testing. We also show that the worst-case voltage droop in the proposed PDN is significantly lower compared to a baseline PDN design (in both functional and test modes). We quantify the impact of voltage droop on yield loss and show that the proposed PDN design reduces yield loss.

In order to evaluate the proposed framework, we carry out the complete M3D design flow and PDN optimization for two processor benchmarks, namely Leon3 and OpenRISC. We also place and route a baseline design with uniformly spaced minimum-width rails in each metal layer. Our results show that the PDN design obtained using the proposed approach significantly increases the reliability of at least 40% of wire segments in the PDN. We also carried out dynamic simulation to obtain the worst-case power-supply droop when a functional workload is executed on Leon3 for the proposed PDN and the baseline. We observed that the worst-case voltage droop is 50.5% lower for the proposed PDN compared to the baseline PDN.

The rest of the chapter is organized as follows. Section 5.1 describes the related prior work. Section 5.2 presents the electrical and reliability models for the PDN. Section 5.3 describes the proposed PDN optimization flow. We report simulation results in Section 5.4. Finally, Section 5.5 concludes the chapter.

5.1 Related Prior Work on M3D Power Delivery

An M3D IC can be partitioned at the core-level, block-level, gate-level, and transistor-level. Significant research efforts are being directed towards gate-level design partitioning since transistor-level design partitioning requires extensive redesign of standard cells [44] and block-level design partitioning does not fully exploit the benefits of M3D. In [81], a complete RTL-to-GDSII flow for gate-level M3D was proposed and the reduction in power compared to 2D designs was shown. However, the power delivery challenges were not considered in that work.

A system-level PDN model for analyzing IR drop and Ldi/dt drop in M3D ICs was presented in [43]. In addition, frequency-domain and time-domain analysis was carried out for a full-chip die model. However, this work only describes some guidelines for PDN design and does not address reliable power delivery. A methodology to optimize signal, power, and thermal interconnects in TSV-based 3D IC based on the design of experiments was proposed in [?]. However, this work considers only inter-die interconnects during optimization, and critical challenges associated with the reliability of PDN such as EM and SM were not considered. For M3D ICs, in addition to optimization of inter-tier interconnects, tier-specific optimization needs to carried out to address EM and SM challenges.

5.2 PDN Model

PDN design depends on the placement and routing of standard cells since power MIVs cannot be placed over active areas. This can lead to irregular power MIV placement, thereby causing significant IR drop in the bottom tier and increasing the susceptibility of non-bottom tiers to EM. In addition, thermo-mechanical stress in the uppermost metal layer during fabrication can lead to voids or exacerbate EM, thereby impacting PDN reliability. In order to evaluate these issues, we need to model the PDN and its reliability.

5.2.1 Electrical Model

The electrical model of the die-level PDN consists of the parasitics of the metal wires that constitute the PDN in the top and bottom tiers, and the parasitics of power MIVs. The die-level model can be extended with models for C4 bumps, package, and PCB to obtain a system-level model for PDN. The C4 bumps and power lines in the package and PCB are modeled as a series connection of resistors and inductors. From [82] and [43], we obtain the values of these resistors and inductors: (i) $R_{C4} = 1\ m\Omega$ and $L_{C4} = 10$ pH; (ii) $R_{PKG} = 10\ m\Omega$ and $L_{PKG} = 100$ pH; (iii) $R_{PCB} = 5\ m\Omega$ and $L_{PCB} = 1\ \mu H$.

5.2.2 Thermal Model

In order to build a thermal model, we make the following assumptions. We assume that a heat sink is placed below the bottom tier. Therefore, heat generated in the upper tiers has to be transferred through the PDN and silicon layers. Layers with low thermal conductivity, such as SiO_2 and isolation layers, act as thermal barriers.

We also assume that the temperature at the heat sink is 27°C. Since the switching activity of a cell contributes to heat generation, we extract the switching activity of all the instances in our design. The C4 bumps and the entire PDN are modeled as thermal resistances.

5.2.3 EM Model

Failures due to EM and SM occur due to the formation of voids in a voltage-rail segment. These voids arise over time; therefore, EM signoff has to be carried out as a part of PDN design in order to ensure reliable operation over the lifetime. We obtain the worst-case current density for each voltage-rail segment using simulations and estimate the mean time to failure (MTTF) based on the current-density values. To facilitate our discussion of the EM model, we use the following notations: (i) T is the chip operational temperature; (ii) σ_T is the thermal stress in a metal line confined in inter-metal dielectric when it is cooled from zero-stress temperature T_{ZS} to T; (iii) L is the length of the metal line; (iv) D_a is the effective metal atomic diffusivity; (v) eZ is the effective charge of migrating metal atoms; (vi) B is the effective bulk elasticity modulus; (vii) Ω is the metal atomic lattice volume; (viii) ρ is the metal resistivity; (ix) j is the current density; (x) k_B is the Boltzmann constant.

The time for void nucleation due to EM in a metal line can be expressed as a function of the stress σ_T induced in that line [83]. The void nucleation time is defined as an instant in time when the stress at the cathode end of the metal line reaches critical stress (σ_{crit}) and expressed as:

$$t_{nuc} = \frac{L^2 k_B T}{2D_a B\Omega} \ln \left\{ \frac{\frac{eZ\rho jL}{2\Omega}}{\sigma_T + \frac{eZ\rho jL}{2\Omega} - \sigma_{crit}} \right\} \tag{5.1}$$

where $D_a = D_0 exp(-(E_D - \Omega^* \sigma_{crit})/k_B T)$. Here, E_D is the activation energy for vacancy diffusion and D_0 is the exponential prefactor. Note that this equation is employed only if the Blech limit [84] is not satisfied, i.e., $(j \times L) \geq \frac{\Omega\delta\sigma^i}{eZ\rho}$.

For each voltage-rail segment in the bottom tier, we obtain the worst-case current density using Monte-Carlo simulation and estimate the MTTF using (5.1). We use (5.1) since metal lines develop stress due to high-temperature processing of the top layer. Table 5.1 presents the values of the parameters used in (5.1). We estimate the MTTF of the PDN in the bottom tier as the MTTF of the weakest link in that tier.

For the top tier, we estimate the MTTF using Black's equation (shown below) since (5.1) is applicable only when an additional stress is induced. Therefore, we express the MTTF for a metal line in the top tier as:

$$\mathrm{MTTF_{use}} = \mathrm{MTTF_{stress}}(\frac{j_{stress}}{j_{use}}^{-n} exp\{E_a(T_{stress} - T_{use})/k_B T_{use} T_{stress}\}), \qquad (5.2)$$

where $\mathrm{MTTF_{stress}}$ is the MTTF in the accelerated-stressed condition. Here, T_{stress} is 600 K, j_{stress} is 3 MA/cm^2, and E_a is 0.86 eV [83]. The current density in a voltage-rail segment is obtained using simulation and the MTTF of the top tier is estimated as the MTTF of the weakest link in that tier.

5.3 PDN Optimization

We employ GP to intelligently explore the design space to optimize the resources dedicated for power delivery in order to minimize the impact on signal routing and ensure reliable operation of PDN. GP-based design space exploration (DSE) has been shown to be effective for architectural-level design optimizations [85]. Given

Table 5.1: Parameters used in the EM model.

Parameter	Value
E_D	0.65 eV
σ_{crit}	47.5 MPa
σ_T	46 MPa
Z	10
B	7.6×10^9 Pa
D_0	6.7×10^{-12} m^2/s

Input: Features ϕ for PDN designs in D, design space D, subset of design space d, performance metrics θ for PDN designs in D, function to evaluate quality of approximation Q, fitness function ψ, GP parameters χ and μ, maximum iterations Max
Output: Best approximation function f_{opt} for the mapping between features ϕ and performance metrics θ
Initialization: initial approximation function F_0, training set $Z \leftarrow \emptyset$, iteration $k \leftarrow 0$, an intial set of N randomly-generated approximation function F_0

Generate training data: For each design PDN_i in d, add ($\phi(PDN_i)$, $\theta(PDN_i)$ to Z
for each approximation function f_i in F_0 **do**
    Evaluate $Q(f_i)$ and $\psi(f_i)$
end for
while $\max\{\psi(f_0), \psi(f_1), \ldots, \psi(f_N)\}$ does not converge, where $f_i \in F_k$ or $k < Max$ **do**
    Select $(1-\chi)$ approximation functions from F_k and insert in F_{k+1}
    Select χ approximation functions from F_k, pair them up, produce an offspring, and insert in F_{k+1}
    Select μ approximation functions from F_{k+1} and randomly mutate them
    for each approximation function f_i in F_{k+1} **do**
        Evaluate $Q(f_i)$ and $\psi(f_i)$
    end for
    $k \leftarrow k+1$
end while
return $\max\{\psi(f_0), \psi(f_1), \ldots, \psi(f_N)\}$, where $f_i \in F_k$

FIGURE 5.1: Algorithm for PDN optimization using GP-based DSE.

a sample of collected performance metrics, approximation functions to predict the global performance of all candidate designs are built. As design spaces grow, it is more efficient to employ predictive approximation functions rather than a standard search algorithm. We employ GP to create appropriate approximation functions, and fit them to the collected performance metrics. Predictive models are built in GP-based DSE using the simulation results of a small sample of designs picked in advance.

The first step in the proposed optimization algorithm is to identify key features ϕ to represent a PDN and set the design space D. For the PDN-optimization problem, design features such as MIV count, MIV distribution, and resources for power delivery in the each tier can be used. The appropriate ranges for each design feature can be defined to set D. A small subset d of the design space is chosen and the performance metrics θ are extracted for that subset of design points. For the PDN-optimization problem, performance metrics such as maximum IR drop, MTTF, and total wire length can be used.

GP is then used to find the best approximation function for the mapping between features ϕ and performance metrics θ. The algorithm is structured such that accurate candidate approximation functions are more likely to survive and be recombined. The candidate approximation functions are polynomial equations represented as expression trees. An individual node in a tree represents a design feature. The accuracy of a candidate approximation function is decided based on the quality of approximation Q and fitness function ψ.

$$Q(f_i) = \frac{\sum_{n=1}^{|d|} |f_i(\phi(PDN_n) - \theta(PDN_n)|}{\sum_{n=1}^{|d|} |f_i(\phi(PDN_n)}, \ \psi(f_i) = Q(f_i) - c * p^2 \tag{5.3}$$

The quality $Q(f_i)$ of approximation function f_i is obtained by calculating the sum of absolute differences between the approximation function's predictions and the collected performance metrics for all sample design points. Note that $\psi(f_i)$ is the fitness function, where c is a constant used to control the size penalty and p is the number of tuning parameters. The fitness function is used to penalize lengthy expression trees and keep the equations more compact.

5.4 Simulation Results

5.4.1 Simulation Setup

Simulations were performed on the Leon3 and OpenRISC processor benchmarks to evaluate the proposed technique. This benchmark was implemented using the Arizona State Predictive PDK (ASAP) 7 nm technology library. We used Cadence RTL compiler to carry out synthesis and Cadence Innovus to carry out placement and routing. We used six metal layers of the metal stack from ASAP 7 nm library for routing the benchmark.

We implemented the PDN before we placed and routed the standard cells using

Cadence Innovus. We then carried out RC extraction for that benchmark. We use the extracted RC model to carry out PDN IR-drop and reliability analysis using Cadence Voltus. We used a Python script to extrapolate the current density values obtained from Voltus to MTTF. This Python script implements the model described in Section 5.1.2. We implemented the GP-based DSE algorithm in Python. The code was run on a 64-bit Linux Server with a quad-core Intel Xeon 2.53 GHz CPU and 12 GB memory.

5.4.2 Design Flow

First, the design is partitioned into two tiers using a modified version of Shrunk-2D [80]. We start with a placed and routed 2D design that satisfies timing constraints. From this design, only the placement information is retained. Next, we scale the placement coordinates of each standard cell in the design by $1/\sqrt{2}$ since the footprint of an M3D design is half the footprint of the corresponding 2D design. This results in the overlap of standard cells. The standard cells are then partitioned into two tiers to remove the overlap by defining regular partitioning grids, and performing an area-balanced global min-cut [80]. We scale the placement coordinates instead of standard cell dimensions, as proposed in [80], since this method does not require modification to standard cell and technology LEF, which can be time-consuming and error-prone.

In order to carry out PDN design and signal MIV insertion using existing tools, we duplicate the metal stack in the technology LEF [29]. The original metal layers now constitute the metal stack for the bottom tier, and the duplicated metal layers constitute the metal stack for the top tier. In addition, the standard cell LEF is also duplicated, and the pins in the duplicated LEF are modified to the top tier metal layers. Next, we build the PDN with the full metal stack before inserting signal MIVs. This avoids the placement of signal MIVs at the location of power MIVs.

The dimensions of the voltage-supply rails are similar to the ones defined in the ASAP 7 nm technology library.

We then determine the location of signal MIVs by routing the design with a duplicated metal stack, and obtaining the location of vias that connect the uppermost metal layer of bottom tier with lowermost metal layer of top tier. More details about signal MIV insertion can be found in [29]. Note that duplication of the metal stack will lead to cell overlap during placement, but there will be no overlap during routing. However, we can reuse the placement of standard cells obtained from the 2D baseline. Next, we carry out initial routing, timing characterization, clock tree optimization, and timing-driven routing for each tier using a netlist corresponding to that tier and signal MIV locations. We then obtain placed and routed designs for each tier. These designs are then merged to a final M3D design, and timing analysis is subsequently carried out.

5.4.3 PDN Optimization

We considered four features to represent the PDN design space: (i) power and ground MIV count; (ii) maximum power and ground MIV area per placement tile; (iii) percentage of routing resources in the top tier that is allocated for power delivery; (iv) percentage of routing resources in the bottom tier that is allocated for power delivery.

We next define the design space. We restrict the power and ground MIV count to be between 1% and 5% of the total number of MIVs between two tiers. We restrict the maximum power and ground MIV area to be between 1% and 5% of the area of a single placement tile. We also restrict the total routing resources allocated for power delivery in the top and bottom tier to be between 5% and 25% of the total routing resources. We have chosen these ranges based on the values reported for these parameters [31, 43, 44].

We run simulations for 81 PDNs designs. These designs are obtained by setting the above four features to their min, max, and median values. For each design, we obtain the MTTF and worst-case IR drop. Using the above features and MTTF/IR drop values for 81 PDN designs as input, we run the genetic-programming algorithm to obtain the approximation functions. We then combine the two approximation functions into a single desirability function. We then carry out a min-search on the resulting function to obtain the features of the optimal PDN design. For the Leon3 benchmark, we obtained the features of the optimal PDN design as: (i) MIV count: 2.75%, (ii) MIV density: 2%, (iii) total routing resources for power delivery in the top tier: 11%, and (iv) total routing resources for power delivery in the bottom tier: 15%.

5.4.4 PDN Reliability

We compare the reliability of the PDN design obtained using the proposed approach with a baseline PDN designed by uniformly spaced minimum-width rails in each metal layer. We then carry out signoff-quality RC extraction for the baseline and the proposed PDN designs. We use the extracted RC model to carry out PDN reliability analysis using Cadence Voltus. The thermal stress-aware electromigration model for M3D is given as input to Voltus for reliability analysis.

Fig. 5.2 presents the current density histogram for the baseline and proposed PDN designs for Leon3. We observe that the proposed PDN has a higher percentage of wire segments with current density less than 0.125×10^9 A/m^2 compared to the baseline PDN. Therefore, the PDN design obtained using the proposed approach significantly increases the time-to-failure of at least 40% of the wire segments in the PDN. The corresponding figure for OpenRISC was 34%.

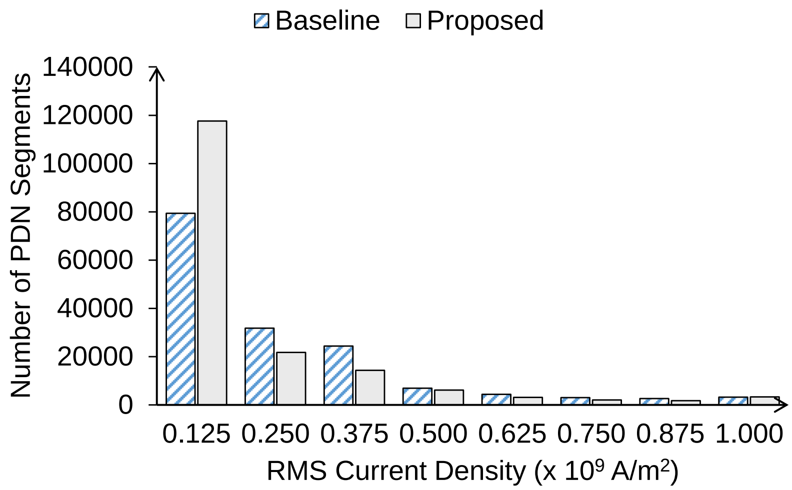


FIGURE 5.2: Current-density histogram for baseline and proposed PDN designs for Leon3 benchmark.

5.4.5 Voltage-Droop Analysis

We also developed a framework to carry out dynamic simulation using Cadence Voltus. This framework relies on simulation waveform dumps of realistic workloads. In the proposed framework, we carry out a dynamic simulation of PDN to obtain the voltage at every node in the design for a specified time window in the workload. Due to tool limitations, the length of the simulation window is limited to 1000 time steps. We executed three workloads from the MiBench benchmark suite, namely basicmath, qsort, and crc32 on Leon3. We obtained the signal activity from the execution of each workload in the form of a simulation dump (VCD file) generated by performing a post-synthesis simulation in ModelSim. We read the simulation waveform dump during dynamic simulation to obtain the switching activity for the gates in the design.

In order to analyze voltage droop during testing, we also carry out dynamic simulation of transition-delay patterns generated using the netlist obtained after the place-and-route step. We write out patterns in STIL format from test-pattern generation tool. We then convert the STIL patterns to Verilog testbench and carry out gate-level simulation using that testbench. We dump a VCD file when we carry out the gate-level simulation. This VCD file contains the switching activity for gates in the design while executing the pattern set.

We first carry out stand-alone thermal analysis of an M3D design to obtain a thermal map. We then carry out a dynamic voltage-droop analysis to obtain on-chip power distribution, which is then used as input for the thermal analysis. This iterative process is repeated until the change is voltage droop and temperature is less than 5%. Fig. 5.3 presents the temperature distribution and voltage at instance with maximum droop obtained by carrying electrical and thermal co-simulation on a PDN design with the following features. We observe that the peak chip temperature is 94 °C and worst-case power-supply drop is 0.1 V, where the nominal supply voltage is 0.7 V. Such high level of power-supply noise is undesirable. We therefore study the worst-case power-supply noise for PDN designs obtained by varying power/ground MIV count and MIV density.

Fig. 5.4(a) presents the supply voltage at the gate with maximum droop for three PDN designs. The MIV count in these three PDN designs is 1%, 2.5%, and 5%. We

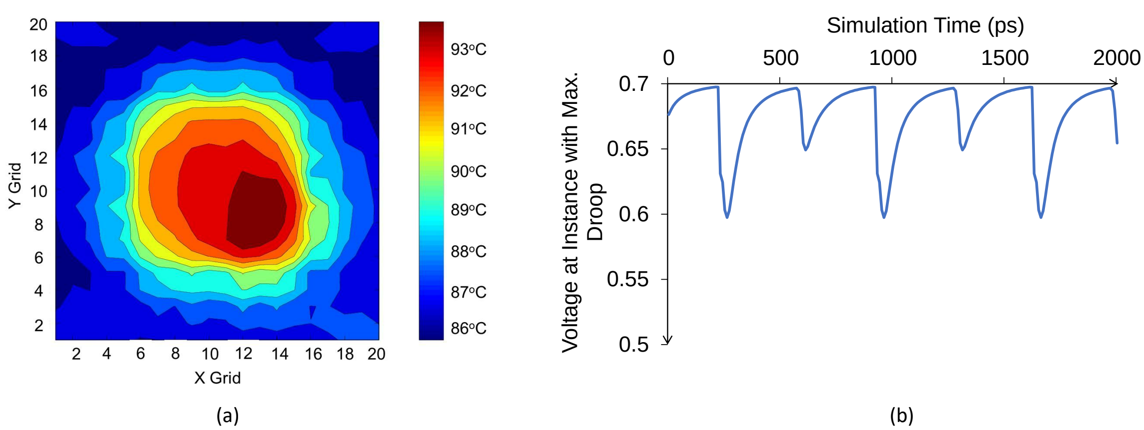
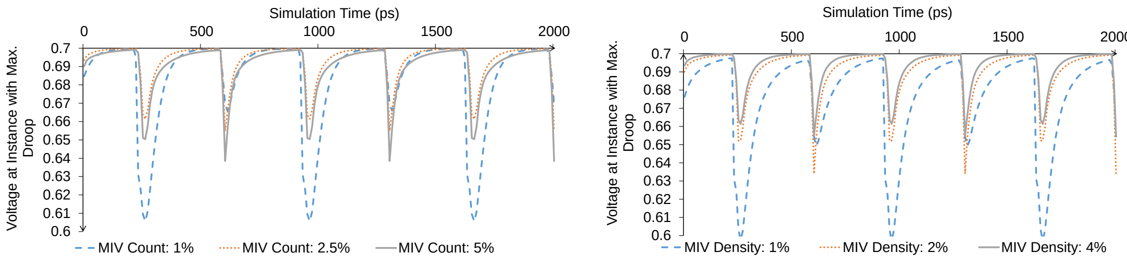


FIGURE 5.3: Results obtained from dynamic electrical and thermal co-simulation: (a) 2D lumped temperature distribution; (b) voltage at instance with maximum droop.



(a) (b)

FIGURE 5.4: Plots of voltage at gate with maximum droop obtained by varying: (a) MIV count; (b) MIV density.

also restrict the power and ground MIV area to 2% of the area of a single placement tile. We observe that the worst-case power-supply drop is 0.092 V when the MIV count is restricted to 1% of total number of MIVs. We observe a significant reduction is the worst-case power-supply drop when the MIV count is increased to 2.5% and 5%. This reduction in the worst-case power supply drop is due to the decrease in the effective resistance of the PDN when additional power and ground MIVs are introduced. We also observe a slight increase in the worst-case power-supply drop when the MIV count is increased from 2.5% to 5%. This can be attributed to the increase in signal wirelength due to an increase in the routing congestion caused by the introduction of additional vias. An increase in the signal wirelength leads to higher switching power; thereby, increasing the load on the PDN and offsetting the improvement in the voltage droop due to the introduction of additional vias.

Fig. 5.4(b) presents the supply voltage at the instance with maximum droop for three PDN designs. The MIV density in these three PDN designs is 1%, 2.5%, and 5%. We also restrict the power and ground MIV count to 2.5%. We observe that the worst-case power-supply drop is 0.1 V when the MIV density is restricted to 1%. We observe a significant reduction is the worst-case power-supply drop when the MIV density is increased to 2.5% and 5%. This reduction in the worst-case power-supply drop can be attributed to the decrease in routing congestion when MIV density is

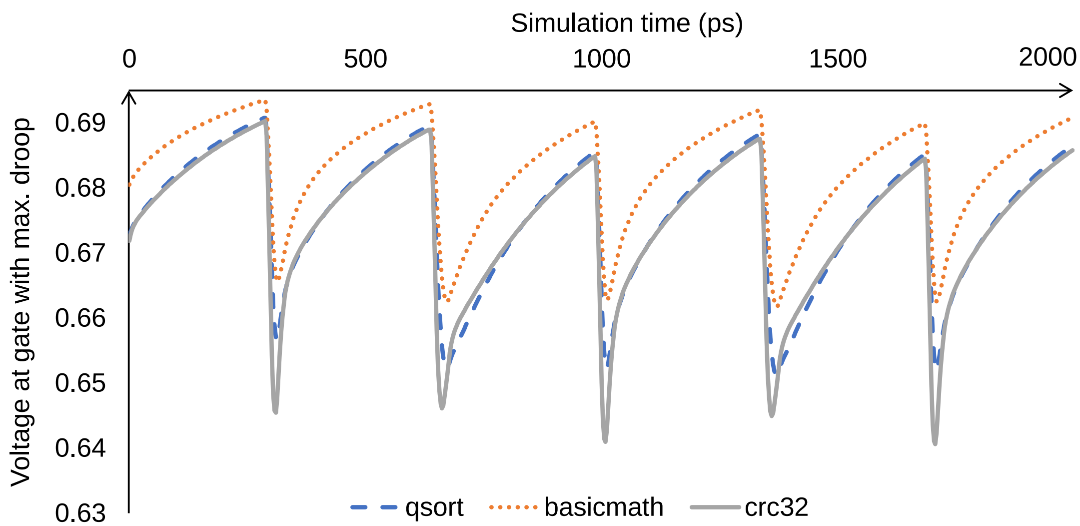


FIGURE 5.5: Voltage at gate with maximum droop for Leon3.

increased; thereby, leading to a reduction in the signal wirelength and load on the PDN.

Fig. 5.5 presents the supply voltage at the gate with maximum droop for three MiBench workloads on Leon3. We observe that the worst-case power-supply drop is 45 mV, 37 mV, and 59 mV for qsort, basicmath, and crc32, respectively. Fig. **??** compares the supply voltage at the gate with maximum droop for the proposed PDN design and the baseline when qsort is executed on Leon3. We observe that the worst-case voltage droop is 91 mV for the baseline PDN. Therefore, the proposed PDN design reduces the worst-case voltage droop by 50.5%. The corresponding figure for OpenRISC is 40.2%

Fig. 5.7 presents the supply voltage at the gate with maximum droop during scan shift and capture phases of test application for Leon3. The power budget during scan shift or capture limits the number of flip-flops that can toggle at the same time. As expected, the worst-case power-supply droop reduces as we reduce the power budget for both scan shift and capture. We also observe that this is more significant during scan shift. This is expected since high switching activity during scan shift can cause

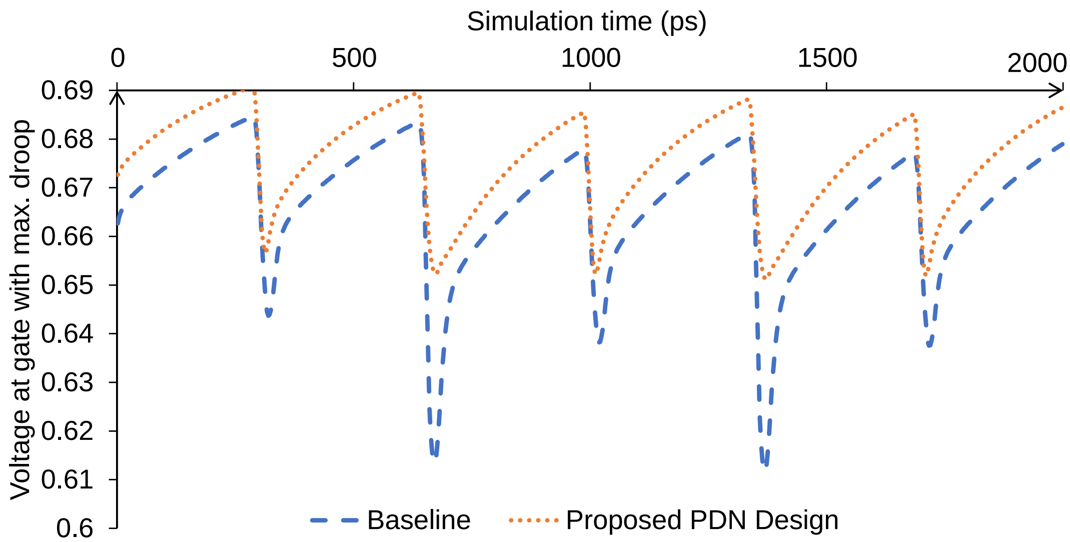
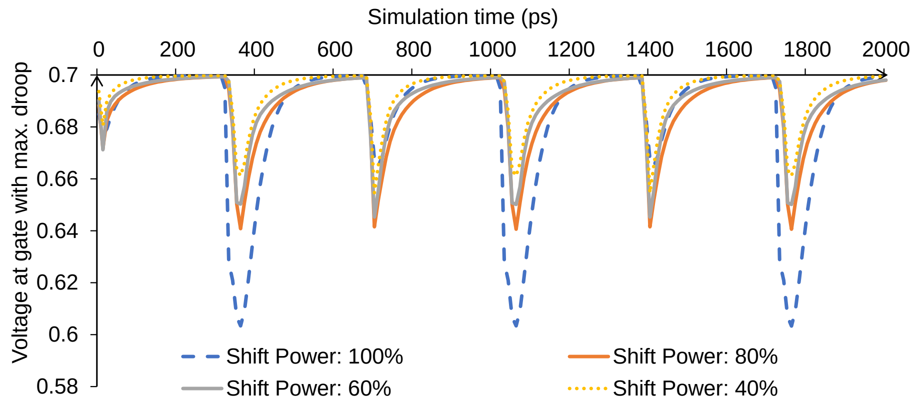


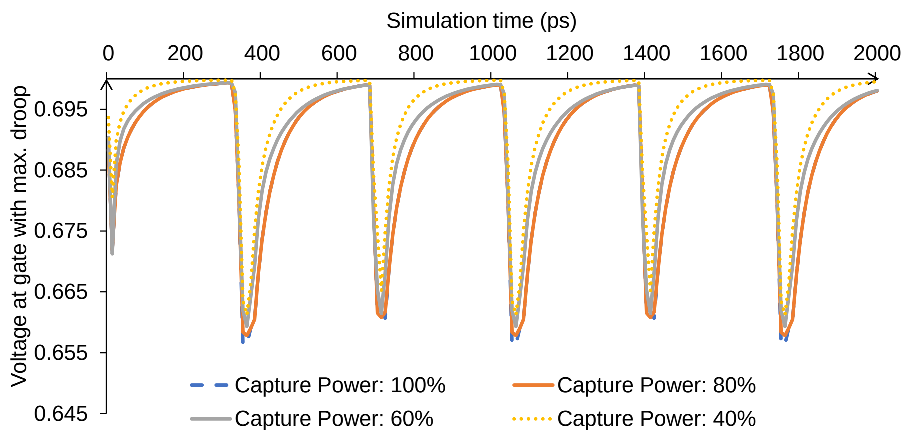
FIGURE 5.6: Voltage at gate with maximum droop for the proposed PDN design and baseline when qsort is executed on Leon3.

a significant droop in supply voltage. On the other hand, we do not observe a significant change during capture; Leon3 is a single-clock design and the number of flip-flops toggling per pattern is much lower than the specified budget.

Fig. 5.8 compares the supply voltage at the gate with maximum droop for the proposed PDN design for Leon3 with the baseline when test patterns generated using 60% power budget are applied. The worst-case voltage droop is 73 mV and 59 mV for the baseline PDN during scan shift and capture, respectively. On the other hand, the worst-case voltage droop is only 55 mV and 41 mV for the proposed PDN design during scan shift and capture, respectively. The corresponding numbers for OpenRISC are 65 mV and 44 mV for the baseline PDN, and 51 mV and 33 mV for the proposed PDN design. Therefore, the proposed PDN design also reduces the worst-case voltage droop during testing.



(a)



(b)

FIGURE 5.7: Voltage at gate with maximum droop for Leon3 for different power budgets during: (a) scan shift; (b) capture.

5.4.6 Analysis of Yield Loss

We next quantify the impact of voltage droop on yield for our PDN design and the baseline. We use a statistical model to quantify the robustness of the design to voltage droop during scan-based testing. A design is robust to voltage droop during testing if all the scan cells (SCs) in the design are robust to voltage droop. An incorrect value captured at a SC can cause the chip to fail at the tester.

The robustness of a SC can be assumed to consist of two statistically independent

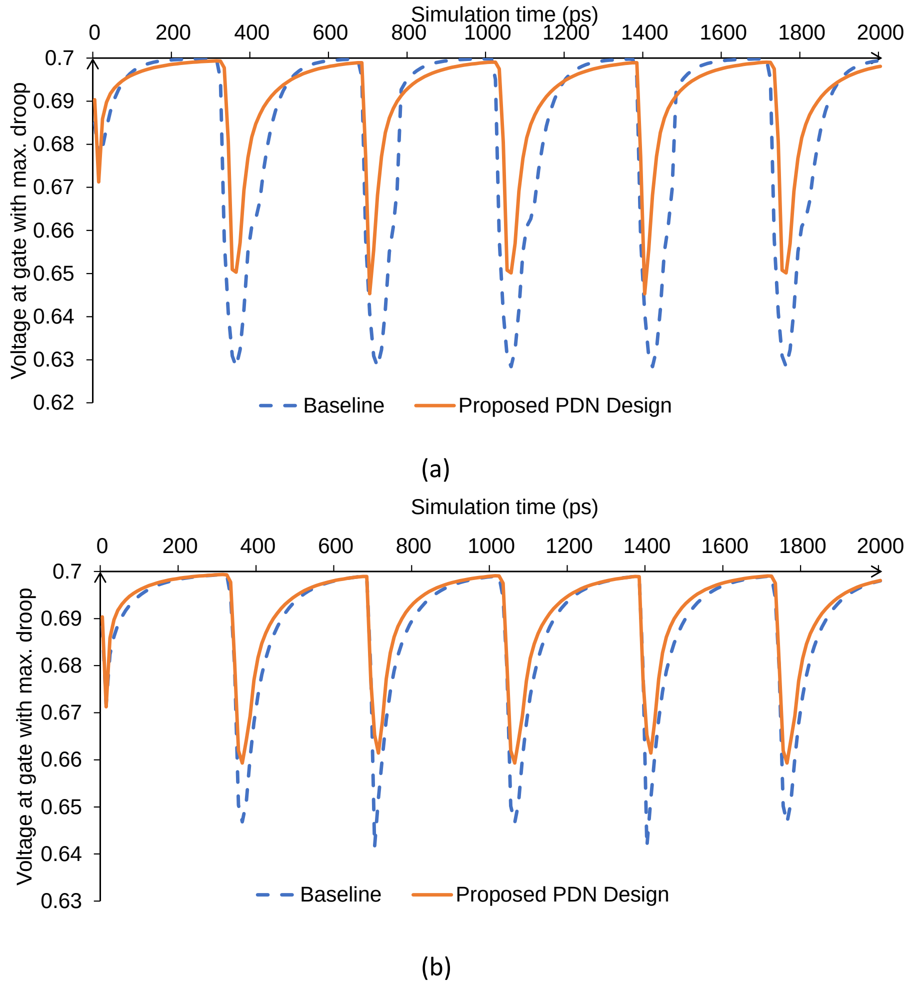


FIGURE 5.8: Voltage at gate with maximum droop for Leon3 for the proposed PDN design and baseline for patterns generated using 60% power budget during: (a) scan shift; (b) capture.

events: (i) robustness to voltage droop during scan shift; (ii) robustness to voltage droop during capture. Therefore, the probability that an incorrect value is captured at SC_i, $P'(SC_i)$, can be defined as: $P'(SC_i) = 1 - P_S(SC_i) \times P_C(SC_i)$, where $P_S(SC_i)$ (or $P_C(SC_i)$) is the probability that a correct value is captured at SC_i during scan shift (or capture).

To facilitate our discussion of the statistical model, we use the following terms: (i) shift path starts at the Q output of one SC and ends at the SI input of the next SC in the scan chain; (ii) capture/data path starts at the Q output of one SC and ends at D input of the next SC. Let us consider the design presented in Fig. 5.9(a). For SC_4 in this design, the path from Q of SC_3 to SI of SC_4 is the shift path for SC_4, and the paths from Q of SC_1 and SC_2 to D of SC_4 are capture paths for SC_4.

Under the assumption that all the gates in the shift path get the same supply voltage v, we can obtain the delay distribution of the shift path ending at each SC in the design. Since delay varies linearly with the supply voltage as a first-order approximation [86], the delay distribution for the shift/capture path ending at SC_i can be approximated as shown in Fig. 5.9(b). Let $D_S(v)$ denote the delay distribution of the shift path ending at SC_i. Let V_{crit} denote the supply voltage at which the timing margin on the shift path becomes zero. The probability that a correct value can be captured at SC_i during scan shift can be expressed as: $P_S(SC_i) = \int_{V_{crit}}^{V_{Supply}} D_S(v)$. d(v).

Let $D_C(v)$ denote the delay distribution of the capture path with the smallest slack for SC_i, where v is supply voltage to all the gates in the fan-in cone of SC_i. Let V_{crit} denote the supply voltage at which the timing margin on the capture path becomes zero. The probability that a correct value is captured at SC_i can be expressed

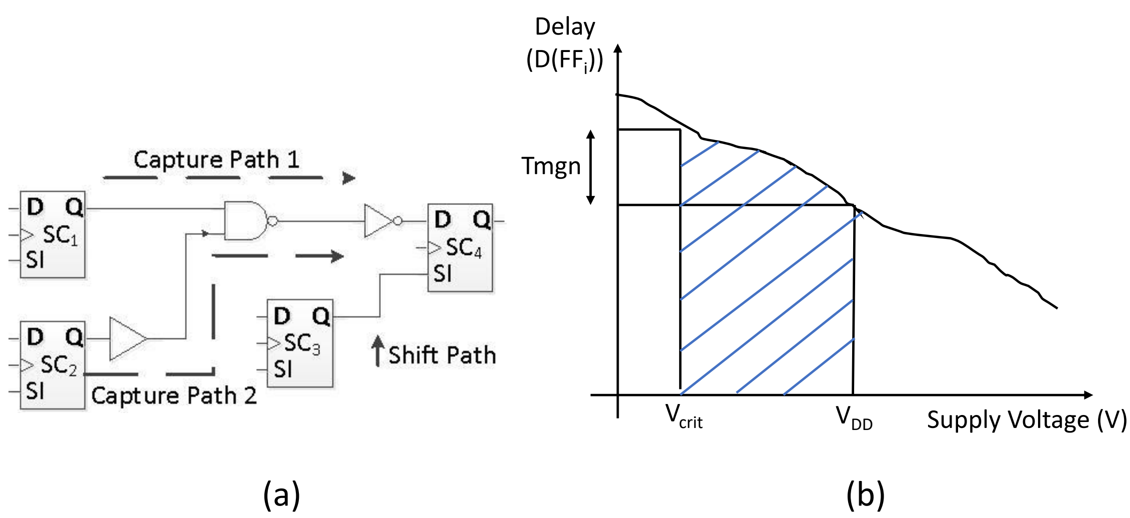


FIGURE 5.9: Illustration of: (a) shift and capture path for SC_4; (b) delay distribution for shift or capture path.

as: $P_C(SC_i) = \int_{V_{crit}}^{V_{Supply}} D_C(v)d(v)$.

From the above equations, we express the probability that an incorrect values can be captured at scan SC SC_i as:

$$P'(SC_i) = 1 - (\int_{V_{crit}}^{V_{Supply}} D_S(v)d(v)) \times (\int_{V_{crit}}^{V_{Supply}} D_C(v)d(v)) \tag{5.4}$$

For a design with N SCs, we define the yield loss, YL, as: $YL = \sum_{i=1}^{N} P'(SC_i)$. This metric indicates the likelihood of a good chip failing on the tester. Smaller values of YL are clearly desirable.

In order to obtain YL for the proposed design and the baseline, we first obtain the timing of all the paths in the design using Synopsys PrimeTime. We also obtain the supply voltage at every SC for the proposed design and baseline using Cadence Voltus and assume that all the gates in the shift and capture paths observe the same supply voltage. We then obtain the probability that an incorrect value is captured during shift and capture for all SCs using (5.4). We then obtain YL for the proposed design and the baseline as 8124.6 and 12100.2, respectively. The corresponding YL numbers for OpenRISC are 11033.4 and 14503.7. We therefore conclude that the proposed PDN design reduces yield loss significantly.

5.5 Conclusions

We have presented an optimization approach for reliable power delivery in M3D ICs. We have also analyzed the voltage droop during testing and compared it with that observed during functional operation. We have quantified the impact of voltage droop during testing on yield loss. Our results show that the proposed design significantly increases the reliability of at least 40% of the wire segments in the PDN. In addition, our results show that the proposed PDN design significantly reduces the worst-case voltage droop and yield loss due to voltage droop compared to a baseline PDN.

6

Conclusions

It is becoming increasingly difficult to sustain device scaling in an economically viable manner due to challenges associated with interconnect scaling, lithography of small features, and process variations. The semiconductor industry is therefore exploring 3D integration to reduce power consumption with smaller device footprint, and higher performance. M3D integration is an emerging technology in which transistors are processed layer-by-layer on the same wafer. It is attractive for 3D integration because it has the potential to achieve higher device density compared to TSV-based 3D stacking. Sequential integration of transistor layers requires high-density vertical interconnects, known as the ILVs. Due to the above benefits, there is growing interest towards the adoption of M3D. However, test challenges for M3D have remained largely unexplored. In this dissertation, we have targeted some of the challenges related to testing of M3D ICs and proposed solutions to them.

This dissertation presents the first comprehensive test solutions for emerging M3D integrated circuits. This research has led to theoretical insights, significant academic and industrial collaborations, simulations results using advanced process design kits, and a set of test and DfT solutions. Novel DfT solutions have been presented to carry out modular testing of M3D ICs. We have also addressed some of the key challenges related to reliable power delivery and voltage droop during testing. We have also made advances in modeling electrostatic coupling and wafer-bonding defects.

Chapter 2 addresses analysis and fault modeling of defects that are new to M3D ICs. We quantified the impact of electrostatic coupling and wafer-bonding defects on the threshold voltage of a top-layer transistor in an M3D IC. In addition, we showed that wafer-bonding defects can lead to a change in the resistance of ILVs, and in some cases, lead to an open in an ILV or a short between two ILVs. We also studied the impact of these defects on path delays and on the effectiveness of delay-test patterns for large benchmarks. Our results showed that the timing characteristics of an M3D IC can be significantly altered due to coupling and wafer-bonding defects if the thickness of its inter-layer dielectric is less than 100 nm.

Chapter 3 presented a new DfT solution for M3D ICs based on dedicated test layers. The proposed DfT solution enables testing of a partial assembly as well as a full assembly with all layers. It consists of a set of cooperating test layers, one for each interface between two functional layers. We described the operating modes and configurations supported by the test layers. We also derived a cost model accounting for the additional steps involved in the fabrication of an M3D IC. We then compared the cost of the proposed DfT solution with the cost of a potential test solution based on P1838. Our results have shown that the proposed solution is more cost-efficient than the P1838-based solution for a wide range of ILV density. We have also presented a test scheduling and optimization technique for wafer-level testing of M3D ICs.

Chapter 4 presented an ILV BIST solution for M3D ICs since conventional scan-based test pattern generation do not provide the diagnostic resolution to diagnose and characterize ILV faults. In the proposed ILV BIST solution, interface-register cells in a test layer are stitched into a TRC using their functional outputs and the ILVs.

We showed that the proposed solution detects all hard opens and shorts in the ILVs. We validated the detection of all hard opens and shorts using HSpice simulations. We also proposed an artificial neural network-based framework to estimate the size of ILV defect and showed that the prediction accuracy of the proposed framework is extremely high.

Chapter 5 described an optimization approach for reliable power delivery in M3D ICs. We have also analyzed the voltage droop during testing and compared it with that observed during functional operation. We have quantified the impact of voltage droop during testing on yield loss. Our results show that the proposed design significantly increases the reliability of the wire segments in the PDN. In addition, our results show that the proposed PDN design significantly reduces the worst-case voltage droop and yield loss due to voltage droop compared to a baseline PDN.

6.1 Future Directions

Due to complexity associated with fabrication, testing of emerging devices/technologies is of particular concern. Existing test solutions are inadequate for emerging technologies since they do not consider technology-specific details. Therefore, defects that are new to emerging technologies are not screened unless tests that consider technology-specific details are included as a part of manufacturing test. New tools are to analyze, model, and screen manufacturing defects, and DfT solutions to enable testing, defect isolation, and yield enhancement are required. Moreover, with the emergence of new technologies, new avenues for test-cost optimization are available. This research has opened up the following new directions.

- *Test Pattern Generation for Technology-Specific Delay Faults*

  We have seen in our research that scaling down of the ILD thickness in M3D ICs leads to electrostatic coupling between device layers, and ILD defects impact circuit timing. Some of these defects can significantly impact the performance of an M3D IC. Therefore, there is a need to generate a test-pattern set that can be used to detect delay faults in M3D designs. These patterns can be used during the characterization phase to fix process issues that are causing delay faults.

Although several variability-aware test-pattern generation methods exist, these are unlikely to be effective for M3D ICs. For example, the delay distributions for the gates obtained by considering variations in process parameters are not as useful for generating test patterns for ILD defects. Therefore, a test-pattern generation framework that can be used to generate a small set of patterns to screen ILD defects is needed. In order to select a diverse fault set, we need to take into account the physical location, logic depth, and slack on longest path for each fault.

- *Test Solutions for M3D ICs Partitioned at the Transistor Level*

  In the transistor-level design partitioning approach, the P-channel and N-channel transistors are placed on different layers, thereby allowing the fabrication process to be optimized separately for each layer. The performance boosters for each channel type, e.g., channel material, substrate orientation, channel strain, raised source and drain strain, etc., can be optimized separately. For example, high performance CMOS gates were designed by stacking of III-V nFETs over SiGe pFETs.

  Since the channel material, substrate orientation, channel strain, etc., used for the n-channel and p-channel devices can be completely different, the fabrication process, environment, and the number of process steps can be different for them. Therefore, the defects that arise during the fabrication process can be different for the layer with n-channel devices and that with p-channel devices.

Some of these defects arise due to random particle contaminants while others may be systematic in nature. In order to understand the defects and the failure mechanisms corresponding to those defects, we need to test each device layer separately. Such testing will enable defect isolation, thereby enabling yield enhancement.

Bibliography

[1] P. Batude, M. Vinet, B. Previtali, C. Tabone, C. Xu, J. Mazurier, O. Weber, F. Andrieu, L. Tosti, L. Brevard, B. Sklenard, P. Coudrain, S. Bobba, H. B. Jamaa, P. E. Gaillardon, A. Pouydebasque, O. Thomas, C. L. Royer, J. M. Hartmann, L. Sanchez, L. Baud, V. Carron, L. Clavelier, G. D. Micheli, S. Deleonibus, O. Faynot, and T. Poiroux, "Advances, challenges and opportunities in 3D CMOS sequential integration," in *IEEE International Electron Devices Meeting (IEDM)*, pp. 7.3.1–7.3.4, Dec 2011.

[2] X. D. et al., "Fabrication Cost Analysis and Cost-Aware Design Space Exploration for 3-D ICs," *IEEE Trans. CAD*, vol. 29, no. 12, 2010.

[3] P. Batude, T. Ernst, J. Arcamone, G. Arndt, P. Coudrain, and P. E. Gaillardon, "3-D Sequential Integration: A Key Enabling Technology for Heterogeneous Co-Integration of New Function With CMOS," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 2, pp. 714–722, Dec 2012.

[4] P. Batude, C. Fenouillet-Beranger, L. Pasini, V. Lu, F. Deprat, L. Brunet, B. Sklenard, F. Piegas-Luce, M. Cass, B. Mathieu, O. Billoint, G. Cibrario, O. Turkyilmaz, H. Sarhan, S. Thuries, L. Hutin, S. Sollier, J. Widiez, L. Hortemel, C. Tabone, M. P. Samson, B. Previtali, N. Rambal, F. Ponthenier, J. Mazurier, R. Beneyton, M. Bidaud, E. Josse, E. Petitprez, O. Rozeau, M. Rivoire, C. Euvard-Colnat, A. Seignard, F. Fournel, L. Benaissa, P. Coudrain, P. Leduc, J. M. Hartmann, P. Besson, S. Kerdiles, C. Bout, F. Nemouchi, A. Royer, C. Agraffeil, G. Ghibaudo, T. Signamarcheix, M. Haond, F. Clermidy, O. Faynot, and M. Vinet, "3DVLSI with CoolCube process: An alternative path to scaling," in *IEEE Symposium on VLSI Technology*, pp. T48–T49, June 2015.

[5] L. Brunet, P. Batude, F. Fournel, L. Benaissa, C. Fenouillet-Beranger, L. Pasini, F. Deprat, B. Previtali, F. Ponthenier, A. Seignard, C. Euvrard-Colnat, M. Rivoire, P. Besson, C. Arvet, E. Beche, O. Rozeau, O. Billoint, O. Turkyilmaz, F. Clermidy, T. Signamarcheix, and M. Vinet, "(Invited) Direct Bonding: A Key Enabler for 3D Monolithic Integration," *ECS Transactions*, vol. 64, no. 5, pp. 381–390, 2014.

[6] I. Radu, G. Gaudin, W. van den Daele, F. Letertre, C. Mazure, L. D. Cioccio, T. Lacave, F. Mazen, P. Scheiblin, T. Signamarcheix, and S. Cristoloveanu, "Novel low temperature 3D wafer stacking technology for high density device integration," in *IEEE European Solid-State Device Research Conference (ESSDERC)*, pp. 151–154, Sept 2013.

[7] Z. Or-Bach, *CHIPS 2020 VOL. 2: New Vistas in Nanoelectronics*, pp. 51–91. Springer International Publishing, 2016.

[8] D. H. Kim, K. Athikulwongse, M. B. Healy, M. M. Hossain, M. Jung, I. Khorosh, G. Kumar, Y. J. Lee, D. L. Lewis, T. W. Lin, C. Liu, S. Panth, M. Pathak, M. Ren, G. Shen, T. Song, D. H. Woo, X. Zhao, J. Kim, H. Choi, G. H. Loh, H. H. S. Lee, and S. K. Lim, "Design and Analysis of 3D-MAPS (3D Massively Parallel Processor with Stacked Memory)," *IEEE Transactions on Computers*, vol. 64, pp. 112–125, Jan 2015.

[9] P. Coudrain, D. Henry, A. Berthelot, J. Charbonnier, S. Verrun, R. Franiatte, N. Bouzaida, G. Cibrario, F. Calmon, I. O'Connor, T. Lacrevaz, L. Fourneaud, B. Flechet, N. Chevrier, A. Farcy, and O. Le-Briz, "3D Integration of CMOS image sensor with coprocessor using TSV last and micro-bumps technologies," in *IEEE Electronic Components and Technology Conference (ECTC)*, pp. 674–682, May 2013.

[10] P. Batude, T. Ernst, J. Arcamone, G. Arndt, P. Coudrain, and P. E. Gaillardon, "3-D Sequential Integration: A Key Enabling Technology for Heterogeneous Co-Integration of New Function With CMOS," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 2, pp. 714–722, Dec 2012.

[11] A. W. Topol, D. C. L. Tulipe, L. Shi, D. J. Frank, K. Bernstein, S. E. Steen, A. Kumar, G. U. Singco, A. M. Young, K. W. Guarini, and M. Ieong, "Three-dimensional integrated circuits," *IBM Journal of Research and Development*, vol. 50, pp. 491–506, July 2006.

[12] J. Rearick, "New Paradigm Shift in 3-D Design and Testing." *ITC 3D Test Workshop*, 2015.

[13] S. Kannan, R. Agarwal, A. Bousquet, G. Aluri, and H. S. Chang, "Device performance analysis on 20nm technology thin wafers in a 3D package," in *IEEE International Reliability Physics Symposium*, pp. 4C.4.1–4C.4.5, April 2015.

[14] ITRS, 2013.

[15] K. Arabi, K. Samadi, and Y. Du, "3D VLSI: A Scalable Integration Beyond 2D," in *ACM International Symposium on Physical Design*, pp. 1–7, 2015.

[16] S. Bobba, A. Chakraborty, O. Thomas, P. Batude, T. Ernst, O. Faynot, D. Z. Pan, and G. D. Micheli, "CELONCEL: Effective Design Technique for 3-D Monolithic Integration targeting High Performance Integrated Circuits," in *Asia and South Pacific Design Automation Conference (ASP-DAC)*, pp. 336–343, Jan 2011.

[17] Y.-J. Lee and S. K. Lim, "Ultrahigh Density Logic Designs Using Monolithic 3-D Integration," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Dec 2013.

[18] "2015 IEEE 3D Test Workshop." `http://www.pld.ttu.ee/3dtest/`.

[19] S. K. Goel and A. Mehta, "Circuit And Method For Monolithic Stacked Integrated Circuit Testing," Apr. 2015. US Patent No. 9,110,136.

[20] T. Naito, T. Ishida, T. Onoduka, M. Nishigoori, T. Nakayama, Y. Ueno, Y. Ishimoto, A. Suzuki, W. Chung, R. Madurawe, S. Wu, S. Ikeda, and H. Oyamatsu, "World's first monolithic 3D-FPGA with TFT SRAM over 90nm 9 layer Cu CMOS," in *IEEE Symposium on VLSI Technology*, pp. 219–220, June 2010.

[21] R. Ishihara, M. R. T. Mofrad, J. Derakhshandeh, N. Golshani, and C. I. M. Beenakker, "Monolithic 3D-ICs with single grain Si thin film transistors," in *IEEE International Conference on Solid-State and Integrated Circuit Technology (ICSICT)*, pp. 1–4, Oct 2012.

[22] F. Carta, S. M. Gates, A. B. Limanov, H. Hlaing, J. S. Im, D. C. Edelstein, and I. Kymissis, "Sequential lateral solidification of silicon thin films on low-k dielectrics for low temperature integration," *Applied Physics Letters*, vol. 105, no. 24, 2014.

[23] T. Yonehara, "Epitaxial layer transfer technology and application," in *IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference*, pp. 1–3, Oct 2015.

[24] D. C. Sekar and Z. Or-Bach, "Monolithic 3D-ICs with single crystal silicon layers," in *IEEE International 3D Systems Integration Conference (3DIC)*, pp. 1–2, Jan 2012.

[25] B. Rajendran, A. K. Henning, B. Cronquist, and Z. Or-Bach, "Pulsed laser annealing: A scalable and practical technology for monolithic 3d ic," in *IEEE International 3D Systems Integration Conference (3DIC)*, pp. 1–5, Oct 2013.

[26] K. Garidis, G. Jayakumar, A. Asadollahi, E. D. Litta, P. E. Hellstrm, and M. stling, "Characterization of bonding surface and electrical insulation properties of inter layer dielectrics for 3D monolithic integration," in *IEEE Joint International EUROSOI Workshop and International Conference on Ultimate Integration on Silicon (EUROSOI-ULIS)*, pp. 165–168, Jan 2015.

[27] O. Billoint, H. Sarhan, I. Rayane, M. Vinet, P. Batude, C. Fenouillet-Beranger, O. Rozeau, G. Cibrario, F. Deprat, A. Fustier, J. E. Michallet, O. Faynot, O. Turkyilmaz, J. F. Christmann, S. Thuries, and F. Clermidy, "A comprehensive study of Monolithic 3D cell on cell design using commercial 2D tool," in *IEEE Design, Automation and Test in Europe Conference (DATE)*, pp. 1192–1196, March 2015.

[28] S. Panth, K. Samadi, Y. Du, and S. K. Lim, "High-density integration of functional modules using monolithic 3D-IC technology," in *ASP-DAC*, pp. 681–686, Jan 2013.

[29] S. Panth, K. Samadi, Y. Du, and S. K. Lim, "Placement-Driven Partitioning for Congestion Mitigation in Monolithic 3D IC Designs," *IEEE Trans. CAD*, vol. 34, pp. 540–553, April 2015.

[30] D. H. Kim and S. K. Lim, "Physical Design and CAD Tools for 3-D Integrated Circuits: Challenges and Opportunities," *IEEE Design & Test*, vol. 32, pp. 8–22, Aug 2015.

[31] S. K. Samal, K. Samadi, P. Kamal, Y. Du, and S. K. Lim, "Full Chip Impact Study of Power Delivery Network Designs in Gate-Level Monolithic 3-D ICs," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2017.

[32] L. Peng, S. W. Kim, M. Soules, M. Gabriel, M. Zoberbier, E. Sleeckx, H. Struyf, A. Miller, and E. Beyne, "W2W permanent stacking for 3D system integration," in *IEEE Electronics Packaging Technology Conference (EPTC)*, pp. 1–4, Dec 2014.

[33] V. Dragoi, P. Czurratis, S. Brand, J. Beyersdorfer, C. Patzig, J.-P. Krugers, F. Schrank, J. Siegert, and M. Petzold, "Low Temperature Fusion Wafer Bonding Quality Investigation for Failure Mode Analysis," *ECS Transactions*, vol. 50, no. 7, pp. 227–239, 2013.

[34] J. Rubin, K. Winstel, A. Hubbard, C. Murray, K. Chung, J. Kelly, B. Khan, A. Kumar, and V. Paruchuri, "Essential edge protection techniques for successful multi-wafer stacking," in *IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S)*, pp. 1–2, Oct 2015.

[35] S. W. Kim, L. Peng, A. Miller, G. Beyer, E. Beyne, and C. S. Lee, "Permanent wafer bonding in the low temperature by using various plasma enhanced chemical vapour deposition dielectrics," in *IEEE International 3D Systems Integration Conference (3DIC)*, pp. TS7.2.1–TS7.2.4, Aug 2015.

[36] L. Di Cioccio, I. Radu, G. Gaudin, T. Lacave, F. Baudin, M. Sadaka, and T. Signamarcheix, "(Invited) Low Temperature Direct Bonding 3D Stacking

Technologies for High Density Device Integration," *ECS Transactions*, vol. 58, no. 9, pp. 17–28, 2013.

[37] D. Dussault, E. Liebscher, F. Fournell, and V. Dragoi, "Control of cavitation density through gas and acoustic uniformity in a proximity megasonic pre-bond cleaning system," 2013.

[38] S. Panth, K. Samadi, Y. Du, and S. K. Lim, "Power-performance study of block-level monolithic 3D-ICs considering inter-tier performance variations," in *DAC*, pp. 1–6, June 2014.

[39] R. Wang and K. Chakrabarty, "A design-for-test solution for monolithic 3D integrated circuits," in *ETS*, pp. 1–6, May 2016.

[40] E. J. Marinissen, T. McLaurin, and H. Jiao, "IEEE Std P1838: DfT standard-under-development for 2.5D-, 3D-, and 5.5D-SICs," in *ETS*, pp. 1–10, May 2016.

[41] S. Ravi, "Power-Aware Test: Challenges and Solutions," in *ITC*, pp. 1–10, Oct 2007.

[42] P. Girard, N. Nicolici, and X. Wen, *Power-Aware Testing and Test Strategies for Low Power Devices*. Springer, 2009.

[43] K. Chang, S. Das, S. Sinha, B. Cline, G. Yeric, and S. K. Lim, "Frequency and Time Domain Analysis of Power Delivery Network for Monolithic 3D ICs," in *ISLPED*, July 2017.

[44] S. Panth, K. Samadi, Y. Du, and S. K. Lim, "Tier-partitioning for Power Delivery vs Cooling Tradeoff in 3D VLSI for Mobile Applications," in *DAC*, 2015.

[45] A. Koneru and K. Chakrabarty, "Analysis of electrostatic coupling in monolithic 3d integrated circuits and its impact on delay testing," in *IEEE European Test Symposium (ETS)*, pp. 1–6, May 2016.

[46] A. Koneru, S. Kannan, and K. Chakrabarty, "Impact of Wafer-Bonding Defects on Monolithic 3D Integrated Circuits," in *EPEPS*, 2016.

[47] A. Koneru, S. Kannan, and K. Chakrabarty, "Impact of Electrostatic Coupling and Wafer-Bonding Defects on Delay Testing of Monolithic 3D Integrated Circuits," *J. Emerg. Technol. Comput. Syst.*, 2017.

[48] Y. Sato, S. Hamada, T. Maeda, A. Takatori, Y. Nozuyama, and S. Kajihara, "Invisible Delay Quality - SDQM Model Lights Up What Could Not be Seen," in *IEEE International Test Conference (ITC)*, Nov 2005.

[49] A. L. Loke, Z.-Y. Wu, R. Moallemi, C. D. Cabler, C. O. Lackey, T. T. Wee, and B. A. Doyle, "Constant-current threshold voltage extraction in HSPICE for nanoscale CMOS analog design," in *Synopsys Users Group Conference*, pp. 1–19, 2010.

[50] E. Augendre, M. Vinet, L. Clavelier, and P. Batude, "Integrated Circuit With Electrostatically Coupled MOS Transistors and Method for Producing Such an Integrated Circuit," Oct. 2014. US Patent No. 8,853,785.

[51] S. Vincent, I. Radu, D. Landru, F. Letertre, and F. Rieutord, "A model of interface defect formation in silicon wafer bonding," *Applied Physics Letters*, vol. 94, no. 10, 2009.

[52] P. C. Andricacos, C. Uzoh, J. O. Dukovic, J. Horkans, and H. Deligianni, "Damascene copper electroplating for chip interconnections," *IBM Journal of Research and Development*, vol. 42, pp. 567–574, Sep 1998.

[53] T. Poiroux, O. Rozeau, P. Scheer, S. Martinie, M. A. Jaud, M. Minondo, A. Juge, J. C. Barb, and M. Vinet, "Leti-UTSOI2.1: A Compact Model for UTBB-FDSOI Technologies–Part I: Interface Potentials Analytical Model," *IEEE Transactions on Electron Devices*, vol. 62, pp. 2751–2759, Sept 2015.

[54] O. D. Patterson, H. Wildman, D. Gal, and K. Wu, "Detection of Resistive Shorts and Opens using Voltage Contrast Inspection," in *SEMI/IEEE ASMC Conference*, pp. 327–333, May 2006.

[55] K. Peng, M. Yilmaz, K. Chakrabarty, and M. Tehranipoor, "Crosstalk- and Process Variations-Aware High-Quality Tests for Small-Delay Defects," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 21, pp. 1129–1142, June 2013.

[56] M. Yilmaz, K. Chakrabarty, and M. Tehranipoor, "Test-Pattern Selection for Screening Small-Delay Defects in Very-Deep Submicrometer Integrated Circuits," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 29, pp. 760–773, May 2010.

[57] Mango, C. T. Chao, L. C. Wang, and K.-T. Cheng, "Pattern selection for testing of deep sub-micron timing defects," in *IEEE Design, Automation and Test in Europe Conference (DATE)*, vol. 2, pp. 1060–1065 Vol.2, Feb 2004.

[58] A. Koneru, S. Kannan, and K. Chakrabarty, "A Design-for-Test Solution for Monolithic 3D Integrated Circuits," *International Conference on Computer Design*, vol. 00, pp. 685–688, 2017.

[59] A. Koneru, S. Kannan, and K. Chakrabarty, "A design-for-test solution based on dedicated test layers and test scheduling for monolithic 3d integrated circuits," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2018.

[60] A. Koneru, S. Kannan, and K. Chakrabarty, "Testing monolithic three dimensional integrated circuits," Nov. 2017. US Patent Application no. 15/801,380.

[61] J. Broz, D. Aaberge, S. Hetsugi, G. Gerard, and J.-P. Gibaux, "Advanced Vertical Technologies for Low Damage Probing of Bumps, Pillars, and Pads," June 2017.

[62] T. Nakazawa, D. V. Kulkarni, and O. A. Martin, "IC substrate package yield prediction model and layer level risk assessment by design analysis," *IEEE Trans. Semiconductor Manufacturing*, vol. 29, no. 3, pp. 257–262, 2016.

[63] E. Singh, "Analytical modeling of 3D stacked IC yield from wafer to wafer stacking with radial defect clustering," in *VLSID*, pp. 26–31, Jan 2014.

[64] X. Wu, Y. Chen, K. Chakrabarty, and Y. Xie, "Test-access mechanism optimization for core-based three-dimensional SOCs," in *IEEE International Conference on Computer Design*, pp. 212–218, Oct 2008.

[65] B. Noia, K. Chakrabarty, and E. J. Marinissen, "Optimization Methods for Post-Bond Testing of 3D Stacked ICs," *Journal of Electronic Testing*, vol. 28, pp. 103–120, Feb 2012.

[66] S. Bahukudumbi and K. Chakrabarty, "Test-Length and TAM Optimization for Wafer-Level Reduced Pin-Count Testing of Core-Based SoCs," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 28, pp. 111–120, Jan 2009.

[67] E. J. Marinissen and Y. Zorian, "Testing 3D chips containing through-silicon vias," in *ITC*, pp. 1–11, Nov 2009.

[68] R. Weerasekera, L.-R. Zheng, D. Pamunuwa, and H. Tenhunen, "Extending systems-on-chip to the third dimension: performance, cost and technological tradeoffs," in *ICCAD*, pp. 212–219, Nov 2007.

[69] Z. Or-Bach, B. Cronquist, Z. Wurman, I. Beinglass, and A. Henning, "Precision bonders - A game changer for monolithic 3D," in *SOI-3D-Subthreshold Microelectronics Technology Unified Conference*, 2014.

[70] E. J. Marinissen, S. K. Goel, and M. Lousberg, "Wrapper design for embedded core test," in *IEEE International Test Conference*, pp. 911–920, 2000.

[71] V. lyengar et al., “Test access mechanism optimization, test scheduling, and tester data volume reduction for system-on-chip,” *IEEE Transactions on Computers*, vol. 52, pp. 1619–1632, Dec 2003.

[72] A. Koneru and K. Chakrabarty in *IEEE Press*.

[73] M. Abramovici and M. A. Breuer, “Fault Diagnosis Based on Effect-Cause Analysis: An Introduction,” in *ACM/EDAC/IEEE Design Automation Conference*, pp. 69–76, June 1980.

[74] B. Noia and K. Chakrabarty, “Pre-bond probing of TSVs in 3D stacked ICs,” in *IEEE International Test Conference*, Sept 2011.

[75] K. Shibin, V. Chickermane, B. Keller, C. Papameletis, and E. J. Marinissen, “At-Speed Testing of Inter-Die Connections of 3D-SICs in the Presence of Shore Logic,” in *IEEE Asian Test Symposium*, pp. 79–84, Nov 2015.

[76] S. Makar, “A layout-based approach for ordering scan chain flip-flops,” in *International Test Conference*, Oct 1998.

[77] J. Munkres, “Algorithms for the Assignment and Transportation Problems,” *J. of the Society for Industrial and Applied Mathematics*, 1957.

[78] S. Deutsch and K. Chakrabarty, “Contactless Pre-Bond TSV Test and Diagnosis Using Ring Oscillators and Multiple Voltage Levels,” *IEEE Trans. CAD*, vol. 33, pp. 774–785, May 2014.

[79] W. Zhao and Y. Cao, “New generation of predictive technology model for sub-45nm design exploration,” in *ISQED*, pp. 6 pp.–590, March 2006.

[80] S. Panth, K. Samadi, Y. Du, and S. K. Lim, “Shrunk-2-D: A Physical Design Methodology to Build Commercial-Quality Monolithic 3-D ICs,” *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 36, pp. 1716–1724, Oct 2017.

[81] S. Panth, K. Samadi, Y. Du, and S. K. Lim, “Design and CAD Methodologies for Low Power Gate-Level Monolithic 3D ICs,” in *ISLPED*, 2014.

[82] S. Das, P. Whatmough, and D. Bull, “Modeling and characterization of the system-level Power Delivery Network for a dual-core ARM Cortex-A57 cluster in 28nm CMOS,” in *ISLPED*, pp. 146–151, July 2015.

[83] X. Huang, A. Kteyan, S. X. Tan, and V. Sukharev, “Physics-Based Electromigration Models and Full-Chip Assessment for Power Grid Networks,” *TCAD*, Nov 2016.

[84] I. A. Blech, "Electromigration in thin aluminum films on titanium nitride," *Journal of Applied Physics*, vol. 47, no. 4, pp. 1203–1208, 1976.

[85] H. Cook and K. Skadron, "Predictive design space exploration using genetically programmed response surfaces," in *DAC*, 2008.

[86] A. Ramalingam, S. V. Kodakara, A. Devgan, and D. Z. Pan, "Robust analytical gate delay modeling for low voltage circuits," in *ASPDAC*, Jan 2006.

Biography

Abhishek Koneru received the B.Tech. and M.Tech. degrees in electrical engineering from the Indian Institute of Technology, Kanpur, India, in 2014. He earned his Ph.D. in electrical and computer engineering from Duke University in 2019. He was an intern with Intel Corporation, Santa Clara, CA, USA, TSMC Technology Inc., San Jose, CA, USA, and Apple Inc., Cupertino, CA, USA. His research interests include testing and design-for-testability of monolithic 3D integrated circuits, and machine learning. He was a recipient of Eiffel Excellence Scholarship from the French Ministry of Foreign Affairs and International Development in 2017 and a scholarship from The Indian Central Board of Secondary Education in 2009.