Title

Tackling nanoscale IC failures through noise-aware testing and silicon debugging

Permalink

https://escholarship.org/uc/item/6zj4d9mw

Authors

Chen, Mingjing
Chen, Mingjing

Publication Date

2012

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Tackling Nanoscale IC Failures through Noise-aware Testing and Silicon Debugging

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Computer Science (Computer Engineering)

by

Mingjing Chen

Committee in charge:

Professor Alex Orailoglu, Chair
Professor Chung-Kuan Cheng
Professor Sadik Esener
Professor Ian G. Harris
Professor William E. Howden

2012

The dissertation of Mingjing Chen is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2012

DEDICATION

To my family.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor, Professor Alex Orailoglu, for the great guidance he provided during my Ph.D years. Through multiple drafts, many debates and many long nights, his guidance has proved to be invaluable. His enthusiasm and intelligence in research is always a model for me to follow in my future career.

Many thanks go to the other professors I have met during the years in UCSD, as well as my academic siblings in the Architecture, Reliability, and Testing (ART) group, including Chengmo Yang, Baris Arslan, Garo Bournoutian, Wenjing Rao.

I also want to thank my friends in UCSD, for sharing with me the unforgettable campus life.

I would also like to thank my parents for always encouraging me with their best wishes.

In the end, I would like to thank my wife, Zeyan Li, for always being there. Without her love, support, tolerance and advice, I would never have gone so far.

The text of Chapter 4, is in part a reprint of the material as it appears in *M. Chen and A. Orailoglu, "Scan power reduction in linear test data compression scheme," International Conference on Computer-Aided Design, 2009*; and in *M. Chen and A. Orailoglu, "Scan power reduction for linear test compression schemes through seed selection," IEEE Transactions on VLSI.* The dissertation author was the primary researcher and author of the publications [16] and [15].

The text of Chapter 5, is in part a reprint of the material as it appears in *M. Chen and A. Orailoglu, "Cost-effective IR-drop failure identification and yield recovery through a failure-adaptive test scheme," Design, Automation and Test in Europe, 2010*; and in *M. Chen and A. Orailoglu, "Examining timing path robustness under wide-bandwidth power supply noise through multi-functional-cycle delay test," submitted to IEEE Transactions on VLSI.* The dissertation author was the primary researcher and author of the publications [17] and [13].

The text of Chapter 6, is in part a reprint of the material as it appears in

M. Chen and A. Orailoglu, "Diagnosing scan chain timing faults through statistical feature analysis of scan images," Design, Automation and Test in Europe, 2011; in *M. Chen and A. Orailoglu, "Diagnosing scan clock delay faults through statistical timing pruning," Design Automation Conference, 2011*; and in *M. Chen and A. Orailoglu, "On diagnosis of timing failures in scan architecture," IEEE Transactions on CAD*. The dissertation author was the primary researcher and author of the publications [18], [19] and [14].

VITA

2002	B. S. in Microelectronics, Tsinghua University, Beijing
2005	M. S. in Electronic Science and Technology, Tsinghua University, Beijing
2006-2010	Teaching Assistant, Department of Computer Science and Engineering, University of California, San Diego
2005-2012	Research Assistant, Department of Computer Science and Engineering, University of California, San Diego
2012	Ph. D. in Computer Science (Computer Engineering), University of California, San Diego

PUBLICATIONS

Journal papers

M. Chen and A. Orailoglu, "Scan power reduction for linear test compression schemes through seed selection," *IEEE Trans. on VLSI*, accepted

C. Yang, M. Chen and A. Orailoglu, "Squashing code size in microcoded IPs while delivering high decompression speed," *Design Automation for Embedded Systems*, vol. 14, no. 3, pp. 265-284, 2010

M. Chen and A. Orailoglu, "On diagnosis of timing failures in scan architecture," *IEEE Trans. on CAD*, accepted

M. Chen and A. Orailoglu, "Examining timing path robustness under wide-bandwidth power supply noise through multi-functional-cycle delay test," submitted to *IEEE Trans. on VLSI*

Conference papers

M. Chen and A. Orailoglu, "Diagnosing scan clock delay faults through statistical timing pruning," *IEEE/ACM Design Automation Conference*, 2011

M. Chen and A. Orailoglu, "Diagnosing scan chain timing faults through statistical feature analysis of scan images," *IEEE/ACM Design, Automation and Test in Europe*, 2011

M. Chen and A. Orailoglu, "VDDmin test optimization for overscreening minimization through adaptive scan chain masking," *IEEE VLSI Test Symposium*, 2010

M. Chen and A. Orailoglu, “Cost-effective IR-drop failure identification and yield recovery through a failure-adaptive test scheme,” *IEEE/ACM Design, Automation and Test in Europe*, 2010

M. Chen and A. Orailoglu, “Scan power reduction in linear test data compression scheme,” *IEEE/ACM International Conference on Computer-Aided Design*, 2009

C. Yang, M. Chen and A. Orailoglu, “Squashing microcode stores to size in embedded systems while delivering rapid microcode accesses,” *CODES+ISSS*, 2009

M. Chen and A. Orailoglu, “Flip-flop hardening and selection for soft error and delay fault resilience,” *IEEE International Symposium on Defect and Fault Tolerance in VLSI Systems*, 2009

M. Chen and A. Orailoglu, “Deflecting crosstalk by routing reconsideration through refined signal correlation estimation,” *ACM Great Lakes Symposium on VLSI*, 2009

M. Chen and A. Orailoglu, “Test cost minimization through adaptive test development,” *IEEE International Conference on Computer Design*, 2008

M. Chen and A. Orailoglu, “Circuit-level mismatch modeling and yield optimization for CMOS analog circuits,” *IEEE International Conference on Computer Design*, 2007

M. Chen and A. Orailoglu, “Improving circuit robustness with cost-effective soft-error-tolerant sequential elements,” *IEEE Asian Test Symposium*, 2007

M. Chen, H. Haggag and A. Orailoglu, “Decision tree based mismatch diagnosis in analog circuits,” *IEEE VLSI Test Symposium*, 2006

ABSTRACT OF THE DISSERTATION

Tackling Nanoscale IC Failures through Noise-aware Testing and Silicon Debugging

by

Mingjing Chen

Doctor of Philosophy in Computer Science (Computer Engineering)

University of California, San Diego, 2012

Professor Alex Orailoglu, Chair

The continued device scaling trend and the aggressive integrated circuit design style have shifted the major device failure mechanism from stuck-at fault types to marginal failures induced by timing uncertainty and signal noise. The production test methodologies currently employed by industry, however, are still based on the traditional structural test schemes that focus on the detection of permanent defects, failing to account for emerging failure mechanisms in nanometer scale designs. The inability of current test methodologies in adapting to the failure mechanism shift imposes critical challenges to the IC providers, mainly observed as significant product quality degradation and yield loss. To make things worse, the marginal failures result in highly ambiguous failure syndromes, invalidating traditional assumptions employed in silicon debugging. The degraded test quality and yield, combined with inaccurate failure diagnosis, lead to a lengthened design-fabrication-debugging cycle needed for ramping up the yield and quality for final production, significantly slowing down the time-to-market and boosting the overall product cost.

Maintaining high quality yet low cost production test for nanometer scale integrated circuits necessitates a comprehensive examination of marginal failure scenarios while minimizing yield loss. Reducing the time-to-market cycle relies on an accurate identification of marginal failure locations and causalities to pinpoint the design and fabrication weaknesses that have gross quality impact. The challenges, though, are the resolution to the paradox between overscreening and underscreening that are simultaneously taking place in today's industrial testing practice, and the extraction of sensible diagnostic signals from highly ambiguous fault behaviors of marginal failures. The presented thesis work overcomes these challenges through the proposition of an innovative marginal failure aware test and diagnosis scheme, capable of thoroughly targeting the functional mode failure scenarios with a low cost structural test platform and the accurate identification of failure-induced feature change in large volume test data. A comprehensive production ramp-up flow, constructed based on the proposed test and diagnosis schemes, is furthermore presented to guide the silicon debugging, test optimization, and yield/quality learning activities, so as to minimize the time-to-market.

From a technical point of view, this thesis work analyzes the power ground noise in functional and testing modes and its impact on circuit timing robustness, with a focus on the differentiation of the functional mode timing failures from the pure testing mode ones, thus enabling a clear decomposition of the noise treatment strategies for different operation scenarios. A set of tightly-coupled approaches, including 1) noise resilience in testing related circuitry for overscreening minimization, 2) approximation of worst-case functional mode noise in structural testing for marginal timing failure detection, and 3) diagnosis of noise-induced timing failure diagnosis in scan paths and scan clock trees for design optimization, are presented to attain the overall goal of high yield, low test escape rate, and fast silicon re-spin. These techniques are developed with the consideration of enabling a seamless adaptation of industrial flows by delivering maximal compatibility to mainstream design-for-testability architectures and testing platforms employed in nanometer scale designs. The successful incorporation of these techniques will significantly expedite the silicon production ramp-up process with highly reduced risk and cost.

Chapter 1

Introduction

Manufacturing testing and failure analysis have been utilized for decades in semiconductor industry as the most important approach for verifying and improving the quality of integrated circuit (IC) products. The fast device integration trend as projected by Moore's law, although strongly pushing the advancement in circuit performance, inevitably increases the difficulty in maintaining high device quality in production. One of the most affected aspects in the testing domain consists of the testing and diagnosis of digital circuit failures, as the device scaling and integration are most aggressively performed in digital circuits. The ever-increasing transistor density, the diminishing pin-to-transistor ratio, and the high process variability in small devices all contribute to a highly elevated failure rate in today's digital integrated circuits. On the other hand, the fierce competition in the digital product marketplace strongly forces the IC vendors to reduce the defective device rate measured as *Defective Parts Per Million* (DPPM). The development of a robust and efficient test and failure analysis methodology, therefore, becomes a vital task that determines the success of digital IC products.

Traditionally, the digital testing and diagnosis approaches have mainly focused on the detection of gross manufacturing defects. Fault models based on gross defects, such as stuck-at faults and transition faults, have been utilized to develop test application schemes and test generation algorithms. Fault coverage, consequently, has constituted a natural metric in the industry for measuring the test quality. For a long period of time, such a scheme, as a stand-alone process separated from the chip design practice, was able to guarantee a very low DPPM, as chips fabricated in old technology nodes were mainly sensitive to gross fabrication defects rather than second-order effects such as process variability and design weakness.

As device feature sizes scale to the nanometer range, the testing and failure analysis of digital circuits have started to tightly couple with the design process during product development, as a number of non-traditional and design-related failure mechanisms have gradually become the dominating factors that impact product quality. The shift in failure mechanisms significantly degrades the quality of standard testing and failure analysis flow, resulting in a low yield and high test escape rate. This in turn slows down the time-to-market, as the IC vendors typically have to perform many more re-spins before ramping up to high volume production. The increased difficulty in performing correct testing and diagnosis furthermore lengthens the cycle between re-spins, as more effort and time are needed to collect the test escape data and identify the failure causality. Given the early time-to-revenue pressure, the learning and optimization of the test and diagnosis methodologies constitute a key task during the industrial product development cycle.

1.1 Need for design information in test and diagnosis

1.1.1 Failure mechanism shift

A semiconductor device can fail the test due to multiple reasons. The failure can possibly be the result of a design weakness, a defect in the fabrication process, or even a false alarm due to an inappropriately applied test. Over a long period of time, the fabrication defects have been considered the major failure mechanisms, since designs with old technology nodes were relatively simple and the design houses typically had rather robust verification strategies to rule out potential design risks. The scaling of the device feature size pushes the utilization of much more aggressive design styles for performance enhancement. This, unfortunately, increases the difficulty in verifying the design robustness under various operation conditions, resulting in a number of reliability concerns.

From a design perspective, a number of power and timing related changes can be evidently observed as a result of technology scaling. First of all, the technology node change is typically accompanied by the design parameter adjustment such as higher clock frequency and lower operation voltage, which results in an increased power demand and rapidly reduced design margins. Secondly, the device integration trend significantly increases the difficulty in routing the signal and power ground nets. To provide enough routing flexibility, more metal layers have to be incorporated in physical design. However, the increase in routing space barely matches the needs. Such a routing space shortage mainly impacts global signals such as power ground nets and clocks, as a strong degradation in routing balancing and an increased routing length can be observed on their distribution networks. Thirdly, the incorporation of multiple clock and power domains in the circuit further increases the design verification complexity as the impact of the inter-domain variability is even harder to model and predict.

The aggressive voltage scaling trend significantly reduces the noise margin in VLSI circuits. On the other hand, excessive power density in nanometer chips debilitates the delivery of sufficient current by the power supply network, causing significant power ground noise in the circuit. Current supply voltage variations are directly implicated in high timing uncertainties. It has been reported that a $10 - 15\%$ voltage drop can induce a $20 - 30\%$ increase in gate delay [47]. This constitutes a critical reliability challenge for modern VLSI systems designed with very tight timing slacks.

The current commercial design flow handles power ground noise through appropriate power budgeting and timing closure at a lower voltage corner, say, 10% lower than the nominal operation voltage. However, the noise amplitude in modern designs typically ranges from tens to even hundreds of millivolts, resulting in circuits frequently operating in a situation that is far from the range where timing is closed in design. As a result, marginal timing failures induced by power ground noise become one of the dominating failure mechanisms in nanometer scale circuits, even for designs that have been strictly verified before tape-out [2].

In contrast to traditional failure types such as stuck-at and transition faults, the marginal failures as outlined above exhibit a set of highly distinct characteristics. The manifestation of such failures is inherently a probabilistic behavior, as multiple factors such as voltage and temperature fluctuation can disturb timing in unpredictable ways. The manifestation condition of marginal failures is also much more complicated than the traditional ones, as the activation and observation of marginal failures not only rely on appropriate logic values on the associated signal nets, but necessitate a certain noise environment to be established. Moreover, the syndromes of marginal failures are highly ambiguous and unpredictable. For example, the device delay variation induced by power ground noise can manifest itself as intermittent and hybrid timing failures. The extra delay imposed on critical paths can possibly raise setup-time violations, whereas hold-time violations also have high likelihood of taking place if the clock on the launching flip-flops are relatively delayed. It is expected that these issues will become even more severe when the fabrication technology scales to 28nm and below, as the design timing margin in these technology nodes can possibly reduce to tens of picoseconds, far less than the noise induced timing variation.

1.1.2 Breaking through structural test limitation

Ideally, a test methodology should examine exactly all the cases that can occur during the functional operation of a chip, no more, no less. An exhaustive functional test program would be the best approximation of such an ideal test; nevertheless, it is a highly impractical option due to the difficulty in developing such a program and the prohibitive cost of applying such a test. Therefore, structural test has been developed as an alternative to functional test. Instead of examining the chip from a functional perspective, the structural test checks whether the modeled faults on the internal nodes of a chip can be correctly activated and observed, so as to flag any failures that contaminate the signal values. Such a test idea, especially the scan-based test, enables easy *automated test pattern generation*

(ATPG), rapid test application as well as high fault coverage, thus being utilized in almost the entire semiconductor industry as the sole solution for production test.

Being a model-based test system, the structural test shares the same shortcomings with any other test system, that is, it can only approach the ideal test goal, but never perfectly match what is desired. The relationship between a structural test and an "ideal" functional test can be illustrated in the Venn diagram shown in Figure 1.1. The two circles in this graph denote the failure coverage of functional and structural tests, respectively. The common region **A** stands for the failures that are covered by both tests, whereas regions **B** and **C** denote the failures that only manifest themselves in functional test or structural test. If a failure can be detected by both tests, then it is a true functional failure that can be flagged during production screening. Chips with such failures should definitely be discarded to maintain product quality. If a failure can only be detected by the functional test, it will escape from the production screening, thus resulting in an increased DPPM. On the other hand, if a failure only manifests itself in the structural test, then it is from a functional perspective a redundant failure which will be unnecessarily failed by production screening. Discarding chips with solely such failures would result in yield loss. Apparently, a good structural test plan needs to maximize region **A** and minimize regions **B** and **C**.

A further investigation on the failure characteristics reveals that the failure distribution in the aforementioned Venn diagram is skewed. The manifestation of traditional fault types such as stuck-at and transition faults only requires the faults being logically sensitized and propagated for observation. Such a fault manifestation condition can be easily fulfilled in both functional and structural tests. Therefore, the majority of traditional faults is distributed in the common region **A** of the Venn diagram. Marginal failures, such as noise-induced timing violations, though necessitate a much more complex manifestation condition, as a certain noise profile on specific timing paths needs to be developed for fault activation. Since the noise distribution and the status of timing paths in test mode can differ significantly from those of the functional mode, the marginal failures have a high

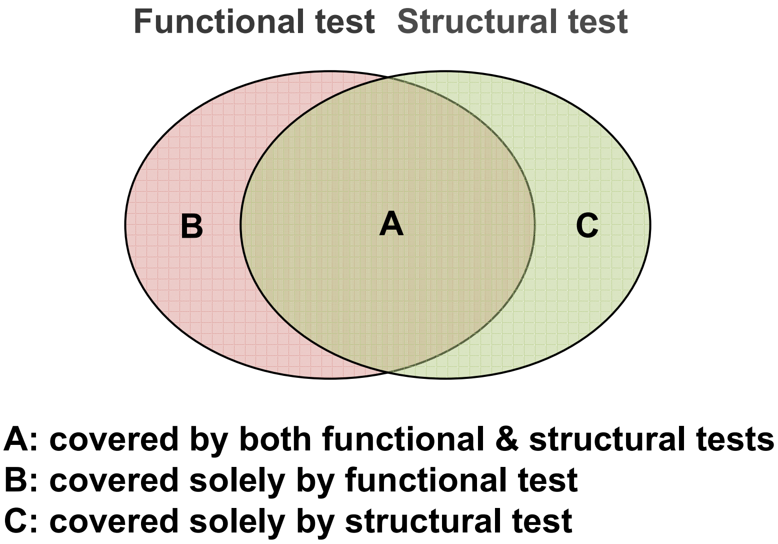

Functional test
Structural test
B
A
C
A: covered by both functional & structural tests
B: covered solely by functional test
C: covered solely by structural test

Figure 1.1: Discrepancy between structural test and “ideal” functional test

likelihood to be distributed in either region **B** or **C** of the Venn diagram, thus constituting the major reason for test escape or yield loss.

In old technology nodes where the design margins are large, the signal noise and the associated timing uncertainty can be reasonably well tolerated in both the testing and functional modes, thus resulting in a rather small area of regions **B** or **C**. This ensures a high test quality and low yield loss cost for chips fabricated with these technologies. The failure mechanism shift that occurs with device scaling, as outlined previously, results in the marginal failures being the dominant failure scenarios for nanometer scale designs. This significantly increases the area of regions **B** and **C**, leading to a highly elevated DPPM and yield loss.

Maintaining high test quality and low test cost forces the test methodology to adapt to the failure mechanism shift. Yet such a change is subject to a strong practical constraint that needs to be enforced for current and prospective IC technologies, that is, the compatibility of the test methodology innovation with the structural test platform needs to be guaranteed in order to reuse the expensive hardware and software infrastructure currently employed in the test floor. Such a goal can be achieved if the structural test can approximate the functional mode behavior so as to detect functional-mode-only failures and avoid the manifestation of test-mode-only failures. The approximation of functional operation, though, requires the understanding of design knowledge that is tightly coupled with the behavioral and physical level behavior of the circuit. From a behavioral design perspective, the differentiation between functional-mode and test-mode paths can help determine the noise optimization targets for different timing paths in the circuit. The estimation of noise profile and the resulting timing impact can be attained through the use of physical design information such as the clock tree structure, power distributed network, and the layout parasitics. A coherent ATPG and test application scheme, guided by the aforementioned design knowledge, is needed to model the circuit behavior in different modes and generate/apply such tests.

1.1.3 Exploring design perspectives in diagnosis

The failure mechanism shift not only raises the need for test methodology adaptation, but also degrades and even invalidates the traditional failure analysis techniques. With the high power ground noise and the associated timing uncertainty in nanometer scale designs, multiple faults with mixed timing violation types can exist. The intermittent fault manifestation, and the possible interaction between multiple faults, result in a highly ambiguous fault syndrome, making it almost impossible to create clear fault dictionary for diagnosis. Furthermore, the syndromes of marginal faults can be highly conflicting with the ones of gross defects. Therefore, traditional diagnosis approaches based on gross fault assumptions can lead to incorrect conclusions in a marginal failure scenario.

The key to resolving a diagnostic problem is the establishment of a small set of valid failure hypotheses and the extraction of strong signals that can differentiate distinct hypotheses. The design knowledge, especially the physical design information, is needed once again to attain these goals in the ambiguous diagnosis space induced by marginal failures. For example, the statistical timing information can be utilized to estimate the failure likelihood of different circuit nodes in the volume diagnostic data, thus filtering out the ambiguity induced by the randomness in individual failure manifestation. Using the design information to guide diagnosis not only avoids the search of irrational failure scenarios, but provides information regarding the criticality of valid failure hypotheses. This enables the designers to focus on the most yield-critical design weakness during re-spin, significantly expediting the time-to-revenue.

1.2 Challenges to be addressed

The development of the aforementioned marginal failure aware test and diagnosis scheme encounters a number of challenges.

1.2.1 Paradox between overtesting and undertesting

Due to the difficulty of pre-silicon verification of marginal failures, at-speed test is incorporated as a standard procedure and plays an important role in today's manufacturing test flow in order to screen out any timing-related failures. Nonetheless, due to the discrepancy between the test mode and functional mode outlined previously, this test potentially yields misleading results from a product quality point of view. More specifically, two major disadvantages can be observed for conventional at-speed test.

- The test operations, especially the scan steps, result in highly non-functional transitions of circuit states, resulting in a large deviation of the noise profile from the functional operation. In scan mode, the density of device toggling activities is typically much higher than that of functional mode, potentially leading to an overtesting of good parts. This in turn causes yield loss of the products.

- The noise profile of modern chips is a compound behavior of high, middle and low frequency noises introduced by different components of the design. The single at-speed capture cycle can only target the high speed noise of the chip, as it terminates before the middle and low frequency noise has time to fully develop. On the other hand, the regular operation of the chip needs to execute through a sequence of functional cycles, which gives sufficient time for full noise development. As a result, current at-speed test can also lead to under-testing of marginal parts, as it fails to examine the worst-case noise situation. This introduces a risk of test escape and increases the DPPM of the products.

Delivering high quality test for noise-induced timing failures necessitates the resolution to the paradox between two somewhat conflicting requirements: avoiding over-testing incurred by excessive non-functional noise and reducing test escapes incurred by the lack of functional noise coverage. This raises a number of challenging questions for production test development. Which set of timing paths mainly contributes to overtesting? Which set of timing paths has not been examined thoroughly during at-speed testing? How to organize the design information during test generation? How to transform the test patterns to change the noise level on a target timing path? How to account for other ATPG constraints, such as test compression and compaction, during test pattern transformation? All these issues need to be clearly addressed in order to develop a test program that can fulfill both requirements.

1.2.2 Noise estimation and control in test generation

Developing appropriate tests for marginal failures necessitates accurate estimation of power ground noise. Traditional noise aware test generation techniques mainly utilize the weighted switching activity (WSA) metric for noise estimation. This metric has a high correlation with the current amplitude drawn in each cycle, thus being reasonably accurate in estimating the amplitude of high frequency noise. However, as a cycle-based current metric, it does not take into consideration the resonance effect of the noise, thus being incapable of estimating the middle frequency noise which spans multiple cycles.

In nanometer scale ICs which typically operate within a high clock frequency range, the development of the worst-case noise in functional operation requires multiple at-speed cycles. An accurate noise estimation framework, considering the noise accumulation effect across multiple cycles, is thus needed to guide test generation. In contrast to the simple WSA metric, this estimation necessitates the consideration of multiple factors, such as the noise resonance period, the characteristics of the power distribution network, the positions of the cells in the layout, and so on. Creating a mathematical model that accurately quantifies the impact of the aforementioned factors is a critical technical challenge that needs to be overcome.

An even more challenging issue is the development of the desired noise level during test generation. For example, delivering the highest test quality necessitates the approximation of the worst-case noise on the targeted functional-mode path. Yet there are multiple degrees of freedoms during test generation that can impact the ultimate noise level. Firstly, a test cube contains a large set of unspecified bits. Different filling strategies of these unspecified bits will result in a completely distinct noise profile in the circuit. Secondly, the number of at-speed cycles used in the test determines the length of the noise accumulation process, thus also having a strong impact on the noise profile. A test generation engine needs to examine all these variables in order to identify the test pattern that generates the expected noise level.

Both the noise estimation and test generation processes face the challenge of high computational cost. The search of the most appropriate test pattern is in fact an intertwined process of noise estimation and test transformation. The large number of unspecified bits in the test cube results in a huge search space, whose size exponentially increases with the number of scan cells in the design, not to mention the need for examining the impact of each test pattern candidate across multiple clock cycles. If dynamic noise simulation is used during the search process, the overall search time cost would become prohibitive. Improving the efficiency of noise estimation and test transformation methodology with negligible impact on accuracy, thus becomes imperative for this test scheme to be practically applicable in the industrial flow.

1.2.3 Extracting clear diagnostic information

Most traditional diagnosis approaches rely on logic tracing and simulation to establish a mapping between the fault hypotheses and syndromes. The effectiveness of such a somewhat deterministic strategy highly depends on the uniqueness and the strictness of the assumed fault behavior. If the fault behavior can be strictly simulated, accurate fault information can be derived by analyzing the logic information extracted from the scan patterns. The permanent fault assumption defines a highly strict behavior model. Therefore, the utilization of the logic analysis strategy on such models delivers a good diagnostic resolution.

Traditional methodologies approach the diagnosis problem using both positive and negative information. They not only use the observed failure syndromes to identify the range where the fault locates, but also utilize the correct portion of the circuit output to exclude certain regions being the fault candidate. However, the failure mechanism shift has invalidated the fundamental assumptions of traditional approaches. For nanometer designs, marginal defects introduced by design and process weakness have become the dominating factor for yield loss, thus constituting one of the top priorities in silicon debugging. The intermittent manifestation and the multitude of such faults result in a highly ambiguous fault behavior, significantly increasing the difficulty in diagnosis. The inherent ambiguity of the failure mechanism observed in nanometer designs reduces the information that can be extracted through logic analysis, leading to a degraded diagnosis resolution, as detailed subsequently. One contributing factor to the ambiguity consists of the unpredictable fault manifestation. Under this situation, a correct circuit output does not necessarily mean the absence of faults. In fact, it can be due to the quietness of the fault in an individual diagnosis run. Another source of ambiguity stems from the interaction between multiple faults. Due to the fault interaction, the fault effect of individual defects can possibly be canceled. This once more shows that the absence of failure syndromes cannot be used for fault exclusion, significantly reducing the information that can be extracted through logic analysis.

The degraded effectiveness of logic analysis results in a large ambiguity space that requires further pruning. Overcoming this challenge necessitates solutions to the following fundamental questions:

- How to extract extra information to guide the pruning of the ambiguity space?

- How to explore the large pruning space in a computationally efficient manner?

1.3 Contributions of This Thesis

To address the industrial need for a fast-to-revenue silicon verification plan in current and forthcoming decades, this thesis work focuses on the development of a marginal-failure-aware test and failure analysis framework. Multiple innovative techniques, which resolve the challenges outlined in the last section and provide the theoretical underpinnings of the proposed framework, are presented in this thesis. The contributions of this work are summarized as follows.

- A scan-mode noise minimization technique for scan architecture utilizing linear compression circuitry, the most widely used DFT architecture in industry. This methodology is capable of identifying the noise-friendly compression seeds of a linear compression system in the face of the strong ATPG constraints imposed by the linear compression scheme, thus minimizing the overtesting risk induced by the excessive scan-mode noise and timing uncertainty. A concurrent compression/compaction framework is furthermore presented based on the noise minimization technique to ensure a noise-safe scan process with negligible impact on the test volume compression ratio.

- A multi-functional-cycle delay test scheme to detect the marginal timing failures under worst-case functional noise situation. This test scheme applies a pre-calculated number of at-speed functional cycles, with the last cycle sensitizing the target path. This strategy not only enables the worst-case noise development, but results in a more "functional" circuit state transition in the capture cycle, thus creating a noise profile that is highly similar to the true functional one. This test delivers higher delay test quality at a lower risk of false alarm, thus benefiting the product cost from both the yield improvement and the DPPM reduction perspectives.

- A failure analysis methodology, capable of diagnosing both permanent and marginal timing failures in scan circuitries. This diagnosis technique closely approximates the behavior of the realistic failure mechanisms observed in silicon, and presents a new perspective of understanding and analyzing the syndrome of the marginal failures. The diagnostic results not only pinpoint the physical region of the failures, but indicate the relative criticality of each failure, thus providing a strong guidance to the design optimization and re-spin task.

- A fast-to-revenue silicon optimization framework, capable of coordinating the test optimization, failure analysis and design re-spin tasks for early production. This framework explores the possibility of performing failure-aware design verification using the diagnostic information before each tape-out. Such a highly guided process can significantly reduce the number of re-spins.

All the aforementioned techniques are developed with the consideration of enabling algorithmic automation and delivering compatibility with the mainstream hardware/software infrastructure so as to ensure a seamless embedding of these techniques into the industrial flow. The successful utilization of these techniques in industry will enable a low cost yet high quality production with a highly reduced time-to-revenue cycle for current and next generation integrated circuits.

1.4 Roadmap

The rest of the thesis is organized as follows. Chapter 2 reviews the state-of-art and analyzes the limitations of existing solutions of handling noise in testing and diagnosis. Chapter 3 presents an overview of the envisioned marginal-failure-aware test and failure analysis framework, focusing on the utilization of design knowledge in guiding test optimization and diagnostic feature extraction. Chapter 4 introduces a scan-mode noise minimization technique which reduces the yield loss due to overtesting of test-mode-only paths. Chapter 5 focuses on the under-testing issue and presents a technique to detect functional timing path failures induced by wide-bandwidth power ground noise. The collaborative use of techniques in Chapters 4 and 5 engenders a comprehensive test plan that maximally approximates the effectiveness of an “ideal” functional test in a low cost structural test platform. Chapter 6 presents a diagnosis methodology, targeting the marginal timing failures in the scan architecture. Chapter 7 presents a flow of integrating the proposed test and failure analysis approaches into a fast-to-revenue silicon verification framework, with a special focus on the failure-aware design verification guided by diagnostic information. Finally, Chapter 8 summarizes the proposed framework and subsequently outlines a set of possible future research directions.

Chapter 2

Related Work

While the failure mechanisms of the VLSI circuits become increasingly complex, the development of innovative test and diagnosis methodologies for addressing these emerging failures constitutes a focal point in the VLSI test community. Researchers from both industry and academia have started to address this issue from various perspectives.

The research practice on this topic is mainly driven by the industrial observation and needs. First of all, it has been observed in almost all industrial designs that the scan activity induces an exceedingly high power and noise level due to the high toggling rate, resulting in a large number of scan chain timing failures. Therefore, researchers have focused on the development of scan power/noise reduction techniques. Various approaches, ranging from test software optimization to hardware design modification, have been proposed to tackle this problem. In addition to the scan issue, the noise-induced timing failures in the capture phase of the structural testing have become an increasingly critical problem, especially in high frequency designs. The noise treatment in the capture phase is even much more complicated than that in the scan phase, as the determination of the proper noise level and distribution in the capture phase is a highly controversial issue. A number of techniques, with possibly conflicting perspectives, have been proposed to handle the noise as well as the timing failures in capture phase. Some approaches focus on the minimization of capture noise to avoid false alarms, whereas the others aim to increase the noise so as to detect more faults. The need for controlling noise during testing has also motivated the research on test noise estimation. A variety of noise estimation models, at distinct abstraction levels, has been proposed to guide the ATPG process. The effort in ATPG and DFT optimization outlined above mainly helps improve test quality and reduce yield loss. Nonetheless, the improvement of the design and fabrication quality necessitates a deep understanding of the root causality of the emerging failures in the chip. This has motivated a large amount of research work devoted into the diagnosis of timing failures, especially the ones occurring in scan circuitries. These techniques aim to identify the locations of the failures, thus providing clues about possible design and fabrication weakness.

Although the previous efforts exhibit a certain level of effectiveness in addressing some representative marginal failures observed in a few application domains, they fall short of resolving the fundamental challenges faced by the industry, namely, the discrepancy between functional operation and structural test, and the complicated syndrome of hybrid failure manifestation. The rationales and assumptions behind these approaches are typically based on the empirical knowledge of certain designs and extreme situations[1]. These rather simplified assumptions fail to deliver satisfactory test and diagnosis quality in real silicon, thus leading to a much lengthened time to revenue. A detailed review is presented in the remaining parts of this chapter to clearly illustrate the limitations of the current state-of-art.

2.1 Noise estimation in ATPG

Controlling noise effect in test generation requires an accurate estimation of the noise waveform. The noise estimation models proposed in the literature can be mainly grouped into two categories, namely, the switching activity based models and simplified voltage-drop models. This section provides an in-depth overview of these models from the perspectives of efficiency and accuracy.

[1] For example, most diagnosis approaches assume ideal timing fault models with deterministic fault manifestation behavior, a situation that typically cannot occur in real silicon.

2.1.1 Switching activity based model

It has been widely observed that the switching activity and the noise amplitude exhibit a positive correlation. The switching activities of transistors cause current to be drawn from the power supply or to the ground, resulting in a rapid change of the instantaneous current. This in turn results in a sharp droop of the power supply voltage or a strong bounce of the ground voltage. Given this observation, the intensity of the switching activity has been utilized as a noise metric in a large number of works.

The basic switching activity model simply counts the number of toggles in each clock cycle. Although computationally highly efficient, this model only captures the first order effect of switching activity on noise generation, thus being highly inaccurate. This simple model has been utilized in a number of early power-aware ATPG and DFT techniques [32, 22, 25, 75, 110]. In order to improve the model accuracy, the load capacitance of each toggling gate has been incorporated into the model so as to reflect the impact of the load on the amplitude of the current. Based on this idea, the weighted switching activity (WSA) model has been developed and widely utilized in the test literature [76, 77, 107, 105, 80, 72, 4]. This metric assigns distinct weights to the switching activities of different gates, with the weight being the number of gate fan-outs plus one. The weight value to some extent tracks the current contribution of the toggling gate and its loads, thus approximating the realistic situation more accurately. Although the use of the WSA metric provides a computationally efficient way of estimating the current demands and noise intensity in global power ground nets, it fails to consider the noise distribution and fluctuation in local regions of the layout. In fact, the noise amplitudes in distinct layout regions can skew significantly as a result of skewed switching activity and different power distribution network design. To provide the regional noise information, the layout-aware weighted-switching activity metric [56, 67, 116], which considers the noise effect of each local regions in the layout, has been proposed by several works that focus on the noise optimization in certain timing critical areas.

The switching activity based models in general provide reasonably accurate noise estimation at very low computational cost. Especially under low clock frequency situation where the cycle-based dynamic noise constitutes the major noise type, these metrics exhibit a very high correlation with the noise amplitude. Therefore, this approach is highly preferable in estimating the noise in low frequency modes such as the scan phase of structural testing. Nonetheless, these metrics fail to account for the possible noise accumulation across multiple cycles, thus being highly inaccurate in a high clock frequency mode such as the at-speed capture phase of structural testing.

2.1.2 Voltage-drop model

The switching activity based models provide an indicator of the noise amplitude, but fail to estimate the noise waveforms. For certain applications such as noise-aware timing simulation, it is necessary to attain the profile of the noise as a function of time, thus necessitating a voltage based noise estimation model. A number of voltage-drop estimation models have been proposed in the literature [51, 50, 27, 100, 5, 101, 12, 11, 60, 7]. Despite the diversity in the specific forms, these models share the similar principle of estimating the voltage variation in an extracted RLC (Resistance, Inductance and Capacitance) model of the circuit. The circuit model characterizes the silicon parameters at different levels, including:

- The package RLC information such as the leads, the ball grids, and the power planes
- The RLC information of redistribution layers (RDL)
- The RLC information of power switches
- The RLC information of the on-chip power distribution network (PDN)
- The RC information of device parasitics
- The RC information of intentionally added decoupling capacitance

With the extracted circuit model, the supply voltage waveform as a result of the device toggling current can be attained through an analytical computation or simulation. These approaches in general provide more comprehensive noise information compared to the switching activity based models, delivering more flexibilities in noise optimization during the ATPG process. The estimation accuracy highly depends on the quality of the circuit models and the simulation/computation algorithm. The computational cost of these approaches is typically much higher than the switching activity based estimation, necessitating the ATPG algorithm to intelligently utilize the noise information so as to minimize the number of noise estimation runs.

2.2 Handling noise in testing

The noise estimation techniques outlined in the previous section enable the researchers to tune the noise level in order to fulfill certain testing goals. Because of the strong industrial interest, the treatment of noise during ATPG and DFT design becomes a focal point of the testing research when the technology nodes enter the nanometer range. An appreciable number of techniques, with possibly conflicting perspectives and goals, have been proposed in the literature to address this issue. One category of techniques focuses on the minimization of test mode noise in order to avoid yield loss, whereas others aim to increase the noise level to guarantee the detection of noise related failures. Both opinions come with their own rationales that might be related to the specific characteristics of chip design, application domain, and quality requirement. Nonetheless, these experience-based optimization strategies fail to provide a coherent theoretical framework for guiding the noise handling in large designs that face complex failures. In this section, we provide a review of these techniques and illustrate their limitations.

2.2.1 Test noise minimization

Test noise minimization, as a main strategy for reducing overtesting, has been extensively studied in the literature. A set of approaches utilizes special

ATPG algorithm to attain the noise reduction goal, whereas the other methodologies focus on the development of DFT architectures that can restrain the noise generation.

ATPG for noise reduction

This category of approaches exploits the flexibility of the ATPG process in order to generate test patterns capable of reducing the test noise in addition to meeting classic ATPG goals.

A number of techniques have been proposed to process the *don't cares* in the test cubes to reduce the transition densities during test mode [9, 113, 107, 106, 62, 104, 24, 109, 66, 103, 70, 102, 57]. These techniques aim to identify an appropriate 0/1 assignment of the *don't care* bits in test cubes to reduce switching activity during test application, thus lowering the power ground noise. Wen et al. [107], for example, propose an X-filling method for reducing the number of transitions in scan flip-flops during capture mode. This approach increases the matching level between the launch cycle and the capture cycle flip-flop states by performing a set of matching heuristics. Several extended approaches [70, 106], following a similar basic strategy, are proposed to improve the effectiveness of this methodology.

The basic X-filling approaches only consider the fault activation and propagation constraints during ATPG, which is a highly impractical assumption. Since almost all modern industrial designs heavily use test compression techniques, the X-filling strategy of an ATPG process needs to maximally fulfill the constraints for both the noise reduction and test compression goals. Several research works have been proposed to resolve this challenge. The work in [62] proposes a capture power reduction technique dedicated to the nonlinear encoding compression scheme proposed in [104]. The work in [109] performs X-filling instead of DFT insertion to reduce test power. The X-bits are sequentially filled in a greedy manner according to their power impact. Linear constraint propagation is performed during the X-filling process to guarantee the compressibility of the filled test cubes. The technique proposed in [66] improves the aforementioned X-filling technique by filling the X values according to the power impact of the free bits in the seeds and performing a post-processing adjustment of the filled test cube. A low power compression technique based on scan chain partitioning is proposed in [103]. This technique clusters scan cells that have similar test data distribution, thus enabling an easy encoding of the test data within the same scan chain partition and reducing the toggling activities during scan.

In addition to the standard X-filling approaches, researchers have proposed several other ATPG ideas for test power and noise reduction. The technique proposed in [41] identifies an input control pattern at the primary inputs of a full-scan circuit. During test application, the input control values freeze a large number of switching activities at the immediate logic load of the scan flip-flops, thus reducing the power and noise in the combinational part of the circuit. Sankaralingam et al. propose a static compaction technique to minimize the test power dissipation [83]. Appropriately selecting the merging order of test cubes during test compaction reduces both average and peak power of the final test set. A similar approach is proposed in [56], which guides the test compaction algorithm using a layout-aware switching activity metric in order to meet a pre-defined test noise threshold in every local region of the chip layout. The noise threshold employed in these techniques is typically determined based on the designers' experience, which might be error-prone for today's large designs.

The aforementioned software-based approaches can be easily adopted into the ATPG algorithm without incurring hardware overhead. Nonetheless, the successful utilization of these approaches necessitates the resolution of the challenge imposed by various ATPG constraints which possibly limit the effectiveness in noise reduction.

DFT architecture for noise reduction

A set of approaches focuses on reducing test power through scan architecture modifications [25, 88, 30, 110, 111, 82, 108, 9, 3]. Sinanoglu et al. propose a technique of performing bijective transformations on scan patterns through the insertion of simple logic gates into the scan chains [88]. The transformed patterns that traverse through the scan chains have a low level of toggling activities, thus being able to reduce the power and noise in the scan phase with no impact on test application time. The technique proposed in [108] transforms the standard scan architecture to a set of selectable, separate scan paths. In low power testing mode, each scan path is loaded and observed individually through a bypassing circuitry. An adaptor circuit is inserted into the DFT structure to control the adaptation of the scan chains during the scan phase. This technique restrains the scan behavior in local scan paths, and enables the majority of the scan architecture to remain quiet, thus significantly reducing the overall toggling activities in the scan chains. Lee et al. propose an interleaving scan architecture based on adding delay buffers among the scan chains [58]. This scheme results in staggered toggling activities within the scan chains. Although the average power consumption remains almost unchanged, the peak power and noise can be reduced with the skewed toggling activities.

In addition to scan chain modification, techniques based on scan data or clock suppression have also been proposed to attain power and noise reduction [24, 84, 8, 71, 10, 99, 112, 31]. In [24], a DFT design is proposed to reduce the test power in a linear compression environment by inserting gating circuitry to generate sequences of constant values in the test cube. Sankaralingam et al. propose a scan architecture that can disable the clocks of a subset of scan chains [84]. Since the fault detection process typically requires very few scan chains to be activated, the clock of a large number of scan chains can be disabled during test application. This not only reduces power and noise in the logic circuitry, but lowers power in the clock tree as well. A similar approach is proposed in [8], which gates off the clock of partial scan chains during test application. Some researchers also look into the utilization of a clock gating approach in reducing the power ground noise in capture mode. The technique proposed in [112] selectively gates off for each test pattern the clock paths of circuit modules that have no impact on the activation and observation of the targeted faults. Furukawa et al. propose a technique which utilizes the clock-gating and X-filling in a collaborative manner to reduce the toggling activity in at-speed capture cycle [31].

The hardware-based techniques typically are quite effective in power and noise reduction as they encounter much fewer ATPG constraints compared to the software-based ones. Nonetheless, these techniques necessitate either the insertion of special DFT circuitries or the modification of traditional test application schemes. The resultant hardware overhead and design difficulty restrain the application of these techniques.

Technical limitations

The noise minimization approaches outlined previously are able to reduce the marginal failure rate during production test, thus significantly improving the yield. However, a blind minimization of the noise can result in the design weakness escaping from the test, as the test fails to examine the normal noise level encountered during functional operation. This leads to a degraded test quality and an increased DPPM. To avoid this unpalatable undertesting issue, the test minimization process needs to intelligently differentiate the test mode noise from the functional one and focus on the noise reduction in test-only area. None of the aforementioned techniques have looked into this issue, thus failing to deliver a satisfactory solution to production test plans that require high test quality.

2.2.2 Noise failure detection

Given the undertesting risk induced by the noise minimization techniques, another set of methodologies, in contrast, focuses on the maximization of power ground noise in order to reduce test escapes [79, 80, 68, 65, 67].

The work in [79] provides a silicon case study of an Intel microprocessor design, and shows that regular at-speed test is incapable of detecting the worst-case noise in functional operation. The technique proposed in [80] presents an algorithm to generate test patterns that result in high switching activities at targeted wires so as to stress the circuit with higher voltage droop. In [68], layout information is utilized to guide the IR-drop maximization in the neighborhood of the target devices. The techniques in [65, 67] extract functional constraints of the circuit and embed them into the search for a test pattern with maximal IR-drop so as to approximate the functional operations more closely.

The utilization of these techniques to some extent improves the test coverage on marginal failures induced by design weakness, potentially delivering higher test quality. Nonetheless, these techniques share the same shortcomings as the noise minimization techniques, that is, there is no differentiation between test mode and functional mode behavior. In all probability, the increased noise level can worsen the timing margins of non-functional paths, thus resulting in the good parts failing the test. As a result, the test quality improvement is attained at the cost of highly increased yield loss. Moreover, since these techniques are all based on the conventional single-capture cycle schemes and fail to consider the impact of the noise effect that spans multiple at-speed cycles, their capability to approximate the true functional noise profile and detect noise-induced failures still remains highly limited.

2.3 Scan architecture failure analysis

The improvement on the design and test flow needs to be accompanied by the innovation in failure analysis techniques, as any optimization on the design and test side needs to be based on the silicon information learned during failure analysis. As a result of intensive test mode noise, marginal failures in scan architecture have become one of the most dominating and complex failure mechanisms for failure analysis, and have drawn increasing research attention. The fault diagnosis flow for a scan architecture can be typically decomposed into two stages, namely, a scan chain diagnosis step that provides the information about the scan cell failure locations and probabilities, and a clock buffer pruning step to identify the faulty clock buffers based on the failing scan cell information. In this section, we provide an overview of the previous work for both steps.

2.3.1 Scan chain diagnosis techniques

A considerable number of approaches have been proposed in the literature to address the scan chain diagnosis problem. A comprehensive survey of this area has been provided in [46]. In general, these approaches can be categorized into three groups.

- **Tester-based diagnosis:** Techniques in this category utilize physical failure analysis equipment to identify the failing locations in the scan chain [26, 90, 92, 40, 52, 73, 49]. The work proposed in [26] uses electron-beam probing to detect incorrect toggling patterns in the scan chain. Approaches in [90, 92], identify fault sites based on the detection of light emission of off-state leakage current. Another leakage current based technique utilizes the I_{DDQ} test platform and special scan patterns to capture the abnormal current induced by incorrect toggling behavior in scan chains [40]. Techniques in [52, 73] program the voltage and clock of the ATE to the specific condition necessary for fault triggering and then change the test environment to the opposite condition for fault observation. These approaches typically provide good diagnostic quality. Nonetheless, they necessitate a time-consuming debugging process and possibly expensive equipment, thus being confined to a limited application range.

- **Hardware-modification based diagnosis:** A number of techniques incorporate special scan chain designs to enhance the diagnosability [85, 86, 29, 74, 94, 61]. The approach proposed in [85] connects the output of each scan cell to the a scan cell that belongs to a different scan chain, in order to observe the content of failing scan cells from the fault-free chains in the diagnosis mode. Although this technique enhances the scan chain diagnosability, it almost doubles the routing overhead in the scan architecture, which significantly constrains its applicability. The scan architecture proposed in [29] inserts between scan cells XOR gates that are controlled by an external signal. Appropriate programming of the control signal can guarantee the diagnosis of the rightmost stuck-at fault in the scan chain. A technique is proposed in [74], which adds a set/reset capability to the scan outputs of the flip-flops so as to detect the stuck-at fault between each pair of adjacent scan cells. A similar approach is proposed in [86], which, in addition to the set/reset capability, adds the value-flipping capability to the scan cells to further enhance the diagnosability. The use of global set/reset signals, though, significantly increases the routing congestion as well as the design complexity. The approach in [94] partitions the scan chain into multiple segments, thus enabling the bypassing of the segment that contains the failing scan cells. These techniques reduce the algorithmic complexity of scan chain diagnosis, yet at the cost of extra hardware overhead and increased design complexity.

- **Software-based diagnosis:** This category of techniques aims to identify fault sites through algorithmic analysis of the failing data collected during regular scan operation [91, 44, 37, 48, 42, 43, 64, 34, 35, 63, 89, 98, 36, 20]. Such techniques typically incur much lower cost compared to the tester/hardware-based approaches, thus having a wider application in general designs. Due to this reason, the proposed work focuses on developing innovative techniques in this domain.

Despite the differences in technical details, the software-based diagnosis techniques mainly share a set of fundamental ideas. Most of the software-based techniques target permanent faults in the scan chain. One group of such approaches identifies the candidate faulty cells through fault simulation of test patterns [37, 48, 42, 89]. The observed values on the fault-sensitive bits of the test pattern can indicate the upper and lower bounds of each fault site. Another set of techniques focuses on the creation of a comprehensive fault dictionary by examining the effect of all possible fault hypotheses, in order to enable a look-up table based fault diagnosis process [34, 35]. Some researchers also propose the generation of special diagnostic patterns, such as single-excitation patterns, to examine each scan cell independently [64, 63]. Good diagnostic results can be attained with these techniques under the permanent fault assumption. Nonetheless, such ideal assumptions barely match the realistic failures observed in today's VLSI circuits.

Only a few proposals have been presented to address intermittent fault diagnosis. The approach proposed in [44] employs a Bayesian decision model to identify the location that has the highest fault probability. However, this approach can only be applied to scan chains with a single fault. The approach in [43] utilizes the signal probability computation to search for the fault locations that maximally account for the observed pattern; assumptions made in advance regarding the fault manifestation probability though degrade somewhat the diagnostic quality. Another technique [98] uses error count as an indicator of fault locations. But this technique provides no information about fault types and manifestation probabilities. In order to gain a deeper understanding of the increasingly complex failure conditions seen in today's large designs, a comprehensive methodology is needed to analyze scan chains with mixed fault types and various manifestation probabilities.

2.3.2 Scan clock fault diagnosis techniques

The fault diagnosis on global scan signals, such as the scan clock, is a relatively new topic in the testing literature, and only a few approaches have been proposed recently. In [69], a technique is proposed to characterize and test the clock faults, but no diagnosis issue is addressed in this paper. The work in [23] proposes a method which identifies the defects on a scan enable tree by examining the effect of adjusting the scan enable deassertion timing with respect to the scan clock. Nonetheless, this approach mainly relies on capture-cycle fault manifestation, and therefore cannot be extended to the fault diagnosis in the scan clock tree. The technique proposed in [45] identifies a set of candidate faulty buffers for each scan chain failure through logic tracing. The candidate sets of different scan chain failures are subsequently intersected to identify a minimum set of common buffers that account for all the scan chain failures. In scan designs with multiple scan cell failures, this technique might result in a large set of candidate buffers. The approach proposed in [59] performs fault simulations subsequent to the logic tracing, and ranks the fault candidates based on the matching level between the simulated and actual scan chain failures. The fault simulation step improves the diagnostic resolution at reasonable computational cost, as permanent fault simulation can be accurately performed for each individual pattern in a relatively short time. Nonetheless, it is not applicable to intermittent timing fault diagnosis as the individual pattern based simulation employed in this work is unable to account for the statistical behavior of the timing failures. The simulation of the intermittent timing failures is much more expensive than permanent faults, as statistical timing analysis on the relevant clock paths is necessitated to estimate the fault manifestation probability. Due to these reasons, a new set of approaches, capable of maximally constraining the number of fault hypotheses and pruning the hypothesis space with the least amount of simulations, is sorely needed for silicon debugging in an intermittent fault scenario.

Chapter 3

Design-guided, noise-aware test and diagnosis framework overview

Addressing the challenges outlined in Section 1.2 to account for the failure mechanism shift necessitates the construction of a test and diagnosis framework that is capable of identifying and detecting the true functional-mode marginal failures with the help of design knowledge, yet still utilizing the cost-effective structural test and failure analysis infrastructures. A *design-guided, noise-aware test and diagnosis framework* is therefore proposed in this thesis. During test development, the proposed framework uses design information to approximate the manifestation conditions of functional-mode failures and avoid the generation of false alarms. During failure analysis, the design information is utilized to filter out unrealistic failure hypotheses and create strong signals regarding failure location, type and criticality. This enables efficient, parallel optimization in both design and test, thus resulting in a fast-to-revenue product development cycle.

3.1 Yield and quality co-optimization

Achieving early time-to-market is one of the key factors that determine the success of a semiconductor product. Attaining this goal necessitates the yield and test quality requirements of volume production to be fulfilled in a short time frame. Nonetheless, the complexity of modern integrated circuits results in a rather lengthened production ramp-up time, typically ranging from months to even years, which constitutes a bottleneck of the product development cycle.

Traditionally, the production ramp-up flow mainly deals with the silicon defects due to immature fabrication processes, as shown in the lower half of Figure 3.1. Foundries typically deliver parts of different process corners to ATE for test characterization. The systematic failures observed on the tester would be examined by a failure analysis process, with the diagnostic conclusions fed back the foundries to guide the fabrication process learning. With the failure mechanism change, the production ramp-up process, while still addressing the traditional fabrication issue, starts to shift its focus onto the design and test optimization side. This aspect of production ramp-up flow targets on fixing the design marginalities and addressing the associated soft failures in test. Nonetheless, both design and test encounter critical challenges:

Design challenge The production ramp-up phase has a very limited set of degrees of freedom in performing design modification. The effective utilization of the limited optimization budget relies on the accurate identification of the design weakness.

Test challenge The most important and sensitive testing goals, yield and test quality, are seemingly conflicting, making the co-optimization of them highly difficult.

Addressing these challenges relies on the understanding of the failures from a design perspective. Given this observation, the proposed test and diagnosis framework utilizes design knowledge to guide the test development and failure analysis in order to achieve a co-optimization of yield and test quality. The high level principles of this framework are illustrated in the upper half of Figure 3.1. With the information about the functional behavior and the physical design of the circuit, the test program can be optimized to approximate the functional behavior of the circuit, thus performing a more focused examination on the marginal failures that can manifest themselves during functional operation. Similarly, by examining the design implications on failure behavior, the failure analysis flow can identify

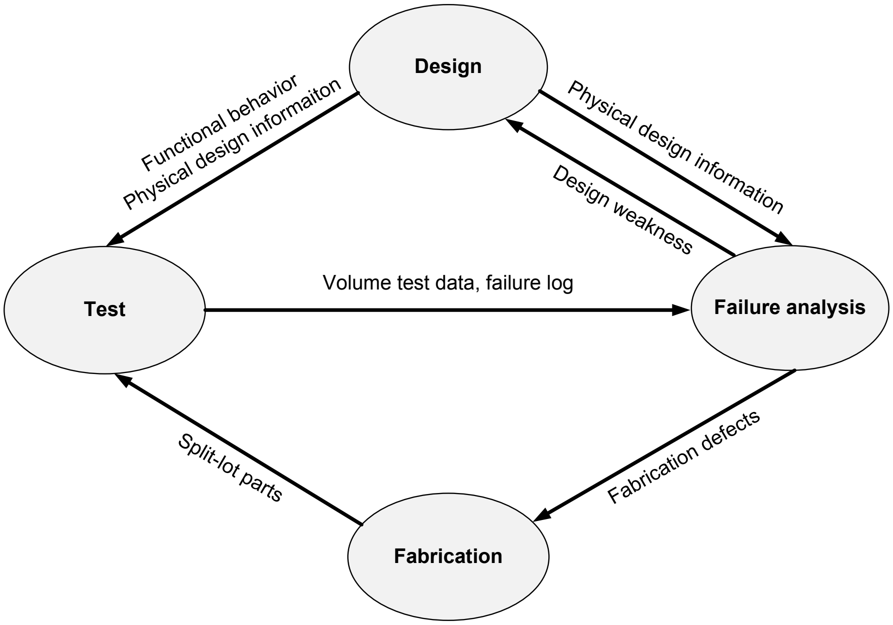


Figure 3.1: Design-guided test and diagnosis framework

more physically realistic fault hypotheses to draw more accurate diagnostic conclusions. The advantages of utilizing design information in test and diagnosis are briefly summarized as follows.

- Functional behavior information helps differentiate the functional-mode timing paths from the test-mode ones, thus enabling the test generation to focus on the marginal failure detections on functional paths through the use of distinct noise handling policies for these two sets of paths.

- Power distribution network information helps the extraction of a realistic circuit model for power ground noise estimation, thus enabling a more accurate execution of the noise handling policy during test generation.

- Path layout information helps identify the timing critical paths that are most prone to marginal failures. Given the limited test time budget, focusing on critical path failure detection during test generation delivers the highest test quality.

Table 3.1: Impact on yield, test quality, and time-to-revenue

	Design-guided test optimization	**Design-guided failure analysis**
Yield	Reduce overtesting in test-mode paths	Accurate failure identification for design fix
Test quality	Detect worst-case noise failures in functional-mode paths	——
Time-to-revenue	Less ATE data collection time Faster test pattern tuning	Quick pruning of failure hypotheses Fewer re-spins & faster turn-around

- Clock tree information helps identify the root cause of marginal timing failures, as the failures can be resulting from the clock skews induced by the delay variation in clock buffers.

- The timing slack information helps estimate the robustness of different timing paths, thus enabling the utilization of failure manifestation probability as an observed syndrome to prune out the unrealistic failure hypotheses during diagnosis.

Utilizing the design-guided test and failure analysis methodology proposed in this thesis enables the co-optimization of yield and test quality during production ramp-up, the challenge that traditional approaches fail to resolve. From the test perspective, the mitigation of test-mode marginal failures minimizes the overtesting effect, thus significantly improving the yield. At the same time, higher coverage and worst case noise generation in functionally critical paths improve the test quality and reduce the test escapes. In addition to the contribution of test optimization, the yield and test quality improvement is also gained from a more efficient design optimization enabled by the proposed flow. The design-guided failure analysis is able to pinpoint not only the locations of the design weakness, but the failure type and criticality. This enables the designers to wisely spend the optimization margin to effectively fix the design issue, thus reducing the number of re-spins during production ramp-up. Meanwhile, the localized design change resulting from

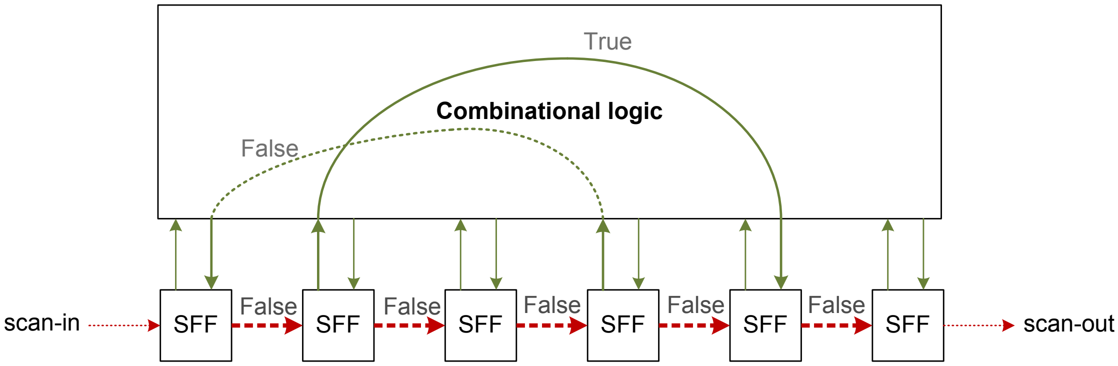


Figure 3.2: Timing path differentiation

the accurate failure analysis also reduces the difficulty and effort in fixing the design weakness, thus resulting in a shorter turn-around time between re-spins. The impact of the proposed flow in yield, test quality, and time-to-revenue is summarized in Table 3.1.

3.2 Embedding functional view in structural test

In production test, one major goal consists of examining whether the circuit can operate correctly at the clock frequency defined by the circuit specification. In the nanoscale circuit era where the power ground noise plays an important role in circuit timing, such a check needs to be performed with the consideration of noise impact. The timing check is a rather "functional" target as the speed of a circuit is tightly coupled with its functionalities. Nonetheless, the examination of such a "functional" target has to be performed in a structural test platform due to various constraints. The discrepancy between functional and test modes, as outlined in Chapter 1, results in inaccurate and even misleading test conclusions. The incorporation of functional perspectives in the structural test development can effectively minimize the discrepancy, thus improving both the yield and test quality. This basic idea constitutes the fundamental principle of the test development methodology proposed in this thesis.

To concretely illustrate the advantages of the proposed test development methodology, we herein provide an overview of the noise and timing handling strategy utilized in the proposed test scheme, focusing on the extraction and embedding of functional perspectives.

The discrepancies between functional and test modes can be observed in multiple design levels. One of the major discrepancies that highly impact the test results consists of the difference in the timing paths being sensitized in these two modes, as shown in Figure 3.2. In test mode, each test pattern is firstly shifted into the flip-flops through the scan chains. A capture operation is then performed to apply the test pattern to combinational logic and capture the test response. Finally the test response is shifted out through the scan chain for test comparison. It can be seen from the test application process that the scan phase sensitizes a large number of timing paths that are dedicated to scan operations. Since these scan paths (red paths in Figure 3.2) are disabled during functional operation, they can be considered as functionally false paths whose timing robustness has no impact on chip functionality. As a result, the manifestation of any marginal timing failure on these paths purely increases the yield loss, with no contribution to the test quality. Therefore, an ideal test scheme needs to mitigate the marginal failures in scan paths so as to guarantee the correct loading/observation of test stimuli/responses, as doing so can significantly improve yield with no degradation on test quality. The exceedingly high scan-mode power ground noise induced by intensive scan toggling activity has been observed as the major reason for the marginal timing failures in scan chains. Therefore the minimization of scan-mode noise becomes one of the most efficient strategies for reducing overtesting and improving yield.

In the capture phase, the scan paths are disabled and the capture paths (the green paths in Figure 3.2) that pass through the combinational logic are sensitized. Compared to the scan phase, the behavior of the capture phase is more similar to the functional operation but still not completely identical. On the one hand, this phase does sensitize a large number of true functional paths (illustrated as the solid lines in Figure 3.2), thus being able to examine their timing robustness. On the other hand, since the scan pattern sets the circuit to a non-functional state, the capture cycle also enables a lot of functionally false paths (illustrated as the dotted lines in Figure 3.2). In the interest of test quality, the capture cycle should examine the maximum noise condition of the functional paths, as this condition represents the worst-case timing that these paths can encounter during functional operation. Yet to improve yield, the noise on the false paths should be minimized to reduce false alarms. It is almost impossible to attain both goals simultaneously, as the functional and false paths can be physically close or even partially overlapped in chip layout. Nonetheless, if the test scheme can tune the capture phase in such a way that the state transition in the at-speed capture cycle approximates the functional operation, the dilemma between overtesting and undertesting can be maximally resolved. When the circuit enters near-functional state transitions, the worst-case noise can then be developed on the timing-critical paths to perform the most strict timing check. Meanwhile, the chance of creating a false alarm in such a strict timing check is negligible, as the number of false paths is minimized in near-functional state transitions. The development of near-functional state transitions, the differentiation of functional paths from non-functional ones, and the identification of timing critical paths, all require the embedding of the functional perspectives into the structural test development.

In addition to the aforementioned behavioral level perspectives, the physical level view of the design also provides important information for test optimization. Traditional structural test schemes only apply one clock cycle in capture phase to sensitize faults. Yet in high speed circuit, multiple clock cycles are typically needed to develop the worst-case noise in functional operation, as the accumulation of the middle and low frequency noise requires a relatively long period of time. Such a discrepancy results in the traditional test scheme being incapable of approximating the worst-case functional-mode noise condition, thus leading to test escapes. To attain high test quality, the capture phase needs to approximate the noise accumulation process occurring in functional operation. The estimation of the noise accumulation time needed for worst-case noise development, however, constitutes a critical technical challenge. The resolution to this problem necessitates an investigation of the impulse response of the circuit, as this inherent characteristic determines the die-down time of the noise. The physical level information, such as the power distribution network design, the decoupling capacitance, and the circuit parasitics, is therefore needed to create a noise estimation model for guiding the test development.

In light of the aforementioned principles, the test development methodology proposed in this thesis performs concurrent optimization on both scan and capture phases. Fine-grained noise handling strategies, guided by the functional perspectives, are applied to the scan and capture phases so as to attain sophisticated optimization goals on both functional and non-functional timing paths. An innovative multi-functional-cycle capture scheme, capable of approximating the functional operation in structural test, is utilized to develop the worst-case noise profile in functional paths. The combination of the structural test application and the near-functional fault detection enables yield and test quality co-optimization at very low test cost.

3.3 Failure hypothesis pruning using design information

In addition to effective test optimization, fast production ramp-up also relies on accurate failure analysis. As outlined in Chapter 1, traditional failure analysis fail to deliver satisfactory diagnosis results in nanoscale circuits, with the major challenge being the unpredictable behavior of marginal failures. This problem is especially significant in scan circuitry, as both the scan chains and the scan clock trees have exceedingly high toggling rate, easily creating intermittent timing violations with hybrid failure types.

The scan process can be considered as a unidirectional movement of a data stream at a constant speed. Since the manifestation of scan chain timing failures statistically changes the speed of the movement, the scan data will be partially moved forward or backward depending on the types of the timing violations. This enables us to extract evident failure features from a relatively abundant set of scan data. More specifically, the abnormal phase movement[1] of the scan data inherently

[1] The phase is the relative positions of the observed scan data with respect to their positions

reflects the failure location and type, whereas its occurrence probability signifies the magnitude of the failures. The statistical significance of such information can effectively suppress the noise created by the random manifestation seen in individual scan patterns, thus enabling an accurate identification of the failing scan cells and the associated timing violation types.

The failing scan cell information provides clues for further investigation into the scan clock trees, as the the failures in the scan cells can be an indirect result of delay faults in their clock paths. This unfortunately opens a huge search space, as the number of clock buffers in modern designs is typically quite large, not to mention the complicated interaction of multiple faults. Nonetheless, two levels of design information can be utilized in the clock tree examination to quickly prune the failure hypotheses. Since the timing violations in scan cells are in nature the results of asymmetric delay variations in the clock paths, the topological structure of the clock tree, coupled with the information regarding scan cell timing violation types, enables a logic tracing of candidate range of faulty clock buffers. This logic level analysis quickly prune out unrealistic fault assumptions and narrows down to a small set of clock buffers, as a large portion of the clock buffers are not even on the reachable paths of the failing scan cells. Once a relatively small set of candidate faulty buffers is identified, the timing information of the clock buffers, with the consideration of the noise impact on delay variations, can be further utilized to derive the timing violation probability of the associated scan cells. This physical level pruning enables a clear identification of the failure hypotheses that closely match the observed syndromes, thus helping the designers accurately plan their optimization budget in re-spins.

The utilization of the design information enables the diagnosis process to perform statistical and logic analysis in an intricate manner so as to leverage the benefit of each. For example, the timing violation type information extracted from statistical scan data analysis enables an accurate deterministic tracing of the clock tree, which further reduces the computational effort of statistical timing analysis needed for failure probability estimation. The merger of logic and statistical

in the expected ones.

analysis not only elevates the diagnostic accuracy, but significantly improves the efficiency of the diagnosis flow.

The aforementioned principles have motivated the design-guided failure analysis methodology proposed in this thesis. This approach closely approximates the realistic failure mechanisms observed in silicon, capable of identifying complex failure combinations with hybrid timing violation types and pinpointing their root causality in scan clock trees. The incorporation of this approach in the production ramp-up phase can significantly expedite the re-spin cycle and reduce the number of re-spin iterations.

Chapter 4

Overtesting avoidance in scan mode

After introducing the global picture of the design-guided, noise-aware test and failure analysis system, we proceed to look into the underpinning techniques that support this framework. As mentioned before, one fundamental challenge induced by the discrepancy between the functional operation and the structural test consists of the overtesting in non-functional circuitries. More specifically, the dynamic noise induced scan failures constitute the major source of yield loss, as the abnormally intensive toggling activity during the scan phase can easily result in the noise amplitude exceeding the design margin. The mitigation of scan-mode noise can significantly recover yield with no impact on test quality, as all paths impacted by scan-mode noise are false paths in functional mode.

In this chapter, we focus on providing a solution to the most industrially-important yet most challenging yield recovery issue, namely, the noise mitigation in scan architectures, using linear compression circuitries. On-chip test compression schemes based on XOR networks have been widely employed in large industrial scan designs, due to their utmost compression ratio and simple decompression mechanism. Nonetheless, such schemes specify *don't cares* through a linear mapping of seed vectors, which precludes the possibility of preprocessing the original test cubes for scan-mode noise reduction. The consequent dynamic noise level can quite possibly be highly elevated. Drastic noise reduction yields can nonetheless be attained through a judicious seed selection strategy, if one considers that multiple legal seeds with highly divergent consequent noise impact may typically exist in a linear compression scheme.

The proposed technique identifies the noise-friendly seeds through a highly guided system relaxation process. A mathematical transformation is proposed to embed the compression constraints into a linear system constructed over the independent bits of the seed, which not only resolves the constraint conflicts between compression and noise reduction, but drastically reduces the seed space to be searched. Such a low noise compression scheme exploits solely the flexibility in the seed space, necessitating no alterations in the decompression hardware.

We further propose an optimization in the ATPG flow consisting of a concurrent compaction and compression scheme. The upgraded ATPG flow ensures through controlled cube compaction the delivery of at least a minimal number of seed options for each of the cubes while further privileging a balanced distribution of seed options for each cube. This in turn leads to a concurrent reduction in both the average and peak scan noise among all cubes. The proposed scheme has little impact on the fault dropping efficiency, thus guaranteeing the same level of test volume reduction as traditional compression techniques, as confirmed by the experimental results provided.

In this chapter, we first provide a brief review of the linear compression scheme and outline the proposed idea of selecting compression seeds for noise reduction. The mathematical and algorithmic framework of this noise-aware compression technique is presented subsequently. In Section 4.4, the noise-aware compression technique is extended and generalized to a concurrent compression/compaction ATPG flow with a focus on noise and test volume co-optimization.

4.1 Linear compression based scan architecture

Figure 4.1 illustrates the scan test architecture using an XOR network as its on-chip decompression hardware. In each scan cycle, the XOR network takes as

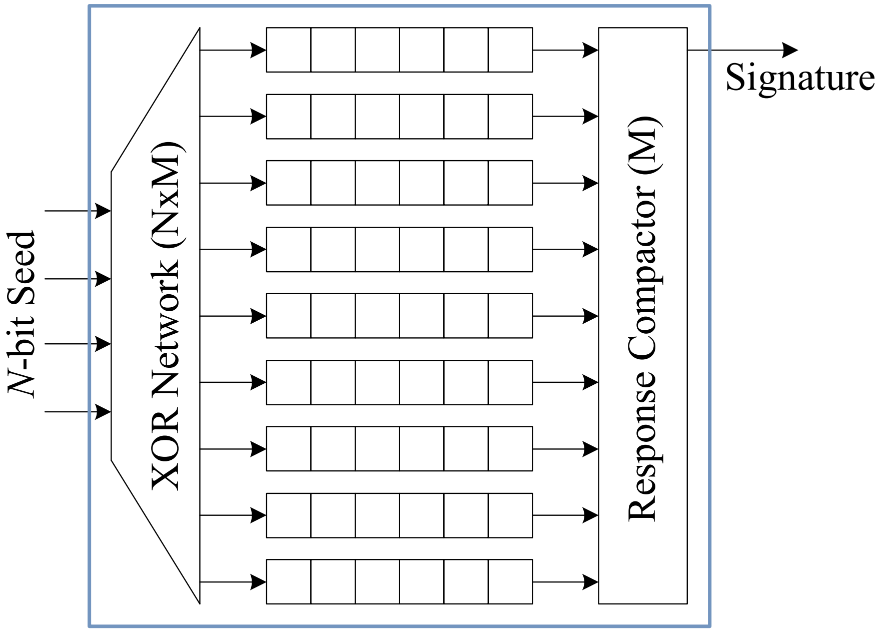
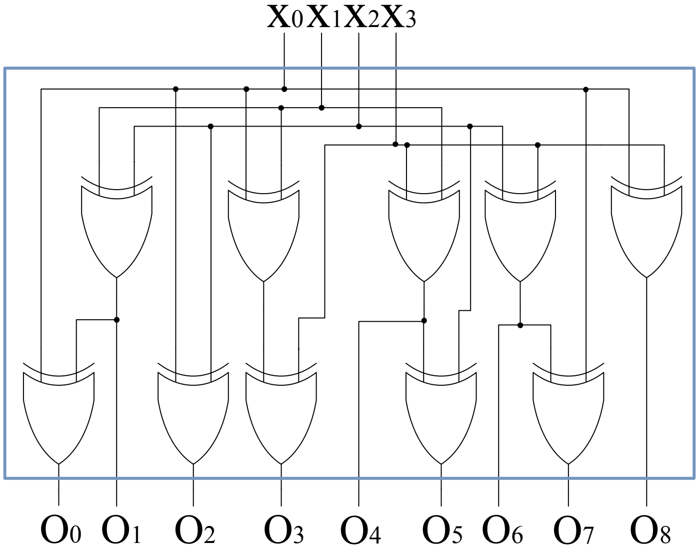


Figure 4.1: Scan architecture using XOR network-based decompression hardware

inputs an N-bit seed from the ATE and decodes it to an M-bit test slice. All bits of a test slice are generated in parallel and are applied to the actual scan chains simultaneously. The scan-out values of internal scan chains are compacted in parallel using an XOR-based response compactor [111] or a multi-input signature register (MISR). The decompression process is independent of the response-compaction mechanism, thus enabling the designers to choose the most appropriate compaction hardware according to the needs. The fixed-length to fixed-length decompression mechanism offered by the XOR network based decompression scheme eliminates the need for complicated synchronization methodologies between ATE and the decompression hardware, making it highly practicable for a variety of industrial applications.

The decoding mechanism of the XOR network is essentially a linear transformation from the seed to the test slice. Each bit of the test slice is a linear combination of multiple seed bits. Thus the compression process consists of the identification of an appropriate seed vector that can successfully generate all the specified bits of the original test slice through the linear transformation. Test slices

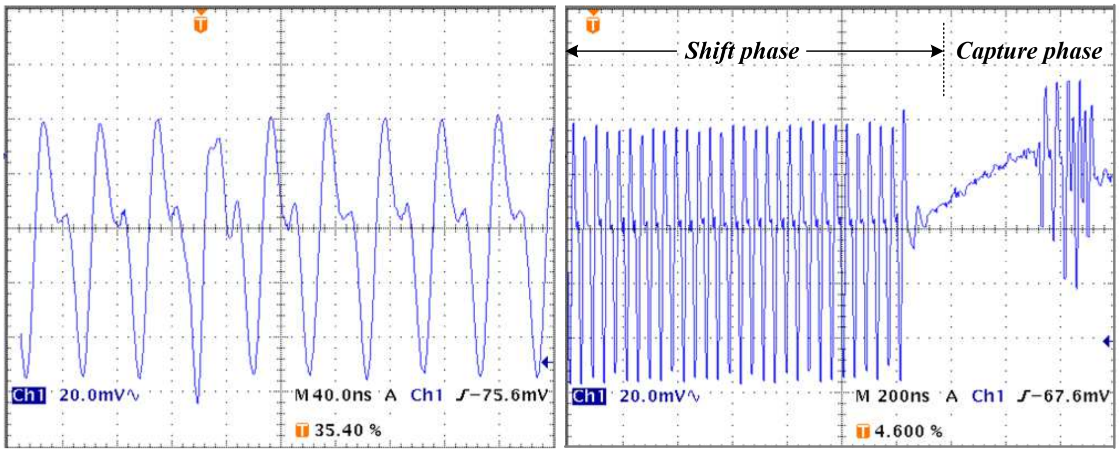


$$\begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_0 \\ X_1 \\ X_2 \\ X_3 \end{bmatrix} = \begin{bmatrix} O_0 \\ O_1 \\ O_2 \\ O_3 \\ O_4 \\ O_5 \\ O_6 \\ O_7 \\ O_8 \end{bmatrix}$$

Figure 4.2: Example XOR network and its matrix representation

with few specified bits can be reconstructed from very short seeds, thus delivering appreciable compression ratios. An illustrative XOR network is shown in Figure 4.2, which can decompress a 4-bit seed to a 9-bit test slice. The aforementioned linear transformation can be represented as a matrix multiplication over the finite field $GF(2)$, in which the addition operations are defined as the XOR operations, as shown in Figure 4.2.

The seed for a particular test slice is identified by solving the set of linear equations corresponding to all the specified bits in the test slice. The set of linear equations to be solved varies as the specified bit sets of each test slice are typically distinct. The linear compression scheme can be embedded into the test generation flow to enable a concurrent compression and compaction. The ATPG algorithm starts by generating an unfilled test cube for a fault in the fault list. Every newly generated test cube is compacted with the previous cubes as long as the resulting cubes are still compressible using the XOR network. At the point when a compacted cube reaches a predefined compression threshold, a seed is deemed to have been identified for this particular cube. The *don't cares* in the cube are then completely specified through the linear transformation effected on the seed. Fault



(a) Dynamic IR-drop in shift cycles

(b) Shift IR-drop vs. Capture IR-drop

Figure 4.3: High dynamic IR-drop during scan phase

simulation with the resulting fully specified test cube is then performed to drop additional faults. The outlined compaction/compression scheme is iterated until a target fault coverage is reached.

4.2 Noise-aware test compression

Although the linear compression scheme delivers significant benefits in test volume reduction, the manner of filling *don't cares* through linear transformations yields highly random transition patterns between adjacent test slices, thus resulting in appreciable dynamic noise during scan phase. It has been widely observed that the scan phase typically exhibits far higher power ground noise than the capture phase, which not only causes heat and reliability issues during testing, but significantly increases the risk of overscreening good chips. Figure 4.3(a) illustrates the scan mode dynamic IR-drop measured from an industrial chip. In each scan cycle, the power supply voltage experiences a sudden surge followed by a recovery process with a small overshoot. During the capture phase, the state transitions are constrained by the circuit logic, thus typically inducing fewer flip-flop togglings and a smaller dynamic IR-drop, as shown in Figure 4.3(b).

Since scan-mode noise is one of the limiting factors for scan-based testing, a noise-aware test compression technique is necessitated for large industrial scan designs. It is difficult to envision how traditional scan power/noise reduction techniques, such as *adjacent filling* [9], could be used in conjunction with the compression constraints, as the aggressive consumption of *don't cares* necessitated by these techniques would result in the failure of compression by generating inconsistent linear equation sets.

To solve this issue, one needs to explore the flexibility in the seed identification process during compression. One observation is that a large portion of test slices can be constructed from multiple alternative seeds. Since a seed is a solution to the linear system determined by the specified bits of the test slice, the size of the seed space strongly hinges on the number of constraints introduced in the linear system. More concretely, if the rank of the coefficient matrix of a consistent linear system constructed from an N-input XOR network is R, a total of 2^{N-R} solutions corresponding to seeds will exist. A natural upper bound on the rank of the coefficient matrix consists of the number of specified bits in the test slice. It has been widely observed that the number of specified bits in test cubes is quite small (typically fewer than 5% [96]), and the distribution of the specified bits across different test slices is typically non-uniform. Only the few test slices with the highest densities of specified bits are subject to a tight constraint induced by the XOR network, with the rest being loosely constrained, thus providing a high seed selection flexibility. Therefore, a large optimization space can be expected for most test slices under compression.

The scan noise impact of distinct seeds exhibits a high variation; therefore, it is possible to attain noise reductions by selecting the noise-optimal seed during the compression process. As illustrated in Figure 4.4, assuming the XOR network in Figure 4.2 is employed as the decompression hardware, two sets of seeds are identified for the original test cube. The test cube generated from Seed 1 contains a total of 17 transition patterns in adjacent scan flip-flops, whereas the one generated from Seed 2 has only 9 transition patterns. Selecting Seed 2 for compression delivers a 50% reduction almost in scan toggling activity with no impact on the

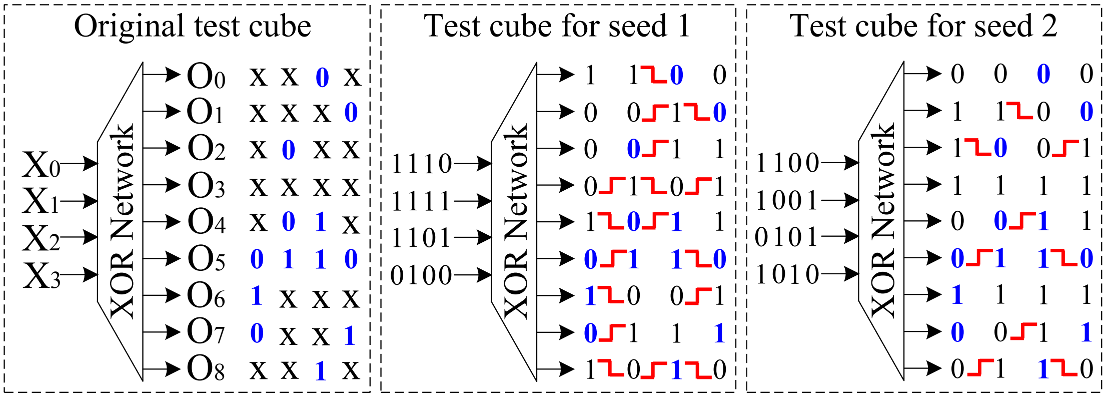


Figure 4.4: Noise reduction through seed selection

compression ratio whatsoever, thus enabling significant reduction in power ground noise given the high correlation between dynamic noise and toggle activity.

Applying seed selection may impact fault dropping in unpredictable ways. Yet while specific instances may display benefits or deterioration, the overall effect is typically neutralized, resulting in negligible changes to the test set size, as can be observed in the experimental results.

Achieving the proposed noise-aware test compression scheme necessitates an accurate yet computationally-efficient methodology for the identification of the noise-optimal seed for each test cube. We present the algorithmic framework for the proposed technique in the following section, and discuss the strategy of embedding it into the ATPG flow in Section 4.4.

4.3 Algorithmic framework

Scan-mode dynamic noise is induced by the test pattern transition between adjacent test slices (columns of scan flip-flops) in the test cube. Consequently, attaining scan noise reduction necessitates the maximal match of each test slice with its neighbors. Technically, slice matching can be performed in an iterative manner from one end of the test cube to the other end. Each matching step

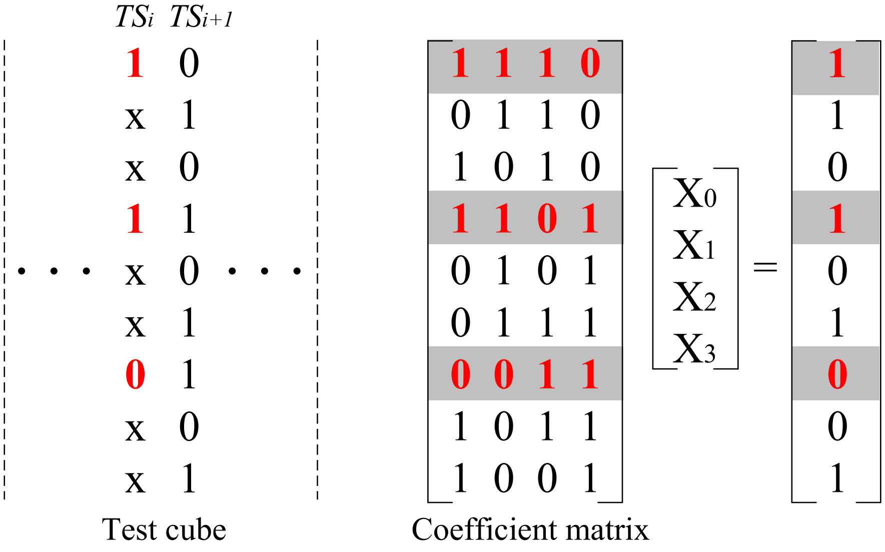


Figure 4.5: Matrix representation for slice matching

completely specifies one test slice, which in turn is used as the reference value for the next step.

4.3.1 Problem formulation

To apply the high-level slice matching idea without impacting compressibility, one needs to investigate the constraints imposed by the compression scheme. With no loss of generality, let us denote the two slices to be matched as TS_i and TS_{i+1}, respectively, with TS_{i+1} being completely specified in the previous matching step.[1] The maximal matching between TS_i and TS_{i+1} in the test compression context requires the identification of a seed for TS_i, which fulfills the following two conditions.

1. All the specified bits in TS_i must be completely generated from the seed to guarantee the correctness of compression.

[1] In the proposed methodology, the slice matching process traverses the slices starting from the rightmost end of the test cube. Although the right-to-left processing order might slightly deviate from the optimal one, it significantly reduces the computational complexity for slice matching. More importantly, such a processing order gives priority to the test slices at the right end of the test cube as the togglings in these slices contribute more noise during the scan operation, thus delivering a near-optimal noise reduction.

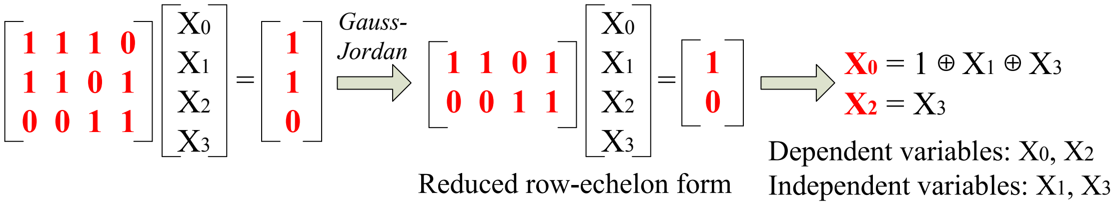


Figure 4.6: Extracting independent bits of the seed

2. For the unspecified bits, the filling values generated from the seed should maximally match the corresponding values in TS_{i+1}.

Such a formulation can be represented by a linear equation system, as shown in the example in Figure 4.5. The linear functions corresponding to the specified bits (the three functions in the gray area) assume exactly the values in TS_i, whereas the functions corresponding to the *don't care* bits are expected to yield the values in TS_{i+1}. The two test slices can be perfectly matched, resulting in a near-zero scan noise contribution, if the linear system thus generated has a solution. In all probability though, an inconsistent system with no solution of zero scan power is to be expected. In this case, a seed for TS_i needs to be identified which completely satisfies all the equations corresponding to the specified bits and maximally satisfies the remaining ones.

4.3.2 Seed space transformation

If the linear equation system representation for the slice matching problem is inconsistent, directly solving this problem is not only computationally expensive but mathematically difficult to handle, as it necessitates the manipulation of a coefficient matrix of a size $M \times N$ and the fulfillment of the strict constraints induced by Condition 1 outlined above. To resolve these challenges, we introduce a mathematical transformation which maps the linear system under examination to a much smaller system constructed over only the independent bits of the seed, with the constraints of Condition 1 implicitly embedded, thus enabling a more efficient mathematical treatment of the original problem.

$$\begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{1 \oplus X_1 \oplus X_3} \\ X_1 \\ \mathbf{X_3} \\ X_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} X_1 \\ X_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}$$

Figure 4.7: Constructing a reduced system in independent bit space

The constraints induced by Condition 1 can be extracted from the equations corresponding to the specified bits in TS_i. To attain this goal, an intermediate subsystem consisting of only these equations can be constructed and converted into the *reduced row-echelon* form using Gauss-Jordan elimination [93]. Figure 4.6 illustrates the matrix conversion for the slice matching example given in Figure 4.5. In the *reduced row-echelon* representation, the coefficient matrix columns where the leading 1 of each row appears correspond to dependent variables (X_0 and X_2 in this example) and the remaining columns correspond to independent ones (X_1 and X_3). The constraints induced by these equations can thus be extracted by representing the dependent bits as functions of independent bits, as shown in the figure.

By substituting these functions for the corresponding dependent bits, the extracted constraints can further be embedded into the remaining functions (the ones corresponding to the unspecified bits in TS_i) to construct a reduced linear system over the independent bit space. The slice matching problem thus reduces to the identification of an independent bit pattern which maximally satisfies the equations in the reduced system. Figure 4.7 illustrates the reduced system construction process for the example given in Figure 4.5. As all the dependent variables are eliminated from the representation, the coefficient matrix of the reduced system is much smaller than the original one, thus engendering a much more efficient matrix manipulation in the subsequent seed searching phase. The implicit embedding of the constraints additionally simplifies the subsequent mathematical treatment for

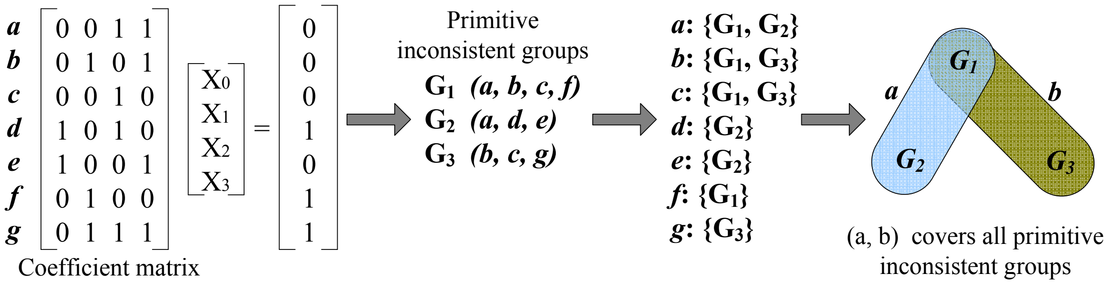


Figure 4.8: Primitive inconsistent group

the slice matching problem.

4.3.3 Noise-optimal seed identification

As the number of valid seeds increases exponentially as a function of the independent bit space dimension, computationally-efficient techniques for the search for the noise-optimal seed, that is, the seed that maximally satisfies the reduced linear system, become highly desirable. Since the seed for a consistent system can be trivially attained through the use of Gauss-Jordan elimination, we focus in this section on the more typical and challenging case of maximally solving a reduced system that is inconsistent.

Basic idea: system relaxation

By definition, an inconsistent system fails to satisfy at least one equation, independent of the seed being used. Dropping a set of equations may transform the inconsistent system to a consistent one but raises the issue of identifying the minimal set of equations to be dropped so as to deliver the maximal possible satisfaction of the original inconsistent system. To help make correct dropping decisions that improve the consistency of the relaxed system, it is important to examine the root cause of linear system inconsistency.

If the column vector in the right hand side of the equation system is incorporated into the coefficient matrix as the last column, an *augmented matrix* can be constructed. A comparison between the ranks of these two matrices indicates that no solution exists for this system if and only if the rank of the augmented matrix exceeds the one of the coefficient matrix. Such a rank mismatch is induced by the linear dependencies among one or more subgroups of rows (equations) in the coefficient matrix. Dropping equations that do not belong to the linearly dependent groups does not improve the system consistency at all and is therefore useless from the system relaxation perspective, as such dropping decisions fail to reduce the rank mismatch in those groups. A successful system relaxation process needs to identify all linearly dependent groups that are inconsistent and convert them to consistent ones by dropping a minimum number of equations. To help attain this goal, we introduce the concept of a *Primitive Inconsistent (PI) group.*

Definition 1. *A group of M linear equations is a* ***primitive inconsistent group****, if it satisfies the following conditions.*

1. *The linear system constructed with these equations is inconsistent.*
2. *Any subsystem of this linear system is consistent.*

Wasteful and unnecessary equation dropping decisions can be avoided if the system relaxation process is constrained to drop purely the equations that belong to primitive inconsistent groups, as the dropping of any such equation suffices to convert the associated primitive inconsistent groups to consistent ones.

An inconsistent system typically contains multiple primitive inconsistent groups. The system shown in Figure 4.8 consists of 7 equations which form a total of three primitive inconsistent groups. As can be observed, distinct primitive inconsistent groups might have shared equations. If we list for each equation the set of primitive inconsistent groups that contains it, the *optimal system relaxation strategy* can be identified by solving a set cover problem which searches a minimum number of equations whose associated sets cover all the primitive inconsistent groups. As shown in Figure 4.8, dropping equations a and b can convert the system to a consistent one as these two equations cover all the three primitive inconsistent groups.

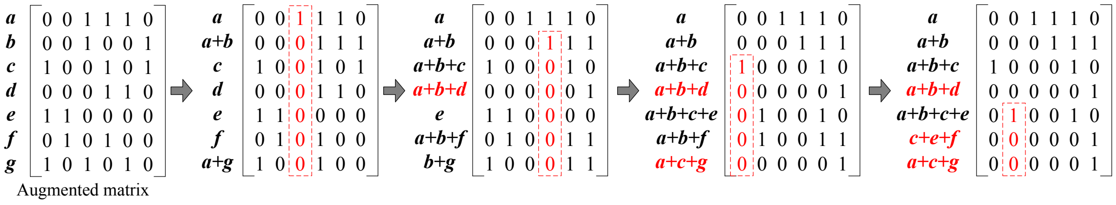


Figure 4.9: PI group identification through forward propagation

Primitive inconsistent group characteristics

Identifying primitive inconsistent groups based on Definition 1 is exceedingly expensive computationally, as a total of $\sum_i C_i^M$ combinations of equations need to be examined in the worst case for a group of M equations. Nonetheless, as shown subsequently, the primitive inconsistent group exhibits certain algebraic characteristics. The proposed technique identifies primitive inconsistent groups by searching for such algebraic characteristics through linear transformations on the augmented matrix of the original system, which significantly reduces the computational cost and speeds up the identification process. The algebraic characteristics of the primitive inconsistent groups can be summarized as in the following lemmas.

Lemma 1. *The coefficient rows of any subset of equations in a primitive inconsistent group are linearly independent.*

Proof. Without loss of generality, let us assume that the primitive inconsistent group contains M equations, E_1 through E_M. We denote the rows of its coefficient matrix as c_1 through c_M, and the rows of its augmented matrix as a_1 through a_M.

Let us assume the coefficient rows of a subset of equations are linearly dependent. Then within this subset of equations we can always find a group of equations E_{i_1} through E_{i_k} ($k < M$), whose coefficient rows c_{i_1} through c_{i_k} fulfill the condition $c_{i_1} + c_{i_2} + ... + c_{i_k} = \overrightarrow{0}$. Consequently, the summation of the rows of the augmented matrix of this set of equations, a_{i_1} through a_{i_k}, would result in the coefficient part[2] of the sum vector being all zeros. Hence there are two cases

[2] Since the augmented matrix is constructed by appending the value column of the equations to the right hand side of the coefficient matrix, the coefficient part of any row in an n-column

to examine depending on the value of the rightmost bit of the row sum.

In the first case wherein the rightmost bit of the row sum is 0, the summation of a_{i_1} through a_{i_k} yields a null vector. Therefore, a_{i_1} can be represented by the remaining equations as $a_{i_1} = a_{i_2} + a_{i_3} + ... + a_{i_k}$. Hence the original inconsistent system E_1 through E_M reduces to a smaller inconsistent system E_{i_2} through E_{i_M}, which contradicts the definition of a primitive inconsistent group.

In the second case wherein the rightmost bit of the row sum is 1, equations E_{i_1} through E_{i_k} already form an inconsistent group which is smaller than the original inconsistent group. This contradicts similarly the definition of a primitive inconsistent group.

Since the two cases discussed above cover all the possibilities yielded by the assumption and they both lead to contradictions, the assumption can hold under no circumstances. Therefore, it is proven that the coefficient rows of any subset of equations in a primitive inconsistent group are linearly independent. □

Lemma 2. *The row sum of the augmented matrix of a primitive inconsistent group is a unit vector. The coefficient part of its row sum is all zeros and the rightmost bit of the row sum is 1.*

Proof. Let us again assume that a primitive inconsistent group consists of M equations. The inconsistency constraint results in the rank of its coefficient matrix being less than M. Therefore, the row vectors of its coefficient matrix must be linearly dependent. Since it has been shown in Lemma 1 that any subset of equations are linearly independent, fulfilling the dependency requirement of the complete set necessitates the satisfaction of the condition $c_1 + c_2 + ... + c_M = \overrightarrow{0}$. Given this, the coefficient part of the row sum of the augmented matrix must be a null vector. If the rightmost bit of the row sum is 0 too, the group becomes a consistent one, which conflicts with the assumption. Therefore, the only possible value of the rightmost bit of the row sum is a 1. □

Lemma 2 describes an important characteristic of primitive inconsistent

augmented matrix is defined as the leftmost $n - 1$ bits of that row.

groups that can be efficiently checked at a quite low computational cost. Utilization of this characteristic significantly speeds up the identification of primitive inconsistent groups, as any group that fails to fulfill the condition outlined in Lemma 2 can be directly excluded from consideration. An effective matrix transformation technique is subsequently proposed to enable the construction of multiple inconsistent subgroups that exhibit the characteristics shown in Lemma 2, significantly reducing the search complexity. The transformation process is furthermore constrained in such a way that all the inconsistent subgroups identified by checking the Lemma 2 characteristics are indeed primitive ones, enabling the utilization of set cover algorithms to filter out the minimum set of shared equations among these groups as the dropping targets.

Primitive inconsistent group identification

The key for primitive inconsistent group identification is the construction of equation groups that exhibit the characteristic of Lemma 2. If the augmented matrix of the system is transformed in such a way that the leading 1's of different rows are in distinct columns, an inconsistent group would always have one special row being a unit vector whose coefficient part is all zeros[3]. The subgroup of equations whose row sum yields this special row must be inconsistent. Based on this observation, we propose a transformation algorithm to generate a matrix with the aforementioned characteristics. The proposed transformation is attained by performing a sequence of row-elementary operations on the matrix. Nonetheless, it differs from the traditional row-elementary operation based transformation techniques such as Gauss-Jordan elimination as only a forward propagation of the rows in the matrix is performed. As detailed subsequently, imposing this constraint during the matrix transformation process guarantees the primitivity of the identified inconsistent groups.

Figure 4.9 presents an illustrative example for the proposed matrix transformation process. The algorithm examines the rows of the matrix from top to

[3] If none of its rows shows such a characteristic after the transformation, the system would be a consistent one as each row (equation) has at least one independent variable.

bottom. For each row under examination, the transformation process identifies the column that contains the leading 1 of the row. If the leading 1 of the row under examination is in the last column, the process skips this row and proceeds to examine the leading 1 of the next row. Otherwise, it determines whether the identified column contains additional 1's in the positions below the row under examination. If any such additional 1 is found, the transformation process eliminates it by adding the row under examination to the row containing this additional 1. In the example given in Figure 4.9, the leading 1 of the first row is in the third column, and the second and seventh rows also have a 1 in the third column. These two additional 1's are eliminated by adding the first row to the second and seventh rows respectively. Such an elimination process is repeated for every row under examination. The entire matrix transformation process is outlined in Algorithm 1. After the transformation, the leading 1's of different rows, if they are not the last bits of the corresponding rows, would lie in distinct columns of the transformed matrix. It should be noted that only forward propagation of the rows is needed to attain the goal of distributing the leading 1's in distinct columns. Compared to the traditional approach such as Gauss-Jordan elimination, this not only reduces the computational cost of the transformation process, but delivers an additional benefit in terms of guaranteeing the primitivity of inconsistent group construction, as detailed subsequently.

If the transformed matrix contains special rows wherein the coefficient part is all zeros and the rightmost bit is 1, this system must be an inconsistent system. The subgroups whose row sums yield these special rows constitute inconsistent subgroups within this system. In the example given in Figure 4.9, three inconsistent subgroups can be identified, namely $\{a, b, d\}$, $\{a, c, g\}$, and $\{c, e, f\}$. Hence the system relaxation process should be constrained to only drop equations from these groups. Nonetheless, as previously discussed, the primitivity of these groups must be guaranteed so as to avoid the dropping of equations that fail to improve the consistency of the system. The proposed transformation process does provide such a desired property, which can be proven based on the following observations.

Observation 1: During the transformation process, no equation can under any

Algorithm 1 *Matrix transformation for primitive inconsistent group identification*

```
for $i = 0$ to $M - 1$ do
  Identify the column $C$ that contains the leading 1 of row $i$
  if Column $C$ is not the last column of the matrix then
    for $j = i + 1$ to $M - 1$ do
      if Row $j$ has a 1 in column $C$ then
        Add row $i$ to row $j$
      end if
    end for
  end if
end for
```

circumstances participate in the row sum computation of rows above it.

Observation 1 is a direct consequence of the forward propagation constraint imposed on the transformation process.

Observation 2: If the transformation process generates, by summing up a subgroup of equations, a row wherein the coefficient part is all zeros, the equation that is located in the lowest row among this subgroup would under no circumstances participate in the row sum computation of any other rows.

It has been shown in Observation 1 that no equation can be added back to the rows above it. Hence the only case that needs to be examined further is whether it can be added to the rows below it by forward propagation. When the lowest row in this subgroup is examined, all the other equations in this subgroup have already been added to it by the transformation process, as the forward propagation of rows is performed in a top-to-bottom manner. Therefore, the lowest row when examined would already be transformed to the row sum of this subgroup: a row vector whose coefficient part is all zeros. In this case, the transformation process will skip this row and directly proceed to the examination of the next row. Hence the equation corresponding to this row would never be propagated forward to the rows below it. Since neither backward nor forward propagation of this equation is possible, it would not participate in the row sum computation of any other rows.

The transformation characteristic described in Observation 2 ensures the primitivity of the identified inconsistent groups, as formally stated in the following theorem.

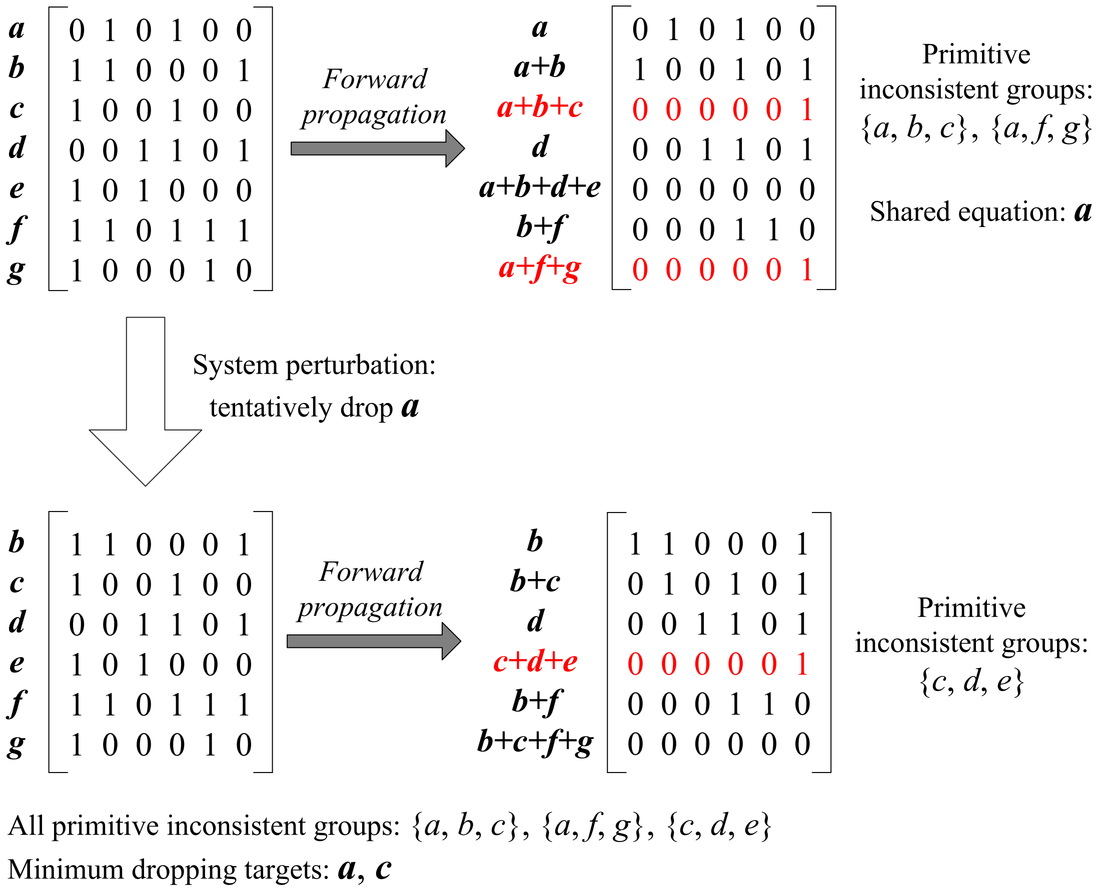
Theorem 1. *If the transformed matrix contains row vectors wherein the coefficient part is all zeros and the rightmost bit is 1, then the subgroups whose row sums generate these vectors are all primitive inconsistent groups.*

Proof. Let us assume the proposed matrix transformation process generates a row vector wherein the coefficient part is all zeros and the rightmost bit is 1, attained by summing up the rows of equations e_{i_1} through e_{i_k}, where $i_1 < i_2 < ... < i_k$. Apparently this group of equations forms an inconsistent group.

Let us assume this group is not primitive. Then there must be a smaller primitive inconsistent group contained within it. Therefore, we can always partition this group into two disjoint subgroups: e_{p_1} through e_{p_m} and e_{q_1} through e_{q_n}, where the row sum of the first subgroup is a null vector and the row sum of the second subgroup has an all-zero coefficient part followed by a single 1. With no loss of generality, let us further assume that e_{p_m} is the lowest row among the first subgroup, and e_{q_n} is the lowest row among the second subgroup. According to Observation 2, such equations would not propagate to the rows above or below them. Consequently, the transformation process would never place these two equations into the same group, which contradicts the assumption. □

The proposed technique is able to identify multiple primitive inconsistent groups in one transformation run, thus enabling system relaxation through the dropping of the shared equations. The system shown in Figure 4.9 contains three primitive inconsistent groups, which can all be identified in just one transformation run. Dropping the shared equations a and e eliminates all the inconsistent groups, thus relaxing the system to a consistent one.

Attaining the minimum dropping set necessitates analyzing all the primitive inconsistent groups simultaneously. Nonetheless, since a certain set of equations



$$\begin{array}{c|cccccc} \boldsymbol{a} & 0 & 1 & 0 & 1 & 0 & 0 \\ \boldsymbol{b} & 1 & 1 & 0 & 0 & 0 & 1 \\ \boldsymbol{c} & 1 & 0 & 0 & 1 & 0 & 0 \\ \boldsymbol{d} & 0 & 0 & 1 & 1 & 0 & 1 \\ \boldsymbol{e} & 1 & 0 & 1 & 0 & 0 & 0 \\ \boldsymbol{f} & 1 & 1 & 0 & 1 & 1 & 1 \\ \boldsymbol{g} & 1 & 0 & 0 & 0 & 1 & 0 \end{array} \xrightarrow{\textit{Forward propagation}} \begin{array}{c|cccccc} \boldsymbol{a} & 0 & 1 & 0 & 1 & 0 & 0 \\ \boldsymbol{a+b} & 1 & 0 & 0 & 1 & 0 & 1 \\ \boldsymbol{a+b+c} & 0 & 0 & 0 & 0 & 0 & 1 \\ \boldsymbol{d} & 0 & 0 & 1 & 1 & 0 & 1 \\ \boldsymbol{a+b+d+e} & 0 & 0 & 0 & 0 & 0 & 0 \\ \boldsymbol{b+f} & 0 & 0 & 0 & 1 & 1 & 0 \\ \boldsymbol{a+f+g} & 0 & 0 & 0 & 0 & 0 & 1 \end{array}$$

Primitive inconsistent groups: $\{a, b, c\}, \{a, f, g\}$

Shared equation: $\boldsymbol{a}$

System perturbation: tentatively drop $\boldsymbol{a}$

$$\begin{array}{c|cccccc} \boldsymbol{b} & 1 & 1 & 0 & 0 & 0 & 1 \\ \boldsymbol{c} & 1 & 0 & 0 & 1 & 0 & 0 \\ \boldsymbol{d} & 0 & 0 & 1 & 1 & 0 & 1 \\ \boldsymbol{e} & 1 & 0 & 1 & 0 & 0 & 0 \\ \boldsymbol{f} & 1 & 1 & 0 & 1 & 1 & 1 \\ \boldsymbol{g} & 1 & 0 & 0 & 0 & 1 & 0 \end{array} \xrightarrow{\textit{Forward propagation}} \begin{array}{c|cccccc} \boldsymbol{b} & 1 & 1 & 0 & 0 & 0 & 1 \\ \boldsymbol{b+c} & 0 & 1 & 0 & 1 & 0 & 1 \\ \boldsymbol{d} & 0 & 0 & 1 & 1 & 0 & 1 \\ \boldsymbol{c+d+e} & 0 & 0 & 0 & 0 & 0 & 1 \\ \boldsymbol{b+f} & 0 & 0 & 0 & 1 & 1 & 0 \\ \boldsymbol{b+c+f+g} & 0 & 0 & 0 & 0 & 0 & 0 \end{array}$$

Primitive inconsistent groups: $\{c, d, e\}$

All primitive inconsistent groups: $\{a, b, c\}, \{a, f, g\}, \{c, d, e\}$

Minimum dropping targets: $\boldsymbol{a}$, $\boldsymbol{c}$

Figure 4.10: Iterative PI group identification through system perturbation

would not be incorporated in the row sum computation, not all the subspaces of the system can be explored in one transformation run. If a primitive inconsistent group happens to exist in the unexplored space, it would not be identified by the current transformation run. The system shown in Figure 4.10 contains three primitive inconsistent groups, whereas the first transformation run can only identify two of them. To identify the additional primitive inconsistent groups, one needs to perturb the system so as to enable the transformation process to search the unexplored space. An efficient way of system perturbation consists of eliminating all the groups identified so far by tentatively dropping their shared equations. If additional primitive inconsistent groups exist, the remaining system would be still inconsistent, thus forcing the next transformation iteration to pinpoint new inconsistent groups. As shown in Figure 4.10, the third primitive inconsistent group can be identified by applying the same transformation technique to the relaxed system after tentatively dropping equation a.

Algorithm 2 $Noise-optimal\ seed\ identification$

```
Construct linear equation system for slice matching
if system consistent then
  Any solution is a noise-optimal seed
else
  Map the system to the independent bit space
  while system inconsistent do
    Identify primitive unsolvable groups through forward propagation
    Add the identified groups into the primitive inconsistent group list
    Identify the minimum set of shared equations using set cover algorithms
    Tentatively drop the identified equations from the original system
  end while
  Solve the relaxed system and generate a seed
end if
```

An iterative system relaxation process can thus be performed to identify the *minimum* set of dropping targets, as summarized in *Algorithm* 2. The algorithm takes the augmented matrix of the original system as input. Each transformation run identifies as many additional primitive inconsistent groups as possible and incorporates them into the primitive inconsistent group list. After each transformation run, the set cover algorithm is applied to identify the minimum set of shared equations among all the primitive inconsistent groups identified so far. Tentatively dropping these equations perturbs the system, giving rise to the identification of new primitive inconsistent groups. Such an iterative relaxation process terminates when the system is finally converted to a consistent one. It should be noted that only the shared equations identified in the last iteration are actually dropped. The equation dropping occurring in all previous iterations is "*tentative*", with the goal being the perturbation of the system. In any iteration, the equations are dropped from the original equation list, with the dropping decisions of the previous iterations being ignored. Since the last iteration is able to consider all the primitive inconsistent groups concurrently, an *optimal system relaxation* solution is thus ensured. Solving the relaxed, consistent system delivers the compression seed that minimizes scan-mode noise.

The computational complexity of the seed search process is determined by the speed of identifying primitive inconsistent groups. For a linear system consisting of m variables and n equations, the proposed transformation performs at most $\sum_{i=1}^{n} i * m = n(n+1)m/2$ single-bit binary additions, which bounds the computation time for each iteration of primitive inconsistent group identification. As at least one new primitive inconsistent group will be identified in each iteration leading to the dropping of additional equations from the system, the number of iterations required for identifying all primitive inconsistent groups is typically less than n, indicating that the total number of computational operations is less than $n^2(n+1)m/2$. In practical cases, since the proposed transformation technique is able to identify a large number of such groups in one transformation run, only a few iterations are needed for identifying the *optimal compression seed.*

4.4 Concurrent compaction/compression for low noise scan

The attainment of full fault coverage with linear compression schemes necessitates the incorporation of such schemes into the ATPG flow wherein concurrent compaction and compression are performed. However, traditional schemes perform a greedy compaction on the unfilled test cubes as long as the linear compression on the resulting cube is feasible. This strategy would result in a highly unbalanced rank distribution among linear equation systems for the cubes being compressed, which in turn causes a high skew in the seed space size among the cubes, thus degrading the effectiveness of the noise-aware compression technique.[4]

To maximally exploit the noise reduction benefit of the proposed scheme, it is essential to uniformly provide sufficient seed space for each cube being compressed. This requires the ATPG algorithm to perform more informed target selection and constrain the extent of cube compaction. We thus propose a *noise-aware compaction/compression flow* based on these principles.

It is worth noting that a newly generated cube can usually be compacted with multiple cubes as long as the resulting cube passes the compressibility check. Therefore, it is possible to select the compaction candidate which impacts the noise reduction potential the least. Since the system rank of the resulting cube can be attained during the compressibility check, the proposed compaction strategy will always select the compaction candidate which yields the minimum rank in the linear system of the resulting cube. Such a strategy significantly decreases the skew in the rank distribution among the cubes being compressed, thus resulting in a more balanced scan noise reduction among different cubes.

To ensure a sufficient number of candidate seeds for each cube, we further impose a threshold in the system rank of the compacted cube. If the system rank of a cube exceeds the threshold, the cube will be relayed to the compression phase, and no further compaction on it is necessitated. The rank constraint precludes

[4] The size of the seed space is a function of the rank of the coefficient matrix of the linear system, as has been discussed in Section 4.2.

Algorithm 3 *$Noise - aware\ ATPG$*

```
$TL \leftarrow \emptyset$
$CL \leftarrow \emptyset$
$FL \leftarrow$ fault list
while $FL \neq \emptyset$ do
  Select a fault $f$ and generate a test cube $C$ for $f$
  Perform compatibility and compressibility check to identify compaction candidates for $C$
  if Compaction candidates exist then
    Identify the candidate $C'$ that leads to minimum rank in the merge result, and merge $C$ into $C'$
    if rank($C'$) $\geq$ threshold then
      Noise-aware compression on $C'$ to completely specify it, and delete $C'$ from $CL$
      Fault simulation on $C'$ followed up by dropping all detected faults from $FL$
      $TL \leftarrow TL + C'$
    end if
  else
    $CL \leftarrow CL + C$
  end if
end while
Compress remaining cubes in $CL$
Reverse fault simulation to drop redundant tests from $TL$
Output $TL$
```

the possibility of aggressive compaction shrinking the seed space. On the other hand, the imposition of the rank threshold might slightly increase the number of tests due to the reduced compaction level. However, such an increase is usually negligible, as a large portion of faults covered by the uncompacted cubes can be fortuitously dropped through the fault simulation of previous test cubes which have been completely specified by the compression phase. This observation has been confirmed by our experimental results.

The proposed ATPG flow is summarized in *Algorithm* 3. When an unfilled test cube is generated for a fault in the fault list (FL), the algorithm first tries to compact it with previously generated unfilled cubes stored in the cube list (CL). The selection of the compaction candidate is based on the rank balancing principle outlined above, and the compaction process is constrained by the rank threshold. The cubes that exceed the rank threshold are compressed with a focus on scan noise reduction. The resulting fully specified tests are added onto the test list (TL), and fault simulations on them are performed to drop all the faults detected. This process iterates until no detectable faults remain in the fault list, thus providing a complete fault coverage.

4.5 Experimental results

The proposed scheme has been implemented for effectiveness validation. Our implementation uses the Atalanta [54] ATPG tool as the engine for unfilled test cube generation. The HOPE [55] simulator is utilized for fault simulation. The concurrent compaction and compression algorithm is implemented using the C programming language.

Since the power ground noise in scan mode has been observed to be highly correlated with the dynamic power, we examine the effectiveness of the proposed technique using the dynamic power metric that can be easily quantified using established models [33]. We first compare the proposed technique to our prior noise/power reduction technique [16], as this technique is subject to exactly the

Table 4.1: Benchmark circuits

Circuit	# PI	# FF	# Scan chain
s13207	62	638	160
s15850	77	534	160
s35932	35	1728	160
s38417	28	1636	160
s38584	38	1426	160
FPU	187	13803	160

same compression constraints thus enabling an apples-to-apples comparison which directly illustrates the effectiveness of the proposed idea. The linear compression technique in [6] is used as a comparison baseline. The noise-aware compression technique in [16] and the proposed technique are compared to the baseline approach, and their impact on dynamic scan power and compression ratio are reported. To perform a fair comparison, we use the same decompression hardware, which consists of a 33-to-160 XOR-network, in all three schemes. In line with well-established practice in VLSI test literature, we experiment on some of the more aggressive ISCAS89 benchmarks. We furthermore extend the scope of experimentation by incorporating an examination of our technique on an industrial design, namely a floating-point unit (FPU). The general information about these circuits and the characteristics of the scan architecture employed are provided in Table 4.1.

To explore the impact of the compaction level on dynamic scan power and compression ratio, the noise-aware compression process is repeated multiple times for each circuit, with various rank thresholds applied during the compaction process.

The average scan power comparison is presented in Table 4.2. The second column of this table presents the rank threshold specified for each run. The average scan power of the three schemes under comparison is summarized in columns 3, 4, and 6, respectively. It can be observed that significant power reduction can be delivered by both noise-aware compression schemes compared to the baseline. The proposed scheme results in appreciable improvement over the technique in [16], as the proposed compression process minimizes the number of linear equations that

Table 4.2: Average scan noise comparison

Circuits	R_{th}	Comp. [6]	Noise-aware comp. [16]		Proposed	
		P_{ave}	P_{ave}	%redc. [6]	P_{ave}	%redc. [6]
s13207	11	322	197	38.8	181	**43.8**
	15		195	39.6	184	**42.8**
	19		200	38.0	187	**41.8**
	23		194	39.8	190	**40.8**
	27		215	33.1	196	**39.0**
s15850	11	239	144	39.9	133	**44.4**
	15		165	31.0	150	**37.3**
	19		157	34.0	152	**36.3**
	23		167	30.1	156	**34.6**
	27		168	29.7	163	**31.8**
s35932	11	905	623	31.2	610	**32.6**
	15		661	26.9	645	**28.7**
	19		687	24.1	660	**27.0**
	23		706	22.0	687	**24.1**
	27		715	21.0	695	**23.2**
s38417	11	871	594	31.9	576	**33.9**
	15		610	30.0	588	**32.4**
	19		614	29.5	606	**30.5**
	23		625	28.3	613	**29.6**
	27		641	26.4	623	**28.5**
s38584	11	783	502	35.8	481	**38.6**
	15		510	34.9	492	**37.2**
	19		537	31.3	502	**35.8**
	23		553	29.4	533	**32.0**
	27		572	26.9	539	**31.1**
FPU	11	7104	4632	34.8	4546	**36.0**
	15		4803	32.4	4630	**34.8**
	19		4890	31.2	4651	**34.5**
	23		4996	29.7	4755	**33.1**
	27		5185	27.0	4779	**32.7**

can not be satisfied, thus minimizing the toggling between adjacent scan slices. It can be furthermore observed that a lower rank threshold in general results in a higher power reduction, as less aggressive compaction is performed in this case, leading to a higher flexibility in seed selection during compression. Although the effectiveness of the proposed scheme decreases at higher rank thresholds, a $20\%-40\%$ reduction in scan power can still be attained even with highly aggressive compaction.

Table 4.3 presents a comparison in terms of peak scan power. The rank threshold impacts the peak power in a manner similar to its effects on the average power. It can be observed that the proposed scheme delivers appreciable peak power reduction through the entire range of the rank threshold.

The compression ratio of each scheme is further examined in Table 4.4. The volume of the compacted test set generated by Atalanta [54] is used as the baseline volume (i.e., the uncompressed data volume) for computing the compression ratio that is defined in the following equation.

$$\frac{Uncompressed\ Data\ Volume}{Compressed\ Data\ Volume} \tag{4.1}$$

The compression ratios (CR) for the traditional linear compression scheme and the two noise-aware compression schemes are shown in columns 3, 4 and 6, respectively. Columns 5 and 7 present the compression ratio loss of the noise-aware compression schemes compared to [6]. It can be observed that the compression ratio attained by the noise-aware compression schemes in general increases with the rank threshold, as more aggressive compaction would be performed with a higher rank threshold, potentially resulting in fewer tests. The compression ratios attained by the two noise-aware compression schemes are quite close. The traditional compression scheme [6] typically yields a higher compression ratio compared to the noise-aware compression schemes, as it performs the most aggressive compaction. Nonetheless, the difference is quite small when the rank threshold of the noise-aware compression process is set to a high value. For all of the benchmarks examined, the proposed scheme delivers appreciable power reduction at the cost of a very small compression ratio loss for rank thresholds surpassing 19.

In order to further illustrate the power reduction capability of the proposed

Table 4.3: Peak scan noise comparison

Circuits	R_{th}	Comp. [6]	Noise-aware comp. [16]		Proposed	
		P_{max}	P_{max}	%redc. [6]	P_{max}	%redc. [6]
s13207	11	351	270	23.1	261	**25.6**
	15		273	22.2	270	**23.1**
	19		286	18.5	279	**20.5**
	23		304	13.4	295	**16.0**
	27		319	9.1	304	**13.4**
s15850	11	271	218	19.6	202	**25.5**
	15		224	17.3	210	**22.5**
	19		218	19.6	205	**24.4**
	23		227	16.2	211	**22.1**
	27		236	12.9	234	**13.7**
s35932	11	995	717	27.9	706	**29.0**
	15		754	24.2	727	**26.9**
	19		773	22.3	741	**25.5**
	23		793	20.3	780	**21.6**
	27		801	19.5	789	**20.7**
s38417	11	930	670	28.0	659	**29.1**
	15		701	24.6	680	**26.9**
	19		712	23.4	694	**25.4**
	23		729	21.6	709	**23.8**
	27		743	16.6	720	**22.6**
s38584	11	844	618	26.8	596	**29.4**
	15		631	25.2	608	**28.0**
	19		650	23.0	613	**27.4**
	23		674	20.1	651	**22.9**
	27		704	16.6	671	**20.5**
FPU	11	8377	6552	21.8	6371	**23.9**
	15		6740	19.5	6490	**22.5**
	19		6891	17.7	6584	**21.4**
	23		6904	17.6	6750	**19.4**
	27		7018	16.2	6803	**18.8**

Table 4.4: Compression ratio comparison

Circuits	R_{th}	Comp. [6]	Noise-aware comp. [16]		Proposed	
		CR	CR	$CR_{loss}\%$	CR	$CR_{loss}\%$
s13207	11	7.13	6.49	8.98	6.43	9.87
	15		6.98	2.17	6.76	5.21
	19		7.30	1.47	6.93	2.86
	23		6.90	3.21	6.90	3.21
	27		7.30	1.47	7.00	1.82
s15850	11	10.40	8.78	15.61	8.59	17.40
	15		9.26	10.98	9.13	12.19
	19		9.89	4.91	9.99	3.97
	23		10.09	3.01	9.84	5.37
	27		10.90	-4.81	11.02	-5.95
s35932	11	6.29	5.17	17.85	4.99	20.68
	15		5.36	14.80	5.36	14.80
	19		6.03	4.15	6.16	2.11
	23		7.06	-12.21	6.29	0
	27		6.73	-6.99	6.43	-2.24
s38417	11	19.19	14.59	24.0	15.14	21.13
	15		15.61	18.67	15.31	20.22
	19		16.57	13.63	17.00	11.43
	23		17.52	8.73	18.15	5.42
	27		18.23	4.99	18.92	1.41
s38584	11	12.11	10.99	9.23	10.83	10.55
	15		11.16	7.86	11.37	6.10
	19		11.64	3.89	11.78	2.75
	23		11.73	3.13	11.73	3.13
	27		12.16	-0.44	11.92	1.57
FPU	11	26.48	24.83	6.22	24.42	7.77
	15		25.08	5.26	25.00	5.58
	19		25.65	3.12	26.25	0.88
	23		25.79	2.61	26.15	1.23
	27		26.82	-1.29	27.12	-2.40

Table 4.5: Comparison to schemes in [66] in average noise

Circuits	Shift-aware X-Filling [66]	X-Filling +Flipping [66]	Comp. [16] w/ $R_{th} = 11$	Proposed comp. w/ $R_{th} = 11$
	% P_{ave} redc.	% P_{ave} redc.	% P_{ave} redc.	% P_{ave} redc.
s13207	17.9	20.6	39.2	43.8
s35932	32.4	50.9	31.2	32.6
s38417	29.2	30.8	31.9	33.9
s38584	18.0	19.8	35.8	38.6
ave.	24.4	30.5	34.5	37.2

Table 4.6: Comparison to schemes in [66] in peak noise

Circuits	Shift-aware X-Filling [66]	X-Filling +Flipping [66]	Comp. [16] w/ $R_{th} = 11$	Proposed comp. w/ $R_{th} = 11$
	% P_{peak} redc.	% P_{peak} redc.	% P_{peak} redc.	% P_{peak} redc.
s13207	12.9	15.0	23.1	25.6
s35932	16.6	18.3	27.9	29.0
s38417	17.7	20.1	28.0	29.1
s38584	13.0	14.9	26.8	29.4
ave.	15.1	17.1	26.5	28.3

technique, we provide herein a comparison to the two most effective scan power reduction schemes [66] that are based on sequential linear compression schemes, as summarized in Tables 4.5 and 4.6. The proposed scheme in general delivers higher percentage of peak and average power reductions. While the power reduction advantage of the proposed scheme may be influenced by the differing initial compression ratios involved in the comparison, it is noteworthy that our technique can be utilized on top of any initial linear compression technique to deliver appreciable power improvement as it is able to perform global search within the solution space of the linear equation system.

Finally, we provide a comparison to a state-of-the-art scan noise reduction technique [103] not based on linear compression schemes. The peak power and total power reduction over the random X-fill scheme is reported in [103]. Since the total power value cannot be directly mapped to average power[5], we perform a

[5] The conversion from total power to average power depends on the number of shift-cycles which is not reported in [103].

Table 4.7: Comparison to schemes in [103]

Circuits	Scan partitioning [103]	Comp. [16] w/ $R_{th} = 11$	Proposed comp. w/ $R_{th} = 11$
	% P_{peak} redc.	% P_{peak} redc.	% P_{peak} redc.
s13207	24.0	23.1	25.6
s15850	30.0	19.6	25.5
s38417	17.0	28.0	29.1
s38584	13.0	26.8	29.4
ave.	20.8	24.4	27.4

Table 4.8: Algorithm execution time in minutes

Circuits	ATPG	Compression	Total
s13207	0.24	1.77	2.01
s15850	0.32	1.26	1.58
s35932	1.84	0.91	2.75
s38417	2.77	3.61	6.38
s38584	1.24	3.53	4.77
FPU	38.50	32.28	70.78

comparison solely in terms of peak power reduction, as shown in Table 4.7. It can been seen that the proposed technique delivers higher peak power reduction for all benchmarks except for s15850. The advantage of the proposed technique is more evident in large benchmarks such as s38417 and s38584, which indicates a better applicability of the proposed technique in large industrial designs.

The computation time of the ATPG process and the proposed noise-aware compression scheme is reported in Table 4.8. Since the seed searching is performed for each slice of the test cube, the compression time is in general proportional to the product of the number of test patterns and the scan chain length. It can be further observed that, when the benchmark size increases, the compression time increases at a lower pace than the ATPG time, indicating a good scalability of the proposed compression technique. For large SOC designs consisting of multiple embedded testable cores, the proposed technique can be applied for each core independently and the computation for different cores can be performed in parallel on distributed servers in an enterprise platform, enabling the application of the proposed technique to industrial designs.

4.6 Conclusions

Modern industrial scan designs are confronted with the challenge of excessive test volume and noise, necessitating the concurrent resolution of the test compression and noise reduction problems. The somewhat divergent optimization criteria impose strong constraints in the compression phase, significantly increasing the difficulties in identifying the optimal compression results.

A noise-aware linear compression scheme is proposed in this chapter, which exploits the flexibility in the seed space and identifies noise-optimal seeds during the compression process. The proposed scheme extracts and embeds the compression constraints imposed by the XOR network through a mathematical transformation, thus enabling a constraint-free and computationally-efficient seed searching process in a highly reduced space. The search for the noise-optimal seeds is then accurately modeled as a problem of maximally solving a possibly inconsistent linear system over the reduced space. A novel linear system pruning strategy is employed to identify the optimal solution through a highly guided system relaxation process. Optimization strategies for postprocessing the results to approximate a global optimum are also proposed.

The proposed compression technique is further embedded into a noise-aware ATPG flow wherein a controlled compaction strategy is performed to balance the noise reduction potential across different test cubes. The proposed scheme necessitates no alteration to current linear compression hardware, introduces negligible hardware and timing overhead to the design, and imposes no constraints on the ATE control/synchronization mechanism; thus it can be easily incorporated into a variety of industrial applications. Compared to the traditional compression schemes, the concurrent compaction and compression scheme proposed in this chapter delivers appreciable average and peak scan noise reduction while incurring negligible impact on the compression ratios.

The text of Chapter 4, is in part a reprint of the material as it appears in *M. Chen and A. Orailoglu, “Scan power reduction in linear test data compression* *scheme," International Conference on Computer-Aided Design, 2009*; and in *M. Chen and A. Orailoglu, "Scan power reduction for linear test compression schemes through seed selection," IEEE Transactions on VLSI*. The dissertation author was the primary researcher and author of the publications [16] and [15].

Chapter 5

Detecting functional-mode marginal failures

The mitigation of scan-mode power ground noise mainly helps minimize the yield loss due to the overtesting of non-functional paths in the circuit. Yet the development of a production test plan also needs to fulfill the other side of the equation, that is, the undertesting of functional paths also needs to be avoided to maintain an extremely low DPPM level. Circuits designed and fabricated with nanometer scale technology, especially the ones with long functional paths, are sensitive to power ground noise across a wide frequency range, thus necessitating a strict examination of circuit robustness against noise during manufacturing test.

Conventional at-speed testing techniques possibly result in the escape of marginal timing failures as they are unable to account for the impact of middle and low frequency noise on circuit timing. To address this challenge, we propose, in this chapter, a novel multi-functional-cycle test scheme that targets the noise-induced failures on critical paths of the circuit. The proposed technique explores the noise profile of at-speed functional cycles and approximates it in delay testing through the application of multiple capture operations, thus maximally detecting the timing failures that potentially take place under the worst-case functional mode noise. The noise impact of individual devices on the critical paths is characterized through simulations on the power mesh model extracted from the circuit layout. This enables a computationally efficient yet SPICE-accurate estimation of the compound noise profile of the test pattern through the linear superposition of individual ones. Guided by this noise estimation technique, a test pattern transformation flow is proposed to maximize the noise in pseudo-functional test operations.

In this chapter, we first present a technical overview of the proposed methodology of approximating functional noise profile in structural testing. The multi-functional-cycle test scheme is then introduced in Section 5.2. The power ground noise estimation framework is detailed in Section 5.3. Section 5.4 explores a test pattern transformation technique that approximates the worst-case functional noise in testing mode.

5.1 Handling noise in delay testing

5.1.1 Higher or lower noise? A dilemma for testing

As can be seen from the literature review, the treatment of capture cycle noise is a somewhat controversial topic. One category of research focuses on the mitigation of the power ground noise in the capture cycle to reduce false alarms, whereas another set of research focuses conversely on the maximization of the noise to reduce test escapes. Both opinions come with their own rationales that might be related to the specific characteristics of chip design, application domain, and quality requirement. Putting aside the difference in the detailed treatment strategies, it can be seen that these opinions inherently share a common observation, that is, the test mode noise profile significantly deviates from that of the functional mode, thus failing to approximate the functional mode failure mechanism.

The scan phase typically has very high toggling density as the associated state transitions are completely non-functional. The resulting strong power ground noise leads to a high failure probability on the scan paths. Since the noise profile of the scan phase is completely different from that of the functional mode and the scan paths are disabled during functional operation, the noise-induced failures in scan mode can be considered as false alarms. Therefore, the mitigation of scan mode noise can effectively reduce yield loss without impacting the test quality, as

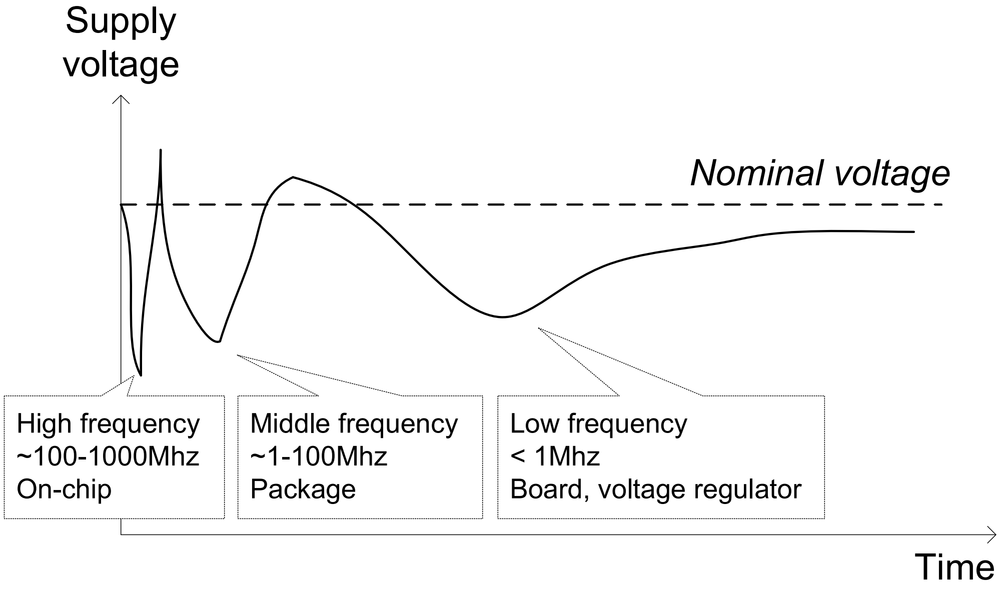


Figure 5.1: Power ground noise frequency at different circuit levels

discussed in a number of previous works. In industrial practice, a slow scan clock is typically employed to reduce the noise effect.

The capture phase noise, though, is much more intricate and requires more subtle treatment during testing. On the one hand, the capture operation exercises the functional timing paths of the circuit. Therefore, examining these paths under a noise situation helps detect the noise failures that can occur during functional operation. On the other hand, the initial circuit state at the capture cycle is still quite different from the true functional ones, potentially resulting in a certain level of deviation in the noise profile. The generation and application of delay test patterns need to take both factors into account so as to deliver high test quality without impacting yield.

To illustrate the noise behavior in at-speed clock cycles, we show in Figure 5.1 the resonant frequency of the RLC network at various circuit levels. As shown in the figure, the power ground noise of a circuit system exhibits a very wide bandwidth. The on-chip inductance and capacitance typically result in high frequency noise ranging from 100Mhz to 1Ghz. This type of noise dies down rapidly and typically does not span multiple clock cycles. The inductance and capacitance increase significantly at the package level, resulting in a much slower resonance effect than the on-chip level. Therefore, the noise frequency at this level typically ranges from 1Mhz to 100Mhz. The board level has the largest inductance and capacitance, thus resulting in slow frequency noise that is less than 1Mhz. Since modern circuits typically operate at very high clock frequency, both the middle and low frequency noise require multiple functional cycles to die down. The noise profile during circuit operation is a compound effect of noise over the entire bandwidth. Research work reported by Intel [79] has shown that multiple at-speed cycles are needed to generate the worst-case noise in a circuit. For illustrative purposes, we show in Figure 5.2 the silicon characterization results reported by Intel. In this characterization practice, the power supply response is measured for a burst of 30 at-speed cycles. A total of two voltage dips are encountered during the 30 cycles, with the first one representing the worst-case noise[1]. Because of the impact of the middle and low frequency noise, a total of 5 clock cycles are needed for the noise to fully develop to the worst-case.

The delay test scheme currently employed in industry typically has one slow launch cycle followed by only one at-speed capture cycle, as shown in Figure 5.3. Since the single capture cycle does not give enough time for the full development of the noise in high speed circuits, the capture edge might see too little voltage drop, thus resulting in optimistic testing results. For high performance circuits, overlooking marginal timing failures might lead to a high customer return cost. A novel test scheme, which approximates the functional operation while creating the worst-case noise situation, is needed to improve the delay test quality.

5.1.2 Approximating functional mode noise in testing

The aforementioned analysis has shown the importance of considering the wide-bandwidth noise effect during testing. Since the low frequency noise is caused by the motherboard design quality which is independent of the chip design itself,

[1] The amplitude of the first voltage droop is 13.5% of the nominal voltage. For 65nm designs with a nominal voltage of around 1.2V, the voltage droop can be as high as 160mv.

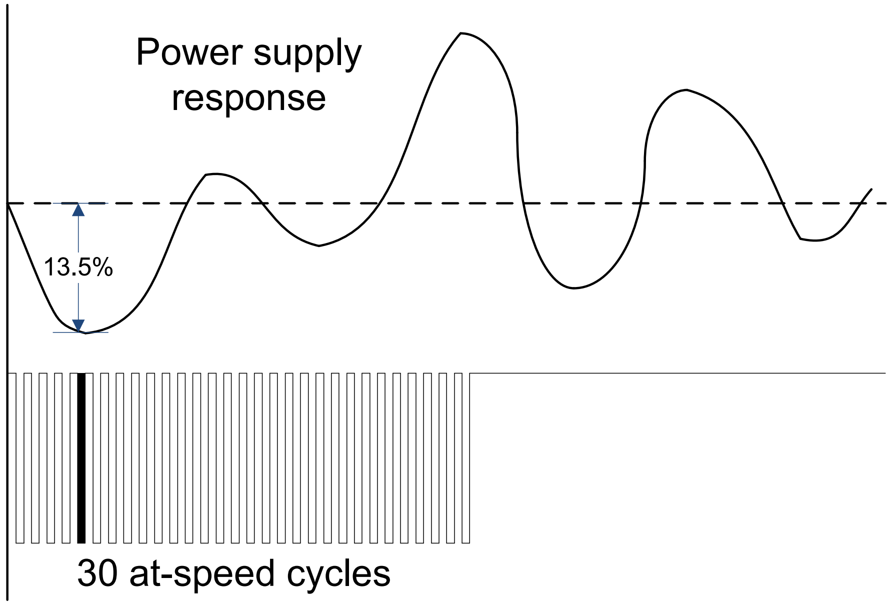


Figure 5.2: Power supply noise during at-speed cycles

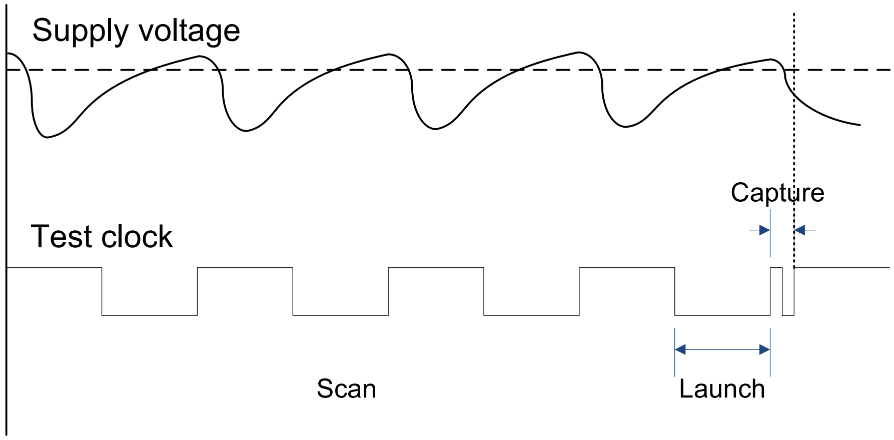


Figure 5.3: Power supply noise during at-speed cycles

it should be excluded from the chip quality criteria and alleviated through the correct tuning of the board decoupling capacitance and power regulator design. The generation of delay test needs to focus on the high and middle frequency noise that is caused by on-chip and package inductance and capacitance. The power mesh resonance frequency of a packaged chip is inherently determined by the RLC model that can be extracted from the chip layout. Hence the number of at-speed cycles that are needed for full noise development can be estimated through simulation. Guided by this information, it is possible for a delay test scheme to apply multiple at-speed functional cycles after the scan-in of a test pattern, with the last at-speed cycle exercising the timing check on the target path. This scheme delivers two main advantages over the conventional one.

1. The sequence of at-speed cycles triggers the full development of the voltage drop. Since the number of necessary at-speed cycles can be predicted, this scheme is able to approximate the worst-case noise that is determined by the circuit resonance characteristics, delivering a better delay test quality.

2. During the at-speed cycles, the circuit performs a sequence of regular state transitions. Therefore, the state transition occurring at the last at-speed cycle is expected to be much more “functional” than that of the first one. This reduces the risk of a false alarm triggered by a non-functional state transition, and potentially yields a noise profile that is closer to the true functional one.

Such a multi-functional-cycle delay test can be generated for the most critical paths in the circuits and applied during the package-level screening of the test program. The need for multi-cycle state transitions imposes a certain level of overhead for test generation and application. Nonetheless, since the noise frequency at the package level can typically die down within 100ns, it is highly probable to reach the voltage dip within less than 10 at-speed cycles, as shown in the characterization results from Intel. Therefore, the overhead can be controlled at a very minimal level, while delivering a significant improvement of the delay test quality.

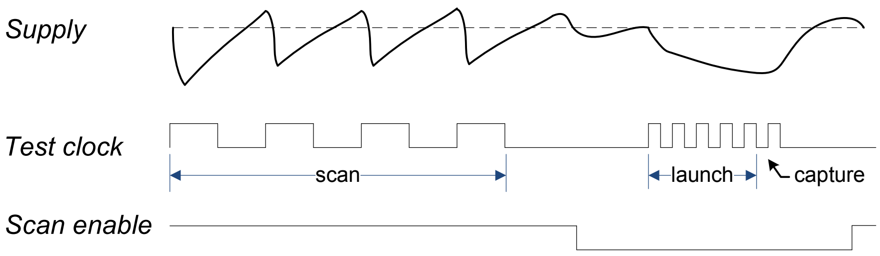


Figure 5.4: Test clock diagram of proposed test scheme

The technical details of the proposed methodology are presented in subsequent sections.

5.2 Multi-functional-cycle delay test

5.2.1 Test application scheme

The proposed multi-functional-cycle delay test scheme is illustrated in the test clock diagram shown in Figure 5.4. The proposed test application process mainly consists of a slow scan phase and an at-speed functional phase. The test stimuli are shifted into the scan chain using slow scan cycles so as to mitigate the power ground noise during the scan phase. An idle period is inserted between the scan phase and the functional phase to ensure that the scan noise dies down before the start of the functional phase and that the scan-enable signal has sufficient time to toggle[2]. The functional phase consists of a sequence of N at-speed functional clock cycles, with the first $N-1$ cycles launching the toggling behavior on the target path through consecutive functional state transitions and the last cycle capturing the path toggling for path delay checking.

The proposed delay test scheme differs from conventional ones mainly in the launch phase. Conventional schemes have only one at-speed clock cycle, namely,

[2] It is highly difficult to attain an at-speed scan-enable signal in large scan designs, as the scan-enable signal is a global signal that needs a large distribution network. The proposed clock scheme eliminates the need for an at-speed scan-enable signal, thus reducing the timing closure difficulty in back-end design.

the capture cycle. The launch cycle in conventional schemes is typically a slow scan cycle (launch-off-shift) or slow functional cycle (launch-off-capture), neither of which is capable of creating the noise profile encountered in functional operation. In the proposed scheme, the burst of launch cycles triggers at-speed state transitions that approximate the functional operations. This process not only gives rise to the full development of a functional mode noise profile, but makes the circuit state more similar to functional ones through the consecutive state transitions. As a result, a highly stressful yet "functional" noise situation can be created for the capture cycle to enable the maximal detection of noise-induced timing failures on the target path.

5.2.2 Test generation flow adaptation

The proposed test application scheme raises a number of challenges in the test generation flow. First of all, the use of multiple functional cycles necessitates the generation of test cubes that fulfill the fault excitation conditions of a few cycles before the capture cycle. This can be resolved through the use of a sequential ATPG engine. More importantly, the ATPG process needs to make appropriate decisions on a number of key parameters in order to ensure the generation of the worst-case noise in the capture cycle. There are two levels of flexibilities that need to be investigated during test generation.

1. **The number of at-speed functional cycles** This parameter is one additional dimension of ATPG flexibility introduced by the proposed scheme. The approximation of the worst-case functional noise necessitates the accurate identification of the functional sequence length that is needed for full noise development.

2. **The filling of unspecified values in the test cube** Distinct fillings of the unspecified values in the test cube lead to completely different sequences of the circuit states during the functional cycles, thus resulting in a highly varying noise profile in the functional phase. The ATPG algorithm needs to search in the X-filling space and identify the strategy that maximizes the

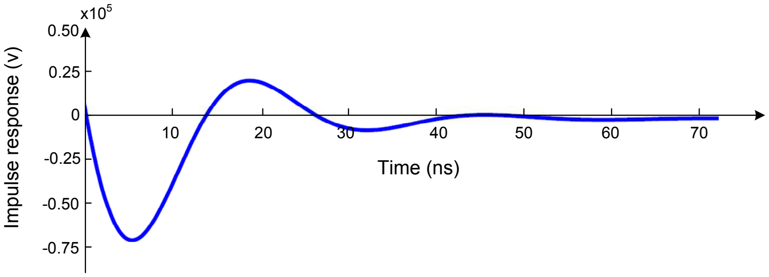


Figure 5.5: Impulse response

noise.

The aforementioned flexibilities significantly enlarge the search space that the test generation process needs to traverse, thus imposing critical challenges on the efficiency of the ATPG flow. Among all the factors that contribute to the computational cost, the noise estimation step constitutes the most critical one, not only because the estimation of multi-cycle noise impact relies on expensive simulation, but because this complex analysis needs to be repetitively performed multiple times during test generation. The minimization of simulations without impacting estimation accuracy thus becomes a key optimization goal. The development of such a noise estimation technique constitutes one of the focal points of this work.

5.3 Power ground noise estimation

The search for delay test patterns that create worst-case noise necessitates resolutions to two noise estimation problems, one being the determination of the number of functional cycles needed for full noise development, the other being the noise waveform estimation for various X-fillings of test cubes.

5.3.1 Noise development time analysis

The noise waveform of a power distribution network (PDN) is inherently the response of the PDN system to the external current stimuli [81, 28]. In VLSI circuits, the current stimuli are mainly generated from the toggling of gates and flip-flops. If we denote the current stimuli as a function of time, $i(t)$, the noise as a function of time can be attained by calculating the convolution between the current stimuli and the system impulse response, as shown in Equation 5.1.

$$v(t) = \int_0^t z(\tau) i(t-\tau) d\tau \tag{5.1}$$

where $z(\tau)$ is the impulse response of the PDN system. This equation explains from an analytical perspective why the full noise development can span multiple clock cycles. Figure 5.5 illustrates an impulse response waveform of a PDN. The resonance frequency and die-down speed of the impulse response are completely determined by the inherent property of the system[3]. The convolution process described in Equation 5.1 can be conceptually considered as a weighted sum of the current amplitude at different time points, with the weight being the impulse response. Consequently, the current at $t - \tau$ would contribute to the noise at t, as long as τ is less than the die-down time T of the impulse response. This behavior results in a noise accumulation time window of size T, that is, the noise amplitude at any time point t is an accumulated effect of the current waveform in the time window $[t - T, t]$. The current waveform that occurs earlier than $t - T$ has no contribution to the noise at t, as its noise impact already dies down before t.

The aforementioned analysis indicates that the die-down time T is sufficient for the full noise development during the functional phase of the proposed test scheme, since T is the size of the maximal noise accumulation window. It should be noted that the impulse response resonates between the undershoot and overshoot regions, consequently generating positive and negative portions during the noise accumulation process. As a result, the die-down window T only specifies the possible time range of the worst-case noise occurrence. The worst-case noise

[3] The impulse response keeps attenuating over time. It is considered as having “died down” when the amplitude falls short of a certain threshold.

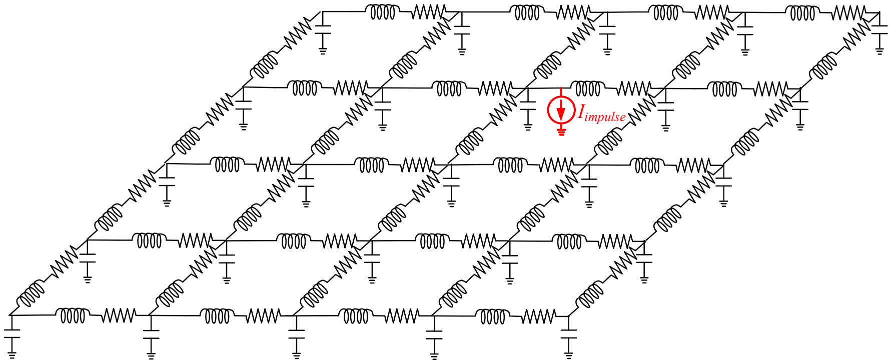


Figure 5.6: Simulation model for impulse response

can occur at any time point within this window, depending on the specific noise accumulation process.

The die-down time T can be accurately estimated through SPICE simulation. As shown in Figure 5.6, the simulation model consists of a PDN RLC model extracted from the chip layout using a parasitic extraction tool and an impulse current source applied to the nodes that map to the layout location of the target path. The simulated voltage waveform on the node clearly shows the impulse response die-down time of this system. According to empirical data, the impulse response die-down time at the package-level typically ranges from a few to tens of nanoseconds.

Given the impulse response die-down time T and the functional clock period p, it can be calculated that the worst case noise would occur within one of the first $N = T/p$ functional cycles. To generate the worst-case noise in the capture cycle, the test generation process needs to examine a total of N different candidate test application scenarios, each with a distinct number of functional cycles. For each candidate test application scheme, the maximum noise in the capture cycle needs to be estimated. The test application scheme that delivers the worst case noise is then selected to be incorporated into the delay test flow.

5.3.2 Noise profile creation through linear superposition

The ATPG process needs to be guided by an efficient noise estimation strategy in order to generate the pattern that creates the worst-case noise profile. The weighted switching activity (WSA) metric has been employed in a number of conventional noise-aware ATPG approaches for noise estimation. Nonetheless, this single-cycle based metric is unable to account for the multi-cycle noise accumulation process, the key effect considered in the proposed delay test scheme. An accurate multi-cycle noise waveform can be attained by simulating the voltage response of current stimuli, albeit at the cost of a nontrivial simulation time. An innovative noise estimation framework is needed to provide an accurate multi-cycle noise profile at a highly constrained simulation cost.

One widely observed characteristic of the power ground noise consists of the highly localized noise effect [67, 68, 95], that is, the power ground noise on a timing path is mainly impacted by the behavior of cells that are physically close to the path. Based on this observation, the proposed flow focuses on estimating the noise impact of cells that are within a certain *Effect Range* of the target path. The identification of the *Effect Range* from layout has been discussed in previous work [67, 68, 95]. Typically this region is defined as a rectangle area that is within a few neighboring standard cell rows and a certain range along each row. The cells within the *Effect Range* contribute to the noise by drawing current from the power supply or feeding current to the ground, depending on whether the cell performs a rising or falling toggling. Therefore, if we model the behavior of these cells as piece-wise-linear current sources and connect them to the RLC model nodes that map to their layout locations, their noise impact on the target timing path can be established by simulating the voltage waveforms of the nodes that correspond to the path.

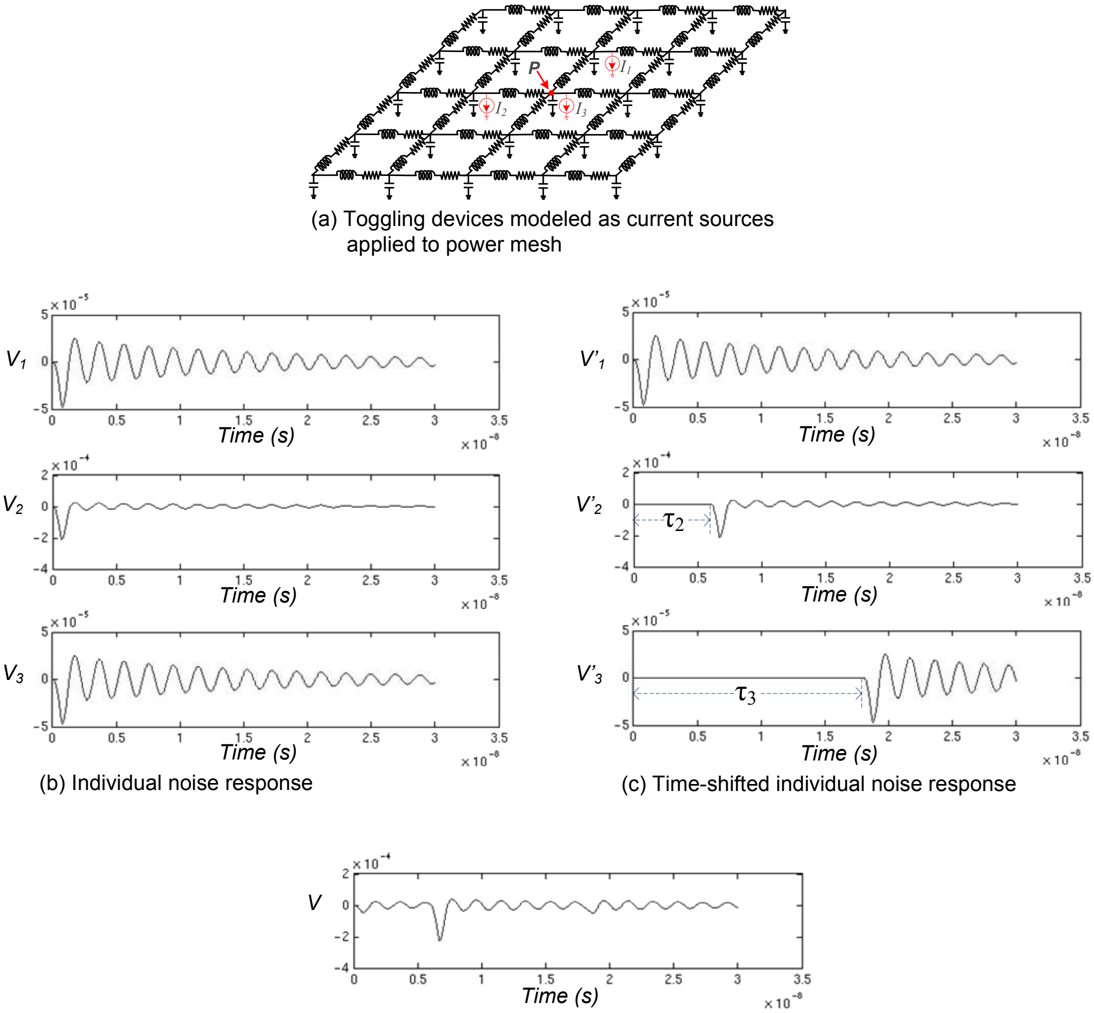
Performing localized noise estimation helps reduce the computational cost by limiting the number of cells to be considered. Nonetheless, if the estimation flow completely relies on simulation to attain the noise profile under different cell behaviors, the computational cost would still be prohibitive, even within a small *Effect Range*. More specifically, if an *Effect Range* contains M cells and each cell

i has C_i different current waveforms, a total of $\prod_{i=1}^{M} C_i$ simulations are needed to cover all possible noise profiles on the target node. Such an exponentially increasing simulation cost makes it highly impractical to perform on-the-fly noise simulation during test generation.

A solution to this challenge can be offered by realizing that the PDN system of a chip is inherently a linear system, which satisfies the linear superposition property. More specifically, given two current sources $i_1(t)$ and $i_2(t)$, if we denote their individual noise impact on a linear system H as $v_1(t) = H\{i_1(t)\}$ and $v_2(t) = H\{i_2(t)\}$, the following additive property always holds for any scalar values α and β.

$$\alpha v_1(t) + \beta v_2(t) = H\{\alpha i_1(t) + \beta i_2(t)\} \tag{5.2}$$

This property enables the accurate estimation of the impact of multiple cells through a quick superposition of the individual cell impacts. In actual circuit operation, cells in different positions of the timing path toggle at distinct time points because of the signal propagation delay. Moreover, the toggling activities of the cells across multiple cycles need to be estimated in an overall noise profile. From the noise simulation perspective, the timing difference of toggling activities indicates that the current stimuli of different current sources cannot be applied simultaneously. Instead, each current stimulus is associated with an initial delay in the time domain. The linear superposition process needs to also account for the time-shifted noise response of each individual current source. The proposed timing-aware noise superposition strategy is illustrated in Figure 5.7. As shown in part (a) of the diagram, three toggling devices are modeled as current sources and mapped on the RLC model of the PDN system, and their noise impact on node $\boldsymbol{P}$ is investigated. The individual noise responses of the three current sources on node $\boldsymbol{P}$ are shown in part (b), which can be denoted at $v_1(t)$, $v_2(t)$ and $v_3(t)$. Assuming that current stimuli i_2 and i_3 are delayed for τ_2 and τ_3 respectively, the noise response would be shifted by the delay time accordingly, as shown in Figure 5.7(c). More specifically, the noise waveforms induced by these two delayed



(a) Toggling devices modeled as current sources applied to power mesh

(b) Individual noise response

(c) Time-shifted individual noise response

(d) Superposed noise response

Figure 5.7: Noise estimation through superposition

current stimuli can be expressed as:

$$v_2'(t) = \begin{cases} v_2(t - \tau_2), & \text{if } t \geq \tau_2 \\ 0, & \text{if } t < \tau_2 \end{cases} \tag{5.3}$$

$$v_3'(t) = \begin{cases} v_3(t - \tau_3), & \text{if } t \geq \tau_3 \\ 0, & \text{if } t < \tau_3 \end{cases} \tag{5.4}$$

The superposed noise waveform $v(t)$, as expressed in Equation 5.5, is shown in Figure 5.7(d).

$$v(t) = \begin{cases} v_1(t), & \text{if } t < \tau_2 \\ v_1(t) + v_2(t - \tau_2), & \text{if } \tau_2 \leq t < \tau_3 \\ v_1(t) + v_2(t - \tau_2) + v_3(t - \tau_3), & \text{if } t \geq \tau_3 \end{cases} \tag{5.5}$$

The example shown in Figure 5.7 explains the reason for noise development across multiple functional cycles: the noise waveforms of stimuli in later cycles are added to the waveforms of earlier stimuli before they die down.

The use of the superposition property significantly reduces the computational cost of noise estimation, as simulation is performed solely for the noise characterization of each individual cell. The number of simulations needed for characterization can be expressed as $\sum_{i=1}^{M} C_i$, which follows a linear relationship with the number of cells. All the characterization simulations can be performed in parallel at the pre-processing stage before the ATPG process. The noise estimation step during ATPG only needs to perform simple additive computations on the time-shifted versions of the characterized noise waveforms, thus being highly efficient.

The piece-wise-linear current stimulus associated with each cell can be pre-characterized through SPICE simulation on standard cells. Since the current stimulus waveform is a function of the number of switching transistors, the proposed technique characterizes for each standard cell multiple versions of current stimuli, each representing a different toggling condition of the cell. During the ATPG process, the noise estimation process chooses the appropriate noise waveform for superposition according to the cell toggling condition induced by the test pattern.

The amount of time-shift during noise superposition can also be pre-characterized by performing a static timing analysis on the targeted cells. The combination of precise pre-characterization and computationally-efficient superposition yields a SPICE-accurate yet fast noise estimation process that enables efficient ATPG.

5.4 Test pattern transformation for noise failure detection

The proposed ATPG flow performs guided test pattern search to identify the one that approximates the worst-case functional mode noise, as summarized in Algorithm 4. The proposed algorithm first performs pre-characterizations of the noise waveforms induced by individual cell toggling and estimates the maximal number of functional cycles needed for noise development. Candidate test application schemes with different number of functional cycles are examined subsequently. For each candidate test application scheme, the proposed technique generates a delay test cube for the target path and performs simulated-annealing search for the X-filling strategy that maximizes the capture cycle noise. Finally, the test application scheme that delivers the worst-case noise is selected and the associated test pattern is added to the test set.

The high-level ATPG flow shown in Algorithm 4 employs a simulated-annealing based X-filling engine, as detailed in Algorithm 5. This algorithm searches in the unspecified bit space to identify the X-filling patterns that maximize the capture cycle noise.

During the X-filling search, the noise estimation technique outlined in Section 5.3 is employed to calculate the cost function value. As discussed previously, the SPICE simulation cost is drastically reduced by the utilization of noise superposition technique. Nonetheless, the superposition process introduces a large number of addition operations when the noise curves of individual devices are characterized with a high resolution. For example, a total of 3×10^4 additions are needed for just superposing two 30ns curves characterized with the time-step of 1ps.

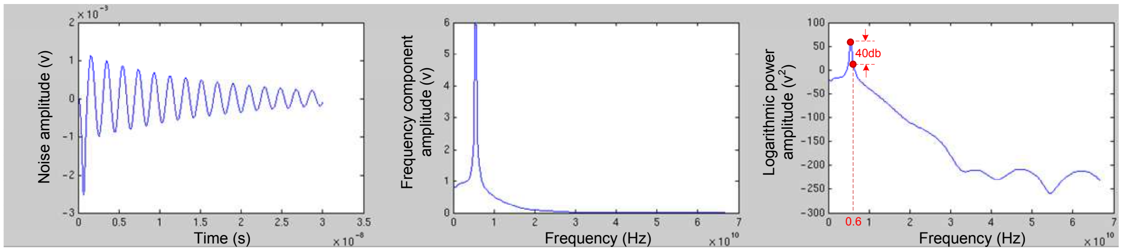
Algorithm 4 *Path delay test generation for worst − case noise approximation*

```
Identify the $Effect\ Range$ of the target path from layout
Characterize the current stimuli of standard cells through SPICE simulation
Characterize the noise waveform for each current stimulus through simulation
Characterize the amount of time shift of each noise waveform through static timing analysis
Estimate the number $N$ of cycles that are equivalent to the impulse response die-down time
for $i = 1$ to $N$ do
  Generate a test cube for the $i$-functional-cycle scheme through sequential ATPG
  Identify the X-filling that maximizes the noise in the capture cycle through simulated-annealing search
end for
Compare the maximal noise for all $N$ schemes; select the one that delivers the worst-case noise and add the associated test pattern into the test set
```

The computational cost of the superposition step, though, can be significantly reduced through the utilization of the error-tolerance capability of the simulated-annealing process and the appropriate curve sampling strategy. At the beginning phase of the simulated-annealing search, the cost function only needs to guide the movement of the annealing process towards a largely optimal direction. As a result, a relatively higher estimation error of the cost function can be tolerated at the high annealing temperature. Therefore, a relatively coarse-grained noise curve sampling strategy, with wider time intervals between the sampled points, can be utilized to provide a largely accurate yet highly rapid estimation of the superposed noise profile. As the annealing temperature keeps decreasing, more accurate noise estimation is needed to guide the movement of the annealing process towards the global optimum, thus requiring a more fine-grained noise curve sampling strategy. By starting with a loose sampling strategy and gradually increasing the sampling granularity as the annealing process proceeds, one can significantly reduce the overall number of addition operations while still guaranteeing the convergence to a near-optimal result.

Algorithm 5 *$Simulated-annealing\ based\ X-filling\ search$*

```
Random X-filling in the test cube to create an initial test pattern $S_0$
Estimate capture cycle noise $v_0$ induced by $S_0$
$S_{max} \leftarrow S_0$; $v_{max} \leftarrow v_0$
Determine initial temperature $T$
Determine initial sampling frequency $F$ using spectrum analysis
repeat
  while $Iteration\ count < Max\ iterations$ do
    Randomly flip a few X-bits to create a trial test pattern $S$
    Estimate capture cycle noise $v$ induced by $S$ using sampling frequency $F$
    $\Delta v \leftarrow v_0 - v$
    if $\Delta v < 0$ then
      $S_0 \leftarrow S$; $v_0 \leftarrow v$
    else
      Generate a random number $r \in (0, 1)$
      if $r < e^{-\Delta v/T_0}$ then
        $S_0 \leftarrow S$; $v_0 \leftarrow v$
      end if
    end if
    if $v > v_{max}$ then
      $S_{max} \leftarrow S$; $v_{max} \leftarrow v$
    end if
  end while
  $T \leftarrow \alpha T$
  $F \leftarrow min(F_{max}, 2F)$
until No improvement in $m$ consecutive inner loops
return $S_{max}$
```



(a) Characterized device noise curve (b) Frequency spectrum (c) Power spectrum in logarithmic scale

Figure 5.8: Sampling frequency selection based on noise spectrum

Gaining efficiency while maintaining a high accuracy level necessitates the determination of the appropriate sampling frequency at each phase of the annealing process. The sampling techniques that are widely utilized in digital signal processing can be employed to overcome this challenge. According to the Nyquist-Shannon sampling theorem [78, 87], if the spectrum of a time-domain function $x(t)$ contains no frequencies higher than B hertz, it can be completely determined by a series of sample points spaced $1/(2B)$ seconds apart. That is to say, as long as the sampling frequency is at least twice of B, the curve can be reproduced from the samples with no errors. The characterized noise curve of individual devices typically contains harmonics with various frequencies. Nonetheless, since it is a fairly smooth curve whose frequency distribution is determined by the PDN's inherent characteristics, its spectrum typically exhibits a rather concentrated distribution, thus giving rise to the possibility of sampling at a relatively low frequency with negligible estimation errors.

As a discrete time series $v_n, 0 \leq n \leq N-1$, the characterized noise of individual devices can be easily transformed to a spectrum in the frequency domain, through a *discrete Fourier Transform*, as outlined in Equation 5.6.

$$V_k = \sum_{n=0}^{N-1} v_n e^{-i2\pi k \frac{n}{N}}, \qquad {\scriptstyle k=0,\ldots,N-1} \tag{5.6}$$

A number of *Fast Fourier Transform* (FFT) algorithms have been proposed in the literature to perform *discrete Fourier Transform* efficiently. A representative

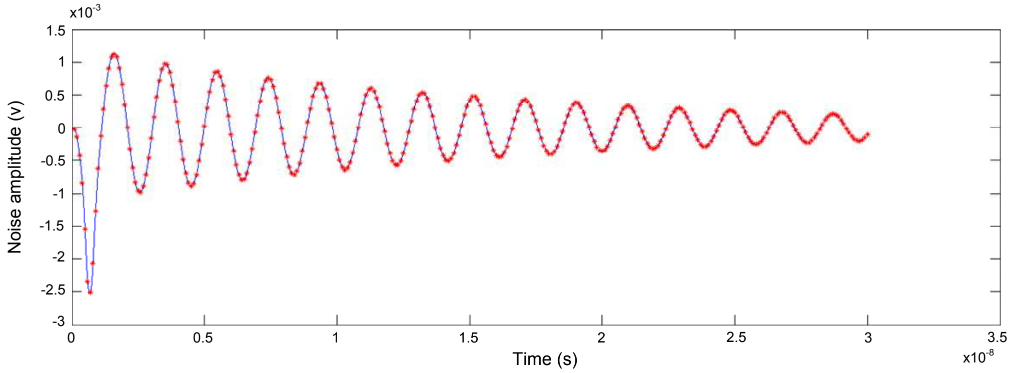


Figure 5.9: Comparison between full-resolution and sampled curves

noise curve induced by one toggling device is shown in Figure 5.8(a). The simulation resolution corresponds to a $1000GHz$ sampling frequency, as a time-step of 1ps is used. The frequency spectrum of the noise curve, calculated by FFT, is presented in Figure 5.8(b). It can be observed that the spectrum is highly concentrated at the neighborhood of $5GHz$, the base frequency of this noise curve. The amplitudes of the Fourier components attenuate rapidly when the frequency moves away from the base one, indicating a rather weak contribution of other frequency components to the noise waveform. Such a spectrum can be approximately considered as having a frequency upper bound B by ignoring the higher frequency components that have weak contributions, thus enabling the utilization of the sampling theorem to compute the appropriate sampling frequency. The contribution of each frequency component is typically quantified by its power, a metric defined as half the squared amplitude of this component. To determine where the spectrum can be safely truncated, a power spectrum diagram can be plotted in the logarithmic scale, as shown in Figure 5.8(c). The logarithmic scale transformation computes the vertical axis of Figure 5.8(c) as $20log_{10}(power\ amplitude)$. Therefore, the value difference between any two points defines the ratio of their power amplitudes via the logarithmic unit of dB. From this plot, it can be easily identified that a more than $40dB$ gap exists between the power of the $6GHz$ component and the peak power, indicating that the impact of ignoring the $6GHz$ component on signal power is less than 0.01%. The impact of frequency components that are higher than $6GHz$ is even more negligible as the power spectrum curve drops rapidly in this region. Therefore, a $40dB$ drop from the peak is selected as the threshold for determining the truncation point B of the spectrum. In the example shown in Figure 5.8, a bound frequency of $6GHz$ indicates a lossless sampling frequency of $12GHz$, which is 83 times lower than the original frequency of $1000GHz$ in the full-resolution curve. Figure 5.9 presents a comparison between the full-resolution curve (in solid line) and the one sampled at $12GHz$ frequency (in bold dots). It can be intuitively seen that the sampled curve precisely tracks the details of the original one using only 1/83 of the original sample points.

The aforementioned analysis shows that using the curves sampled at lower frequency can drastically reduce the number of addition operations in the noise superposition process with very small estimation error. Based on this observation, the proposed simulated-annealing process starts the noise estimation with the Nyquist-Shannon sampling frequency attained through spectrum analysis. When the process enters a new outer-loop iteration and the annealing temperature reduces to a new level, the sampling frequency is increased by a factor of 2, as a higher accuracy level is needed at a lower temperature in order to improve the quality of the solution. The sampling frequency keeps increasing with the temperature reduction, until it reaches the full simulation resolution. Such a dynamically tuned annealing process, combined with the efficient noise estimation, ensures a fast convergence to the near-optimal test pattern.

5.5 Simulation results

We verify the proposed methodology through simulations on the largest ISCAS89 benchmarks. The verilog netlists of these benchmarks are synthesized, their layouts generated, and the RLC models of their power distribution networks extracted. SPICE simulations are performed to characterize the current profiles and the noise profile of individual devices so as to establish a database for noise estimation. All characterization work is performed using a 65nm standard cell library and the 65nm *BSIM4* [1] models are employed as the SPICE model cards. The *PathATPG* tool [115] is utilized to generate the unspecified test cubes for the targeted paths, as this tool supports multi-time-frame expansion during test generation, thus enabling the generation of test cubes that can be directly applied to the proposed multi-functional-cycle test scheme. The processing of the SPICE simulation data is implemented using TCL and MATLAB scripts. The test pattern transformation algorithm for worst-case noise creation is implemented using the C programming language. We evaluate the proposed methodology from the perspectives of accuracy, effectiveness and efficiency, as detailed subsequently.

5.5.1 Accuracy

The capability of the proposed methodology in approximating the worst-case functional noise is determined by the accuracy of the proposed noise estimation technique. Figure 5.10 presents a comparison of the simulated and estimated noise curves for a PDN node randomly selected from s15850. In this plot, the horizontal and vertical axes denote the timeline and noise amplitude, respectively. The solid curve illustrates the actual simulation result, which is attained by inserting current sources to the PDN based on the toggling activity of the test vector and performing SPICE simulation. The dotted curve represents the estimated one, which is generated through the superposition of the individual device noise curves characterized in the preprocessing stage. As can been seen, the two curves match precisely throughout the entire simulation time window, and no perceptible errors can be observed at any time point along the curve.

The matching level between the simulation and estimation results can be quantified through a correlation analysis as shown in the scatter plot in Figure 5.11. In this plot, each dot illustrates the noise at a time point, with the horizontal axis value denoting the estimated noise and the vertical axis value representing the simulated one. It can be seen that the dot distribution forms a 45-degree straight line, indicating a high matching level between the simulation and estimation. If a total of n time points are sampled along the noise curves and if we denote the estimated and simulated noise samples using $e_i, 1 \leq i \leq n$ and $s_i, 1 \leq i \leq n$

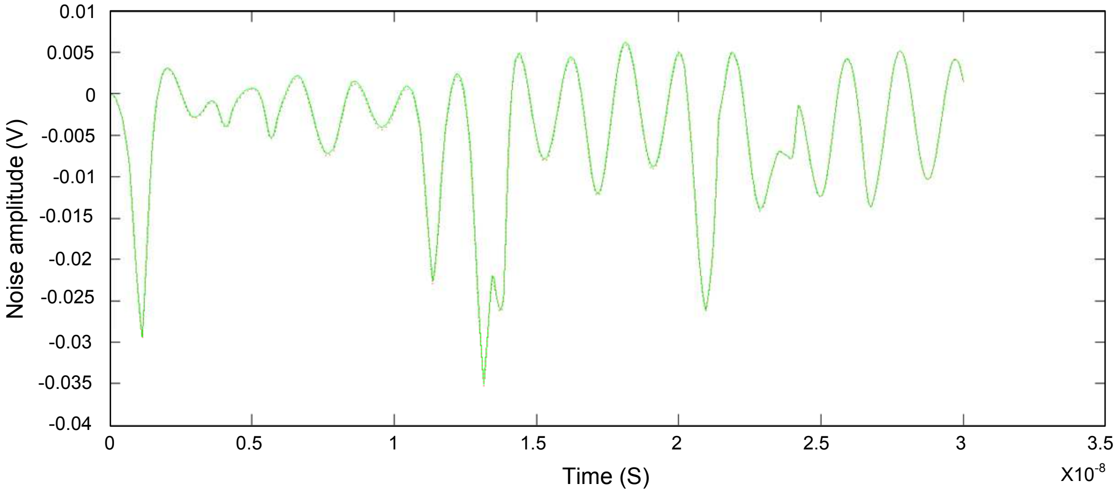


Figure 5.10: Noise curve comparison

respectively, the correlation R_{es} between the simulated and estimated noise curves can be computed using Pearson's correlation metric, as shown in Equation 5.7

$$R_{es} = \frac{\sum_{i=1}^{n}(e_i - \bar{e})(s_i - \bar{s})}{\sqrt{\sum_{i=1}^{n}(e_i - \bar{e})^2 \sum_{i=1}^{n}(s_i - \bar{s})^2}} \tag{5.7}$$

A correlation of 0.9999 can be attained between the simulated noise and the noise curve estimated through the proposed approach.

The aforementioned discussion shows that the proposed technique is capable of delivering a SPICE-accurate noise estimation, thus enabling a precise search for the test patterns that approximate the functional noise situation.

5.5.2 Effectiveness

We further compare the proposed multi-functional-cycle test scheme to conventional single-at-speed-cycle delay test schemes, namely the *Launch-Off-Shift* (LOS) and *Launch-Off-Capture* (LOC) schemes. This comparison mainly focuses on the noise amplitude induced by these test schemes, as it closely correlates with the extra path propagation delay that causes marginal timing failures.

For each benchmark, five critical paths are selected as the targets of testing. Table 5.1 summarizes for each path the simulated die-down time of the impulse

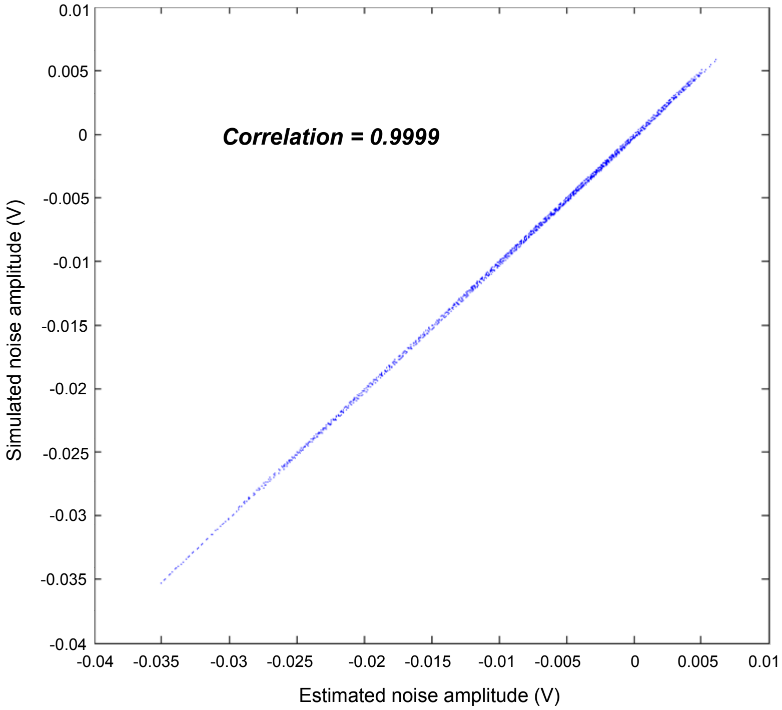


Figure 5.11: Correlation between simulated and estimated results

Table 5.1: Length of capture phase

Circuits		$Path_1$	$Path_2$	$Path_3$	$Path_4$	$Path_5$
s15850	T_{down}(ns)	30.3	33.4	31.8	27.6	32.1
	$\#cycle_F$	3	3	3	2	3
s35932	T_{down}(ns)	42.5	41.9	44.9	44.3	40.6
	$\#cycle_F$	4	4	4	4	4
s38417	T_{down}(ns)	45.8	46.2	46.6	44.5	45.2
	$\#cycle_F$	4	4	4	4	4
s38584	T_{down}(ns)	46.1	38.3	41.0	42.9	42.0
	$\#cycle_F$	4	3	4	4	4

Table 5.2: Average noise amplitude comparison

Circuits	Paths	LOS (mv)	LOC (mv)	Proposed scheme		
				amp.(mv)	$\%incr_{LOS}$	$\%incr_{LOC}$
s15850	$Path_1$	19.5	18.9	24.6	26.2	30.2
	$Path_2$	18.4	19.7	24.8	34.8	25.9
	$Path_3$	21.3	20.6	25.7	20.7	24.8
	$Path_4$	20.5	20.0	23.6	15.1	18.0
	$Path_5$	18.7	19.4	24.3	29.9	25.3
s35932	$Path_1$	37.1	38.4	47.2	27.2	22.9
	$Path_2$	38.1	39.0	49.0	28.6	25.6
	$Path_3$	42.9	41.6	52.1	21.4	25.2
	$Path_4$	41.2	40.4	53.5	29.4	31.9
	$Path_5$	39.4	40.7	51.5	30.7	26.5
s38417	$Path_1$	46.9	45.5	56.9	21.3	25.1
	$Path_2$	40.6	40.8	56.1	38.2	37.5
	$Path_3$	43.2	42.1	57.6	33.3	36.8
	$Path_4$	45.6	47.1	62.2	36.4	32.1
	$Path_5$	44.9	43.6	57.9	29.0	32.8
s38584	$Path_1$	47.6	48.4	59.8	25.6	23.6
	$Path_2$	43.2	44.3	53.7	24.3	21.2
	$Path_3$	41.8	42.8	58.2	39.2	36.0
	$Path_4$	46.5	47.2	66.0	41.9	39.8
	$Path_5$	44.3	43.2	60.7	37.0	40.5

response, and the number of at-speed functional cycles[4] that result in the worst-case capture noise. As clearly confirmed by the result, multiple functional cycles are needed to develop the peak noise.

A detailed comparison on average noise is presented in Table 5.2, and an analogous comparison on peak noise is summarized in Table 5.3[5]. In both tables, the noises induced by the conventional LOS and LOC schemes are reported in Columns 3 and 4. The noise resulting from the proposed approach is reported in Columns 5-7, with Column 5 presenting the noise amplitude, Column 6 report-

[4] A clock period of 10ns is used in simulation.

[5] It should be noted that the average noise reported in Table 5.2 is calculated by averaging the noises on the PDN nodes that map to the target path. While performing the average noise computation, the noise on each individual node is defined as its worst-case noise amplitude within the capture cycle. This average noise metric can effectively reflect the accumulated delay variation on the entire path. The maximal noise reported in Table 5.3 is attained by comparing the noises on different nodes on the path and selecting the worst-case. This metric reflects the localized timing variation on different segments of the path.

Table 5.3: Peak noise amplitude comparison

Circuits	Paths	LOS (mv)	LOC (mv)	Proposed scheme		
				amp.(mv)	$\%incr_{LOS}$	$\%incr_{LOC}$
s15850	$Path_1$	22.8	21.2	26.2	14.9	23.6
	$Path_2$	20.6	22.0	26.1	26.7	18.6
	$Path_3$	23.7	23.4	27.6	16.5	17.9
	$Path_4$	22.1	23.1	25.4	14.9	9.9
	$Path_5$	21.2	20.8	25.9	22.2	24.5
s35932	$Path_1$	39.7	40.5	49.6	24.9	22.5
	$Path_2$	41.2	41.7	51.7	25.5	24.0
	$Path_3$	44.5	44.4	53.8	20.9	21.2
	$Path_4$	44.1	42.9	55.9	26.7	30.3
	$Path_5$	42.5	43.8	52.6	23.8	20.1
s38417	$Path_1$	50.6	48.8	59.4	17.4	21.7
	$Path_2$	44.1	43.7	58.5	32.7	33.9
	$Path_3$	46.0	44.5	61.2	33.0	37.5
	$Path_4$	49.6	50.3	64.4	29.8	28.0
	$Path_5$	50.2	47.6	61.7	22.9	29.6
s38584	$Path_1$	50.9	51.4	63.6	25.0	23.7
	$Path_2$	47.2	47.8	58.2	23.3	21.8
	$Path_3$	45.3	46.6	62.3	37.5	33.7
	$Path_4$	51.3	51.7	71.9	40.2	39.1
	$Path_5$	48.4	48.1	64.4	33.1	33.9

ing the percentage increase over the LOS scheme, and Column 7 reporting the percentage increase over the LOC scheme.

It can be observed that the proposed test scheme results in an elevated noise level compared to the conventional schemes, thus enabling a more rigorous testing of noise-induced-failures that can occur in functional operation. Although the proposed test pattern transformation technique mainly focuses on maximizing the average noise along the path in order to examine the worst-case accumulated path delay, it also leads to a significant increase in the maximal noise on the path, resulting in the testing of failures due to noise hot-spots. Higher noise can be observed in relatively larger benchmarks, as large circuits typically have higher impedance in their PDN systems. The higher RC constants in larger benchmarks also slow down the noise die-down process, resulting in a more evident noise accumulation effect. As a result, the noise gap between the conventional schemes and the proposed

Table 5.4: Worst-case simulation time in seconds

Circuits	$Path_1$	$Path_2$	$Path_3$	$Path_4$	$Path_5$
s15850	2415.78	2782.58	2507.40	2295.52	2576.93
s35932	3695.07	3609.80	3929.99	3871.63	3435.98
s38417	4236.46	4291.09	4354.78	3982.25	4038.32
s38585	4715.59	4159.29	4313.12	4422.11	4230.34

scheme increases along with the circuit size, leading to an increased test quality in large designs. Moreover, the application of multiple-functional-cycles results in a highly "functional" state transition in the capture cycle, thus effectively mitigating the false-alarms due to non-functional noise profile.

5.5.3 Efficiency

The successful application of the proposed methodology relies on a well-controlled computational cost. The computational time of the proposed approach mainly stems from two portions, namely, the SPICE simulation time and the test pattern transformation time. Expensive SPICE simulations constitute the major bottleneck and thus need to be controlled through an efficient algorithmic flow. The proposed technique constrains the use of simulations in the preprocessing stage where the individual device noise is characterized independently. This not only minimizes the number of simulations but enables a parallel simulation flow, significantly reducing the computational cost. The SPICE simulation cost is outlined in Table 5.4. Since all simulations can be performed in parallel, we report the longest simulation time among all runs performed for each path. For the benchmarks used in our simulation, the worst-case simulation time can be controlled within 2 hours, even with a high time-step resolution of 1ps. For large industrial designs that are difficult to handle with SPICE, a number of fast IR-drop simulation tools, such as Redhawk, can be utilized for characterization.

The test pattern transformation process need not perform any simulation. Its simulation-annealing engine estimates the noise profile through computationally-efficient superposition operations guided by the dynamically tuned sampling strategy, thus enabling a fast search for the worst-case test patterns. Table 5.5 presents

Table 5.5: Test pattern transformation time in seconds

Circuits	$Path_1$	$Path_2$	$Path_3$	$Path_4$	$Path_5$
s15850	33.61	36.66	31.38	40.32	39.49
s35932	49.34	60.86	68.38	59.06	56.40
s38417	67.31	44.97	64.04	52.34	60.07
s38585	54.57	61.08	76.80	56.14	68.09

the CPU time of the test pattern transformation process. It can be seen that this portion of the computational cost is negligible compared to the SPICE simulation.

5.6 Conclusions

In this work, we address the challenge of detecting the noise-induced timing failures in nanoscale designs.

Conventional delay test schemes fail to approximate the actual noise profile that can occur in functional operation, thus leading to degraded test quality. The detection of noise failures necessitates the proper modeling of the noise accumulation effect occurring in the power distribution network of the circuit. Motivated by the observation that the noise accumulation process typically spans multiple at-speed cycles, we propose a novel multi-functional-cycle test scheme. This scheme applies multiple at-speed functional-like launch cycles to allow the full development of the noise, and performs the capture operation at the worst-case noise situation. The application of this scheme to selected critical paths in the circuit enables a much more rigorous examination of the circuit timing robustness under noise situation while reducing the false alarms possibly incurred by non-functional state transitions, thus significantly improving the test quality.

To enhance the ATPG efficiency of the proposed scheme, we further propose a novel noise estimation technique. We have observed that the linearity of the power distribution network lends itself to a fast noise estimation through the superposition of noise curves characterized for individual devices. The utilization of the superposition strategy minimizes the number of SPICE simulations and enables a parallel simulation flow in the noise characterization stage. By leveraging the efficiency of the proposed noise estimation technique, a test pattern transformation process, driven by a simulated-annealing engine, is performed to search for the test patterns that create the worst-case noise profile in the target paths.

The proposed methodology is completely compatible with a standard scan-based design architecture, necessitating no hardware modification whatsoever. Integration of the proposed scheme into an industrial test flow results in a more robust yet economical test plan, significantly reducing the test escape risk and maintaining high product quality.

The text of Chapter 5, is in part a reprint of the material as it appears in *M. Chen and A. Orailoglu, "Cost-effective IR-drop failure identification and yield recovery through a failure-adaptive test scheme," Design, Automation and Test in Europe, 2010*; and in *M. Chen and A. Orailoglu, "Examining timing path robustness under wide-bandwidth power supply noise through multi-functional-cycle delay test," submitted to IEEE Transactions on VLSI.* The dissertation author was the primary researcher and author of the publications [17] and [13].

Chapter 6

Marginal failure diagnosis in scan circuitry

Ramping up silicon yield to production necessitates both the test development and failure analysis being effectively addressed. The techniques discussed in Chapters 4 and 5 contribute to the test development side by enabling a high quality and low yield loss digital testing with a focus on covering functional-mode marginal failures which are difficult to detect using traditional structural test. Yet to enable efficient design optimization and reduce the re-spin cost, the failure analysis process needs to be supported by an effective diagnosis methodology that ascertains the locations and root causes of marginal failures.

As mentioned in Chapter 2, among various failure mechanisms, scan chain timing faults appear increasingly prevalent in nanometer scale designs with high clock frequencies [44]. It has been reported that the scan circuitry occupies around 30% of the chip area [53] and possibly contributes up to 50% of the chip failures [114]. With the high power ground noise level during the scan phase, the operational condition of the chip can easily fall outside the voltage range within which the designers close timing, thus introducing high delay uncertainties in scan chains. As a result, multiple faults with mixed timing violation types can exist in the failing scan chains. The mitigation of such failures necessitates accurate identification of the fault locations and precise understanding of the associated timing violation behaviors.

Scan cell level failure diagnosis enables the effective identification of scan cell defects and weaknesses that directly cause the observed timing failures. Nonetheless, it is also highly possible that the observed scan chain failures are indirect manifestations of faults in the scan clock distribution network. The increased timing uncertainty in clock trees can easily incur a large number of timing failures in fault-free scan chains. In such a scenario, it is essential to investigate down to the clock buffer level during failure analysis so as to provide a comprehensive set of diagnostic information for design and fabrication improvement.

We propose, in this chapter, a methodology capable of diagnosing both permanent and intermittent scan chain timing faults. This approach closely approximates the behavior of the realistic failure mechanisms observed in silicon, and extracts the feature of marginal scan failures from volume diagnostic data. This brings up a new perspective in understanding and analyzing the syndrome of the marginal failures. In addition to the identification of direct scan cell failure causality, we furthermore extend the proposed methodology by incorporating a clock buffer failure analysis, so as to also ascertain the possible indirect failure causality in clock trees. The diagnostic results not only pinpoint the physical region of the failures, but indicate the relative criticality of each failure, thus providing a strong guidance to the design optimization and re- spin task.

We first examine the technical challenges induced by the failure mechanism shift and outline the fundamental ideas proposed in this work. The algorithm that performs the scan cell level diagnosis is detailed in Section 6.2. Section 6.3 presents the technique that performs a further analysis based on the scan cell failure information and derives the potential faulty clock buffers.

6.1 Scan Timing Failure Diagnosis Overview

The fault manifestation mechanism of timing failures in scan circuitry is illustrated in Figure 6.1. The discrepancy between the expected and the observed scan-out data constitutes the fault syndrome that can be collected from the tester.

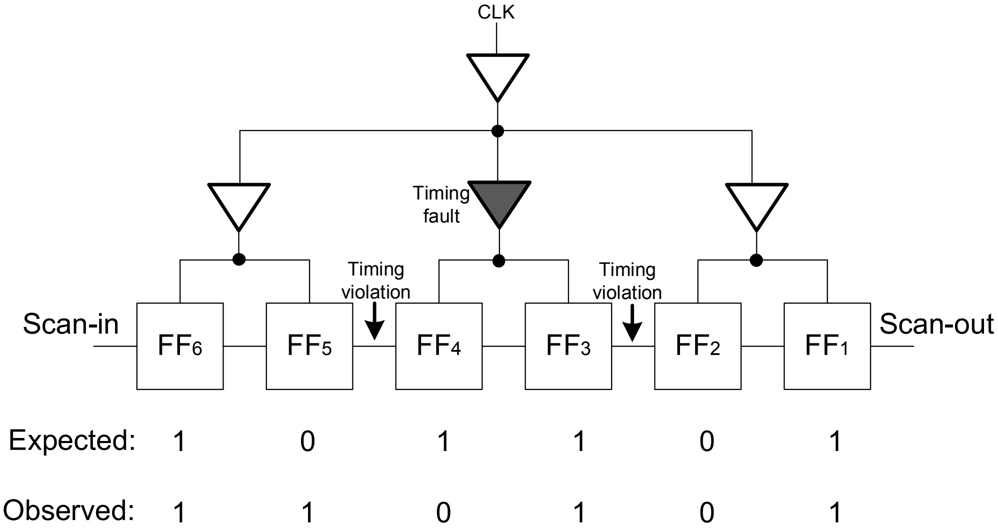


Figure 6.1: Scan timing fault manifestation

Such a fault syndrome can either be directly caused by the timing defects in the scan chain or be an indirect result of the timing defects in the scan clock. For the former case, the identification of failing scan cells is sufficient for ascertaining the root causality of the failure, whereas for the latter case, one needs to investigate the possible faulty clock buffers so as to understand the design/process weakness. Therefore, it is highly important for the diagnosis flow to investigate both scenarios and provide information regarding both the failing scan cells and the possible faulty clock buffers. Developing such a comprehensive diagnostic flow constitutes the focus of this work.

6.1.1 Technical challenges

Traditional diagnosis approaches typically assume deterministic fault behavior which enables the creation of a fault dictionary through accurate fault simulation. Nonetheless, the failure mechanism shift resulting from the design and fabrication change has invalidated the traditional assumptions. The intermittent manifestation and the multitude of marginal failures result in highly ambiguous fault behavior, significantly increasing the difficulty in diagnosis. The inherent ambiguity of the failure mechanism observed in nanometer designs reduces the

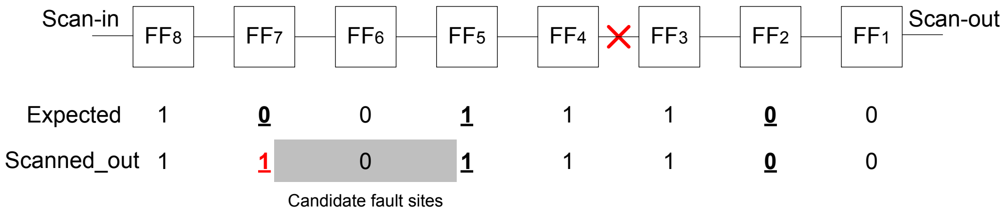


Figure 6.2: Incorrect diagnosis of intermittent faults

information that can be extracted through logic analysis, leading to a degraded diagnostic resolution, as detailed subsequently.

One contributing factor to the ambiguity consists of the unpredictable fault manifestation. Figure 6.2 illustrates the impact of intermittent failure manifestations on scan cell diagnosis accuracy. Let us assume that an intermittent hold-time fault exists between scan cells 4 and 3. The three scan bits highlighted with underlines display sensitivity to hold-time faults. After scanning out the pattern, the sensitive bit 7 is corrupted by the fault, whereas the sensitive bits 5 and 2 match the expected values. The traditional bound analysis approach based on the permanent fault assumption would identify a candidate fault range between cells 7 and 5, which fails to cover even the actual fault site. Apparently, the intermittent fault invalidates the logic reasoning for identifying the right bound[1] of the fault range, as the absence of the failure cannot be utilized to exclude an intermittent fault. This highly degrades the effectiveness of logic analysis, as the entire range between cells 7 and 0 has to be considered as fault candidates.

Another source of ambiguity stems from the interaction between multiple faults. Figure 6.3 shows a fault interaction case that can occur in the scan clock tree. In this example, the two scan cells are driven by faulty buffers with delay defects, thus being prone to timing violations. Nonetheless, the two faulty buffers impose similar extra delays on the clock distribution path, ending up maintaining a correct relative timing between these two scan cells. As a result, the fault effect of individual defects is canceled by the fault interaction. This once more shows that the absence of failure syndromes cannot be used for fault exclusion, significantly

[1] The bound that is closer to scan-out

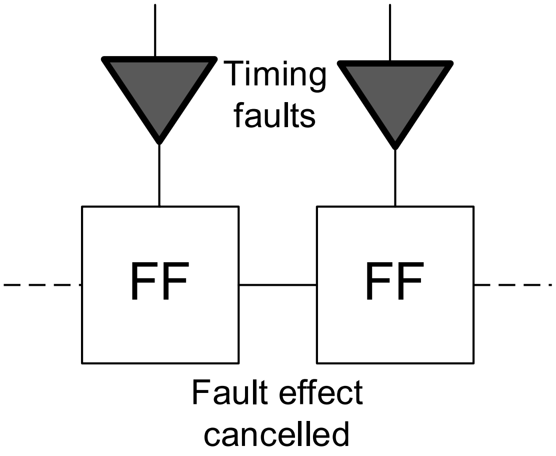


Figure 6.3: Scan clock fault interaction

reducing the information that can be extracted through logic analysis.

6.1.2 Proposed approach

If we view the scan process as the movement of a data stream, the manifestation of scan chain timing failures statistically moves the scan data forward or backward. Therefore, two pieces of extra information can be extracted from a relatively abundant set of scan data, one being the abnormal phase movements of scan data, the other being the occurrence probability of abnormal phase movements. The abnormal phase movement inherently reflects the failure location and type, whereas its occurrence probability signifies the magnitude of the failures. The statistical significance of such information can effectively filter out the noise created by the random manifestation seen in individual scan patterns, thus enabling an accurate identification of the failing scan cells and the associated timing violation types. After attaining the information of failing scan cells, a topological analysis can be further performed to trace the clock buffers that can result in the identified scan failures and establish the minimum set of fault hypotheses. The impact of each fault hypothesis on failure probabilities is examined through statistical timing analysis, and the ones that closely approximate the observed failure probabilities are identified as the most probable faults.

It should be noted that the proposed technique does not rely solely on statistical information. In fact, it performs statistical and logic analysis in an intricate manner so as to leverage the benefit of each. For example, the timing violation type information extracted from statistical analysis enables an accurate deterministic tracing of the clock tree, which significantly reduces the candidate faulty buffer set.

The merger of logic and statistical analysis not only elevates the diagnostic accuracy, but significantly improves the efficiency of the diagnosis flow. By performing topological analysis, one can distribute non-interacting fault candidates into different groups and prune each individual group through statistical timing analysis. Such a partitioning strategy reduces the computational complexity by orders of magnitude without impacting the optimality of the results, as different groups are independent of each other in terms of fault manifestation. During the pruning of each group, a branch-and-bound strategy with highly guided branch selection and bounding heuristics is utilized to further improve the efficiency.

The technical details of the aforementioned methodology are presented in subsequent sections.

6.2 Scan cell timing failure diagnosis

6.2.1 Preliminaries

A set of preliminary information related to the proposed work is outlined in this section.

Scan cell fault models

The proposed technique analyzes the scan chain timing failures based on two fault models, namely the setup-time violation and hold-time violation. Since a timing fault can cause incorrect toggles in two possible directions, we assign to each fault type two parameters that define the manifestation probabilities of rising and falling faults, respectively. Table 6.1 provides a summary of the proposed fault

Table 6.1: Fault model

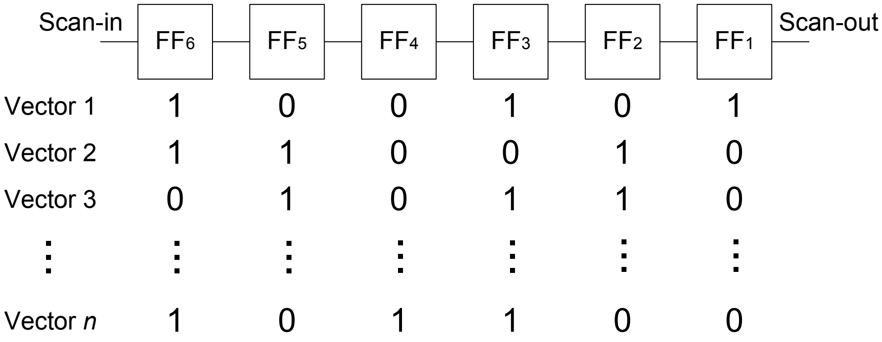
Fault type	Manifestation probability
Setup-time violation	Slow to rise: $M_{sr}, 0 \leq M_{sr} \leq 1$
	Slow to fall: $M_{sf}, 0 \leq M_{sf} \leq 1$
Hold-time violation	Fast to rise: $M_{fr}, 0 \leq M_{fr} \leq 1$
	Fast to fall: $M_{ff}, 0 \leq M_{ff} \leq 1$

model. Compared to traditional models which typically assume unidirectional or equal probability for both directions, the proposed model approximates the realistic fault behaviors better, as it is able to reflect the asymmetric manifestation probabilities induced by nonideal design and fabrication.

Application scheme of diagnostic patterns

The failing scan chains can be easily identified by performing a flush test [46]. Hence the proposed work mainly focuses on the fine-grained diagnosis down to the scan cell level after attaining the failing scan chain information.

As outlined in the introduction, the proposed technique identifies the faulty cells by analyzing the scan pattern image of the failing scan chain. Therefore, the proposed diagnostic scheme first applies a number of timing violation immune vectors to the chip-under-diagnosis in order to create the scan image. The application of these vectors follows the standard scan-based test application scheme, namely, the *scan-in→capture→scan-out* flow. The timing violation immune vectors consist of all 0's or all 1's in the failing scan chains and random patterns in fault-free scan chains. The application of such vectors introduces no errors in the scan-in and capture phases. Hence the expected scan-out patterns can be attained without ambiguity through logic simulation. As the capture step creates in the failing scan chains toggling patterns that are sensitive to timing faults, the captured responses are no longer immune to the scan chain faults and would be corrupted during scan-out. The difference between the observed scan-out patterns and the expected ones provides information regarding the faulty cells that would be identified during the post-analysis step.



	FF_6	FF_5	FF_4	FF_3	FF_2	FF_1
Vector 1	1	0	0	1	0	1
Vector 2	1	1	0	0	1	0
Vector 3	0	1	0	1	1	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Vector *n*	1	0	1	1	0	0

Figure 6.4: A scan image formed by n scan vectors

Scan chain images

Applying multiple diagnostic vectors would generate a set of distinct responses in each failing scan chain. The union of these response patterns forms an image of the associated failing scan chain. Formally, the image of a scan chain can be represented in a matrix form, with each column denoting a scan cell and each row corresponding to the response of a diagnostic vector. Figure 6.4 presents an illustrative scan image of a scan chain. Several images can be defined to facilitate the analysis of the failure syndromes.

Expected image: The expected image of a scan chain is formed by the correct response values before the scan-out step introduces any errors.

Observed image: The observed image of a scan chain is formed by the response values that are actually observed at the scan-out pin of this chain.

6.2.2 Fault location and type identification

Basic idea

The overall impact of scan chain timing faults can be extracted by individually analyzing the behavior of each type of timing violation. The manifestation of a hold-time fault would speed up by one cycle the propagation of the scan value that passes through the fault site as the input of the affected scan cell toggles too fast to meet the hold-time constraint, whereas a setup-time violation delays the propagation of the scan value by one cycle due to the slow toggling of the affected signal. From a scan image perspective, such a speedup or delay effect can be interpreted as a forward or backward movement of the affected portion of the scan image. For instance a scan sequence of "101001011" can be converted to "X10100111" by a hold-time fault between the second and third bits from the right. It can be observed that the phase of the sub-sequence "101001" is moved one bit forward. If a portion of the scan image passes through multiple faults during the scan-out stage, its phase would be moved back and forth multiple times, as the impact of scan chain faults is accumulative. Such an accumulative distortion effect transforms the expected image to the observed image. If we compare these two images by traversing them from the scan-out to the scan-in, the detection of any change of the phase skew between them signifies the existence of a new timing fault. The type of the timing fault can furthermore be identified by examining the direction of the phase movement.

To precisely extract the phase movement information, we propose to monitor the column-wise correlation between the expected and observed images, as any phase movement would strongly change the correlation levels of the scan image columns under comparison. In a fault-free case, the expected image exactly matches the observed one, thus leading to a full correlation between the columns with the same index. However, when a column, i, of the expected image passes through a hold-time (setup-time) fault, the correlation between it and column $i-1$ $(i+1)$ of the observed image would significantly increase as a result of the phase movement behavior. Such an abrupt change of the column correlation provides clear signals for identifying phase movement.

As the columns of scan images are essentially binary vectors, the metric proposed in [97] is employed to estimate the column-wise correlation. Such a metric has been widely utilized in resolving pattern recognition problems and has proven highly effective in assessing binary vector similarity. Formally, let S_{ij} $(i, j \in \{0, 1\})$ be the number of occurrences of bit pair (i, j) at the corresponding positions of two binary vectors. Then the correlation between the two vectors is defined as follows.

c_1	c_2
1	0
0	1
0	0
1	1
0	0
1	0
1	1
0	0

$S_{00} = 3$

$S_{01} = 1$

$S_{10} = 2$

$S_{11} = 2$

$\mathrm{corr}(c_1, c_2) = 0.26$

Figure 6.5: Column-wise correlation computation

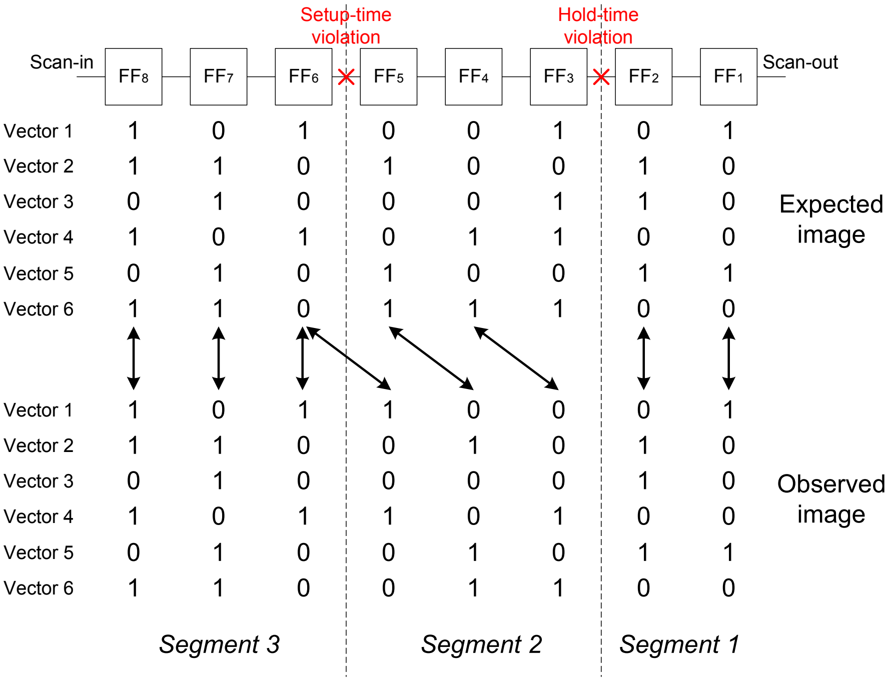
$$corr = \frac{S_{11}S_{00} - S_{01}S_{10}}{((S_{01}+S_{00})(S_{01}+S_{11})(S_{10}+S_{00})(S_{10}+S_{11}))^{1/2}} \tag{6.1}$$

Figure 6.5 shows an example of column correlation computation. The aforementioned binary vector correlation metric shares similar properties with the correlation coefficient of random variables. It has a value range of $[-1, 1]$, with the value 1 indicating a perfect positive correlation and -1 denoting the correlation between complementary vectors. If the two vectors under comparison are independent and sufficiently random, the correlation between them would be around 0, as the four types of bit pairs examined in this metric would have similar occurrence probabilities, resulting in the numerator of the metric hovering around 0.

The principle of detecting phase movement through monitoring column-wise correlation can be generically utilized for both the permanent and intermittent timing fault diagnosis, as discussed subsequently.

Permanent fault diagnosis

A permanent timing fault in the scan chain would result in strict phase movements of affected scan image columns, as the propagation of scan values at the fault site will always be sped up or delayed. This in turn results in an abrupt change of the column-wise correlations. The columns to the right of the fault location are not impacted by the fault during the scan-out stage. Therefore, such columns in the expected image should have a full correlation with the corresponding columns



	FF_8	FF_7	FF_6	FF_5	FF_4	FF_3	FF_2	FF_1
Vector 1	1	0	1	0	0	1	0	1
Vector 2	1	1	0	1	0	0	1	0
Vector 3	0	1	0	0	0	1	1	0
Vector 4	1	0	1	0	1	1	0	0
Vector 5	0	1	0	1	0	0	1	1
Vector 6	1	1	0	1	1	1	0	0

	FF_8	FF_7	FF_6	FF_5	FF_4	FF_3	FF_2	FF_1
Vector 1	1	0	1	1	0	0	0	1
Vector 2	1	1	0	0	1	0	1	0
Vector 3	0	1	0	0	0	0	1	0
Vector 4	1	0	1	1	0	1	0	0
Vector 5	0	1	0	0	1	0	1	1
Vector 6	1	1	0	0	1	1	0	0

Figure 6.6: Phase skews induced by permanent faults

in the observed image. Nonetheless, the fault would impact the columns to the left of the fault site, thus changing the correlation relationship. More specifically, a single hold-time violation would move the affected column i in the expected image one bit forward to the position of $i - 1$ in the observed image, resulting in a full correlation between them. Similarly, a single setup-time violation would result in a full correlation between column i in the expected image and column $i + 1$ in the observed image. From the aforementioned analysis, it can be seen that the position where the correlation relationship starts to change is the fault location. If the correlation change at the fault site signifies a forward phase movement, the fault type can be ascertained to be a hold-time violation. Otherwise, a setup-time violation is indicated by a backward phase movement.

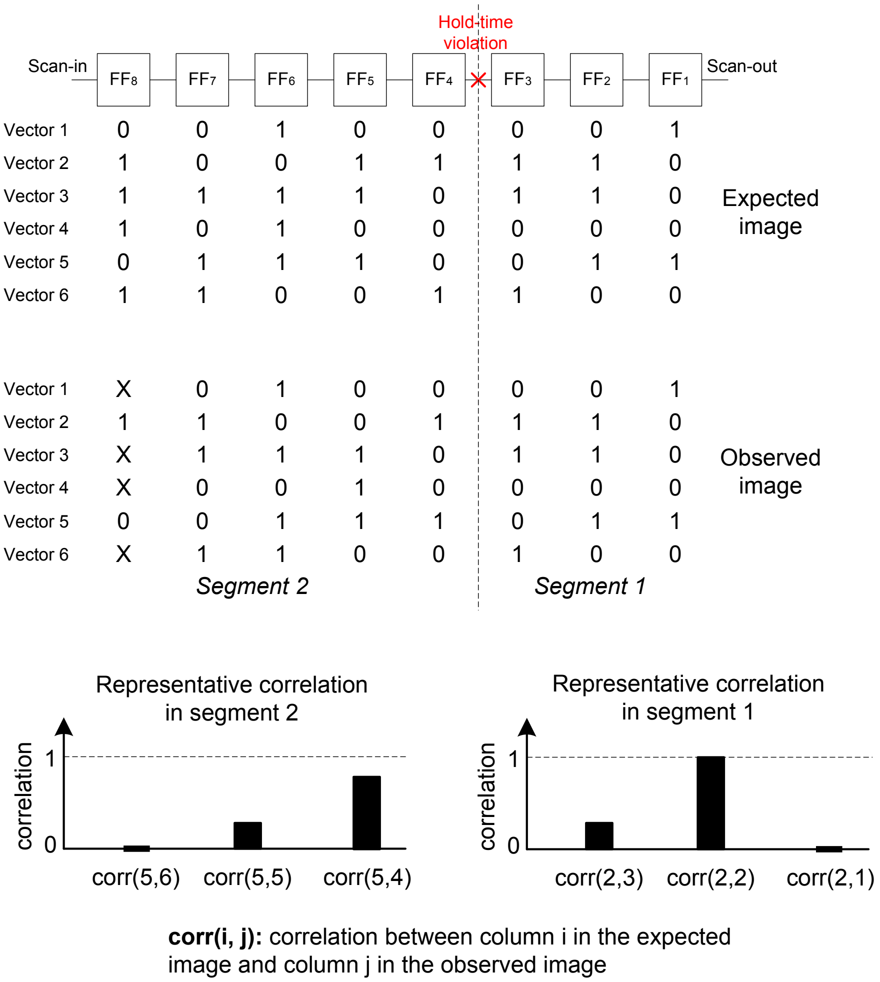
In the case of multiple permanent faults, the accumulative phase movements need to be considered. Figure 6.6 illustrates a double-fault example with mixed fault types. The two fault sites naturally partition the scan image of this faulty scan chain into three segments. During the scan-out stage, these three segments pass through different numbers of faults, thus experiencing distinct fault accumulation effects. Segment 1 passes through no faults, thus exhibiting no phase skew between the expected and observed images. Segment 2 is moved forward by a hold-time fault, thus having a phase skew[2] of 1. The phase skew returns back to 0 in Segment 3, as the setup-time and hold-time faults that it passes through cancel each other's impact on the phase. It can be summarized from the analysis above that a permanent hold-time (setup-time) fault always increments (decrements) the phase skew between the expected and observed images by 1. Hence the fault locations and types can be identified by sweeping these two images from scan-out to scan-in and detecting positions and directions of phase skew change.

Intermittent fault diagnosis

The phase movement behavior is less pronounced in an intermittent fault scenario compared to the permanent fault situation discussed above, as the probabilistic manifestation of such faults only partially moves the columns of the scan image. Therefore, the extraction of phase skew information from the somewhat muted and ambiguous fault syndrome constitutes the main challenge in applying the proposed diagnosis idea to intermittent fault scenarios.

Despite the syndrome ambiguity, an intermittent fault can still sharply change the column-wise correlation of the scan images, as illustrated by the example shown in Figure 6.7. Let us denote the correlation between column i in the expected image and column j in the observed image as $corr(i, j)$. Segment 1 of the scan image exhibits a full correlation in $corr(i, i), 1 \leq i \leq 3$. Segment 2 though exhibits a reduced $corr(i, i)$ but an increased $corr(i, i-1)$ due to the impact of the hold-time violation, signifying a forward phase movement. Because of the intermittent fault manifestation, the columns in Segment 2 are only partially moved, resulting in $corr(i, i-1)$ being less than 1. This in turn indicates that the phase skew between the expected and observed images in Segment 2 is within the value range of $(0, 1)$. Similarly, a setup-time violation can cause a phase skew that resides in the value range of (-1, 0) by increasing $corr(i, i+1)$.

[2] The phase skew is defined to be positive (negative) if the phase of the observed image is ahead of (behind) the expected one.



Expected image

	FF_8	FF_7	FF_6	FF_5	FF_4	FF_3	FF_2	FF_1
Vector 1	0	0	1	0	0	0	0	1
Vector 2	1	0	0	1	1	1	1	0
Vector 3	1	1	1	1	0	1	1	0
Vector 4	1	0	1	0	0	0	0	0
Vector 5	0	1	1	1	0	0	1	1
Vector 6	1	1	0	0	1	1	0	0

Observed image

	FF_8	FF_7	FF_6	FF_5	FF_4	FF_3	FF_2	FF_1
Vector 1	X	0	1	0	0	0	0	1
Vector 2	1	1	0	0	1	1	1	0
Vector 3	X	1	1	1	0	1	1	0
Vector 4	X	0	0	1	0	0	0	0
Vector 5	0	0	1	1	1	0	1	1
Vector 6	X	1	1	0	0	1	0	0

corr(i, j): correlation between column i in the expected image and column j in the observed image

Figure 6.7: Correlation change in intermittent fault scenarios

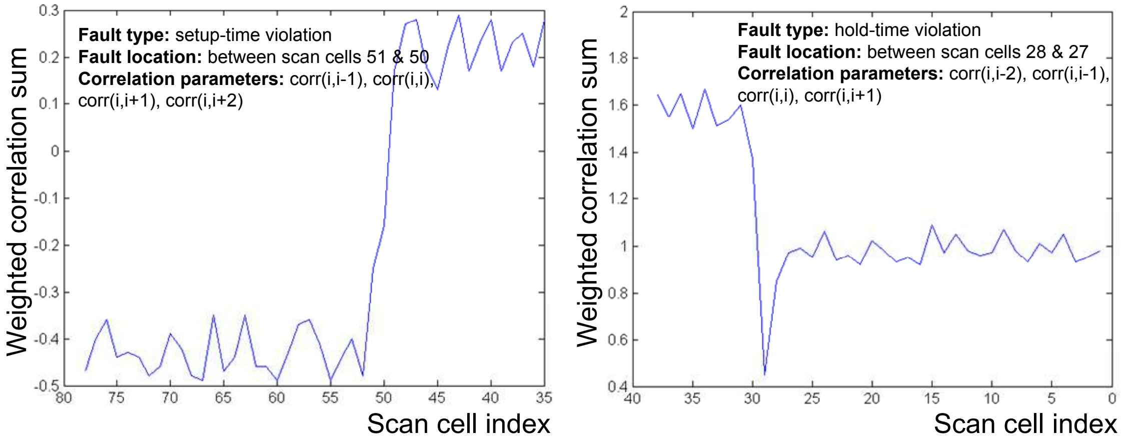


Figure 6.8: Curves of weighted correlation sum

Multiple intermittent faults in the scan chain incrementally change the phase skew of scan images. If the current phase skew is within the value range of $(v, v+1)$, a new timing fault can either increase or decrease the phase skew by at most 1, depending on the fault type and the manifestation probability. Hence the updated phase skew should reside in the value range of $(v-1, v+2)$. Such a phase movement can be clearly reflected by the change of four correlation values: $corr(i, i-v+1)$, $corr(i, i-v)$, $corr(i, i-v-1)$, $corr(i, i-v-2)$, as these four correlation parameters represent the heavy correlations associated with the current and the next phase skews. More specifically, a hold-time fault may increase $corr(i, i-v-1)$ and $corr(i, i-v-2)$, but decrease $corr(i, i-v+1)$ and $corr(i, i-v)$, whereas a setup-time fault would change these correlations the other way around. In light of this observation, a weighted correlation sum is defined to extract the phase movement information, as shown in the following equation.

$$\begin{aligned} S_{corr} &= -2 * corr(i, i-v+1) - 1 * corr(i, i-v) \\ &\quad +1 * corr(i, i-v-1) + 2 * corr(i, i-v-2) \end{aligned} \tag{6.2}$$

Intermittent fault diagnosis can be efficiently performed with the guidance of the weighted correlation sum. If we sweep the scan images from scan-out to scan-in and extract S_{corr} for each column, it can be observed that the S_{corr} value remains relatively stable in scan chain portions that contain no faults. However,

S_{corr}	C1	C2	C3	C4	C5	C6	C7	C8
	1.61	1.50	1.67	1.27	1.15	1.18	1.25	1.14

CUSUM	S0	S1	S2	S3	S4	S5	S6	S7	S8
	0	0.26	0.42	0.74	0.67	0.47	0.30	0.21	0

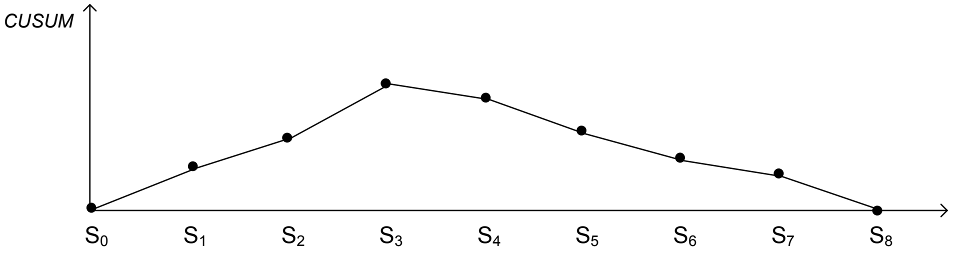


Figure 6.9: CUSUM chart of S_{corr} sequence

a sharp increase (decrease) in S_{corr} would occur at a hold-time (setup-time) fault site, providing clear signals of fault locations and types. Figure 6.8 presents the simulated S_{corr} curves for a scan chain that contains two intermittent timing faults. A sharp change of the curve can be clearly observed at the fault locations, providing high levels of diagnostic resolution.

The *change point analysis technique* [38, 39] is employed to accurately identify the position where the S_{corr} value changes and filter out random fluctuations in the S_{corr} curve. The basic idea of change point analysis is to detect the trend shift of certain statistical parameters such as the mean. The focus on monitoring statistical parameters can effectively differentiate the trend change from random noises. To extract such statistical trend information, the cumulative sum chart (CUSUM) of the S_{corr} value sequence is constructed, as exemplified in Figure 6.9.

Let $c_i, 1 \leq i \leq n$ represent the sequence of S_{corr} values under examination. Then the CUSUM statistics S_i of this sequence is calculated as follows.

$$\begin{aligned} S_0 &= 0 \\ S_i &= S_{i-1} + (c_i - \overline{c}), 1 \leq i \leq n \qquad (6.3) \end{aligned}$$

where $\overline{c}$ is the mean of the sequence, i.e. $\overline{c} = \sum_{i=1}^{n} c_i/n$. It should be noted that the CUSUM statistics outlined above are not the cumulative sums of the sequence.

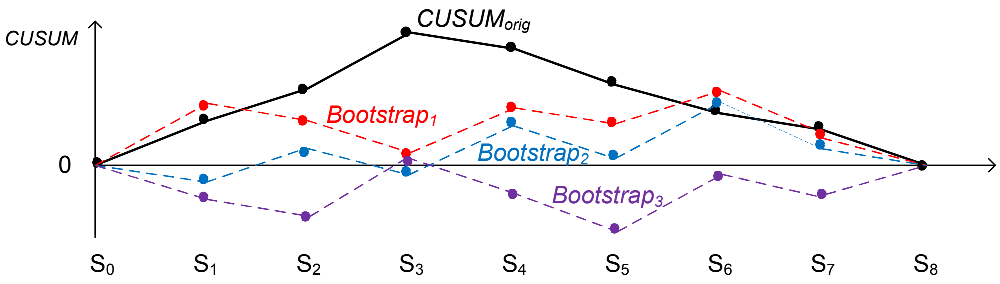


Figure 6.10: Bootstrap analysis

Instead they are the cumulative sums of differences between the sequence values and the average. As plotted in Figure 6.9, CUSUM always starts and ends at zero, as the differences sum to zero.

The directions of the CUSUM plot provide the information about the trend of the S_{corr} sequence. Suppose that within a subsequence the S_{corr} values tend to be above the overall average. Most of the difference values added to the cumulative sum will be positive and the CUSUM will steadily increase. A segment of the CUSUM chart with an upward slope indicates a period where the S_{corr} values tend to be above the overall average. Likewise a segment with a downward slope indicates a subsequence where the S_{corr} values tend to be below the overall average. A sudden change in the direction of the CUSUM indicates a sudden shift or change in the average. Segments where the CUSUM chart follows a relatively straight path indicate a subsequence where the average did not change. As shown in Figure 6.9, the direction of the CUSUM plot changes at position 3, indicating a trend change at this location.

Once the change point is identified, a bootstrap analysis is performed to determine its confidence level. This analysis estimates the confidence level by comparing the magnitude of change that has occurred in the sequence under examination to that of random sequences. If the magnitude of the detected change is evidently larger than a random sequence fluctuation, a high confidence level can be attained for the detected change point. The magnitude of change is estimated as the difference between the maximum and minimum of the CUSUM values, i.e. $S_{diff} = S_{max} - S_{min}$. A large number, N, of bootstrap sequences are constructed by randomly reordering the S_{corr} sequence under examination. Figure 6.10 shows three example bootstrap sequences generated from the original S_{corr} sequence listed in Figure 6.9. The bootstrap sequences thus generated represent the behavior of random data with no obvious trend change. The change magnitude statistics for these sequences, S^1_{diff} through S^N_{diff}, are calculated in the same manner as outlined above. If the original sequence does have an evident trend change within it, its S_{diff} would be larger than that of most randomly reordered sequences where no obvious change is embedded. Hence the percentage of bootstrap sequences whose change magnitudes are less than that of the original sequence provides the confidence level information regarding the likelihood of the detected change being a real one. More specifically, let M denote the number of bootstrap sequences for which $S^{bootstrap}_{diff} < S^{orig}_{diff}$; then the confidence level is defined as:

$$ConfidenceLevel = 100\frac{M}{N}\% \tag{6.4}$$

Using the CUSUM and bootstrap analysis outlined above, the proposed approach sweeps through the failing scan chains to pinpoint all change points with high confidence levels.

6.2.3 Manifestation probability computation

In order to evaluate the criticality of an intermittent fault for cost-effective design optimization, the fault manifestation probability needs to be ascertained. Attaining such information necessitates mathematically modeling the impact of manifestation probability on statistical features that can be observed in the scan images. The distribution of two-bit patterns in each segment of the scan images constitutes a fault-sensitive feature, as the magnitude of distribution distortion is a function of fault manifestation probabilities. Moreover, if the manifestation probabilities of a fault are different in distinct directions, an asymmetric change in the two-bit pattern distribution can also be observed, enabling good characterization of the realistic fault behavior. Therefore, a mathematical framework is proposed in this work to extract the fault manifestation probability information from the

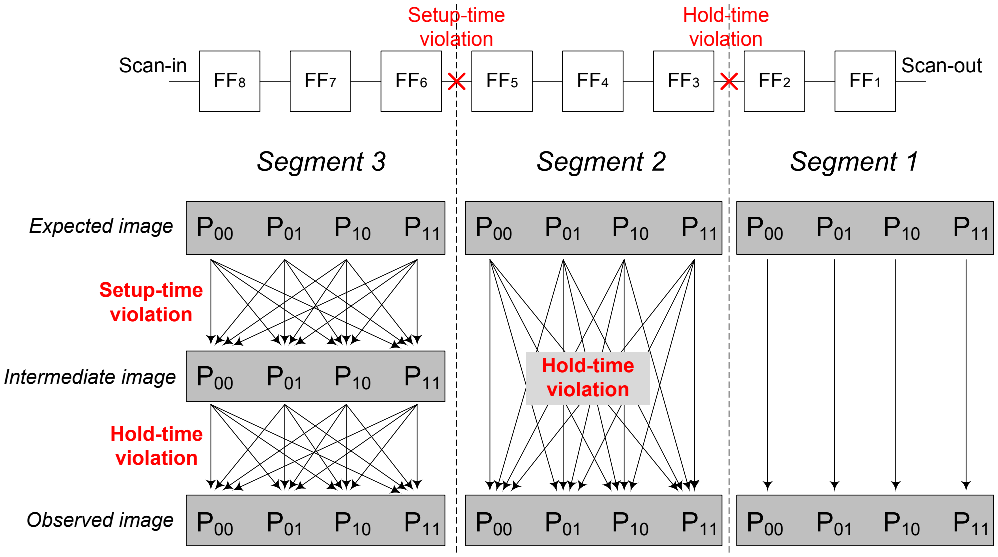


Figure 6.11: Variation of two-bit pattern distribution

two-bit pattern distribution feature.

The manifestation of any intermittent timing fault would distort the two-bit pattern distribution of a scan image by probabilistically transforming one type of two-bit pattern to some other type. If a segment of the scan image passes through multiple faults, such a random distortion process constitutes essentially a Markov chain, as the next distribution resulting from a new fault only depends on the current distribution. Such a process is illustrated in the double-fault example shown in Figure 6.11. Segment 3 of the expected image is first transformed to an intermediate state by the setup-time violation, and then to the observed image by the hold-time violation.

The impact of setup-time and hold-time violations on the two-bit pattern distribution can be modeled as two transition matrices of the Markov chain, as shown in Figure 6.12. These two matrices depict the probability of a fault transforming a certain two-bit pattern to another pattern. For example, in order for a hold-time fault to convert a 00 pattern to 10, the left neighboring bit of this pattern must have a value of 1 and a fast-to-rise type of hold-time violation must manifest itself when this sequence passes through the fault location. This results in a transition probability of $P_1 M_{fr}$. The probabilities for other types of transitions

Transition matrix for hold-time fault

$$H = \begin{array}{c|cccc} & \rightarrow 00 & \rightarrow 01 & \rightarrow 10 & \rightarrow 11 \\ \hline 00\rightarrow & 1-P_1M_{fr} & 0 & P_1M_{fr} & 0 \\ 01\rightarrow & M_{ff}-P_1M_{fr}M_{ff} & 1-M_{ff}-P_1M_{fr}+P_1M_{fr}M_{ff} & P_1M_{fr}M_{ff} & P_1M_{fr}-P_1M_{fr}M_{ff} \\ 10\rightarrow & P_0M_{ff}-P_0M_{fr}M_{ff} & P_0M_{fr}M_{ff} & 1-M_{fr}-P_0M_{ff}+P_0M_{fr}M_{ff} & M_{fr}-P_0M_{fr}M_{ff} \\ 11\rightarrow & 0 & P_0M_{ff} & 0 & 1-P_0M_{ff} \end{array}$$

Transition matrix for setup-time fault

$$S = \begin{array}{c|cccc} & \rightarrow 00 & \rightarrow 01 & \rightarrow 10 & \rightarrow 11 \\ \hline 00\rightarrow & 1-P_1M_{sr} & P_1M_{sr} & 0 & 0 \\ 01\rightarrow & P_0M_{sf}-P_0M_{sr}M_{sf} & 1-M_{sr}-P_0M_{sf}+P_0M_{sr}M_{sf} & P_0M_{sr}M_{sf} & M_{sr}-P_0M_{sr}M_{sf} \\ 10\rightarrow & M_{sf}-P_1M_{sr}M_{sf} & P_1M_{sr}M_{sf} & 1-M_{sf}-P_1M_{sr}+P_1M_{sr}M_{sf} & P_1M_{sr}-P_1M_{sr}M_{sf} \\ 11\rightarrow & 0 & 0 & P_0M_{sf} & 1-P_0M_{sf} \end{array}$$

Hold-time fault manifestation probabilities: M_{fr}, M_{ff}
Setup-time fault manifestation probabilities: M_{sr}, M_{sf}
Signal probabilities: P_1, P_0

Figure 6.12: Transition matrices for timing faults

can be derived in an analogous manner. The transition probabilities are organized in the form of *right transition matrices*, with the rows summing to 1.

In the example shown in Figure 6.11, Segment 2 of the scan image passes solely through the hold-time fault. If we represent the two-bit pattern distributions of expected and observed Segment 2 images as two row vectors, namely, $D_e(Seg_2) = [P^e_{00}\ P^e_{01}\ P^e_{10}\ P^e_{11}]$ and $D_o(Seg_2) = [P^o_{00}\ P^o_{01}\ P^o_{10}\ P^o_{11}]$, then the impact of the fault on Segment 2 can be approximated by a linear transformation.

$$D_e(Seg_2) * H = D_o(Seg_2) \qquad (6.5)$$

Since the pattern distributions of the expected and observed images are known information, the only variables in this equation consist of the fault manifestation probabilities, M_{fr} and M_{ff}, which can be resolved by minimizing the least-square norm of the difference between the two sides of the equation.

$$\begin{aligned} & min\|D_e(Seg_2) * H - D_o(Seg_2)\| \\ & s.t. \;\; 0 < M_{fr} < 1, 0 < M_{ff} < 1 \end{aligned} \tag{6.6}$$

Solving this nonlinear programming problem enables us to identify the probability values that can maximally explain the observed distribution distortion effect in Segment 2. A number of numerical algorithms, such as the *trust region approach* [21], can be utilized to attain solutions with good accuracy.

Once the manifestation probabilities of the rightmost fault are resolved, the proposed scheme furthermore examines the second fault from the right by modeling the distribution variation in Segment 3. In the example shown in Figure 6.11, this segment sequentially passes through the setup-time and hold-time faults. Hence the transformation process can be depicted by the following equation.

$$D_e(Seg_3) * S * H = D_o(Seg_3) \tag{6.7}$$

Since the manifestation probabilities for the rightmost fault are already known, the manifestation probabilities for the second fault, M_{sr} and M_{sf}, can be efficiently computed by solving the nonlinear programming problem constructed with Equation 6.6. Under a multiple fault situation, the manifestation probabilities of all the faults can be efficiently determined by iteratively applying the aforementioned technique.

6.3 Scan clock delay fault diagnosis

The scan cell level diagnosis outlined in Section 6.2 pinpoints the timing defects in the flip-flops and/or interconnects on the scan paths. On the other hand, a defect-free scan path can also exhibit timing faults, as the scan clocks utilized to synchronize these scan cells are distorted by the faults in the clock distribution networks. In such cases, knowledge solely of the failing scan cell information is

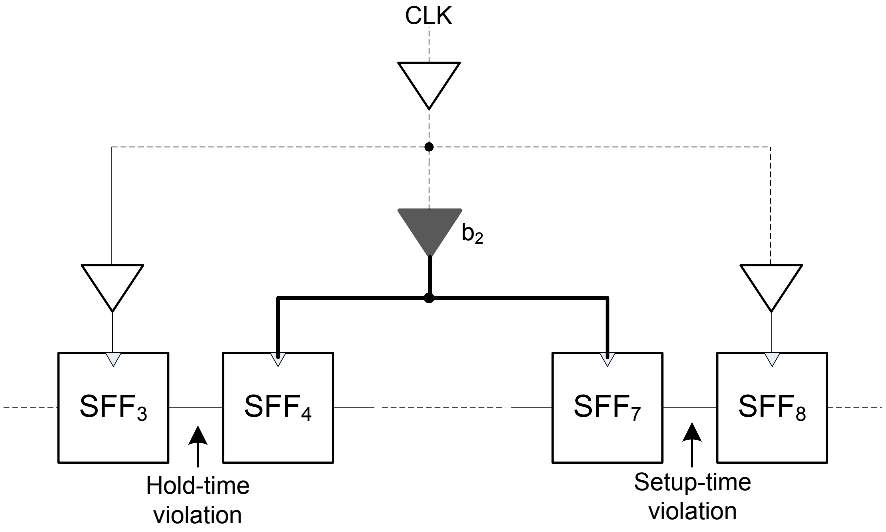


Figure 6.13: Scan chain timing failures induced by clock faults

insufficient for understanding the root failure mechanism. The diagnosis process needs to be extended to the clock buffer level so as to provide accurate information for design and fabrication improvement.

6.3.1 Problem formulation

A set of technical assumptions for the proposed work is outlined in this section.

Clock fault model

As shown in Figure 6.13, a delay fault on buffer b_2 creates a timing skew between SFF_3 and SFF_4. Since the clock of SFF_4 is delayed relative to SFF_3, this fault would lead to a hold-time violation if the clock skew exceeds the hold time slack between these two scan cells. Similarly, the same buffer fault might result in a setup-time violation between SFF_7 and SFF_8.

The proposed technique aims to identify multiple clock buffer faults so as to maximally approximate the realistic failure syndrome which typically contains multiple failing scan cells. As the gross faults are of more interest to designers from the diagnosis perspective, the proposed technique focuses on the diagnosis of such faults by assuming that each scan chain failure is caused by a single clock buffer fault. In contrast to the traditional single-fault assumptions, the proposed single-fault-per-failure assumption allows the combination of multiple clock buffer faults as long as different faults contribute to distinct scan chain failures. Obviously, different scan chain failures can share the same clock buffer as their common causality, as the example shown in Figure 6.13.

Clock buffer delay model

During the design stage, the timing of the scan chains is typically closed at a particular set of PVT corners. Nonetheless, the delay of the clock paths, as a function of the power-ground noise, exhibits a statistical distribution during the actual application of diagnostic patterns, which in turn results in the probabilistic failures observed in scan chains. The proposed technique models the buffer delay as a normal distribution, with the 3-sigma bounds of the distribution corresponding to the min/max voltage corners during the scan operation. The voltage variation can be estimated by power-ground noise analysis tools such as Apache RedHawk.

Diagnostic goal

The statistical delay model enables the quantitative estimation of the expected impact of any fault hypothesis in terms of the resulting scan chain failure locations, types and manifestation probabilities. On the other hand, the actual scan chain failure information can be attained by performing the scan chain diagnosis, as outlined in Section 6.2. Therefore, the ultimate goal of the clock tree diagnosis process consists of the identification of a set of faulty clock buffers that fulfills the single-fault-per-failure constraint and incurs minimum deviation between the expected and observed scan chain failures.

6.3.2 Establishing Fault Hypotheses

The manifestation of clock delay faults on scan chains is subject to two sets of constraints, namely, the logic reachability and the timing relationship. The

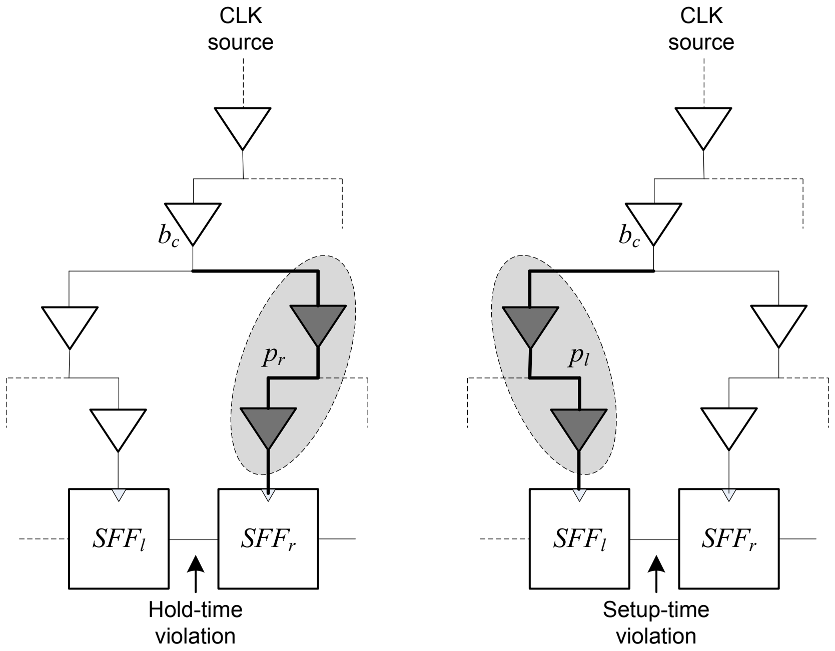


Figure 6.14: Logic pruning based on scan chain failure locations and types

extraction and utilization of these constraints can significantly reduce the space of fault hypotheses, leading to a more efficient diagnosis.

Logic pruning

The logic reachability of a clock buffer to the scan cells constrains the possible locations and types of scan chain failures. Therefore, a small set of candidate faulty buffers can be identified by tracing the clock path back from the scan chain failure syndrome.

The basic idea is illustrated in Figure 6.14. If a hold-time violation is observed between a pair of adjacent scan flip-flops, the clock of the flip-flop on the right, SFF_r, must be abnormally delayed compared to the clock of SFF_l, creating a positive timing skew. Tracing the clock paths of these two flip-flops, it can be easily seen that the two flip-flops share a common path between the clock source to buffer b_c where the clock branches out to two different paths. No delay fault within the common path is able to change the relative timing between the two flip-flops. Therefore, the fault must be located in the fan-out branches of b_c, according to the single-fault-per-failure assumption. Moreover, only the faults in the right branch p_r are able to incur the positive timing skew observed in hold-time violation. Hence buffers in p_r constitute the candidate fault sites for this hold-time failure. On the other hand, if a setup-time violation is observed, the clock fault should reside in the left branch p_l of the clock path so as to reduce the setup timing slack.

For each failure in the scan chain, a set of candidate faulty clock buffers can be identified through the logic pruning idea outlined above. It is important to note that the candidate sets of different scan chain failures can partially overlap as one clock fault might be able to incur multiple scan chain failures.

Fault size estimation

The candidate fault set generated by logic pruning might still contain a large number of unrealistic fault hypotheses. Analyzing the timing impact of the clock faults provides a further criterion for evaluating the matching level between the fault hypotheses and the observed failure behavior.

As discussed in logic pruning, the left and right clock branches, p_l and p_r, can be identified for each pair of adjacent scan cells. The timing skew between these two branches results in the observed scan chain failures. Since the delay of a buffer, b_i, is modeled as a normal distribution, $N(\mu_i, \sigma_i^2)$, the delay of the clock branch, p, can be further modeled as the distribution sum of the buffers along the branch, as shown below.

$$Delay(p) = \sum_{b_i \in p} N(\mu_i, \sigma_i^2) \tag{6.8}$$

Assuming independence among the buffer delay distributions, the clock branch delay can be modeled by another normal distribution, $N(\mu_p, \sigma_p^2)$, with $\mu_p = \sum_{b_i \in p} \mu_i$ and $\sigma_p^2 = \sum_{b_i \in p} \sigma_i^2$. The clock skew distribution can thus be further estimated as the difference between the delay distributions of p_r and p_l.

$$\begin{aligned} skew &= Delay(p_r) - Delay(p_l) \\ &= N(\mu_{p_r} - \mu_{p_l},\ \sigma_{p_r}^2 + \sigma_{p_l}^2) \end{aligned} \tag{6.9}$$

A delay fault of size d in the right branch p_r would change the relative timing between p_l and p_r by increasing the mean of the $Delay(p_r)$ distribution. Thus the skew under the fault condition would be changed to:

$$skew_d = N(\mu_{p_r} - \mu_{p_l} + d,\ \sigma_{p_r}^2 + \sigma_{p_l}^2) \tag{6.10}$$

This results in an increased hold-time violation probability which can be estimated by the following equation.

$$\begin{aligned} P(hold) &= P(skew_d > SLACK_{hold}) \\ &= \int_{SLACK_{hold}}^{+\infty} skew_d \end{aligned} \tag{6.11}$$

Since the scan chain diagnosis step provides the information of hold-time failure probability, the clock delay fault size, d, is the only unknown parameter in this model. Hence solving Equation 6.11 enables the estimation of the fault size that causes the observed hold-time failure. For setup-time violations, a similar mathematical model can be constructed to estimate the delay fault size.

The estimation of fault size not only provides the designers insights regarding the criticality of each fault, but enables further prioritization of different fault hypotheses through timing reasoning. More concretely, a realistic fault hypothesis should not only account for observed failures, but minimize the failure likelihood in scan chain locations where no failures have been observed. The basic idea of timing pruning is illustrated in Figure 6.15. If a hold time failure with a manifestation probability of 0.4 is observed between SFF_3 and SFF_4, buffers b_1 and b_2 are considered fault candidates as they are both in the faulty clock branch. Moreover, a 0.7ns delay defect is needed to incur the observed failure manifestation probability. If b_1 is the actual fault site, a 0.7ns extra delay on it would result in a hold-time violation probability of 0.35 in the location between SFF_{10} and SFF_{11}, where no failure is actually observed. On the other hand, the hypothesis of b_2 being the actual fault results in a setup-time violation in a failure-free location between SFF_7 and SFF_8, with a manifestation probability of 0.1. In this case, b_2 would have a higher likelihood of being the actual fault as it induces less deviation

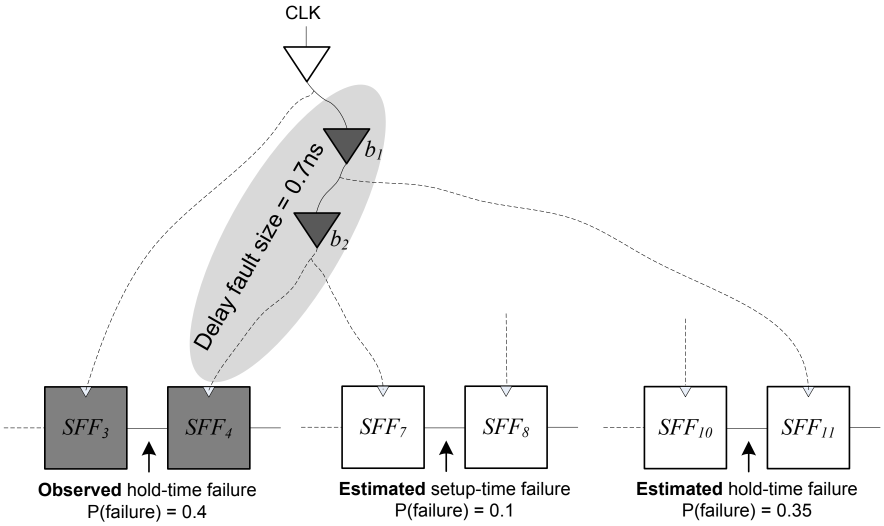


Figure 6.15: Timing reasoning based on fault size estimation

between the estimated and observed failure behavior. Although Figure 6.15 only exemplifies simple fault hypotheses for illustrative purpose, complicated fault hypotheses, such as multiple faults with timing interactions among each other, can be prioritized using exactly the same timing analysis.

6.3.3 Diagnosis Flow

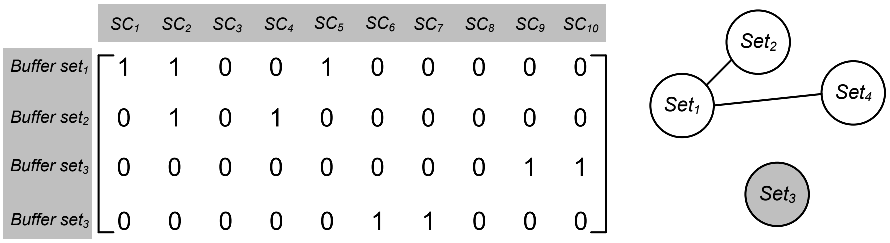
In large scan designs with possibly multiple faults, the timing relationship becomes rather complicated as the interaction between various faults needs to be considered for accurate timing pruning. For instance, two clock faults might cancel each other's impact on a pair of scan cells if they happen to delay both clock branches by a similar amount of time. Exhaustive timing reasoning for every fault hypothesis combination can cover all possible timing interaction effects, yet the computational time proves prohibitive due to the large search space. If n failures are observed in the scan chain and each failure f_i corresponds to a candidate faulty buffer set S_i, a total of $\prod_{i=1}^{n} |S_i|$ fault combinations need to be examined to search for the most feasible fault hypotheses, even under the single-fault-per-failure assumption.

One interesting observation is that the pruning space can be partitioned into several subspaces, as long as the candidate buffers fulfill certain logic relationships. The search for the optimal fault hypothesis for each individual subspace can be performed locally, significantly reducing the computational overhead without impacting the optimality of the overall solution. A branch-and-bound algorithm is employed to perform the pruning within each subspace. A bound estimation metric, which computes the minimum deviation of the expected failure probabilities, is proposed to enable an early pruning of inferior branches of the search tree so as to further improve the efficiency.

Pruning space partitioning

The clock faults can interact with each other through multiple ways. Direct interaction can be found in clock faults which impact the timing of the same pair of scan cells. Two clock faults might also indirectly interact with each other if their direct interactions with other clock faults form a chain that bridges them together. However, if the buffers in two candidate sets exhibit neither direct nor indirect interactions, they can be analyzed independently. The optimal fault hypotheses locally selected from each candidate set would form a globally optimal fault hypothesis combination.

The timing interaction among different candidate sets can be represented with a logic reachability matrix, as shown in Figure 6.16. Each column of the matrix corresponds to a scan cell (SC) in the scan chain, and each row denotes a candidate buffer set. A value of 1 in the matrix position (i, j) denotes that scan cell SC_j is reachable from at least one buffer within Set_i. In this example, sets 1 and 2 directly interact with each other, as they both reach scan cell pairs (SC_1, SC_2) and (SC_4, SC_5). Similarly, sets 1 and 4 directly interact with each other as well. An indirect interaction exists between sets 2 and 4 as set 1 bridges them together. Nonetheless, set 3 exhibits no interaction with any of the other sets. Therefore, the pruning for this example can be performed more efficiently in two disjoint subspaces.



$$
\begin{array}{c|cccccccccc}
 & SC_1 & SC_2 & SC_3 & SC_4 & SC_5 & SC_6 & SC_7 & SC_8 & SC_9 & SC_{10} \\
\hline
\text{Buffer set}_1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
\text{Buffer set}_2 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
\text{Buffer set}_3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\
\text{Buffer set}_3 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0
\end{array}
$$

Figure 6.16: Fault hypothesis space partitioning

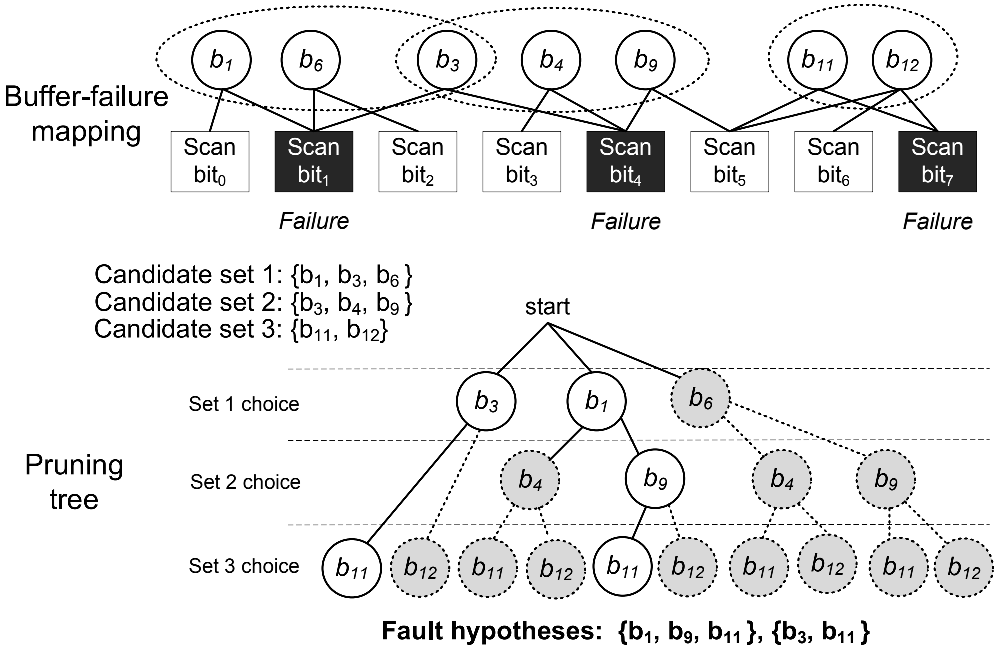


Figure 6.17: Branch-and-bound pruning

Branch-and-bound based pruning flow

A branch-and-bound pruning algorithm is employed to search within each subspace for the optimal fault hypotheses. Based on the single-fault-per-failure assumption, the algorithm selects exactly one fault from each candidate set and incorporates it in the fault hypothesis. This guarantees that all the observed scan chain failures would be covered by the selected faults. Figure 6.17 illustrates a pruning tree structure for a space consisting of three candidate buffer sets.

During the pruning process, the impact of the fault hypothesis under examination on the failure probabilities of various scan chain locations is estimated using the statistical timing model discussed in Section 6.3.2. The deviation between the estimated and the observed failure probabilities is measured by the least-square norm of their differences. During the pruning process, a current minimum deviation is maintained. The proposed branch-and-bound searching algorithm aims to identify the fault hypotheses with minimum deviations, as their expected behavior maximally matches the observed one.

The efficiency of the searching process is mainly impacted by two important factors, namely the order in which the branch is examined and the bounding strategy. The selection of a good searching order enables an early establishment of a near-minimum deviation, and the bounding strategy provides the estimation of the lower bound of the deviation for the node under examination, possibly enabling the pruning of large sub-trees.

The pruning tree is examined in a depth-first-search manner. When a branching point is encountered during the search process, the proposed algorithm selects the branch that examines earlier the buffers shared by multiple failures, as an actual clock defect is likely to create multiple scan failures. Based on this heuristic, the pruning process first searches the leftmost branch of the pruning tree shown in Figure 6.17, as b_3 is driving scan bits 1 and 4 as shown in the mapping between the clock buffers and the failing scan bits. If a tie occurs after applying the first branch selection heuristic, the proposed algorithm gives priority to the buffer that reaches fewer good scan bits and examines the branch associated with such a buffer earlier, as the inclusion of such a buffer in the fault hypothesis results in reduced deviation between the expected and observed syndromes on good scan bits. As shown by the example in Figure 6.17, b_{11} and b_{12} reach the same failing scan bit, i.e. bit 7. But b_{11} is examined before b_{12} in the search tree, as b_{11} only drives one good scan bit (bit 5) whereas b_{12} drives two (bits 5 and 6).

A bounding strategy is also proposed to enable early pruning of inferior fault hypotheses such as the grey area in Figure 6.17. When the pruning process reaches a new node of the search tree, it computes for each affected scan chain location the range of failure probability by taking into account the possible timing interaction with the fault sets in the lower levels of the search tree, i.e., the sets that have not been examined so far by the current fault hypothesis. The failure probability that is closest to the observed one is taken to create a strict lower bound of the deviation for the current fault hypothesis. If such a deviation is still higher than the currently maintained minimum deviation, the subtree below this node would be pruned off as the fault hypotheses associated with this subtree are proven to be inferior. After the pruning process, the remaining leaf nodes of the search tree constitute the set of highly probable fault hypotheses and are ranked based on their deviation magnitudes.

Algorithmic overview

The entire diagnosis flow is summarized in Figure 6.18. As can be seen, the proposed scheme does not rely on special diagnostic instruments or design-for-diagnosis hardware insertion, thus lending itself to a wide range of applications. The diagnosis flow also requires very little involvement of engineering effort, and can therefore be integrated into a highly automated failure analysis framework.

6.4 Simulation results

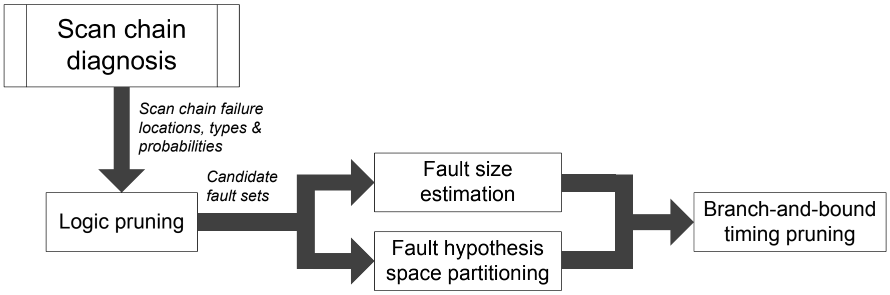

Scan chain diagnosis
Scan chain failure locations, types & probabilities
Logic pruning
Candidate fault sets
Fault size estimation
Fault hypothesis space partitioning
Branch-and-bound timing pruning

Figure 6.18: Diagnosis flow

6.4.1 Scan cell diagnosis results

Scan cell diagnosis simulations have been performed on large ISCAS89 and ITC99 benchmarks and two industrial designs in order to evaluate the effectiveness of the proposed scheme. The two industrial designs employed in our simulation consist of a floating-point unit (FPU) and the top-level logic of a memory system (MEMSYS). The number of scan cells in these two designs is 13803 and 45766, respectively. A total of 20 scan chains are constructed in each benchmark, and 40 scan chains are inserted in each industrial design. Each circuit is assumed to have one faulty scan chain in our simulation. To examine the complicated fault behavior, a set of three timing faults with randomly assigned fault locations and types is injected into the faulty scan chain. For each circuit, 200 randomly generated timing violation immune test stimuli are utilized to create the scan image of the faulty scan chain. Hence the scan image consists of 200 rows. Both the permanent and intermittent fault scenarios have been examined in our simulation.

In the case that permanent faults are injected, the simulation shows that the proposed method always guarantees a perfect diagnostic resolution. This is because the incremental phase movement induced by permanent faults always has a step length of 1. Such a highly regular behavior can be captured by the proposed technique with no ambiguity, leading to a precise identification of the fault locations and types.

The proposed technique also delivers highly accurate diagnostic results for the more challenging case of intermittent faults. We have simulated for each benchmark 100 different three-fault combinations, and a confidence level threshold of 0.90

is used in our simulation to determine the change point interval which represents a candidate fault window. Table 6.2 exemplifies the results of one such simulation run. The diagnosis result (*Diag.*) is compared to the fault injection parameters (*Inj.*) in terms of location, type and manifestation probability[3]. Figure 6.19 presents the statistics regarding the distribution of the diagnostic resolution for the faults examined in all the 100 simulation runs. For most faults, a high diagnostic resolution is delivered, as the resulting candidate fault window typically contains only 2 or 3 scan cells.

In addition to the accurate identification of fault locations, the proposed technique always guarantees the correct identification of the fault type throughout the simulation, as the direction of phase movement exhibits no ambiguity and can be clearly captured.

The estimated fault manifestation probabilities approximate the injected fault quite well, as the error between the injected and estimated probabilities is typically less than 0.1. More importantly, it can be seen that the estimated probability values preserve the relative magnitude relationship of the injected fault manifestation probabilities, providing insights regarding the relative criticality of different faults. Figure 6.20 shows the effectiveness of the proposed estimation technique in preserving the relative magnitude relationship. In this figure, we examine all the simultaneously injected fault pairs and report the percentage of fault pairs whose estimated relative fault magnitude matches the injected one. The horizontal axis of this diagram denotes the difference between the manifestation probabilities of the two faults being examined together. Apparently, the larger the difference, the easier it is to maintain the relative magnitude in estimation, as the estimation error margin increases with the difference. As shown in the figure, a correct estimation ratio of over 60% can be attained even when the difference between the manifestation probabilities of the two faults is less than 0.05. The correct estimation ratio rapidly increases to over 90% when the difference exceeds

[3] The indices of scan cells between which the injected/diagnosed fault resides are listed in the *Location* column. In the *Type* column, HT stands for hold-time fault, and ST for setup-time fault. In the *Prob.* column, the manifestation probabilities of each fault are presented in the form of $P(rising\ fault)$ / $P(falling\ fault)$.

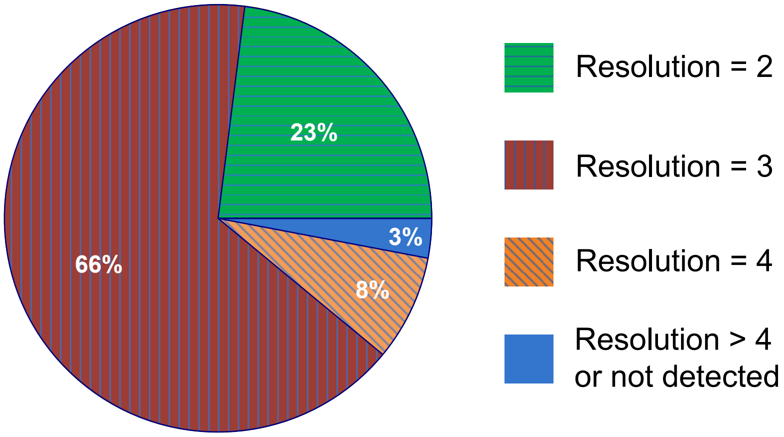


Figure 6.19: Diagnostic resolution distribution

0.2. The high accuracy of the proposed approach in estimating the relative fault magnitudes results in a precise identification of the strong faults as the main factors of the yield loss, thus enabling a more focused debugging and design optimization process.

An interesting observation can be found between the diagnostic resolution and the fault manifestation probability. In general, a better diagnostic resolution can be attained for faults with higher manifestation probabilities, as such faults result in more abrupt phase movement, providing stronger signals that can be detected by the proposed diagnosis technique. Such a trend is shown in Figure 6.21 which summarizes the average diagnostic resolution of faults at distinct manifestation probability regions. The extreme case of such a trend consists of the permanent fault situation where a perfect diagnostic resolution is attained, as outlined above.

Since the scan fault model employed in this work assigns different manifestation probabilities for rising and falling faults, the randomly generated faults used in our simulation are able to cover the generic fault scenarios with asymmetric rising and falling manifestation probabilities. To examine extreme cases, we have further inserted and simulated for each benchmark 10 different three fault combinations where each fault has only one manifestation direction, that is, the manifestation probability of one direction is 0. These faults approximate the behavior of tradi-

Table 6.2: Intermittent fault diagnosis

Circuits	Location		Type		Prob.	
	Inj.	Diag.	Inj.	Diag.	Inj.	Diag.
s13207	11-12	11-13	ST	ST	0.26 / 0.35	0.31 / 0.38
	19-20	18-20	HT	HT	0.45 / 0.43	0.48 / 0.52
	27-28	27-29	ST	ST	0.34 / 0.55	0.40 / 0.51
s15850	8-9	8-9	ST	ST	0.60 / 0.54	0.53 / 0.51
	14-15	14-16	HT	HT	0.31/ 0.49	0.36 / 0.46
	25-26	25-28	HT	HT	0.22 / 0.25	0.25 / 0.30
s35932	22-23	22-23	HT	HT	0.68 / 0.57	0.64 / 0.49
	30-31	29-31	ST	ST	0.27 / 0.30	0.32 / 0.28
	37-38	37-39	HT	HT	0.43 / 0.30	0.46 / 0.37
s38417	19-20	19-21	ST	ST	0.35 / 0.42	0.38 / 0.47
	26-27	27-29	HT	HT	0.24 / 0.68	0.33 / 0.64
	45-46	45-47	ST	ST	0.46 / 0.30	0.42 / 0.34
s38584	13-14	13-15	HT	HT	0.33 / 0.39	0.37 / 0.41
	32-33	32-34	HT	HT	0.22 / 0.30	0.29 / 0.26
	51-52	51-53	ST	ST	0.44 / 0.31	0.47 / 0.39
b17	6-7	6-7	HT	HT	0.67 / 0.60	0.61 / 0.58
	27-28	26-28	HT	HT	0.31 / 0.47	0.36 / 0.44
	41-42	41-43	ST	ST	0.53 / 0.40	0.49 / 0.43
b21	12-13	12-14	ST	ST	0.35/ 0.42	0.38 / 0.45
	20-21	20-21	HT	HT	0.64 / 0.69	0.60 / 0.67
	28-29	27-29	HT	HT	0.42 / 0.39	0.46 / 0.46
FPU	24-25	24-26	ST	ST	0.49 / 0.39	0.45 / 0.43
	102-103	102-105	ST	ST	0.21 / 0.27	0.25 / 0.23
	161-162	160-162	HT	HT	0.46 / 0.30	0.43 / 0.36
MEMSYS	93-94	93-94	HT	HT	0.52 / 0.57	0.58 / 0.61
	441-442	441-443	ST	ST	0.34 / 0.38	0.36 / 0.43
	606-607	606-608	HT	HT	0.31 / 0.50	0.38 / 0.47

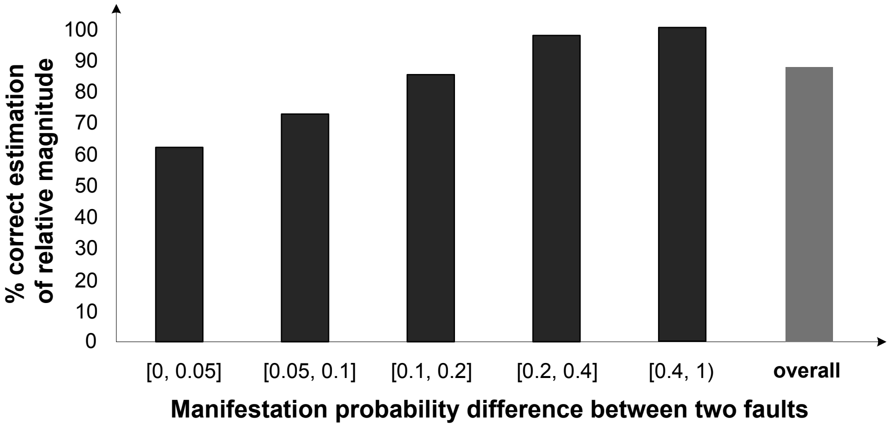


Figure 6.20: Accuracy of relative magnitude estimation

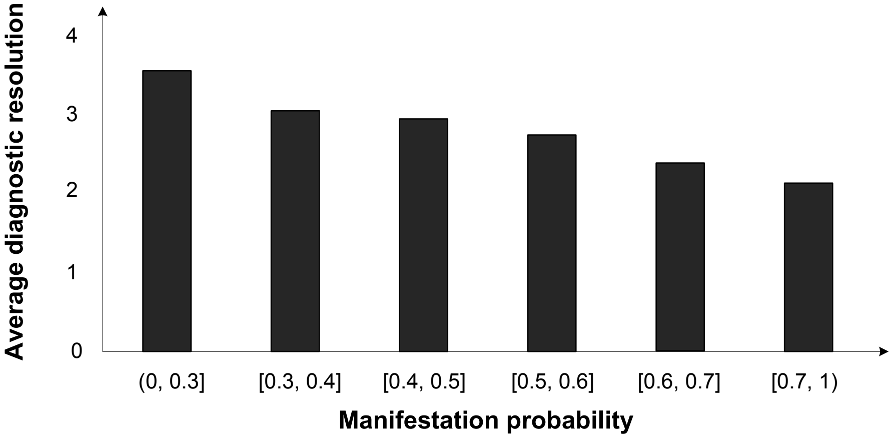


Figure 6.21: Diagnostic resolution distribution in fault manifestation probability

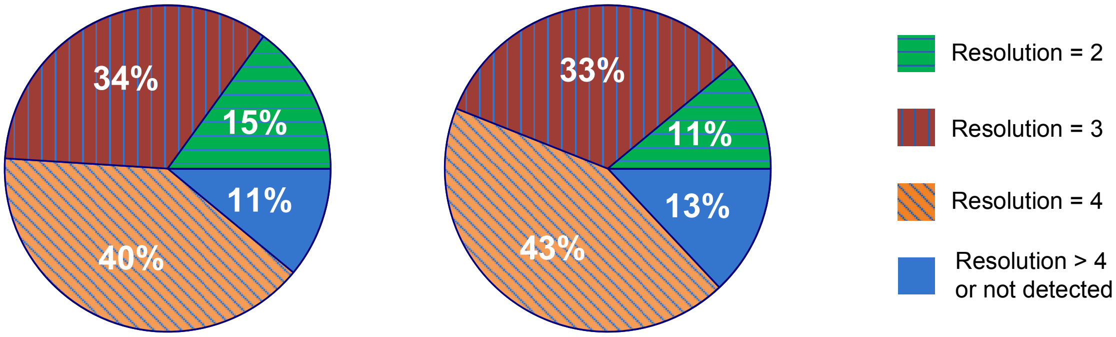


Figure 6.22: Diagnostic resolution distribution of unidirectional faults

tional pattern-dependent fault models such as fast-to-rise/fall and slow-to-rise/fall faults. The diagnostic resolutions of the faults depend on the fault probabilities on their manifestation directions. The size of the candidate fault window in general increases when the fault probability decreases, as the phase movement effect induced by weaker faults is less evident. But as illustrated by the resolution distribution diagram in Figure 6.22, the fault locations of a high percentage of these special faults can still be narrowed down to a window of less than or equal to 4 scan cells.

The computation time of the scan cell diagnosis process is reported in Table 6.3. The second and third columns of the table show the CPU time required for failure location/type identification and the manifestation probability estimation, respectively. The computational complexity of the failure location and type identification process is proportional to the length of the failing scan chains, as the algorithm needs to sweep the scan images to compute the column-wise correlations of the scan data. The computation time required for manifestation probability estimation is mainly determined by the number of faults. Since a three-fault scenario is examined in our simulation, this portion of CPU time remains relatively constant for all the benchmarks. In general, it can be seen that the proposed scan cell diagnosis approach extracts the statistical feature of the scan images in a highly efficient manner, resulting in a very low computational cost.

Table 6.3: CPU time of scan cell diagnosis in seconds

Circuits	Location/type identification	Probability estimation	Total
s13207	0.37	3.51	3.88
s15850	0.44	3.42	3.86
s35932	1.26	3.68	4.94
s38417	1.41	3.71	5.12
s38584	1.43	3.62	5.05
b17	1.37	3.60	4.97
b21	0.41	3.57	3.98
FPU	6.46	3.75	10.21
MEMSYS	20.31	3.85	24.16

6.4.2 Scan clock buffer diagnosis results

Scan architectures of various sizes and their associated scan clock trees are constructed as the simulation benchmarks. Delay faults are randomly injected into the clock buffers, and the failure syndrome in the scan chains is generated through simulation. The proposed scheme takes the simulated failure syndrome as its input and reports a small set of possible fault hypotheses. The reported fault hypotheses are ranked based on their matching level with the observed failure syndrome.

The simulation results are detailed in Table 6.4. A total of eight scan architectures are simulated. Their configurations are listed in column 1. For each scan architecture, the 2-fault and 3-fault scenarios are examined respectively. For each fault scenario, a total of 100 randomly generated fault combinations are injected and diagnosed. The statistical profile of the diagnosis results for the 100 trials is reported in the table to show the average effectiveness of the method. Column 3 lists the average number of reported fault hypotheses as the result of the pruning process. The proposed diagnosis algorithm in general reports only a small set of fault hypotheses in the final results, which confirms that the use of timing reasoning delivers a very high pruning efficiency. Columns 4 through 6 report the percentage of diagnostic trials whose top-ranked fault hypotheses match the injected faults. For example, the very top fault hypothesis reported by 57 percent of diagnostic trials matches exactly the actual faults when two faults are injected in the clock tree of the first scan architecture. The percentage of matching trials increases rapidly

Table 6.4: Scan clock buffer diagnosis results

Scan architecture	Fault scenario	# fault hypo.	% matching trials		
			top 1	top 2	top 3
776 scan cells 258 CLK buffers	2 faults	5	57%	80%	95%
	3 faults	8	52%	74%	93%
1738 scan cells 850 CLK buffers	2 faults	11	49%	74%	90%
	3 faults	12	51%	78%	88%
3948 scan cells 1305 CLK buffers	2 faults	8	55%	76%	92%
	3 faults	12	46%	69%	85%
7953 scan cells 1593 CLK buffers	2 faults	9	52%	73%	89%
	3 faults	15	42%	65%	82%
13821 scan cells 2311 CLK buffers	2 faults	13	54%	76%	85%
	3 faults	17	37%	59%	77%
12612 scan cells 3165 CLK buffers	2 faults	11	50%	68%	86%
	3 faults	14	45%	62%	80%
16090 scan cells 4036 CLK buffers	2 faults	12	48%	66%	83%
	3 faults	17	39%	55%	74%
28302 scan cells 4052 CLK buffers	2 faults	14	45%	64%	81%
	3 faults	16	32%	60%	77%

when the *top 2* or *top 3* fault hypotheses are considered. It can be observed that a high matching percentage is delivered by the proposed scheme for various fault scenarios, which indicates a high likelihood of nailing down the actual faults within a very small group of candidate sets through the application of the proposed technique. It should be noted that, even though some reported fault hypotheses are not completely identical to the injected fault set, they still cover a large portion of the actual faults. For example, when faults are injected in buffers 54, 78 and 213 of the third scan architecture in one trial instance, the *top 3* fault hypotheses reported by the diagnosis process are $\{54, 78, 240, 291\}$, $\{44, 54, 78, 197, 225\}$ and $\{54, 78, 213\}$. Although the *top 2* hypotheses fail to completely match the injected faults in this case, they are still able to flag the majority of the actual fault sites. This observation indicates that the buffers shared by a large number of reported fault hypotheses would have a very high likelihood of being the actual defects. The efficiency of the failure analysis process is significantly improved by focusing on examining such cells.

Table 6.5: CPU time of scan clock buffer diagnosis in seconds

Scan architecture	Fault scenario	Logic pruning	Timing pruning	Total
776 scan cells	2 faults	0.07	4.82	4.89
258 CLK buffers	3 faults	0.09	7.33	7.42
1738 scan cells	2 faults	0.17	10.26	10.43
850 CLK buffers	3 faults	0.24	15.05	15.29
3948 scan cells	2 faults	0.26	15.71	15.97
1305 CLK buffers	3 faults	0.33	19.49	19.82
7953 scan cells	2 faults	0.49	35.04	35.53
1593 CLK buffers	3 faults	0.66	50.72	51.38
13821 scan cells	2 faults	0.65	46.30	46.95
2311 CLK buffers	3 faults	0.82	58.91	59.73
12612 scan cells	2 faults	0.77	72.45	73.22
3165 CLK buffers	3 faults	0.94	107.30	108.24
16090 scan cells	2 faults	1.02	132.07	133.09
4036 CLK buffers	3 faults	1.21	159.64	160.85
28302 scan cells	2 faults	0.98	120.59	121.57
4052 CLK buffers	3 faults	1.27	153.90	155.17

The computation time of the scan clock buffer diagnosis process is reported in Table 6.5. For both the two-fault and three-fault scenarios, we report the CPU time for logic pruning and timing pruning processes, respectively. It can be seen that the logic pruning can finish in one second for most scan architectures used in our simulation, as it only requires a topological tracing of the scan clock trees. The timing pruning is the most time-consuming portion of the diagnosis flow, as it performs computationally expensive timing simulation to prune fault hypotheses. Nonetheless, since we utilize the pruning space partitioning technique and highly guided branch-and-bound search heuristics to significantly reduce the number of simulations, the pruning time cost can still be controlled at a very low level even for large designs with thousands of clock buffers[4], delivering a highly attractive overall performance.

[4] Clock trees of this size are comparable to ones in a large number of industrial designs.

6.5 Conclusions

In this work, we investigate the challenge of diagnosing timing failures in scan architecture. The proposed work examines both the direct and indirect causalities of scan chain timing failures, namely, the timing faults in scan cells and in scan clock buffers, so as to provide a comprehensive set of diagnostic information for design and fabrication improvement.

The methodology proposed in this chapter first performs a scan cell level diagnosis to identify multiple scan chain timing faults with mixed fault types and various manifestation probabilities. A novel approach based on statistically monitoring the phase skew of scan images is proposed to identify the fault locations and types. Column-wise correlations of scan images are extracted to generate strong signals regarding phase movement, thus maximally eliminating the ambiguity induced by probabilistic fault manifestations. A mathematical framework is further proposed to model the impact of intermittent faults as a Markov chain, thus enabling an accurate estimation of the fault manifestation probabilities.

To ascertain the possible indirect failure causalities, we further propose a technique to diagnose the delay faults in scan clock trees that cause the failing scan cells identified by the scan cell level diagnosis. The proposed approach identifies the fault hypotheses that maximally approximate the observed failure behavior based on logic and statistical timing pruning. The incorporation of statistical timing propagation of clock delay faults provides a means for the quantitative estimation of failure manifestation probabilities, thus enabling effective fault pruning by examining the statistical features of intermittent failure syndrome.

The proposed technique is completely compatible with existing scan test schemes, imposing no design overhead whatsoever. Integration of the proposed methodology into current industrial schemes enables efficient silicon debugging, significantly speeding up the design and test optimization cycles.

The text of Chapter 6, is in part a reprint of the material as it appears in *M. Chen and A. Orailoglu, "Diagnosing scan chain timing faults through statistical* *feature analysis of scan images," Design, Automation and Test in Europe, 2011*; in *M. Chen and A. Orailoglu, "Diagnosing scan clock delay faults through statistical timing pruning," Design Automation Conference, 2011*; and in *M. Chen and A. Orailoglu, "On diagnosis of timing failures in scan architecture," IEEE Transactions on CAD*. The dissertation author was the primary researcher and author of the publications [18], [19] and [14].

Chapter 7

A functional flow for fast-to-revenue

After performing an in-depth analysis on the theoretical aspects of the underpinning techniques, this chapter presents a discussion on the industrial applicability of these techniques. A functional flow, focusing on the integration of these techniques into mainstream industrial test and failure analysis infrastructure, is provided to help attain early time-to-revenue.

7.1 Production ramp-up

The production ramp-up stage of a silicon product is an iterative process consisting of test optimization, failure analysis, and design/fabrication optimization. The efficient collaboration of these procedures is critical for the attainment of the overall economic goal. Figure 7.1 illustrates a production ramp-up flow constructed based on the proposed techniques. This process traverses through two major circles, one being the yield debugging iteration, the other one being the quality enhancement cycle.

The yield debugging cycle necessitates the collaboration of test optimization, failure analysis, design optimization, and re-spin. Sample parts fabricated using various process corners are screened by ATE to collect the failure log data. In most typical cases, the yield is relatively low at the beginning stage. Instead

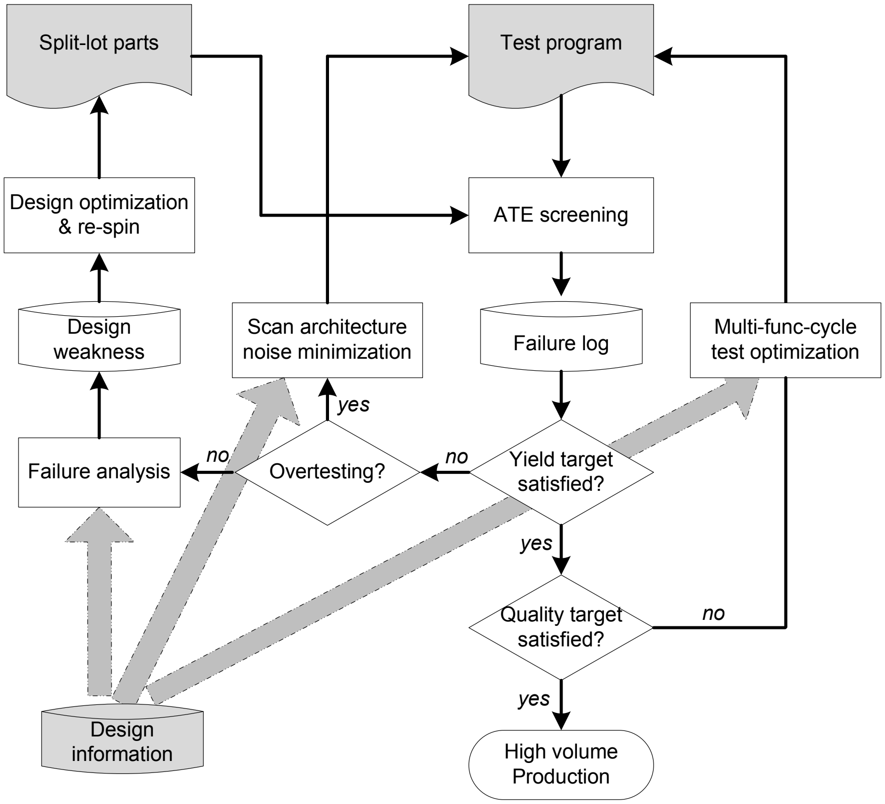


Figure 7.1: Production ramp-up flow

of directly looking into the expensive design re-spin, the most efficient strategy to enhance yield is to first examine whether there is an overtesting phenomenon occurring on the ATE. In case of high scan chain fall-outs being observed on the ATE, a scan noise minimization phase can be applied to the test patterns to reduce overtesting without impacting the test quality. If the test optimization step cannot lower the marginal failure rate down to the acceptable level, there is a high likelihood that certain weak points exist in the design. A failure analysis is thus performed to identify the design weakness and guide the design optimization process. After re-spin, new split-lot sample parts are sent to ATE for further yield debug. This iteration is continued until the yield target is fulfilled.

The test quality enhancement cycle is relatively simpler. The major procedure is to optimize the test program by developing the multi-functional-cycle delay test patterns based on the design information and the post-silicon verification results. This iterative process continues until the test quality target is satisfied.

It should be noted that the proposed techniques perform highly design-specific optimization strategies on different test phases. The clear differentiation between test-mode and functional circuitries in test development resolves the overtesting/undertesting dilemma faced by traditional approaches, thus guaranteeing the rapid convergence of the production ramp-up process.

7.2 Integration implementation

In addition to the high level flow, the integration of the proposed techniques also requires a set of implementation supports and provides a certain level of integration flexibilities. A discussion on various integration issues is provided in this section.

7.2.1 Yield and quality characterization

The efficient execution of the production ramp-up flow necessitates accurate yield and quality characterization to help determine the focus of optimization. The yield characterization can be performed by screening a relatively abundant set of parts from different split-lots. The test quality characterization, though, is a much more complex process. The DPPM number from the customers is the most accurate indicator of the test quality, whereas the collection of such information requires a long period, thus being impractical for time-critical production ramp-up. Therefore, various quality assurance models have been employed in industry for test quality and DPPM estimation. These models are typically constructed based on the test coverages of traditional fault models. Although being quite accurate in estimating the test escape rate due to fabrication defects, they are incapable of evaluating the impact of a design weakness on DPPM. Since this portion of test escape is design related, the statistical information regarding the functional operation is needed to account for the marginal failures during test quality characterization. One promising approach to attain such information consists of performing comprehensive functional verification on sample parts and analyzing the correlation between production test and functional verification results. The production test is required to not only meet the coverage targets for traditional fault models, but also catch all the marginal failures occurring in functional verification. A loose correlation is typically observed at the initial stage of the production ramp-up process. Nonetheless, along with the continuous tuning of the production test, a perfect correlation, with a zero functional failure rate in parts passing the production test, can be achieved, thus assuring a high test quality of the products.

7.2.2 Failure-adaptive test optimization

Higher efficiency can be attained by applying the proposed test optimization techniques in a failure-adaptive manner, as the failure data collected from ATE and functional verification provide insight about the failure distributions. Empirical silicon data have shown that the scan failures can often be test pattern dependent, resulting in only a subset of test patterns inducing scan fall-outs. By processing the ATE log, such test patterns can be easily identified, thus enabling the application of the proposed scan noise minimization technique solely to these patterns. This leaves increased test transformation flexibility for other optimization purposes such as test quality improvement, and saves engineering effort and time in reprogramming the test plan. The development of multi-functional-cycle delay test, on the other hand, can be guided by the functional verification information. If the functional verification results show that a certain circuit module experiences more marginal failures, it is a signal of a more severe functional-mode noise condition in this module. Therefore, more timing paths in this module can be added into the target set of multi-functional-cycle delay test generation, delivering a higher test quality without significantly increasing the test time. This iterative process can be continued until the test quality of all the modules meets the production requirement.

7.2.3 Re-spin verification

The production ramp-up process typically needs to pass through a few iterations of re-spins. Due to the high cost of re-spin, it is highly desirable to perform a verification before taping out the updated design. The proposed failure analysis technique pinpoints a design weakness, enabling a much more focused design optimization. Upon the completion of the design optimization, special examination is needed on the originally weak points. In contrast to the routine verification strategy of signing off the chip at standard PVT (process, voltage, and temperature) corners, one can check the timing correctness of the weak points at the worst-case voltage characterized on silicon. Although the standard cell library typically does not contain the delay information of such a voltage corner, timing analysis can still be performed through the delay extrapolation or interpolation that is supported by a number of commercial tools such as *PrimeTime*. Since the size of the design weak points is typically small, this approach enables the design verification for realistic operation condition at very minimal computational cost, thus ensuring a high level of confidence on the re-spin success.

Chapter 8

Conclusions

When devices scale to the nanometer range, the failing mechanisms of VLSI circuits end up being significantly shifted, with the noise-related marginal timing violations becoming the dominating failures. Such a significant change invalidates a variety of fundamental assumptions of the traditional test and failure analysis approaches, resulting in degraded test quality, reduced yield, and delayed production ramp-up. Among all the challenges raised by the new failure mechanisms, the difficulty in simultaneously attaining high test quality and yield under exceedingly strong power ground noise, and the ambiguous syndrome of marginal failures, constitute the most critical bottlenecks in the post-silicon test and debug stage of product development. The pressing industrial demand for maintaining high yield at low DPPM necessitates innovation in both test development and failure analysis methodologies to adapt to the substantial failure mechanism shift.

In order to resolve the aforementioned challenges, I propose, in this thesis, a comprehensive test and failure analysis framework, capable of precisely detecting and diagnosing the marginal failures under noise condition, thus expediting the time-to-revenue of the silicon products. This thesis work is developed based on a fundamental observation, that is, in contrast to the traditional focus on manufacturing defects, the marginal failures are most frequently the result of design weakness. Therefore, the test quality and the diagnosis accuracy can be significantly improved if the test development and failure analysis are guided by design insights. Various levels of design information, ranging from the behavioral functionality to the physical design, can be utilized to help understand the impact of noise on circuit timing and the root cause of the failures, thus enabling the development of fine-grained noise handling and failure detection strategies. The exploitation of design information, though, needs to be performed in the context of a structural test and failure analysis platform, in order to maintain a low test and debugging cost. Guided by these principles, a set of tightly coupled test optimization and silicon debugging techniques is developed in this thesis work, with a focus on the approximation of the functional operation in structural test and the failure hypothesis pruning using design knowledge.

To attain test quality and yield co-optimization, the production test plan needs to resolve the conflicts between overtesting and undertesting, the problem that traditional approaches fail to address successfully. The proposed test development technique resolves this challenge by differentiating the test-mode and functional-mode timing paths in the circuit using behavioral level design information. On the one hand, the marginal timing failures on test-mode paths have no impact on the functional operation while leading to unnecessary yield loss, and therefore need to be mitigated during testing. On the other hand, the functional paths of the circuit need to be examined under the worst-case noise condition in order to reduce test escapes. To fulfill both requirements, a scan noise minimization technique, combined with a multi-functional-cycle capture scheme, is proposed as the underpinning test optimization methodology. The scan noise minimization technique explores the flexibility in the scan compression phase, and identifies the noise-friendly compression seeds through a refined mathematical and algorithmic framework. The proposed capture scheme utilizes multi-cycle state transitions to minimize the number of non-functional paths being sensitized in the capture phase and develop the worst-case noise profile through noise accumulation. This guarantees the strict timing check on critical functional paths with the consideration of noise impact.

The proposed failure analysis technique focuses on identifying the root cause of marginal timing failures in scan architectures. The hybrid timing violation types in scan chains, compounded by their possibly intermittent manifestations, significantly increase the ambiguity and difficulty in diagnosis. Instead of relying on fault simulation that is incapable of approximating the intermittent fault manifestation, the proposed technique characterizes the impact of timing faults by analyzing the phase movement of scan patterns. Extracting fault-sensitive statistical features of phase movement information provides strong signals for the precise identification of fault locations and types. The manifestation probability of each fault is furthermore computed through a mathematical transformation framework which accurately models the behavior of multiple faults as a Markov chain. The identification of failing scan cells enables a further examination of the possible delay defects in the scan clock buffers, which ascertains the possible root causes of the observed scan chain failures. The proposed scheme characterizes the timing impact of the defective clock buffers by extracting the change in the delay distribution of the clock paths. The active use of the aforementioned design information enables the effective pruning of unrealistic fault hypotheses that would result in highly deviant timing behavior, thus precisely pinpointing the design weakness to guide the design optimization. This approach can significantly shorten the re-spin cycle and reduce the number of re-spin iterations in the production ramp-up process.

In addition to the theoretical contributions, this thesis also delivers a functional framework for seamlessly integrating the underpinning techniques into the current industrial infrastructure. The proposed test development approach maximally approximates the functional operation using a cost-effective scan test application scheme, thus being highly compatible with the regular scan architecture and ATPG flow. The proposed failure analysis technique collects volume data from standard ATE logs, thus enabling a highly automated diagnosis process. The proposed techniques neither require special DFT circuitry nor incur additional hardware overhead, imposing no difficulty in design and verification.

In sum, the successful application of the design-guided, noise-aware test development and failure analysis framework proposed in this thesis, I believe, will continuously deliver high yield and test quality, expedite design re-spin, and guarantee early time-to-market for VLSI circuits in current and future generations.

Bibliography

[1] *http://www-device.eecs.berkeley.edu/bsim3/bsim4intro.html.*

[2] *International Technology Roadmap for Semiconductors (ITRS) 2009 edition: Executive summary.*

[3] R. Adiga, G. Arpit, V. Singh, K. K. Saluja, and A. D. Singh. Modified T-flip-flop based scan cell for RAS. In *Proc. of European Test Symposium*, pages 113–118, 2010.

[4] F. A. Aloul and A. Sagahyroon. Estimation of the weighted maximum switching activity in combinational CMOS circuits. In *Proc. of International Symposium on Circuits and Systems*, 2006.

[5] K. Arabi, R. Saleh, and X. Meng. Power supply noise in SoCs: metrics, management, and measurement. *IEEE Design & Test of Computers*, 24(3):236–244, 2007.

[6] I. Bayraktaroglu and A. Orailoglu. Concurrent application of compaction and compression for test time and data volume reduction in scan designs. *IEEE Trans. Computers*, 52(11):1480–1489, 2003.

[7] R. Bhooshan and B. P. Rao. Optimum IR drop models for estimation of metal resource requirements for power distribution network. In *Proc. of VLSI-SoC*, pages 292–295, 2007.

[8] Y. Bonhomme, P. Girard, L. Guiller, C. Landrault, and S. Pravossoudovitch. A gated clock scheme for low power scan testing of logic ics or embedded cores. In *Proc. of Asian Test Symposium*, pages 253–258, 2001.

[9] K. M. Butler, J. Saxena, T. Fryars, and G. Hetherington. Minimizing power consumption in scan testing: pattern generation and DFT techniques. In *Proc. of International Test Conference*, pages 355–364, 2004.

[10] K. Chakravadhanula, V. Chickermane, B. Keller, P. Gallagher, and P. Narang. Capture power reduction using clock gating aware test generation. In *Proc. of International Test Conference*, pages 1–9, 2009.

[11] Y.-S. Chang, S. K. Gupta, and M. A. Breuer. Analysis of ground bounce in deep sub-micron circuits. In *Proc. of VLSI Test Symposium*, pages 110–116, 1997.

[12] H. H. Chen and D. D. Ling. Power supply noise analysis methodology for deep-submicron VLSI chip design. In *Proc. of Design Automation Conference*, pages 638–643, 1997.

[13] M. Chen and A. Orailoglu. Examining timing path robustness under wide-bandwidth power supply noise through multi-functional-cycle delay test. *submitted to IEEE Trans. on VLSI*.

[14] M. Chen and A. Orailoglu. On diagnosis of timing failures in scan architecture. *IEEE Trans. on CAD of Integrated Circuits and Systems*.

[15] M. Chen and A. Orailoglu. Scan power reduction for linear test compression schemes through seed selection. *IEEE Trans. on VLSI*.

[16] M. Chen and A. Orailoglu. Scan power reduction in linear test data compression scheme. In *Proc. of International Conference on Computer Aided Design*, pages 78–82, 2009.

[17] M. Chen and A. Orailoglu. Cost-effective IR-drop failure identification and yield recovery through a failure-adaptive test scheme. In *Proc. of Design Automation and Test in Europe*, pages 63–68, 2010.

[18] M. Chen and A. Orailoglu. Diagnosing scan chain timing faults through statistical feature analysis of scan images. In *Proc. of Design Automation and Test in Europe*, pages 185–190, 2011.

[19] M. Chen and A. Orailoglu. Diagnosing scan clock delay faults through statistical timing pruning. In L. Stok, N. D. Dutt, and S. Hassoun, editors, *Proc. of Design Automation Conference*, pages 423–428, 2011.

[20] S. Chun, T. Kim, Y. Kim, and S. Kang. An efficient scan chain diagnosis method using a new symbolic simulation. In *Proc. of VLSI Test Symposium*, pages 73–78, 2008.

[21] T. F. Coleman and Y. Li. An interior, trust region approach for nonlinear minimization subject to bounds. *SIAM Journal on Optimization*, 6:418–445, 1996.

[22] F. Corno, P. Prinetto, M. Rebaudengo, and M. S. Reorda. A test pattern generation methodology for low-power consumption. In *Proc. of VLSI Test Symposium*, pages 453–459, 1998.

[23] A. Crouch. Debugging and diagnosing scan chains. *Electronic Device Failure Analysis*, 7(1):16–24, 2005.

[24] D. Czysz, M. Kassab, X. Lin, G. Mrugalski, J. Rajski, and J. Tyszer. Low power scan shift and capture in the EDT environment. In *Proc. of International Test Conference*, 2008.

[25] V. Dabholkar, S. Chakravarty, I. Pomeranz, and S. M. Reddy. Techniques for minimizing power dissipation in scan and combinational circuits during test application. *IEEE Trans. on CAD of Integrated Circuits and Systems*, 17(12):1325–1333, 1998.

[26] K. De and A. Gunda. Failure analysis for full-scan circuits. In *Proc. of International Test Conference*, pages 636–645, 1995.

[27] V. R. Devanathan, C. P. Ravikumar, and V. Kamakoti. Variation-tolerant, power-safe pattern generation. *IEEE Design & Test of Computers*, 24(4):374–384, 2007.

[28] P. Du, X. Hu, S.-H. Weng, A. S. Arani, X. Chen, A. E. Engin, and C.-K. Cheng. Worst-case noise prediction with non-zero current transition times for early power distribution system verification. In *Proc. of International Symposium on Quality Electronic Design*, pages 624–631, 2010.

[29] S. Edirisooriya and G. Edirisooriya. Diagnosis of scan path failures. In *Proc. of VLSI Test Symposium*, pages 250–255, 1995.

[30] M. Elm, H.-J. Wunderlich, M. E. Imhof, C. G. Zoellin, J. Leenstra, and N. Maeding. Scan chain clustering for test power reduction. In *Proc. of Design Automation Conference*, pages 828–833. ACM, 2008.

[31] H. Furukawa, X. Wen, K. Miyase, Y. Yamato, S. Kajihara, P. Girard, L.-T. Wang, and M. Tehranipoor. CTX: A clock-gating-based test relaxation and x-filling scheme for reducing yield loss risk in at-speed scan testing. In *Proc. of Asian Test Symposium*, pages 397–402, 2008.

[32] S. Gerstendorfer and H.-J. Wunderlich. Minimized power consumption for scan-based BIST. *J. Electronic Testing*, 16(3):203–212, 2000.

[33] P. Girard. Survey of low-power testing of VLSI circuits. *IEEE Design & Test of Computers*, 19(3):80–90, 2002.

[34] R. Guo, Y. Huang, and W.-T. Cheng. A complete test set to diagnose scan chain failures. In *Proc. of International Test Conference*, pages 1–10, 2007.

[35] R. Guo, Y. Huang, and W.-T. Cheng. Fault dictionary based scan chain failure diagnosis. In *Proc. of Asian Test Symposium*, pages 45–52, 2007.

[36] R. Guo, L. Lai, Y. Huang, and W.-T. Cheng. Detection and diagnosis of static scan cell internal defect. In *Proc. of International Test Conference*, pages 1–10, 2008.

[37] R. Guo and S. Venkataraman. An algorithmic technique for diagnosis of faulty scan chains. *IEEE Trans. on CAD of Integrated Circuits and Systems*, 25(9):1861–1868, 2006.

[38] D. V. Hinkley. Inference about the change-point from cumulative sum tests. *Biometrika*, 58(3):509–523, 1971.

[39] D. V. Hinkley and E. Schechtman. Conditional bootstrap methods in the mean-shift model. *Biometrika*, 74(1):85–93, 1987.

[40] J. Hirase, N. Shindou, and K. Akahori. Scan chain diagnosis using IDDQ current measurement. In *Proc. of Asian Test Symposium*, pages 153–157, 1999.

[41] T.-C. Huang and K.-J. Lee. An input control technique for power reduction in scan circuits during test application. In *Proc. of Asian Test Symposium*, pages 315–320, 1999.

[42] Y. Huang. Dynamic learning based scan chain diagnosis. In *Proc. of Design Automation and Test in Europe*, pages 510–515, 2007.

[43] Y. Huang, W.-T. Cheng, C.-J. Hsieh, H.-Y. Tseng, A. Huang, and Y.-T. Hung. Intermittent scan chain fault diagnosis based on signal probability analysis. In *Proc. of Design Automation and Test in Europe*, pages 1072–1077, 2004.

[44] Y. Huang, W.-T. Cheng, S. M. Reddy, C.-J. Hsieh, and Y.-T. Hung. Statistical diagnosis for intermittent scan chain hold-time fault. In *Proc. of International Test Conference*, pages 319–328, 2003.

[45] Y. Huang and K. Gallie. Diagnosis of defects on scan enable and clock trees. In *Proc. of Design Automation and Test in Europe*, pages 436–437, 2006.

[46] Y. Huang, R. Guo, W.-T. Cheng, and J. C.-M. Li. Survey of scan chain diagnosis. *IEEE Design & Test of Computers*, 25(3):240–248, 2008.

[47] Y.-M. Jiang and K.-T. Cheng. Analysis of performance impact caused by power supply noise in deep submicron devices. In *Proc. of Design Automation Conference*, pages 760–765, 1999.

[48] Y.-L. Kao, W.-S. Chuang, and J. C.-M. Li. Jump simulation: a technique for fast and precise scan chain fault diagnosis. In *Proc. of International Test Conference*, pages 1–9, 2006.

[49] K. Khusyari, W. T. Ng, N. Jaarsma, R. Abraham, P. W. Ng, B. H. Ang, and C. H. Ong. Diagnosis of voltage dependent scan chain failure using VBUMP scan debug method. In *Proc. of Asian Test Symposium*, pages 271–274, 2008.

[50] A. Kokrady and C. P. Ravikumar. Static verification of test vectors for IR drop failure. In *Proc. of International Conference on Computer-Aided Design*, pages 760–764, 2003.

[51] A. Kokrady and C. P. Ravikumar. Fast, layout-aware validation of test-vectors for nanometer-related timing failures. In *Proc. of VLSI Design*, pages 597–, 2004.

[52] C. L. Kong and M. R. Islam. Diagnosis of multiple scan chain faults. In *Proc. of International Symposium on Testing and Failure Analysis*, pages 510–516, 2005.

[53] S. Kundu. On diagnosis of faults in a scan-chain. In *Proc. of VLSI Test Symposium*, pages 303–308, 1993.

[54] H. K. Lee and D. S. Ha. On the generation of test patterns for combinational circuits. In *Technical Report 12 93, Dept. of Electrical Eng., Virginia Polytechnic Inst. and State Univ.*, 1993.

[55] H. K. Lee and D. S. Ha. HOPE: an efficient parallel fault simulator for synchronous sequential circuits. *IEEE Trans. on CAD of Integrated Circuits and Systems*, 15(9):1048–1058, 1996.

[56] J. Lee, S. Narayan, M. Kapralos, and M. Tehranipoor. Layout-aware, IR-drop tolerant transition fault pattern generation. In *Proc. of Design Automation and Test in Europe*, pages 1172–1177, 2008.

[57] J.-Y. Lee, Y. Hu, R. Majumdar, and L. He. Simultaneous test pattern compaction, ordering and X-filling for testing power reduction. In *Proc. of International Symposium on Quality Electronic Design*, pages 702–707, 2009.

[58] K.-J. Lee, T.-C. Huang, and J.-J. Chen. Peak-power reduction for multiple-scan circuits during test application. In *Proc. of Asian Test Symposium*, pages 453–458, 2000.

[59] K. L. Lee, N. Z. Basturkmen, and S. Venkataraman. Diagnosis of scan clock failures. In *Proc. of VLSI Test Symposium*, pages 67–72, 2008.

[60] M.-S. M. Lee, K.-S. Lai, C.-L. Hsu, and C.-N. J. Liu. Dynamic IR drop estimation at gate level with standard library information. In *Proc. of International Symposium on Circuits and Systems*, pages 2606–2609, 2010.

[61] A. Leininger, M. Goessel, and P. Muhmenthaler. Diagnosis of scan-chains by use of a configurable signature register and error-correcting code. In *Proc. of Design Automation and Test in Europe*, pages 1302–1309, 2004.

[62] J. Li, X. Liu, Y. Zhang, Y. Hu, X. Li, and Q. Xu. On capture power-aware test data compression for scan-based testing. In *Proc. of International Conference on Computer-Aided Design*, pages 67–72, 2008.

[63] J. C.-M. Li. Diagnosis of multiple hold-time and setup-time faults in scan chains. *IEEE Trans. Computers*, 54(11):1467–1472, 2005.

[64] J. C.-M. Li. Diagnosis of single stuck-at faults and multiple timing faults in scan chains. *IEEE Trans. VLSI Syst.*, 13(6):708–718, 2005.

[65] J.-J. Liou, A. Krstic, Y.-M. Jiang, and K.-T. Cheng. Modeling, testing, and analysis for delay defects and noise effects in deep submicron devices. *IEEE Trans. on CAD of Integrated Circuits and Systems*, 22(6):756–769, 2003.

[66] X. Liu and Q. Xu. On simultaneous shift and capture-power reduction in linear decompressor-based test compression environment. In *Proc. of International Test Conference*, 2009.

[67] X. Liu, Y. Zhang, F. Yuan, and Q. Xu. Layout-aware pseudo-functional testing for critical paths considering power supply noise effects. In *Proc. of Design Automation and Test in Europe*, pages 1432–1437, 2010.

[68] J. Ma, J. Lee, and M. Tehranipoor. Layout-aware pattern generation for maximizing supply noise effects on critical paths. In *Proc. of VLSI Test Symposium*, pages 221–226, 2009.

[69] C. Metra, M. Omana, T. M. Mak, and S. Tam. Novel approach to clock fault testing for high performance microprocessors. In *Proc. of VLSI Test Symposium*, pages 441–446, 2007.

[70] K. Miyase, Y. Yamato, K. Noda, H. Ito, K. Hatayama, T. Aikyo, X. Wen, and S. Kajihara. A novel post-ATPG IR-drop reduction scheme for at-speed scan testing in broadcast-scan-based test compression environment. In *Proc. of International Conference on Computer-Aided Design*, pages 97–104, 2009.

[71] E. K. Moghaddam, J. Rajski, S. M. Reddy, and J. Janicki. Low test data volume low power at-speed delay tests using clock-gating. In *Proc. of Asian Test Symposium*, pages 267–272, 2011.

[72] E. K. Moghaddam, J. Rajski, S. M. Reddy, and M. Kassab. At-speed scan test with low switching activity. In *Proc. of VLSI Test Symposium*, pages 177–182, 2010.

[73] F. Motika, P. J. Nigh, and P. Song. *Stuck-at fault scan chain diagnostic method, US patent 7010735.* Patent and Trademark Office, 2006.

[74] S. Narayanan and A. Das. An efficient scheme to diagnose scan chains. In *Proc. of International Test Conference*, pages 704–713, 1997.

[75] N. Nicolici and B. M. Al-Hashimi. Multiple scan chains for power minimization during test application in sequential circuits. *IEEE Trans. Computers*, 51(6):721–734, 2002.

[76] M. Nourani and A. Radhakrishnan. Power-supply noise in SoCs: ATPG, estimation and control. In *Proc. of International Test Conference*, pages 507–516, 2005.

[77] M. Nourani, M. Tehranipoor, and N. Ahmed. Pattern generation and estimation for power supply noise analysis. In *Proc. of VLSI Test Symposium*, pages 439–444, 2005.

[78] H. Nyquist. Certain topics in telegraph transmission theory. *Trans. AIEE*, 47:617–644, 1928.

[79] P. Pant and J. Zelman. Understanding power supply droop during at-speed scan testing. In *Proc. of VLSI Test Symposium*, pages 227–232, 2009.

[80] I. Polian, A. Czutro, S. Kundu, and B. Becker. Power droop testing. *IEEE Design & Test of Computers*, 24(3):276–284, 2007.

[81] H. Qian, S. R. Nassif, and S. S. Sapatnekar. Early-stage power grid analysis for uncertain working modes. *IEEE Trans. on CAD of Integrated Circuits and Systems*, 24(5):676–682, 2005.

[82] R. Sankaralingam, R. R. Oruganti, and N. A. Touba. Adapting scan architecture for low power operation. In *Proc. of VLSI Test Symposium*, pages 35–40, 2000.

[83] R. Sankaralingam, R. R. Oruganti, and N. A. Touba. Static compaction techniques to control scan vector power dissipation. In *Proc. of VLSI Test Symposium*, pages 35–40, 2000.

[84] R. Sankaralingam, N. A. Touba, and B. Pouya. Reducing power dissipation during test using scan chain disable. In *Proc. of VLSI Test Symposium*, pages 319–325, 2001.

[85] J. Schafer, F. Policastri, and R. McNulty. Partner SRLs for improved shift register diagnostics. In *Proc. of VLSI Test Symposium*, pages 198–201, 1992.

[86] J. Schafer, F. Policastri, and R. McNulty. Diagnosis of scan chain failures. In *Proc. of International Symposium on Defect and Fault Tolerance in VLSI Systems*, pages 217–222, 1998.

[87] C. E. Shannon. Communication in the presence of noise. *Proc. Institute of Radio Engineers*, 37(1):10–21, 1949.

[88] O. Sinanoglu and A. Orailoglu. Modeling scan chain modifications for scan-in test power minimization. In *Proc. of International Test Conference*, pages 602–611, 2003.

[89] O. Sinanoglu and P. Schremmer. Scan chain hold-time violations: Can they be tolerated? *IEEE Trans. VLSI Syst.*, 17(6):815–826, 2009.

[90] P. Song, F. Stellari, A. J. Weger, and T. Xia. A novel scan chain diagnostics technique based on light emission from leakage current. In *Proc. of International Test Conference*, pages 140–147, 2004.

[91] K. Stanley. High accuracy flush-and-scan software diagnostic. *IEEE Design & Test of Computers*, 18(6):56–62, 2001.

[92] F. Stellari, P. Song, A. J. Weger, and T. Xia. Broken scan chain diagnostics based on time-integrated and time-dependent emission measurements. In *Proc. of International Symposium on Testing and Failure Analysis*, pages 52–57, 2004.

[93] G. Strang. *Linear Algebra and Its Applications.*

[94] R. C. Tekumulla and D. Lee. On identifying and bypassing faulty scan segments. In *Proc. of North Atlantic Test Workshop*, pages 134–143, 2007.

[95] C. Tirumurti, S. Kundu, S. Sur-Kolay, and Y.-S. Chang. A modeling approach for addressing power supply switching noise related failures of integrated circuit. In *Proc. of Design Automation and Test in Europe*, pages 1078–1083, 2004.

[96] N. Touba. Survey of test vector compression techniques. *IEEE Design & Test of Computers*, 19(4):294–303, 2006.

[97] J. D. Tubbs. A note on binary template matching. *Pattern Recognition*, 22(4):359–365, 1989.

[98] C.-W. Tzeng, J.-S. Yang, and S.-Y. Huang. A versatile paradigm for scan chain diagnosis of complex faults using signal processing techniques. *ACM Trans. Design Autom. Electr. Syst.*, 13(1), 2008.

[99] B. Wang, J. Rajaraman, K. Sobti, D. Losli, and J. Rearick. Structural tests of slave clock gating in low-power flip-flop. In *Proc. of VLSI Test Symposium*, pages 254–259, 2011.

[100] J. Wang, D. M. H. Walker, X. Lu, A. K. Majhi, B. Kruseman, G. Gronthoud, L. E. Villagra, P. J. A. M. van de Wiel, and S. Eichenberger. Modeling power supply noise in delay testing. *IEEE Design & Test of Computers*, 24(3):226–234, 2007.

[101] J. Wang, Z. Yue, X. Lu, W. Qiu, W. Shi, and D. M. H. Walker. A vector-based approach for power supply noise analysis. In *Proc. of International Test Conference*, pages 517–526, 2005.

[102] S.-J. Wang, Y.-T. Chen, and K. S.-M. Li. Low capture power test generation for launch-off-capture transition test based on don't-care filling. In *Proc. of International Symposium on Circuits and Systems*, pages 3683–3686, 2007.

[103] S.-J. Wang, K. S.-M. Li, S.-C. Chen, H.-Y. Shiu, and Y.-L. Chu. Scan-chain partition for high test-data compressibility and low shift power under routing constraint. *IEEE Trans. on CAD of Integrated Circuits and Systems*, 28(5):716–727, 2009.

[104] Z. Wang and K. Chakrabarty. Test data compression for IP embedded cores using selective encoding for scan slices. In *Proc. of International Test Conference*, pages 581–590, 2005.

[105] X. Wen, S. Kajihara, K. Miyase, T. Suzuki, K. K. Saluja, L.-T. Wang, K. S. Abdel-Hafez, and K. Kinoshita. A new ATPG method for efficient capture power reduction during scan testing. In *Proc. of VLSI Test Symposium*, pages 58–65, 2006.

[106] X. Wen, K. Miyase, T. Suzuki, S. Kajihara, Y. Ohsumi, and K. K. Saluja. Critical-path-aware X-filling for effective IR-drop reduction in at-speed scan testing. In *Proc. of Design Automation Conference*, pages 527–532, 2007.

[107] X. Wen, Y. Yamashita, S. Kajihara, L.-T. Wang, K. K. Saluja, and K. Kinoshita. On low-capture-power test generation for scan testing. In *Proc. of VLSI Test Symposium*, pages 265–270, 2005.

[108] L. Whetsel. Adapting scan architectures for low power operation. In *Proc. of International Test Conference*, pages 863–872, 2000.

[109] M.-F. Wu, J.-L. Huang, X. Wen, and K. Miyase. Reducing power supply noise in linear-decompressor-based test data compression environment. In *Proc. of International Test Conference*, 2008.

[110] D. Xiang, S. Gu, J.-G. Sun, and Y.-L. Wu. A cost-effective scan architecture for scan testing with non-scan test power and test application cost. In *Proc. of Design Automation Conference*, pages 744–747, 2003.

[111] D. Xiang, K. Li, J. Sun, and H. Fujiwara. Reconfigured scan forest for test application cost, test data volume, and test power reduction. *IEEE Trans. Computers*, 56(4):557–562, 2007.

[112] B. Yang, A. Sanghani, S. Sarangi, and C. Liu. A clock-gating based capture power droop reduction methodology for at-speed scan testing. In *Proc. of Design Automation and Test in Europe*, pages 197–203, 2011.

[113] J.-L. Yang and Q. Xu. State-sensitive X-filling scheme for scan capture power reduction. *IEEE Trans. on CAD of Integrated Circuits and Systems*, 27(7):1338–1343, 2008.

[114] J.-S. Yang and S.-Y. Huang. Quick scan chain diagnosis using signal profiling. In *Proc. of International Conference on Computer Design*, pages 157–160, 2005.

[115] K. Yang, K.-T. Cheng, and L.-C. Wang. TranGen: a SAT-based ATPG for path-oriented transition faults. In *Proc. of Asia and South Pacific Design Automation Conference*, pages 92–97, 2004.

[116] W. Zhao, J. Ma, M. Tehranipoor, and S. Chakravarty. Power-safe application of transition delay fault patterns considering current limit during wafer test. In *Proc. of Asian Test Symposium*, pages 301–306, 2010.