Modeling, Characterization and Compensation of Performance Variability using On-chip Monitor Circuits for Energy-efficient LSI

Islam A.K.M. Mahfuzul

To
my daughter and my wife

Acknowledgments

The completion of this thesis has been possible by the support from a lot of people. I would like to express my gratitude to everyone who actively and morally supported me in accomplishing this achievement.

My first and deepest gratitude goes to my supervisor Professor Hidetoshi Onodera, who is a role model to me. His great enthusiasm, professionalism, and kindness was the key for my development. Especially his passion for various works always motivated me to explore various problems. His critical way of thinking into a problem made me stronger in thinking. Besides his passion, I will never forget the kindness and patience he offered to me during my most difficult times both in terms of financial and academic. I am really thankful for the supports he provided me in spite of his busy schedule. I believe it is his warm guidance that made it possible to finish this thesis. I am grateful and honored to have him as my adviser. It was really enjoyable and exciting to have meeting with him and discuss various problems.

Next, I would like to express the deepest appreciation to Assistant Professor Akira Tsuchiya for his excellent support. He is an excellent researcher and independent thinker. Advice and comments given by him has been invaluable to me. His vast knowledge always surprises me. I learned from him how to look into a problem critically. He always provided a new dimension to my thinking. I would always discover my weaknesses whenever I discussed with him. I am intellectually indebted to him.

The work of this thesis required test chip design and fabrication. I would like to express my heartfelt gratitude to Professor Kazutoshi Kobayashi of Kyoto Institute of Technology for teaching me various aspects of chip design. The design flow created by him has been a great model for me to follow. It was really enjoyable designing a chip with him in the group. Besides, he provided insightful comments during my undergraduate and masters course which were extremely helpful in developing this thesis.

I am deeply grateful to Associate Professor Tohru Ishihara for giving insightful comments and suggestions. He would always ask questions from the system's perspective which were extremely helpful in my understanding of the scope and application of this work. I am particularly indebt to him for giving me the opportunity to design one of the test chips. His passion for research has been a source of my inspiration.

I owe my deepest gratitude to the thesis supervisors Professor Takashi Sato and Professor Takashi Matsuyama for their insightful and constructive comments which have been a great help to improve the quality of this thesis. I am deeply grateful to Professor Takashi Sato for

reading the thesis thoroughly and providing meticulous comments. Some of his papers have been helpful references to this work. I am particularly indebted to Professor Takashi Matsuyama for his constructive criticisms to this thesis. Those criticisms added a new dimension to my way of thinking. Because of his criticisms, I could look into my thesis from a different angle and improve the thesis.

I am particularly grateful to laboratory Secretary Seiko Jinno for her kind help and assistance during the whole time of my stay. I needed to complete various official procedures and documents for my scholarship and research funds. She always helped me to make the procedures as smooth as possible.

I owe a very important debt to my senior lab member Harukiho Terada who helped me in designing my first test chip. He also taught me the measurement procedures and other knowledge related to measurement. He is a wonderful person and I am blessed to have him as my senior guide in my first year at the lab.

I would like to offer my special thanks to my colleagues Shinichi Nishizawa, Takafumi Miki and Jun Furuta for their support and encouragement. Without their support this thesis would not have been possible. Special thanks to Shinichi Nishizawa for helping in measuring the test chips. He developed the measurement flow which we use in the lab. His enthusiasm, professionalism and willingness to help others encouraged me a lot. I am really blessed to have these wonderful people as my colleagues.

I would particularly like to thank Norihiro Kamae of this lab for his generous support to the test chip designs. We designed several circuits together where he made enormous contribution. Without his contribution this thesis would not have been possible. He gave me constructive comments and helped me debug my circuits which were invaluable to me. His knowledge on hardware and software is unbelievable. He has been an inspiration for me to learn more.

I would like to offer my special thanks to Shuichi Fujimoto with whom I worked together to develop on-chip monitor circuits. He is a very noble and nice person, and adds immense value to the group. He helped me in designing and measuring the test chips. Especially he developed the measurement flow of random telegraph noise which were of extreme help to me.

The chip design required for this work was supported by VDEC (VLSI Design and Education Center). I am deeply grateful to them for giving me the opportunity to design chips and providing various tools for simulation and debugging. I would particularly like to thank Professor Makoto Ikeda of the University of Tokyo for being tolerant and supportive during the design submissions. I would often mess up with the submission deadlines where he kindly guided us to finish submitting our designs.

I owe very important debts to Yuasa International Foundation (ユアサ国際教育学術交流財団) and Honjo International Scholarship Foundation (本庄国際奨学財団) for providing me scholarships during my masters and PhD course. I was facing a very hard time during that time. It was for their support and warm encouragement that helped me keep going.

This work was partly supported by JSPS (Japan Society for Promotion of Science) KAKENHI Grant Number 256432. I owe my deepest gratitude to JSPS for providing the research

fund to complete this work. I owe my deepest gratitude to MEXT (Ministry of Education, Culture, Sports, Science and Technology) for giving me the chance of coming to Japan and providing scholarship for the first 6 years of my Japanese and undergraduate studies.

My first three years of undergraduate study have been in ONCT (Oita National College of Technology) where I learned the basics of engineering. But, most importantly, all the staffs and friends there taught me Japanese culture and language very politely and nicely. Because of their kind support and help, I could master Japanese language and communicate with others in my lab. I would like to show my greatest appreciation to Professor Hirokazu Shimada and Professor Hidenobu Tsurusawa of ONCT for their encouragement and support during my entrance exam to Kyoto University. I would like to offer my special thanks to Associate Professor Yuji Maruki and Professor Teruko Aoki for their warm encouragement and support during my undergraduate study. I always take inspiration and energy by remembering these nice persons.

I would like to thank all my friends and well wishers who constantly gave me mental and intellectual support. My heartfelt appreciation goes to Kenshi Saho of Kyoto University who has been my guide when I was in ONCT. It was because of him that I could enter Kyoto University. His encouragement, suggestions and comments were invaluable to me. Discussions with him has been illuminating and exciting. Special thanks to Yohei Wakisaka for his nice friendship. We would often have discussions on various aspects of technology that would benefit human society.

Finally, I thank all for helping me in achieving this.

A.K.M. Mahfuzul Islam
in Kyoto
January 2014

Abstract

Modeling, Characterization and Compensation of Performance Variability using On-chip Monitor Circuits for Energy-efficient LSI

by

A.K.M. Mahfuzul Islam

Doctor of Philosophy in Informatics

Kyoto University

Professor Hidetoshi Onodera, Chair

This thesis targets energy-efficiency improvement of LSI by reducing variability effect with the help of on-chip monitor circuits. Design of area-efficient on-chip monitor circuits and its usage to compensate variability are discussed. The proposed techniques are all verified with real chip measurement.

LSI has become an integrated part of today's society. Energy-efficient and high reliability of LSI is required to sustain our society as well as meet today's high demand of computing power. Technology and supply voltage scaling have been effecting so far to improve LSI performance. However, due to aggressive technology scaling to create miniaturized devices, variability effect has now become the bottleneck to the advancement of LSI. Intrinsic transistor variability as well as environmental variability such as supply voltage result in large performance variability of LSI. The most severe effect of variability is the stop of supply voltage scaling. In the conventional worst-case design of LSI, large amount of margin is allocated to ensure correct operation. This margin prevents lowering of supply voltage as the amount of margin has become comparable to the required supply voltage. In order to continue supply voltage scaling further, collaboration between process and design technology is needed. This thesis presents a methodology on area and cost-efficient modeling and characterization of transistor variability in digital circuits using on-chip monitor circuits. A topology-reconfigurable monitor circuit is proposed for area-efficient implementation of monitor circuit. Post-silicon tuning technique of performance with adaptive body bias and its effect on energy-efficiency are then investigated. The methods presented in the thesis enables accurate estimation of LSI performance and increases energy-efficiency by incorporating post-silicon tuning into design phase.

First, modeling and characterization methodology of transistor variability is proposed for accurate estimation of LSI performance. Then on-chip digital monitor circuits suitable for parameter extraction are developed. Variability can be categorized into global and local variation. Global variation affects all the devices in the same way. Conventionally, large amount of margin is used to account for this global variation component which increases area, cost and energy. However, their effect can be compensated with on-chip monitoring and compensation techniques after fabrication, thus design margin can be eliminated to improve LSI performance by more than 50%. Local variation is mainly random, thus their effect needs to be accounted during design phase. Statistical and Monte Carlo analysis are performed to estimate LSI behavior under local variation. Local variation varies with device size, therefore trade-off exists between area, energy, reliability and yield. Accurate variation models are required to remove unnecessary margin which increases energy consumption. The amount of global and local variations also vary depending on device type. This thesis develops monitor circuit topologies for estimation of global and local variations for different transistor types. Monitor circuits particularly sensitive to either of the transistor type are developed which enables extraction of individual transistor parameters independently. Simulation and measurement results from a 65-nm test chip shows the validity of the extraction technique and the monitor circuits.

The conventional approach of implementing on-chip monitor circuits to capture both of the global and local variations require huge area, measurement- and implementation-cost. Depending on the characteristic of the target variation for monitoring, different topology of monitor circuit is needed. The number and size of monitor circuits differ as well. Utilizing the monitor circuit topologies developed in the thesis, a topology-reconfigurable monitor circuit architecture is developed for area- and cost-efficient implementation on on-chip monitor circuit. With the single instance of the proposed topology-reconfigurable monitor circuits, monitoring of different kind of variability becomes possible. This is achieved by reconfiguring the circuit topology so that the circuit behavior becomes sensitive to a specific variability source. Large number of samples is obtained by realizing different circuit topologies which enables statistical evaluation of local variation. The topology-configurability enables transistor-by-transistor variation characterization to provide accurate variation models. This can be used to correlate model and hardware to close the gap between estimation and real silicon behavior. The proposed monitor architecture is designed and implemented in a 65-nm process. Measurement results validate the proposed circuit.

Next, an area-efficient on-chip compensation technique using the proposed monitor circuits is developed to compensate global variation. On-chip monitor circuits are used to monitor transistor performance, and transistor body bias is tuned dynamically to adjust transistor performance to the intended value. This creates a feedback system in the chip to adjust transistor performance on the runtime. The small size of the proposed compensation circuit enables implementing it in a fine-grain level. Thus not only the whole chip, but various parts of the chip can be compensated independently to realize optimum operation of each part. The runtime compensation of variability eliminates the need of margin allocated during the design phase.

Thus, large improvement in circuit operation speed and energy can be achieved. The proposed technique is implemented in a 65-nm process. Test chips targeting various process conditions are measured and evaluated. Measurement results show the validity of runtime compensation based on on-chip monitoring.

A universal on-chip monitoring technique is developed which can be used as a bridge between hardware and software. Software level optimization thus becomes possible with the monitored variation information to improve system reliability and energy-efficiency. Various future directions can be derived from the proposed low cost and area-efficient on-chip monitor circuits.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis proposes on-chip monitor circuits by which modeling, characterization and compensation of LSI performance variability can be achieved with low design and implementation cost [1–7]. A universal on-chip monitor circuit is proposed which can be used to monitor device characteristics during runtime with extremely low area [8, 9]. This thesis shows that runtime tuning of LSI performance based on on-chip monitoring to improve energy-efficiency is feasible by measuring real chip performance [10, 11]. This chapter discusses the background and motivation of this research. Literature survey relating LSI design, circuit and architecture techniques to account for variation is presented. Particular focus is put on the role of on-chip monitor circuits.

1.1 Background

Large Scale Integration (LSI) circuits has become an integral part of today's modern information-based society. The role of LSI in our life has expanded from high performance computing to health-care applications. Almost every system now contains LSIs in numbers from several to hundreds. For example, a car has hundreds of LSI to compute power consumption, timing of break, etc. The evolution of internet and smart mobile devices would not have been possible without the tremendous improvement of LSI. In order to build an environment friendly modern society, energy reduction of LSIs has become the most important task. Reliability is another aspect that LSIs must provide to the users. For example, malfunction of LSI in a car may lead to severe accident. Thus, for sustainable and safe society, the following characteristics need to be realized in LSIs.

1. High performance
2. Low energy consumption
3. High reliability
4. Low cost

However, realizing LSIs with the above characteristics has become a huge challenge. In order to design systems that are both energy-efficient and reliable, new design and circuit techniques are required that can adapt to variations and environmental changes as well as user needs. Today data centers and servers consume a significant portion of the total energy consumption. Cooling systems required for high performance chips consume further energy. With the emergence of portable devices, energy consumption in LSI has increased drastically and is expected to increase in the future with high speed internet and video streaming becoming available everywhere. With the drastic increase of energy consuming devices, the term energy-efficient LSI has become synonymous with high performance LSI. Energy-efficient LSI is highly demanded today to realize sustainable and green society. From the user's point of view, energy-efficiency of LSI relates directly to the battery life of a portable device. There are two possible ways to increase the battery life of a device. One is to increase the battery capacity. The other is to reduce the energy consumption. The first option depends on battery technology which has very slow growth compare to the LSI. LSI design paradigm has thus shifted from performance oriented design to energy oriented design. However, in order to support the modern society, high computing capacity is also demanded. Users want their devices to accomplish their tasks as fast as possible. On the other hand, they want to drive their devices as long as possible without connecting to the power supply. Energy-efficient design of LSI is now highly demanded to reduce the total energy consumption while meeting the computing need of the users. It is a huge challenge to design LSI that can give both high performance and lower energy consumption.

The need for higher computing capacity and lower power consumption has been realized by technology scaling till now [12]. Gordon Moore observed the trend of technology scaling and realized the potential of LSI advancement in the coming years. His famous prediction made in 1965 then later came to know as the Moore's law [13]. Moore predicted that the number of transistors on LSI doubles approximately every two years. His prediction has been held true till today. The key feature of technology scaling that made possible both faster and less power consuming LSI is that scaling down the transistor dimensions and supply voltage by a factor k reduces gate delay and power by the same factor. As the technology enters sub-90nm era, scaling of several dimensions has approached the manufacturing limits. The scaling of the gate oxide thickness, for example, has been stopped [14]. The scaling of horizontal dimension is facing its difficulty too because of various physical phenomena such as short channel effect. With several technology innovations, the horizontal scaling is still being continued. However, this aggressive scaling has brought a new problem that is the variation between the manufacturing parts. MOSFET channel length of 22 nm is already being used for commercial products. However, the wavelength of light used in photolithography to pattern millions of MOSFETs onto a chip is still 193 nm. Various innovations such as double patterning are required to print features below 32 nm as the theoretical limitation of minimum gate length that can patterned with 193 nm light is 40 nm [15]. Although device dimensions are rather well-controlled, what varies are, doping locations, grain boundaries, etc. The device parameters can no longer be controlled by the manufacturing process precisely. As a result, difference in performances between

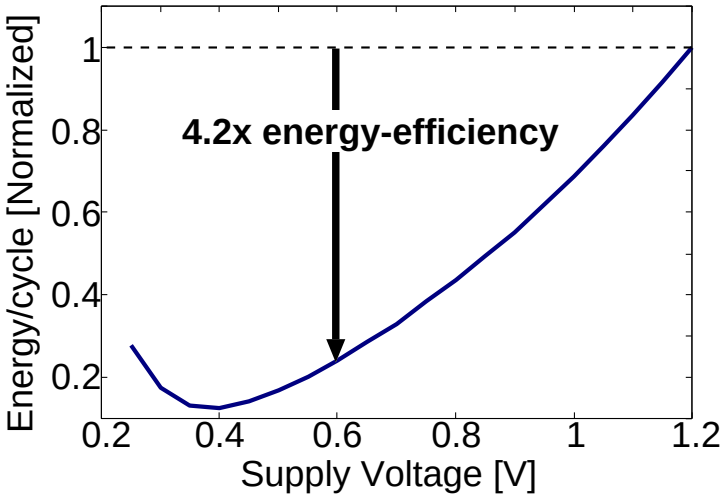


Figure 1.1: Energy per cycle against the supply voltage simulated in a commercial 65-nm process. Energy reaches to the minimum near 0.4 V. Operating at 0.6 V instead of the nominal 1.2 V yield 4 times more energy efficiency.

two transistors has become significant that it is affecting the performance of whole LSI. Among the many problems that this variation causes, the one that is threatening the LSI advancement is the prevention of supply voltage scaling. Supply voltage scaling is stopped since the 90-nm technology because of the random variation in neighboring device performances [16]. Reference [17] predicted the end of supply voltage scaling two decades ago because of the variation in transistor threshold voltage. After 90-nm technology, supply voltage has remained the same at 1.0 V as predicted. Scaling both physical dimensions and power supply simultaneously is required to achieve lower power consumption while maintaining higher operation speed. The stop of supply voltage scaling results in higher power consumption with every new technology node, causing higher temperature in the chips which in turn degrades performance and reliability. Power consumption has now become the limiting factor of today's LSI. Even if we are able to integrate multiple processing units inside a chip, we are not able to operate them simultaneously because of high temperature. Thus, we are now facing a dark silicon era where some parts of the chip have to be powered down [18]. Transistor variability has become the bottleneck for improving LSI performance in scaled technology nodes. Since 32-nm and beyond, new process technology, such as FinFETs [19] and SOI MOSFETs [20] are considered to be the alternatives for the conventional bulk MOSFET. Undoped FinFETs and SOI MOSFETs promise less variability than the bulk counterpart. However, various sources of variability, such as LER (Line Edge Roughness) and metal gate granularity contribute to large variability [21]. Thus, mitigation of device variability with collaboration between design, circuit and system design will be the key for reducing energy consumption while maintaining high computing capacity.

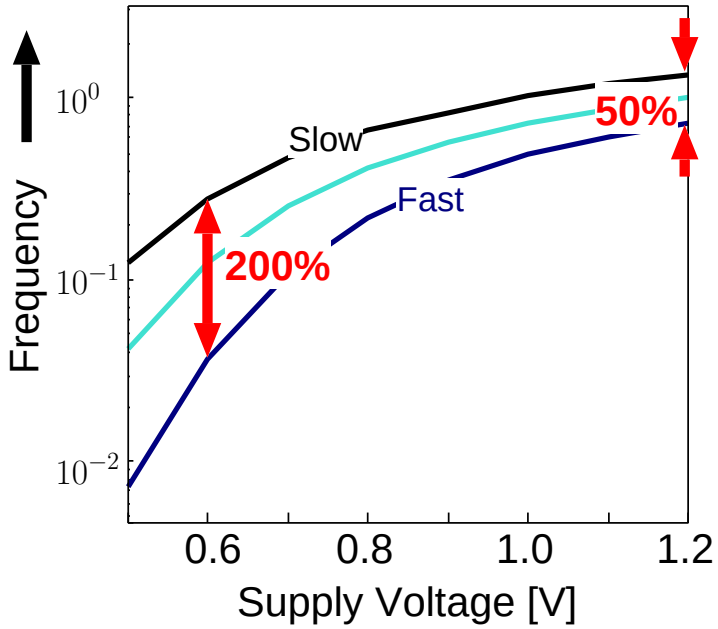


Figure 1.2: Frequency variation between three different process scenarios of fast, nominal and slow. 200% of frequency variation is observed at 0.6 V operation.

1.2 Motivation

The end of supply voltage scaling has pushed the circuit designers to find for new solutions to reduce power consumption. The old way of defining system performance was to use the operating clock frequency as the performance indicator. As power consumption has become a serious issue, the designers have started to look at the system from a different perspective using energy as the key parameter. System throughput instead of pure clock frequency and energy per throughput are the modern specifications of any device. Lowering the supply voltage towards the transistor threshold voltage yields better energy per operation profile. Figure 1.1 plots the change of consumed energy per operation against supply voltage obtained by circuit simulation of a test circuit in a 65-nm process. Reducing the supply voltage to half of the nominal voltage (1.2 V) yields 4-times improvement in energy-efficiency. Energy per operation reaches a minimum point at around 0.4 V which is close to transistor threshold voltage. Because the energy-efficiency is maximum at around the transistor threshold voltage, operating the circuit with supply voltage near the threshold voltage has been termed Near-Threshold Voltage (NTV) operation [22]. Near-threshold voltage operation is considered as one of the key technologies for continuing the technology scaling.

Designing circuits for NTV is challenging as the gate overdrive of transistors that is the difference between gate voltage and threshold voltage is almost zero. This low gate overdrive makes the circuit unstable to environmental changes as well as device characteristic variability. One of the biggest challenges is to deal with large variation in performance. Figure 1.2 shows the operating frequency of a typical LSI for the fastest chip and the slowest chip along with a nominal performance chip for a 65-nm process. At 1.2 V operation, the difference between

the fastest and the slowest is 50% whereas the difference increases to 200% at 0.6 V operation. Thus, the effect of transistor and environmental variability on LSI performance increases drastically with the lowering of supply voltage. In order to ensure reliable circuit operation, large amounts of margins or guard-bands are thus allocated which is very energy-costly.

The International Technology Roadmap for Semiconductors (ITRS) [14] highlights performance variability and reliability management in the next decade as a red brick (i.e., a problem with no known solutions) for design of computing hardware. There are many reports in the literature showing that variation is causing big problem. Intel has reported frequency and power profiles of an 80-core TeraFLOPS processor implemented in a 65-nm process [23]. 28% variation at 1.2 V and 62% variation at 0.8 V between the fastest and slowest cores are observed. Large frequency variation occurs even at the nominal voltage. The effect of device characteristic variation increases drastically with the lowering of supply voltage. The amount of random variation for a 65-nm process is reported to be 190 mV at 5σ level. As we are trying to lower the supply voltage as much as possible, the large amount of random variation will cause circuit failure at low supply voltages. Random variation in transistors cause serious threat to circuit failure and increases the minimum supply voltage (V_{ddmin}) required for correct operation of logic and memory units. In Ref. [24], around 88 mV of increase in V_{ddmin} is reported for a 101-stage inverter chain. The effect of random variation is severe for memory elements such as FFs (Flip-Flop) and SRAMs. Reference [25] reports that V_{ddmin} of FFs prevents the operation at the optimum supply voltage as V_{ddmin} voltage of FFs is much higher than the optimum supply voltage. As a result, the system power supply is set pessimistically to a much higher value to ensure correct operation of all the parts. The amount of margin needs to be reduced for energy-efficiency and on-chip monitor circuits providing variation information can be of extreme helpful to estimate adequate design margin.

Besides the random variation in devices, there are other variation components that affect the circuit performance. Chip temperature can increase to as high as 120° C [26]. Increase in temperature degrade circuit operation speed and increases leakage power. According to ITRS, supply voltage fluctuation is considered to be ±10%. Sudden drop of supply voltage may cause critical timing failure causing system malfunctioning. As shown in Fig. 1.2, some chip can be slow and some chip can be fast. However, fast chips tend to be leaky thus have large energy consumption. The designers thus face a huge challenge to meet both the delay and power constraints as the circuit need to operate correctly under all of the variation scenarios. Device characteristics also degrade over time. Device phenomena such as Negative Bias Temperature Instability (NBTI) is reported to cause 10% of delay degradation in digital circuits for 70-nm process over 10 years [27]. NBTI can cause malfunction in memories [28]. Therefore, during the design phase, digital circuits need to be designed considering all the worst possible cases to make sure it operates at all the conditions. Designing the circuit for the worst possible scenario is energy inefficient as it increases area, power and cost. After the production of the chip, most of the chips may have the nominal or fast operating condition. The chip may face extreme worst-case scenario once in several years. However, as the chip is already designed considering

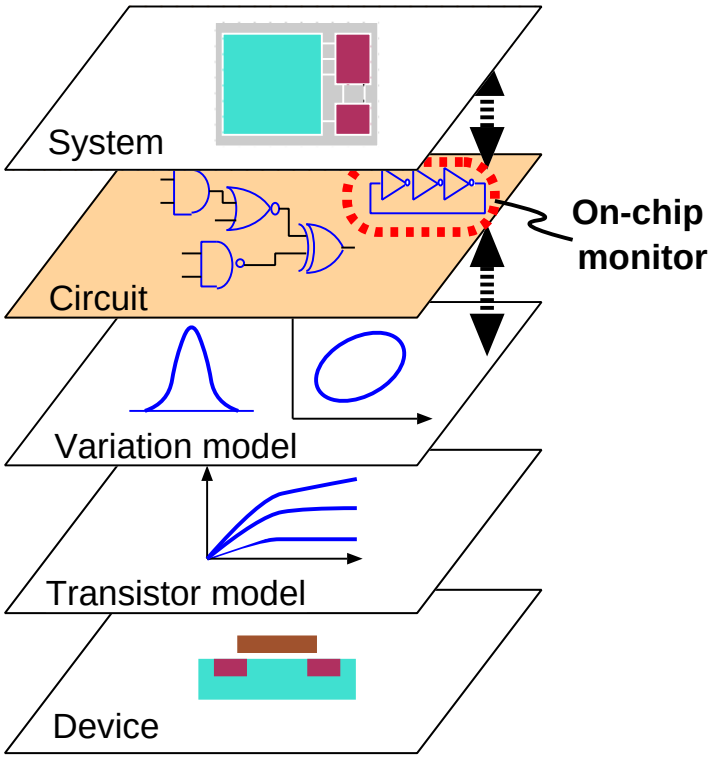


Figure 1.3: On-chip monitor circuits enabling fast characterization of transistor variability and providing hardware information to the system.

the worst-case scenario, we are not able to benefit from the technology scaling. Instead, energy consumption increases as the circuit is not optimized for fast or nominal conditions. Thus, conventional worst-case based design is way too inefficient and new design paradigm incorporating on-chip monitor circuits which can adapt to various changes have become a necessity today.

The state-of-the-art of LSI design is to design the circuits considering fixed threshold voltages of devices. However, due to large variations in environmental parameters such as supply voltage, temperature, circuit activity, dynamic tuning of threshold voltage has become a necessity for energy-efficient operation. Transistor body bias gives the designers an option to tune the threshold voltages during run time. However, without knowing the hardware profile that is the transistor performance, tuning of threshold voltages may not result in energy reduction. Thus, on-chip sensors giving information on the hardware come to play an important role. Effective interface between hardware and software is possible with various kinds of on-chip sensors by which software controlled optimization becomes possible. The future LSI chip will require lots of sensors to monitor transistor performance, temperature, supply voltage etc. not only for energy-efficient and reliable operation but also for cost-effective testing and debugging of LSI.

Figure 1.3 shows a typical design hierarchy of a system on chip. First, transistor models for the target process technology node are given to the circuit designers. This transistor models contain various statistical parameters to simulate the effect of variation on circuit performance. Usually large guardbands are used in this models. Variation models are created based on the manufacturing uncertainties as well as layout of the circuit. To cover all the uncertainties and

guarantee circuit operation, the models tend to be pessimistic. The job of the designers is to design their circuits so that the circuit can operate at all the possible variation scenarios provided by the models. As a result, the circuits tend to be over-designed which result in excessive energy consumption. From the system perspective, the circuits need to operate at various supply voltage and frequency conditions while ensuring correct operation. Another layer of pessimism becomes involved during the choosing of adequate supply voltage for a target operation frequency. Finally, the operating system (OS) also pessimistically assign tasks and wait for the worst-case delay on the transitions between different system operating conditions. So, large amount of energy loss occurs at each layer of the design hierarchy. Combining, the whole system consumes much higher energy consumption which is just for ensuring the correct operation to provide reliable service to the users. Various sensors to communicate between the layers can play a major role in reducing these energy losses drastically. The sensors provide real time information on the hardware which can be used to set the parameters optimally for reliable operation with adequate margin. On-chip monitor circuits can close the gap between real silicon behavior and variation models. On-chip monitor circuits can also provide a mean to the designers to gain more control on their circuit behavior by dynamically controlling the behavior through transistor body bias. Thus, large energy reduction can be possible as the excessive design margins can then be eliminated.

From the above discussions, the past trend of using smaller transistors to achieve higher operating frequency is coming to an end. The new era of LSI scaling is a system-on-a-chip (SoC) approach that combines a diverse set of components using adaptive circuits, integrated sensors, sophisticated power-management techniques, and increased parallelism to build products that are many-core, multi-core, and multi-function [29]. The ability to adapt to the changes in environment and performance as well as self-healing and self-diagnosis mechanism will give us the full benefit of technology scaling. Various tuning mechanism and on-chip sensors are needed to realize flexible circuits that has the ability to adapt. Thus, the future direction SoC design must have capabilities of post-silicon self-healing, self-configuration and error correction. Sensors and various adaptive techniques allowing the system to adapt to variations will play a key role in future SoC design. Among the many sensors, on-chip monitor circuits that provide various information on device and circuit characteristics are of extreme importance today. Effective use of on-chip monitor circuits will play a major role in continuing the advancement of LSI.

1.3 Related Works

1.3.1 Variation-aware Techniques for Energy-efficiency Improvement

Today's LSI design is based on several hierarchies and abstractions as shown in Fig. 1.3. Energy-efficient LSI involves accuracy in presenting the lower hierarchy to the upper. The more we go up the hierarchy, the more amount of energy saving can be achieved with various techniques. So, system and architecture design must consider energy saving as the top most

priority. As a result, it has become the standard to have various kinds of power management techniques in the chip [30, 31].

The operating system (OS) plays a key role in a computing system, especially in real-time systems. However, the conventional way of designing computing systems where the hardware is considered error-free is not valid anymore. As variation in various parts has become dominant, variation-aware design is a must today. Variation can be handled at various layers starting from transistor level to OS level. Ref. [32] proposes a system-level solution that exploits memory power variation through physical address zoning. In order to implement their method, data on variability and hardware support are required. Ref. [33] proposes variation-aware algorithms for application scheduling and power management for multiprocessors to handle process variation. As multi-core and many-core architectures are becoming the standard today, energy-efficient utilization of these resources need cooperation between software and hardware. Variation-aware techniques are required for optimal core allocation [23].

At the system level, large amount of energy can be saved by optimizing system tasks and choosing the suitable supply voltage or operating frequency. DVFS (Dynamic Voltage and Frequency Scaling) is proposed where the supply voltage and operating frequency is varied depending on the workload [34, 35]. DVFS can save large amount of energy for applications where high workload occurs rarely.

Variation can be compensated after the production of a chip with techniques such as adaptive body bias [36–41]. Transistor body bias is used for post-silicon tuning of transistor threshold voltage. Further improvement of energy-efficiency can be obtained for DVFS like architecture by integrating adaptive body bias technique [40, 42, 43]. However, due to large variation, energy-efficient mapping between supply voltage and operating frequency has become a problem. Therefore, various versions of DVFS have been proposed to address variation problem [23, 43–46]. On-chip monitor circuits are required to implement these techniques for higher energy-efficiency.

1.3.2 On-chip Circuits for Performance Monitoring

Ref. [47] uses an on-chip leakage monitor circuit to scan optimal reverse bias voltage for adaptive body-bias circuit. With the increase of variability, deviation between simulated path delay and actual path delay in a manufactured chip become more significant. Therefore, run-time monitoring of timing has become a necessity. Ideally, monitoring of every path in the chip would ensure correct operation but that would require huge area. Several versions error detection circuits to monitor runtime timing failure are reported [48–51]. As a typical LSI contains millions of delay paths, implementing error detection circuitry for each path is not realistic. Conventionally, several critical paths are detected during the design phase and error detection circuits are implemented for those critical paths. However, at low voltage operation where the variability impact increases by multiple times, any path has the potential to become critical. A critical path is the path that has to maximum delay thus determining the maximum operating frequency. Besides, error recovery mechanisms are required for these kind of error

detection circuits which need extra design effort. Many of these techniques are only applicable for processor-like architectures where the system can relay the operation. Many of the general purpose ASICs (Application Specific Integrated Circuit) do not have that function thus detection circuits are not applicable in those circuits. Another critical problem for this approach is that a sudden critical path may be activated depending on the input vector of the circuit. The extra circuitry needed for error detection and recovery consumes large power too which is not desirable.

Instead of monitoring runtime timing error from real paths, representatives of critical paths can be placed in the chip and then monitored to find the maximum delay [52–56]. This approach has small area and power overhead compared to error detection circuits, however ensuring correlation between real paths is the problem. Ref. [52] proposes a distributed critical path timing monitor. It consists of several delay paths. The delay paths consists of different wire loads, pass gates etc. The comparator selects the slowest path thus the operating frequency can be set according to the slowest timing path. Ref. [56] proposes a method to synthesize a stand-alone circuit to represent the aging of critical reliability paths, which are defined as paths that can potentially become critical at some point in time due to aging. Ref. [55] proposes critical path monitor structure which uses several critical path replicas. The outputs of critical path replicas are connected to c-element which detects the slowest path automatically. Ref. [53] proposes path-based ring oscillator which is created from a targeted path. Ref. [54] proposes a methodology to synthesize a representative critical path for post-silicon delay prediction.

Area-efficient on-chip monitoring of power is also required for energy-efficient power management. IBM POWER7 chip uses adaptive energy management systems [30]. Ref. [57] uses activity counters to use as power proxies for runtime monitoring of processor power consumption. On-chip thermal management is proposed to manage temperature and power [58, 59] where temperature sensors are required. Various kinds of temperature sensors are reported [60–63]. Thus, effective use of various kinds of on-chip sensors is essential for energy management of present and future SoC.

1.3.3 On-chip Circuits for Variability Monitoring

Normally, process monitors are placed on the subscribe lines to track process characteristics. However, because of the lack of sufficient numbers, process monitors fail to give us information of variations in detail. In order to measure various variation information, large number of samples need to be measured. The most basic method of characterizing transistor variability is to measure *I*-*V* characteristics of device arrays [64–67]. These methods give us detailed information on the MOSFET characteristics. However, measuring, post-processing and analyzing I–V data are time consuming and relate directly to product cost.

Several post-silicon applications such as post-silicon tuning [68], timing characterization [69] and reliability analysis [70] require accurate variation models. Various methods are reported for fast characterization of variation providing accurate variation models [69, 71–74]. Ref. [71] proposes high speed test structures for in-line process monitoring and model calibra-

tion. Ref. [69] proposes product-representative "at speed" test structures for CMOS characterization. A ring oscillator based test structure for NBTI analysis [72]. Ref. [73] proposes a method to fast characterize MOSFET threshold voltage variation. The method uses one point measurement with respect to a pre-characterized device enabling large improvement in measurement time. Ref. [74] proposes a test structure for the measurement and characterization of layout-induced transistor variation. These methods are useful in providing accurate variation models. However, simple on-chip circuits that can be embedded onto the product chips will provide more accurate information and enable runtime monitoring.

Several on-chip monitor circuits for runtime monitoring of NBTI degradation are reported [72, 75–78]. Ref. [79] proposes on-chip test structure and digital measurement method for characterization of local random variation. Ref. [80] proposes statistical characterization and on-chip measurement methods for local random variability of a process using sense-amplifier-based test structure. Ref. [81] proposes an operational amplifier based monitor circuit for transistor threshold voltage variation. Ref. [82] proposes a method to measure the effects of process variations on circuit performance by means of digitally-controllable ring oscillators. Ref. [83] proposes a standard-cell based on-chip MOSFET performance monitor. On-chip monitor circuits can be used to extract timing information as well. Ref. [84] proposes an on-Chip structure for measuring timing uncertainty functional and test operations. Ref. [85] has shown timing information extraction from oscillation-based test structure.

The need for on-chip monitor circuits that are area and cost-effective, easy to implement, while efficiently providing detailed variation information is increasing. Conventionally, huge efforts are needed to implement various types of on-chip monitor circuits. Area and cost-effective design of monitor circuits is the key. Ability to extract various variation information is another aspect that the monitor circuits must have in order to build accurate variation models. Runtime monitoring capability is also required for runtime tuning of performance.

1.4 Research Goal and Thesis Contribution

Among the many sensors required in a system on a chip, area and energy-efficient monitoring of transistor performance and its variation are key challenges today. The following characteristics of on-chip monitor circuits are required for future LSI.

Digital Digital in nature is important for cost-efficient design and implementation. The monitor circuit need to communicate with other parts, thus the output of the monitor need to be digitized.

Ability to monitor different MOSFETs CMOS circuit consists of pMOSFET network for pull-up and nMOSFET network for pull-down. The overall LSI performance thus consists of pMOSFET and nMOSFET performance. In order to estimate LSI performance accurately, independent variation models for pMOSFET and nMOSFET are required. Independent monitoring of pMOSFET and nMOSFET performance will allow optimum

tuning of LSI performance.

Ability to monitor several variation types LSI performance is suffered from variations of different kinds. Some variations are location-correlated. Some are systematic while some are random. All these different kinds of variability need to be monitored on-chip.

Area-efficiency Area efficiency is an important parameter for fine-grain and distributed implementation on monitor circuits in the chip. Fine-grain implementation will provide accurate models for variations that differ from location to location. Area-efficiency will also reduce cost.

State-of-the-art on-chip monitoring techniques fail to provide all of the above requirements. The goal of this research is to develop universal on-chip monitor circuit and develop guidelines on the usage of on-chip monitor circuits during the design phase as well as in the post-silicon tuning of system parameters. The following advantages can be achieved.

1. Reduce design margin in each layer of design hierarchy by eliminating pessimism
2. Tune system parameters based on the actual hardware profile
3. Provide information for silicon debugging and timing analysis

These can be achieved by the following ways.

- Cost-effective modeling and characterization of variation using on-chip monitor circuits.
- Direct insight into the process information by measuring on-chip monitor circuits. This will help the designers to set adequate supply voltage and operating frequency. This information can also be used for test pattern generation and debugging the silicon.
- Adjust transistor performances to the target values based on actual silicon behavior of MOSFETS. This eliminates the need for worst-case design which will reduce both the area and energy drastically.
- Tune circuit operating condition based on circuit parameters such as temperature, activity, etc. On-chip circuits will provide an interface between hardware and software to effectively optimize the system.

In order to do the above, a methodology on modeling and characterization of MOSFET variation using on-chip monitor circuits is developed. A universal topology-reconfigurable monitor circuit is developed by which area-efficient monitoring of different kinds of variations becomes possible. The small area of the proposed circuit makes the circuit suitable to distribute it across the chip for fine-grain monitoring. The circuit can be used to provide interface between hardware and system as well was model and hardware (Fig. 1.3). An digital runtime performance compensation technique using the proposed on-chip monitor circuits is developed for reducing design margins and improve energy-efficiency. Combining the above achievements, LSI energy-efficiency is expected to increase by multiple times, and LSI manufacturing cost is expected to decrease drastically.

1.5 Thesis Organization

Chapter 2 describes the characteristics of variability that a typical LSI faces today. LSI design methodology and the effect of variation on LSI design are then discussed. Variability effect on LSI performance and various variation-aware design techniques to mitigate variation effect are explained. The need for accurate variation model and on-chip monitor circuits are discussed in the chapter.

Chapter 3 describes a methodology on transistor variability modeling and characterization using on-chip monitor circuits. Parameter estimation techniques for both global and local random variations are proposed. Monitor circuit topologies suitable for parameter estimation are explored and monitor circuits suitable for the estimation are proposed. The proposed circuits are validated using measurement results from test chips fabricated in a 65-nm process. Various corner chips are measured which confirms the validness of the proposed circuits and their applications. Successful extraction of threshold voltage and gate length variation for global and local variation has been performed.

Chapter 4 presents a topology-reconfigurable universal monitor circuit for cost- and area-efficient implementation of on-chip monitor circuit. The proposed monitor has small area thus suitable for distributed implementation onto the chip. Circuit topology and its mechanism along with measurement results from a 65-nm process test chip are discussed here. With a single instance of the proposed universal monitor circuit, measurement and characterization of several kinds of variability have been performed.

Chapter 5 shows a runtime compensation mechanism of LSI performance based on on-chip monitor circuits. A simple and digital built-in self-adjustment scheme of MOSFET threshold voltage is developed. Measurement results from a 65-nm test chip are discussed here. Feasibility of runtime performance compensation is demonstrated.

Chapter 6 summarizes key contributions of the thesis and shows some future guidelines for further improvement of energy-efficiency incorporating on-chip monitor circuits.

Chapter 2

Variability Impact on LSI

In this chapter, the impact of variability on LSI performance will be described. First, various kinds of variability and their characterization method will be described. Then, the present LSI design methodology will be discussed. Particular emphasis will be put on how variability effect is considered during the design phase and how variability increases delay, power and cost. Finally, several design techniques will be explained to mitigate variability effect.

2.1 Characteristics of Variability

Variation is the difference in behaviors between two identical devices. A typical manufacturing process includes many different steps to produce transistors, poly gates, metal interconnections etc. Variation rises from each of these steps. This results in different types of variability with different effects on circuit performance. Categorizing variability is therefore needed to understand variability and their effects on design. Different strategies are needed to deal with different kind of variations. This section discusses on several categories of variations and their effect on circuit.

2.1.1 Static Variation

Static variability are those that occurs during fabrication. These variations do not change over time, hence the term static is used to express them. Variation due to manufacturing process can be divided into Lot-to-Lot (L2L), Wafer-to-Wafer (W2W), Die-to-Die (D2D) and Within-Die (WID) variation [86]. L2L variation is the difference of device characteristics between two lots of silicon. W2W variation is the difference of device characteristics between two wafers of the same lot. L2L, W2W, and D2D variation results in global variation of a die. Therefore, these variations are combined into D2D variation. D2D variation is also called as global or inter-die variation. D2D variation effects all the devices in a die in the same way. One key characteristic of D2D variation is that the amount of variation is same regardless of the device size. For example, two nMOSFETs of 240-nm and 480-nm may suffer from a 40-mV shift in threshold voltage. Because of these characteristic, D2D variation is dealt by analyzing a circuit

considering several extreme process corners. Another characteristic is that different device types may suffer from different amount of D2D variations. For example, pMOSFETs may become faster whereas nMOSFET may become slower. Circuit behavior differs largely depending on the correlation between different device types.

WID variability are those that affect each device within a die. WID variability can be either systematic or random. Systematic WID variability involves location-correlation variation. This kind of variation occurs from the differences in layout pattern and density. Surrounding layout pattern may affect a device's characteristics. Other systematic variation involves gradual slope in device parameters within a die. This kind of variation may occur during the thermal annealing process [87]. The random component is caused from intrinsic atomic level differences thus even two identical devices placed adjacent to each other show different characteristics. WID variability is also called as intra-die or local variation. As the technology scaling continues, WID variation is becoming more significant [88]. In case of large chips, location-correlated systematic component can be as significant as D2D variation as reported in [23]. WID random variation varies with device size [89]. Devices with larger size have less variability. Another characteristic of random variation is that the effect gets reduced with the increase in the number of stages in a path. The amount of variation and its effect on device performance thus affect circuit area and power as device size increases to cope with the increase in variation.

D2D and systematic variations are global for a particular die or an area inside a die and therefore it is possible to detect and compensate these variations after the production of a chip. In the case of static variations, circuit performance can be predicted with statistical analysis using variation models.

2.1.2 Dynamic Variation

Dynamic variation refers to the variation that occurs during the run time of the chip. Examples of dynamic variations include supply voltage (V_{dd}) droops, temperature changes, and transistor aging degradation. V_{dd} droops result from sudden changes in switching activity which causes large current transients in the power delivery system. The droop magnitude and duration depend on the interaction of capacitive and inductive parasitics at the board, package, and die levels with changes in current demand [90]. V_{dd} droops contain high-frequency (i.e., fast changing) and low-frequency (i.e., slow changing) components and occur locally and globally across the die [91].

RTN (Random Telegraph Noise) [92] and NBTI (Negative Bias Temperature Instability) [93] are considered to be the main variation mechanism that effect transistor performance severely. Other phenomenons such as Hot Carrier Effect (HCE), Time-Dependent Dielectric Breakdown (TDDB) also affect LSI reliability. These variations is time dependent.

Temperature variations occur at a relatively slow time scale with local hot spots on the chip. Temperature variation results in leakage variation as well as transistor performance variation. As transistor density is increasing with scaling, power density is also increasing. For high performance chips, external cooling systems are required to suppress the chip temperature.

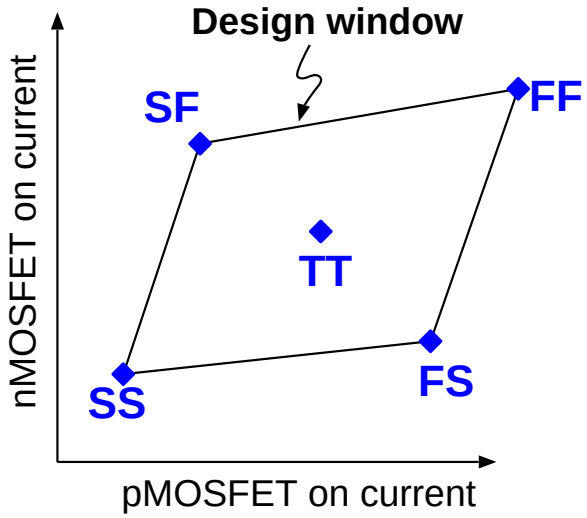


Figure 2.1: Design window defined by the transistor models. The window boundary is set by the models called the corners.

2.2 Variability Impact on LSI

Today's LSI chips contain millions of transistors. Integrating millions of transistors and interconnecting between the transistors is a huge challenge. Automation is needed to design a chip. Various EDA (Electronic Design Automation) tools are used for design and verification. In this section, LSI design flow will be discussed. The accuracy of models that represent real silicon behavior of transistor performance and its variation holds the key for efficient LSI design. The role of models will be explained here.

2.2.1 Variation Model

After the introduction of a new technology node, MOSFET models are built based on detailed *I-V* measurements of MOSFETs. The MOSFET models are then updated periodically to cope with the process characteristic changes over time. Presently, the foundry measures the I–V data of MOSFETs and build models such as BSIM SPICE model. Transistor performance and other physical dimension variations are captured in the statistics of model parameters. In reality, a lot of physical and electrical parameters suffer variations, thus incorporating each of these variations into the model would be too expensive. From the designer's point of view, variations in the key MOSFET parameters such as threshold voltage and gate length are of the concern. Traditionally process variation modeling is targeted for design-time use and guides engineers in the optimization of their chips before silicon fabrication. Thus, the variation models provided by the foundry tend to be pessimistic as all the L2L, W2W and D2D variations are lumped into as D2D variation. These variations are expressed by transistor corner models. Figure 2.1 shows a typical design window for nMOSFET and pMOSFET ON current where the design must meet its specification within this window. Here "TT" refers to typicial pMOSFET and typical nMOSFET condition. "SF" and "FS" refer to slow pMOSFET and fast nMOSFET, and

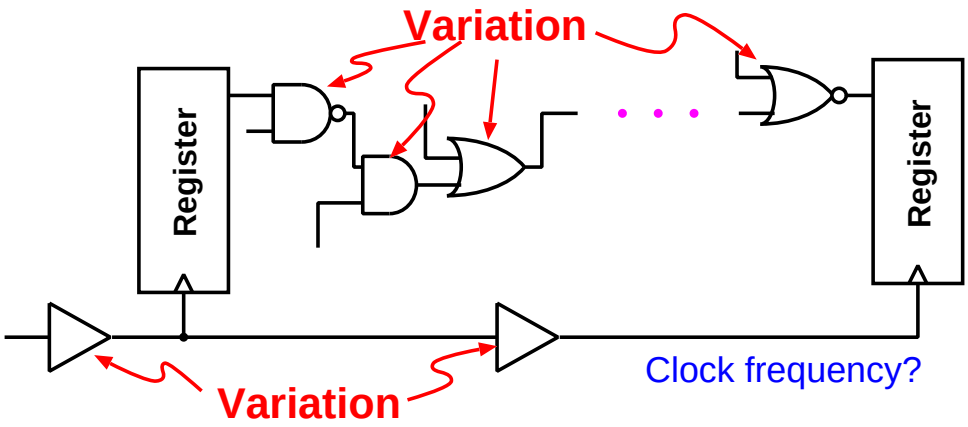


Figure 2.2: A typical synchronous circuit. Registers are clocked by clock signal. Path delays must meet the setup and hold requirements.

vice versa. The window is defined by several points which are called corners. Corner models define these corners in the process space. Random variation is expressed as statistics of model parameters. These models are then used in a Monte Carlo analysis of circuit performance.

2.2.2 Digital Circuit Design

A digital circuit contains millions of transistors, thus automated design flow is required to implement them. Full-custom design where every part of the circuit is designed manually will give us improved performance. However, this design method is not suitable for integrating millions of transistors. In order to automate the design process, various hierarchy and abstractions are used. Synchronous design provides the flexibility of defining abstractions for design automation. Figure 2.2 shows an example of a synchronous circuit. The data are stored in registers. The registers are clocked by a global clock signal. The clock frequency is determined by the worst possible delay between two registers. This delay path is called a critical path. The circuit is described using a Hardware Description Language (HDL). The language supports circuit description at higher level of abstraction, i.e. Register Transfer Level (RTL). The circuit is then synthesized for the target technology process.

Figure 2.3 shows the overall design flow. The design begins with the definition of some specifications for the target circuit or chip. The job of a circuit designer is to realize the specifications for the target technology process. The RTL description of the circuit is then synthesized and mapped to the target process technology. Cell-based design adopts a design flow where the entire circuit is built based on units called cell or gate to help automate the process. The collection of unit cells is called as standard cell library which is also provided by the foundry. The cell library is characterized for various process corners and operating conditions. The characterization results are provided as a form of lookup tables. Delay and power are estimated with the lookup tables during the synthesis process. If the delay and power do not meet the specification, multiple iterations are done while various optimizations are performed. Logic types, gate sizing etc. are the target of optimization. In order to deal with variation, some amount of

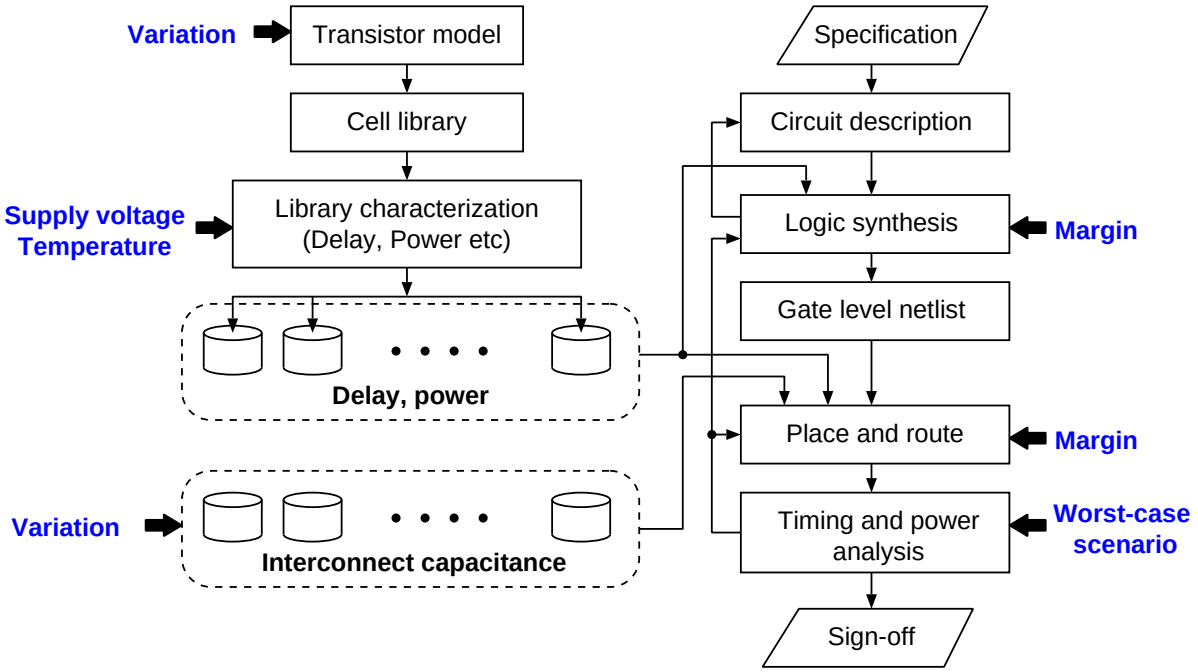


Figure 2.3: A typical design flow of LSI. Margins for variation is considered at various stages of the flow. Worst-case design is way too pessimistic resulting in area, cost and energy overhead.

margin is used in synthesis. This margin complicates the optimization process. The amount of margin is often set pessimistically which causes large area, delay, and energy overhead. The result of synthesis is the gate level description of the circuit. The netlist is then used in the automatic place and route flow. Large amount of margin is set here to ensure correct operation. Interconnect delay also play a major role here. Finally, detailed timing analysis is performed considering worst-case scenarios. If the circuit specification is met then circuit is signed off and the data is sent to the foundry for manufacturing. However, often multiple iterations are performed in between the steps as meeting the specification with large margins and worst-case scenario is difficult. This iteration directly results in the increase of cost.

The conventional design flow for digital circuits uses abstraction of various layers. In fact, one of the reasons for the tremendous improvement of LSI lies in the abstraction between different layers in the design. This has enabled designing LSI with millions of transistors, interconnects etc. However, with the increase of variability and need for energy-efficient computing has imposed great challenges to the conventional design of LSI because of the following reasons. Firstly, routing of clock signal to every component is a challenge. In a synchronous design, the storage elements, typically FFs (Flip-Flop), assume that they receive the clock signal at the same time. In order to route the global clock signal, large amount of resources as well design time is required. Besides, clock signals have to drive large number of gates which consumes lot of power. Variation affects the design of clock network directly. In order to avoid timing failure, clock network is over-designed which causes large energy overhead. Secondly, logic gates in the delay paths are also up-sized to account for variation. In order to prevent hold violation, buffers are inserted into the delay paths. The amount of design margin directly relates to

area and power overhead. Thus, accurate variation models are required to reduce this unwanted design margin as much as possible.

2.2.3 Analog Circuit Design

The conventional approach to design analog circuits is to fine tune each of the design parameters manually. Full custom layout is used where all the layouts starting from transistors to interconnects are done manually. Full custom design helps designers to achieve optimum performance. However, lack of portability for analog designs to a new technology process is a problem. After the introduction of a new technology node, the whole circuit need to be designed from the scratch. Analog circuits with differential operations are affected by mismatch between nominally identical components due to variation. The variations affecting analog performance may be mismatches in transistor threshold voltage, channel length and width, and mismatches in passive components such as resistors and capacitors [86].

2.2.4 Variability Effect on LSI Performance and Cost

Cost

With technology scaling, the number of integration is increasing. In an SoC, several blocks of digital and analog circuits are integrated together. Verifying chip functionally as well as its power and operation speed is challenging. According to the report of ITRS [14], design cost has become the greatest threat to continuation of the semiconductor roadmap.

Delay and Power

Variability causes large delay and power fluctuation. Circuit performance can deviate significantly from simulations because of the presence of variation. For example, a 30% variation in operating frequency and a 5–10 times variation in leakage power can occur in digital integrated circuits if variation problems are not appropriately handled and resolved [94]. Operating the chip with worst-case frequency and supply voltage results in excessive energy consumption. Leakage power is reported to be more susceptible to variability as leakage current has exponential relationship to threshold voltage.

Yield

Yield is the fraction of functional chips that meet the design target over all the chips manufactured. Variability plays a major role in chip performance and test. Failing to encounter variability during the design phase properly may cause severe effect on yield. A slight difference in process variation modeling/extraction may lead to significant yield difference. Especially, variability is reported to cause large yield loss for SRAM which is susceptible to random variation between devices.

Testing

Variability not only degrades the circuit and system performance and increases the power dissipation but also increases design and test cost exponentially. Test of LSI is a must before shipping the products to ensure correct operation. With the emergence of System-on-chip (SoC), testing has become more difficult. An SoC contains several blocks of digital, analog and RF circuits. Testing each part is extremely costly. Often the testing cost surpasses the manufacturing cost. On-chip solutions with the help of sensors can reduce test cost [95]. Delay testing is performed to make sure that the product chip operates at the desired target operating frequency. However, due to process variation and other manufacturing faults, some parts may not achieve the desired frequency, and therefore delays of all the paths must be tested. When some parts of the chip do not achieve the desired frequency, the biggest challenge is to debug the causes of the defect. In case of delay defects, there are mainly two reasons. One is the parametric variation which is often called as process variation and the other is random defect [96].

Reliability

Variability may cause reliability to degrade as chips may malfunction depending on some specific operating condition. Some errors many not be discovered during the test process, and they can be found at the field.

With device scaling, static device variations as well as dynamic variations such as Negative Bias Temperature Instability (NBTI) [93] and Random Telegraph Noise (RTN) [92] have become serious problem. Static variation occurs during the manufacturing process of the device whereas dynamic variation occurs during the run-time. Dynamic variations are difficult to predict and therefore it is difficult to set adequate design margin. In the case of static variations, circuit performance can be predicted with statistical analysis. In the case of dynamic variations, simple statistical analysis is not sufficient. Performance models based on variation characterization are required. In order to develop effective performance models, the nature of dynamic variations and their effects on the circuit behavior needs to be understood and characterized accurately. The effects of both the static and dynamic variations need to be taken into account during the design phase to ensure correct operation. Dynamic variation such as RTN has been considered to be small enough compare to the static random variation. However, with device scaling, RTN induced device variability has become comparable to static device variability [97]. RTN is reported to be causing failures in SRAM [98], flash memories [99] and CMOS image sensors [100]. RTN can cause ring oscillator (RO) frequency fluctuation as much as 10% at low voltage in 40 nm process [101].

2.3 Variation-aware Design Techniques

As variability has started to affect LSI performance severely, the need for variation-aware design is inevitable. Various techniques are proposed to account for variation. This section describes

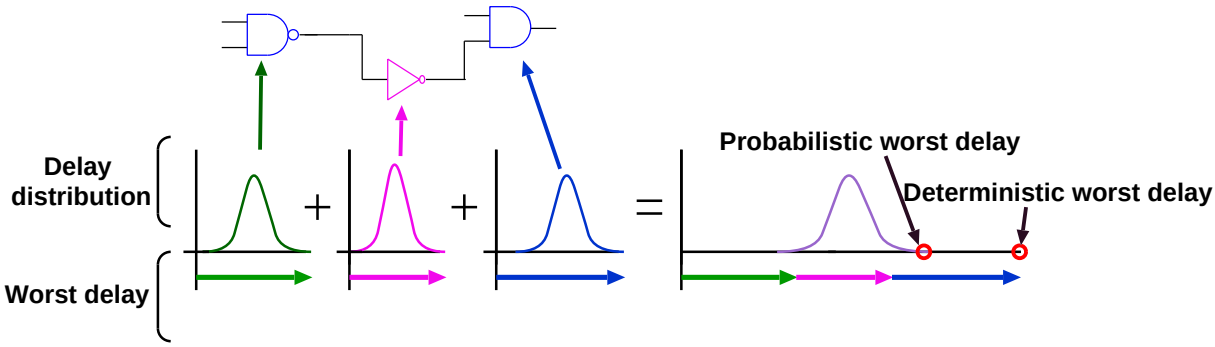


Figure 2.4: STA (Static Timing Analysis) versus SSTA (Statistical Static Timing Analysis). Static timing analysis results in pessimistic delay estimation causing energy overhead.

some of the variation-aware design techniques. The design techniques involves yield improvement, reducing pessimism, and post-silicon optimization.

2.3.1 Design for Manufacturability

Layout pattern has strong effect on variability. This has led to the introduction of design for manufacturability (DFM) techniques where the designers layout the transistors and interconnects following some guidelines. DFM techniques differ from process to process and foundry to foundry. Even within the same process, DFM guidelines may change from time to time. Thus, although creating universal DFM guidelines is difficult, some general design guidelines are considered as must today such as poly gate spacing regularity.

2.3.2 Statistical Design

Conventional timing analysis method considers delay as deterministic. STA (Static Timing Analysis) considers worst-case delay of each gate and calculate the circuit delay by summing the worst-case delays. However, in reality gate delays follow probability distribution thus deterministic result is way too pessimistic. Figure 2.4 shows the difference between deterministic worst delay and statistical worst delay for a circuit. In order to reduce the pessimism, STA considering on-chip variation (OCV) is being used. STA with OCV considers a certain percentage of delay variation for each path regardless of the number of stages. STA with advanced OCV is then being used which considers the effect of number of stages. The amount of percentage delay variation often called OCV coefficient, is need to be set. OCV coefficient is often set pessimistically and does not reflect the actual silicon information.

In order to accurately calculate path delays, SSTA (Statistical STA) is proposed [102, 103]. In SSTA, timing is calculated using sensitivity coefficients and statistical models, thus accurate calculation is possible. However, sensitivity calculation and providing accurate statistical models is difficult which limits the use of SSTA widely.

2.3.3 Error Detection Circuit

Error detection circuits are special circuits to detect timing failure of a FF. Error detection circuits require large area than on-chip monitor circuits as detection circuits are inserted at the path level of real circuits. Error detection circuits provide monitoring of real path delay. They however consume power too for their operation. Detection circuits can be placed in the selective critical paths only to reduce area and power overhead. However, in the presence of large variation, any path has the potential to be critical. Besides, detection circuits are useful only when the circuit can go back in time. Many of the general purpose ASICs do not have that function thus detection circuits are not applicable in those circuits.

Razor I is reported [48] which uses a delay error detecting flip-flop on the critical path of the design to reduce the supply voltage to the point of first failure for a given frequency. It allows reduction in design margins leading to significant energy saving. However, the technique requires additional circuitry like shadow latch and meta-stable detector for error detection.

A Canary FF is proposed [49] which uses a delayed data and a shadow FF along with traditional FF to detect timing error. Since, it compares the data at the output of a FF, it also requires meta-stable detector. Razor II [50] is another flavor of Razor where data transition is checked at the input of a flip-flop. Hence, it does not require a meta-stable detector. However, Razor I and Razor II are used in a processor framework where the corrective action is performed using re-execution of instruction.

Warning sequential schemes are proposed [51] which generates warning signals before any potential error. The benefit of such warning sequences over Razor and Canary FFs is that the scheme can be applied to any ASIC (Application Specific Integrated Circuit). Supply voltage is reduced up to little higher than the point when the warning signal is generated. This can reduce large supply voltage margin. However, this approach has a limitation that a sudden exercise of critical path may violate the timing as the exercise of critical paths depend on the input vectors.

2.3.4 Post-silicon Tuning

Various post-silicon tuning methods are proposed to deal with large variation [104–107]. One way to reduce variability effect after the product of a chip is to add some sort of programmability into the delay. This can be achieved by tuning clock skews by controlling load capacitances for examples [104, 106]. Path delay can be tuned in the similar way [107]. However, this technique requires additional circuitry which increases area. On-chip electrically programmable fuses are required to store the configuration of the programmable delays. Post-silicon tuning methodology can be applied to analog circuits also to reduce random mismatches between devices [105].

2.3.5 Adaptive Body Bias

Figure 2.5 shows a cross-section view of an nMOSFET device. Typically, the body potential of transistor is set to either supply voltage or ground depending on the MOSFET type. However, by

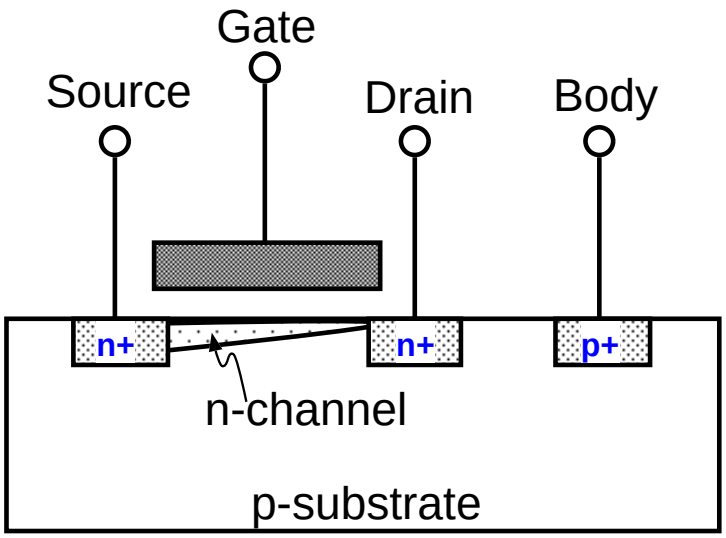


Figure 2.5: A cross section view of nMOSFET. Body terminal can be used as a forth terminal to tune the channel threshold voltage.

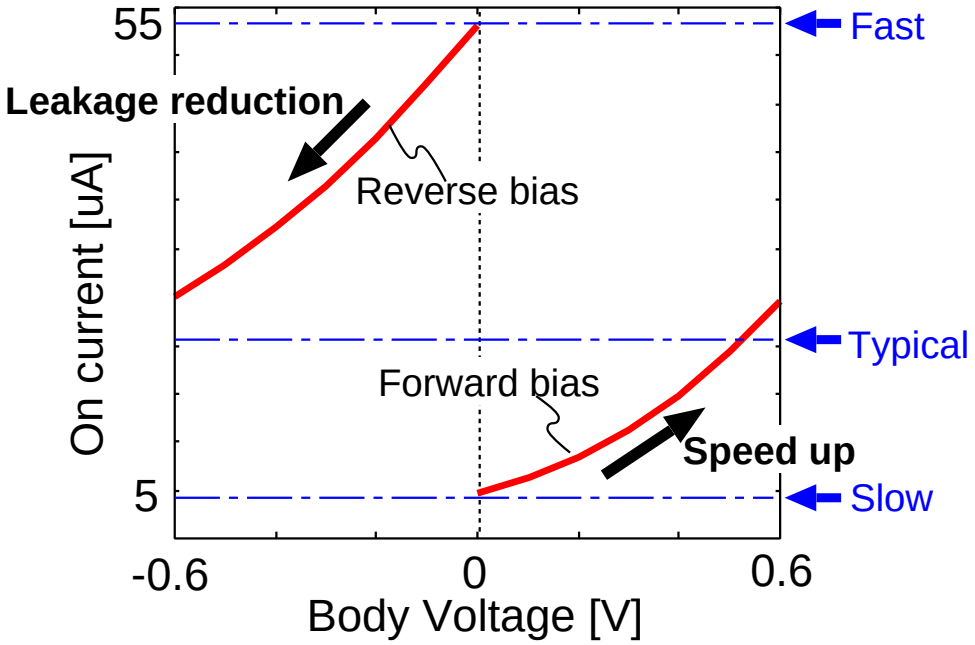


Figure 2.6: Effect of body bias on MOSFET ON current. Applying forward bias to slow devices can speed up the device and applying reverse bias to leaky devices can reduce leakage power.

changing the body potential, channel threshold voltage can be tuned. This gives a new option to the designers. Figure 2.6 shows the change of ON current for an nMOSFET at 0.6 V operation. ON currents for three different worst-case process corners of slow, typical and fast are indicated by arrows. When the chip is in the fast corner, reverse bias can be applied to reduce leakage. Similarly, when the chip is in the slow corner, forward bias can be applied to increase transistor ON current. This way variation effect can be mitigated after the chip production. Applying body bias adaptively can be used to reduce leakage power and improve yield [36, 37, 39]. Leakage power reduction is achieved by applying reverse bias to the chips with large leakage power. Yield improvement is achieved by applying forward body bias to the chip to improve MOSFET ON current. Design-time optimization can be done to further increase the effect of ABB [108]. Body bias can be applied externally to the chip or generated internally inside the chip. On-chip body bias generators are used to generate body voltages [109, 110].

2.3.6 Adaptive Supply Voltage

ASV (Adaptive Supply Voltage) is a variability compensation technique by adjusting the supply voltage to meet the target operation frequency of the chip [37, 111]. ASV is intended for improving yield with the sacrifice in power consumption because supply voltage is increased for chips that does not meet the speed specification. Supply voltage is set to a value where the chip functionality is confirmed after a series of test.

2.3.7 Asynchronous Circuit

An asynchronous circuit is a circuit where the circuit is divided into several functional units and the circuit operates by communicating between the functional units using for example hand-shake protocols. No global clock signal is thus required. Asynchronous design can be one solution to overcome the variability problem that is limiting the performance in a synchronous design. It has been shown that asynchronous operation is generally more robust when compared to synchronous systems [112]. Some examples of self-synchronous designs are reported that show that more robustness can be achieved [113, 114].

2.4 Summary

The continued scaling of MOSFET has been made possible by various innovations in process technology to overcome the shortcomings of present technology. However, the increase of variation in device characteristics has already stopped supply voltage scaling and now threatening the continuation of MOSFET dimension scaling. Process technology itself is not enough to overcome variation problem. Innovations in design and circuit techniques are required to continue LSI advancement. Key challenges are performance optimization with accurate models and prediction between various hierarchies which can be automated for use in CAD tools. On-chip monitor circuits can play a major role in providing accuracy and predictability to the system.

Chapter 3

Variability Modeling and Estimation using On-chip Monitor Circuits

Variation in circuit performance is captured as the difference between real chip measurement and simulated prediction. In order to do further analysis and diagnose circuit failures, the underlying process variation need to be known. This chapter develops an estimation technique that estimates process parameter variation from monitor circuit performances. Monitor circuits suitable for estimation will be explored and a set of monitor circuits will be proposed. Section 3.1 gives an introduction where key advancements of the proposed technique compared to the existing monitoring and parameter estimation techniques are explained. Section 3.2 describes the basic idea behind the proposed estimation technique. Section 3.3 describes the monitor circuit topology used in the thesis and how variation is captured and modeled. Section 3.4 explains the technique in details. Section 3.5 describes several monitor circuit topology and proposes a suitable set of monitor circuits. Section 3.6 validates the proposed monitor circuits by several experiments based on transistor level simulation. The proposed circuits are implemented in a 65-nm test chip. Section 3.7 describes test chip design and measurement procedure. Section 3.8 shows measurement results for D2D global variation for 30 chips. The measured data are then used to extract global variation of V_{thp}, V_{thn} and L. Detailed explanation on the estimation results and their validation are described here. Section 3.9 shows measured distribution of monitor outputs. Standard deviation of V_{thp}, V_{thn} and L are then extracted. Finally, Section 3.11 gives a summary of this chapter.

3.1 Introduction

Process variation needs to be characterized and modeled correctly. Circuit designers use the variation models to predict their circuit's performance. Generally, PCM (Process Control Module) and process monitors are placed at the scribe lines to monitor process variation. The conventional approach is slow as the threshold voltage of each device must be measured individually. Ref. [115] proposes a circuit to measure the threshold voltage shift of a device with respect

to a reference value which allows direct measurement of threshold voltage and thus measurement time and cost is reduced drastically. Device performances measured using DC bias may not correlate with performance in a digital circuit. In a digital circuit, a device operates under switching condition where it goes through several bias conditions over time. Ring oscillators (RO) consisting of standard inverter or NAND gates are considered as representatives of digital circuits and used to monitor variation effect on digital circuits [116]. With the increase of variation, it has become a necessity for product chips to have several monitor circuits which can provide variation information. Variation information from product chips are useful to debug the causes of timing failures in product chips and to estimate the chip performance. For example, Ref. [96] has shown that ROs embedded in the product chip are effective for screening delay defects. On-chip extraction of variation is extremely helpful post-silicon circuit performance prediction and diagnosis.

A conventional approach of digital measurement is to implement ROs consisting of basic inverter and NAND cells. The frequencies of such ROs give us useful information on the process variation to some extent. However, when there is mismatch between pMOSFET and nMOSFET performances, the conventional ROs fail to detect the mismatch. When pMOSFET and nMOSFET move to opposite directions, maximum operating frequency of digital circuit may not correlate to the ring oscillator frequency. Mismatch between nMOSFET and pMOSFET may cause unexpected timing failures. Besides, SRAM yield largely depends on the global and local mismatches between nMOSFET and pMOSFET. For post-silicon diagnosis, the location of the chip in the process space need to be known. Therefore, monitor circuits need to be embedded onto the product chip. This chapter discusses various topologies for monitor circuits. This paper then proposes estimation techniques of pMOSFET and nMOSFET variations from the on-chip measurements of monitor circuits.

Some approaches are proposed to extract process parameter variations from digital circuits such as ROs [117–119]. In Ref. [117], the slew rate of the inverter output is used to monitor rise time and fall time variations separately. In Ref. [118], the theory of pulse shrinking across a buffer ring is used to monitor rise time and fall time of the inverter cell. In Ref. [119], simple inverter structures with different P/N ratio is used to extract variations in pMOSFET and nMOSFET on-currents. This thesis introduces monitor circuits by which variations in process parameters can be estimated.

Authors in Refs. [5, 120, 121] has showed that extraction of different process parameters is possible with modified inverter structures and proper data processing. Extraction of threshold voltage variation from different path delays is proposed in Ref. [122]. In this approach, sensitivities of the monitor circuits are used to extract threshold voltage variations.

This paper proposes an estimation technique to extract process parameter variations based on measurements of multiple monitor circuits. The estimation procedure is similar to the one proposed in Ref. [122]. Key advancements over Ref. [122] are that a) simple inverter based monitor circuits are developed, and b) the estimation technique is extended to estimate both the global and WID random variations. The monitor circuits are designed such that they show

different sensitivity to a particular process parameter. Exploiting the difference in sensitivities, the process parameters are decoupled and expressed as transistor model parameters. The contributions of the thesis are as follows.

1. Monitor process variation directly from on-chip monitor circuits
2. Extract global variation as well as random variation
3. An efficient model-hardware correlation methodology
4. Validation from silicon data

The biggest challenge of this work is to show the validity using silicon measurements as the process variation in real chip is unknown. Body bias has been applied to the chip to emulate global variation in MOSFET threshold voltages to validate the proposed circuits. High correlation has been found in the estimation of global variation corresponding to the body bias values. Estimated WID random variation shows good agreement with measurement results reported in the literature [66]. Device array is being used in their approach whereas on-chip digital monitor circuits are used in this approach. The key contribution of the thesis is that it establishes a systematic technique to estimate global and random variation of process parameters using on-chip monitor circuits.

3.2 Basic Idea

Variation of a particular device dimension is a physical property. Measuring device dimensions on-chip is not realistic. Therefore, variation is observed and analyzed based on device performances. However, measuring device performances directly is expensive and requires large area. On-chip implementation requires easy measurement and implementation. Digital circuits are easy to measure. So, relationship between underlying physical parameter variation and circuit performance variation is complex. However, from a circuit designer's point of view, the term variation is interpreted as the difference between actual measurement from silicon and prediction based on a transistor model. Variability is thus the unpredictability that the designers have to handle during the design phase.

The basic idea of parameter estimation is shown in Fig. 3.1. In the conventional design flow, circuits are designed and then their behavior is simulated using transistor models. Transistor models are created on the basis of $I - V$ measurement data from the wafer. The layout of the designed circuit is implemented on the silicon wafer. Because of the variation in process parameters, characteristic of circuits in real chip deviates from its targeted value. Here, targeted value is the value that the designer makes assumption from SPICE (Simulation Program on Integrated Circuit Emphasis) simulation [123]. Thus, transistor model plays an important role in VLSI design. Circuit designers see the process through the transistor model provided by the foundry. From a designer's point of view, variation is the difference between the prediction they

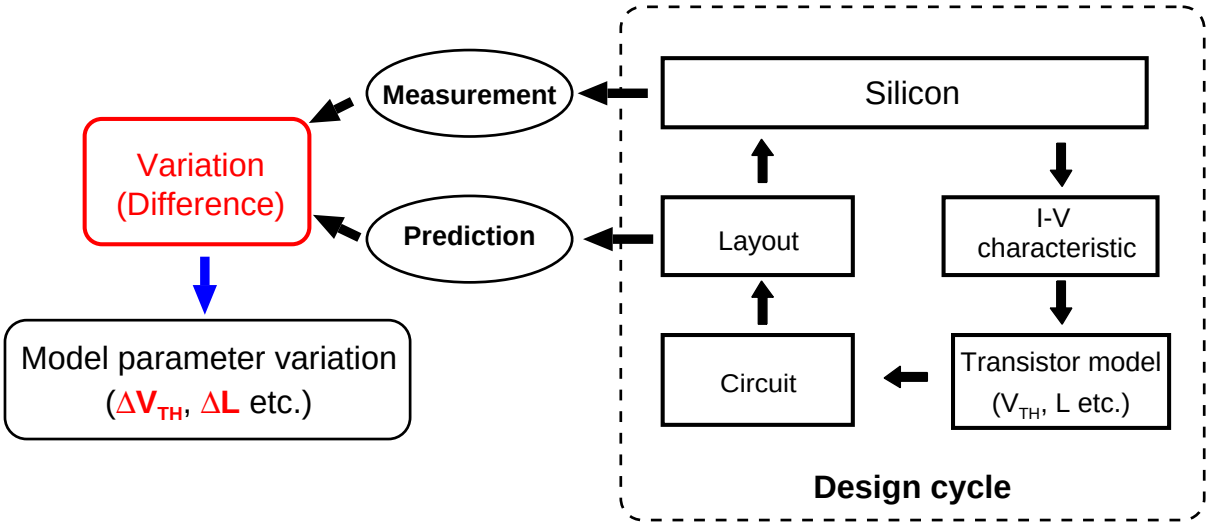


Figure 3.1: Basic idea of the proposed estimation technique. Estimate process parameter variations from the variations in measurements of circuit performance. Use circuit technique and transistor model.

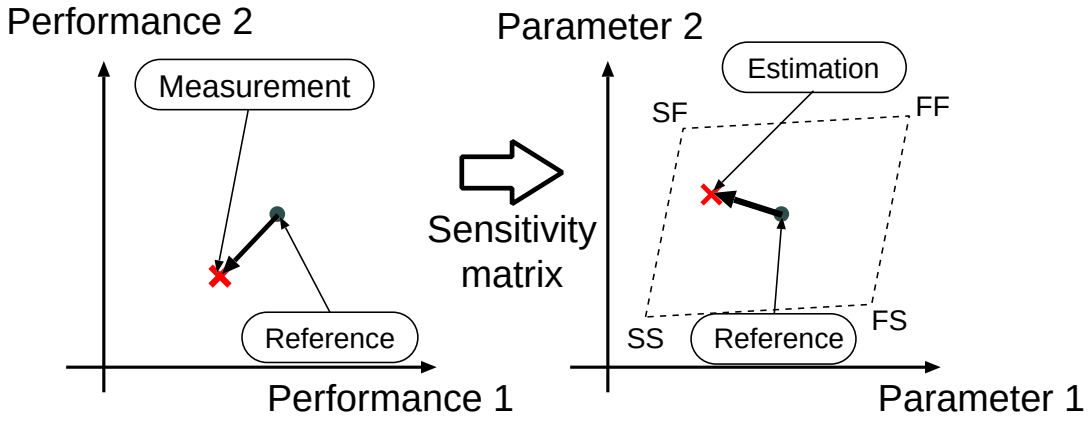


Figure 3.2: Extraction of process parameters from variation-sensitive monitor circuits. Sensitivity matrix relates variations in circuit performances to variations in process parameters. Monitor circuits having different sensitivities to different process parameters are needed. (©2012 IEEE)

make with the transistor models and the real chip performance. The transistor model provided to the circuit designer thus works as an interface between the circuit performances and the process parameters in silicon. Typically, process variation is modeled as statistical variation in key parameters of the transistor model such as threshold voltage and gate length. Depending on the topology and operation, different circuits show different sensitivities to process variation. The sensitivity of a circuit behavior to process variation can be calculated by circuit level simulation. These sensitivities give us useful information on the behavior of circuits under different process variation. In other words, the sensitivity maps the process space onto the circuit behavior space.

If we have several circuits which have different sensitivities to different process parameters, it is possible to extract the amount of variation using sensitivity analysis with reverse calcula-

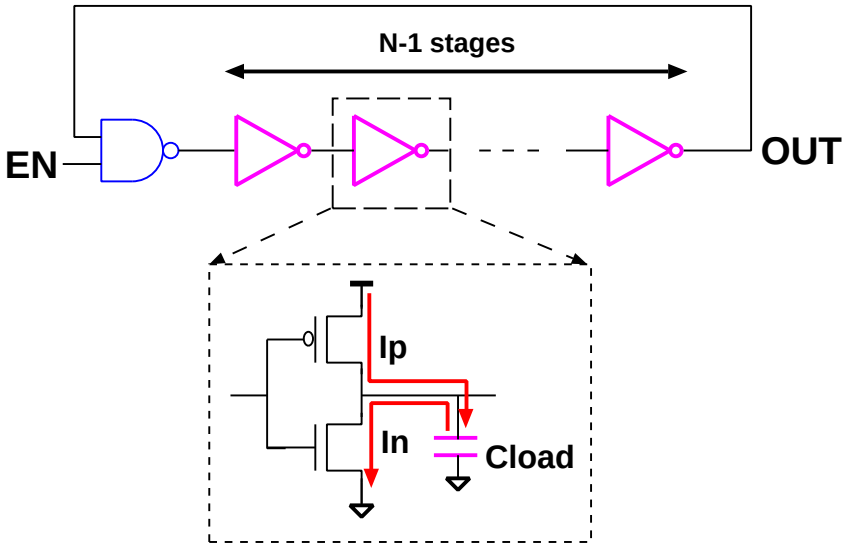


Figure 3.3: An N-staged RO with enable signal.

tion. This concept is illustrated in Fig. 3.2. For simplicity, Fig. 3.2 shows an example of estimation of two parameters from two circuit performances. The values of the process parameters defined in the model are considered to be the reference point. In design, circuit performances are predicted using this transistor model. The idea is to compare the measured performances with the predicted values and estimate the amount of deviation for each process parameter so that predictions get closer to the measured values. So, the differences between measurements and predictions are observable here which is shown in the left graph of Fig. 3.2. Because of variation, the measurement point will vary from chip to chip. Next, performance variation can be mapped to process parameter using the sensitivity matrix. The resulted process parameter domain is shown in the right graph of Fig. 3.2. Design of the sensitivity matrix is most important here to achieve accurate and robust estimation of process parameters. The topology and operation of circuits determine the sensitivity matrix. Therefore, monitor circuits suitable for robust estimation need to be investigated and developed.

3.3 Ring Oscillator as On-chip Monitor

Ring Oscillators (RO) are widely used to monitor process variation because they are easy to implement and measure. ROs can be integrated with other digital circuits. In order to implement the estimation method described in the previous section, ROs having different sensitivity to each process parameter need to be designed. Figure 3.3 shows an N-staged RO with an enable signal. During the oscillation, pMOS and nMOS transistors work in complimentary. The output load of an inverter stage is charged by the pMOS transistor in one cycle and then discharged by the nMOS transistor in the next cycle. So, the period of oscillation can be approximately expressed by using the current equation for pMOS and nMOS transistors using α-power law MOSFET model [124].

$$I_{on} = \frac{\mu C_{ox}}{2}\frac{W}{L}(V_{DD} - V_{th})^{\alpha} \tag{3.1}$$

$$\tau = N\frac{\tau_p + \tau_n}{2} \tag{3.2}$$

$$= \frac{N}{2}\left(\frac{C_L V_{DD}}{I_p} + \frac{C_{load} V_{DD}}{I_n}\right)$$

$$f = \frac{C_{ox}}{N \cdot C_{load} V_{DD} L}\left(\frac{\mu_p}{W_p(V_{DD} - V_{thp})^{\alpha}} + \frac{\mu_n}{W_n(V_{DD} - V_{thn})^{\alpha}}\right)^{-1} \tag{3.3}$$

Here, I_p and I_n are pMOSFET and nMOSFET current respectively. τ_p and τ_n are delays caused by pMOS and nMOS transistors. τ is the period of oscillation. C_{load} is load capacitance of each inverter. μ_p and μ_n are mobility for pMOSFET and nMOSFET respectively. V_{dd} is supply voltage and L is MOSFET gate length. W_p and W_n are pMOSFET and nMOSFET gate width, and V_{thp} and V_{thnn} are pMOSFET and nMOSFET threshold voltage respectively. α is a fitting parameter. Equation (3.3) gives an overall understanding on how RO frequency depends on each of the transistor parameters. Equations (3.3) gives us a guideline on the tunable parameters and their effect on RO frequency. For a conventional inverter topology, the tunable parameters are V_{dd}, L, W_p, W_n and C_{load}. This thesis presents an inverter topology which increases delay sensitivity to V_{thp} and V_{thn} variation by multiple times than the conventional topology. Effect of these tunable parameters on RO frequency sensitivity will be discussed in Section 3.5.

3.3.1 Variability Model for Estimation

A set of process parameters needs to be defined to express the variation effect first. Under process variation, Eq. (3.1) can be written as Eq. (3.4) where β_0 and V_{th0} are values defined by the transistor model.

$$I_{on0} + \Delta I_{on} = (\beta_0 + \Delta\beta) \cdot (V_{DD} - (V_{th0} + \Delta V_{th}))^{\alpha}. \tag{3.4}$$

I_{on0} is the default on-current calculated from transistor model and ΔI_{on} is the variation in chip. Values of $\Delta\beta$ and ΔV_{th} are variations which differ from chip to chip. From Eq. (3.4), at least two parameters are needed to model on-current variation of a single type of MOSFET. For simplification of the model, L is chosen to be a common parameter to model variations in current factors of both MOSFETs as L variation is common for both MOSFETs in standard cells. Majority of the global variation in MOSFET performance is reported to be contributed by MOSFET gate length and threshold voltage variation [26].

Although Eq. (3.3) indicates a non-linearity relationship of RO frequency to parameter variation of V_{thp}, V_{thn} and L, Fig. 3.4 shows that within a small range RO frequency change is almost linear. Suppose ΔV_{thp}, ΔV_{thn} and ΔL are the global variations of those parameters to be estimated and Δf is the corresponding frequency shift that can be measured. If ΔV_{thp}, ΔV_{thn} and ΔL are small, those variations can be related in a linear equation as follows where k_P, k_N and k_L are

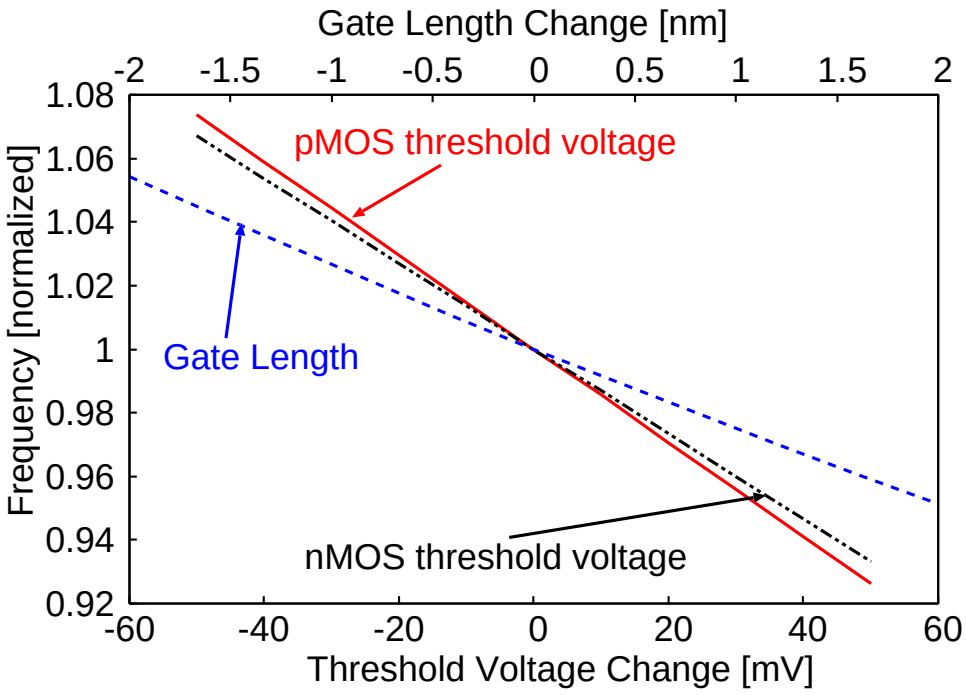


Figure 3.4: Change of frequency according to changes in process parameter values. RO with conventional inverter topology is used.

sensitivity coefficients.

$$\Delta f = f_{\mathrm{m}} - f_0 = k_{\mathrm{p}}\Delta V_{\mathrm{thp}} + k_{\mathrm{n}}\Delta V_{\mathrm{thn}} + k_{\mathrm{l}}\Delta L. \tag{3.5}$$

Here, f_{m} is the measured frequency and f_0 is the reference frequency. f_0 can be calculated using SPICE simulation with resistances and parasitic capacitances (RC) extracted netlist from layout. In order to cancel within-die random effect, RO with large number of stages or average value from many ROs can be used. Sensitivity coefficients can be calculated from SPICE simulation. RC extracted netlist should be used to calculate sensitivity coefficients because parasitic capacitances affect frequency sensitivities. In Eq. (3.5), only three unknown values, the D2D variations of V_{thp}, V_{thn} and L, are there. The problem here is how to get these three unknown values from the RO frequencies.

3.3.2 Capturing Variation

Process variation is categorized into D2D and WID variation. Among the many components of WID variation, the random component is considered to be most significant. A large SoC chip, which contains various circuit blocks, may also suffer from variation that differs from location to location. D2D and location-correlated variation can be considered as global variation as this variation affects all the transistors in the particular location with the same amount. This variation results in a fixed amount of shift in circuit performance which can be detected and compensated during testing and post-silicon measurement. However, measurement of global

variation may contain error due to the random variation effect. In order to reduce the random variation effect, the number of stages in the circuit can be increased or the logic gates can be up-sized. Another way of capturing global variation is to measure large number of instances of the same circuit layout, and then average the measured values.

WID random variation is modeled by Gaussian distribution and standard deviation is used to express the extend of random variation. As random variation is a statistical parameter, sufficient number of samples are required. Implementing multiple instances of the same circuit and measuring performance of each instance will give us a distribution. This distribution reflects the random variation of several underlying process parameters.

3.4 Parameter Estimation Technique

In this section, the proposed parameter estimation procedure is described. Monitor circuits suitable for this technique will be discussed in Section 3.5.

3.4.1 Global Variation

Tuning the design parameters in Eq. 3.3 can realize ROs with different sensitivities. Suppose K number of ROs with various sensitivity vectors are implemented on the chip. K number of linear equations can be built from the K ROs as shown in Eq. (3.6).

$$\begin{aligned} \Delta f_1 = f_{m1} - f_{01} &= k_{p_1}\Delta V_{thp} + k_{n1}\Delta V_{thn} + k_{l1}\Delta L \\ \Delta f_2 = f_{m2} - f_{02} &= k_{p_2}\Delta V_{thp} + k_{n2}\Delta V_{thn} + k_{l2}\Delta L \\ &\vdots \\ \Delta f_K = f_{mK} - f_{0K} &= k_{p_K}\Delta V_{thp} + k_{mK}\Delta V_{thn} + k_{lK}\Delta L \end{aligned} \tag{3.6}$$

For global variation, ΔV_{thp}, ΔV_{thn} and ΔL are same for all transistors and all circuits. Thus, a method like least square method can be used to find the solution for the unknown parameters which best satisfies Eq. (3.6).

Estimation based on the simple linear model of Eq. (3.5) has two potential problems. First, non-linearity in circuit output will affect the estimation accuracy. Second, process variations affect the sensitivity values; thus sensitivity coefficient calculated at the reference condition may be different from the sensitivity coefficient at the estimated condition. In order to overcome these two problems, this thesis proposes an iterative estimation technique where sensitivity coefficients are updated at each iteration; thus correlation between model and hardware can be achieved and non-linearity problem can be overcome too.

In Eq. (3.5), there are three unknown parameters. So, at least three equations are needed to extract these three unknown values. The three equations can be derived from three monitor circuits whose sensitivity vectors form a non-singular matrix. The amount of variation of each

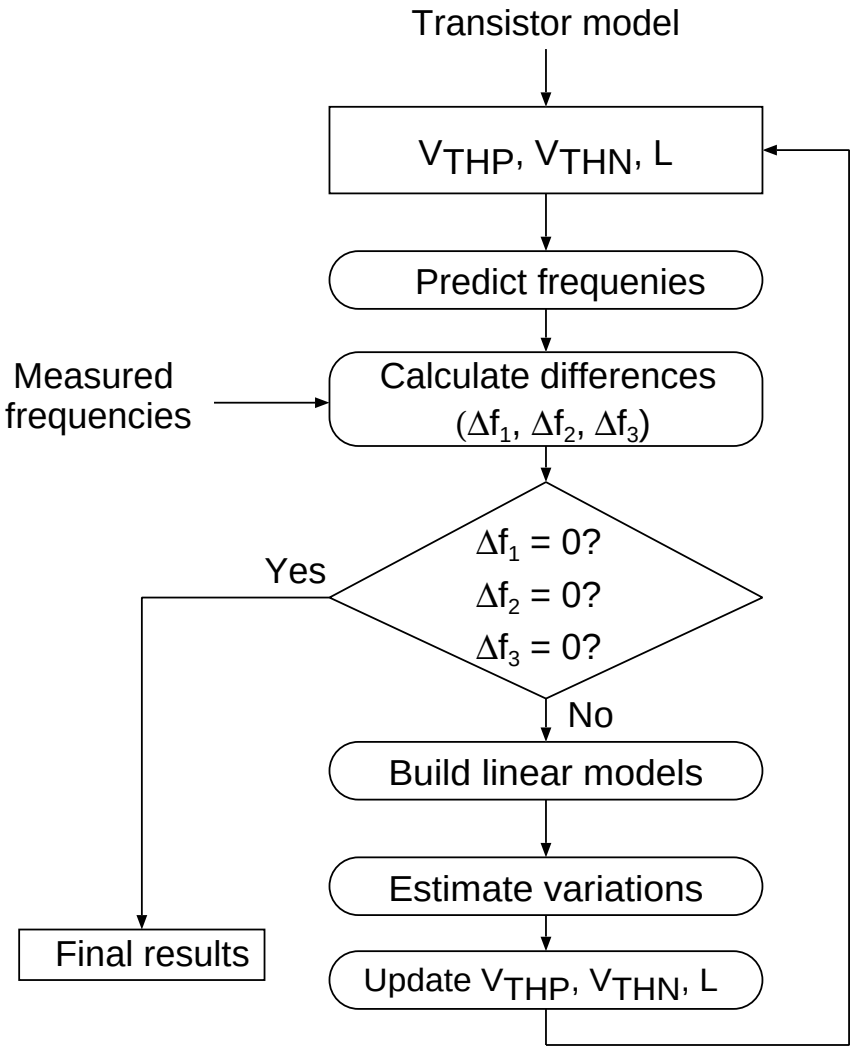


Figure 3.5: Proposed iterative estimation procedure of process parameters. (©2012 IEEE)

process parameter will be estimated from Eq. (3.7).

$$\vec{V} = \mathbf{S}^{-1}\vec{F}, \tag{3.7}$$

where

$$\vec{V} = \begin{pmatrix} \Delta V_{\mathrm{thp}} \\ \Delta V_{\mathrm{thn}} \\ \Delta L \end{pmatrix}, \mathbf{S} = \begin{pmatrix} k_{\mathrm{p1}} & k_{\mathrm{n1}} & k_{\mathrm{l1}} \\ k_{\mathrm{p2}} & k_{\mathrm{n2}} & k_{\mathrm{l2}} \\ k_{\mathrm{p3}} & k_{\mathrm{n3}} & k_{\mathrm{l3}} \end{pmatrix}, \vec{F} = \begin{pmatrix} \Delta f_1 \\ \Delta f_2 \\ \Delta f_3 \end{pmatrix}.$$

Here, vector $\vec{V}$ is the vector for ΔV_{thp}, ΔV_{thn} and ΔL. Matrix $\mathbf{S}$ is the sensitivity matrix and vector $\vec{F}$ is the vector for the frequency shift from the reference value. Vectors $(k_{\mathrm{P1}}, k_{\mathrm{N1}}, k_{\mathrm{L1}})$, $(k_{\mathrm{P2}}, k_{\mathrm{N2}}, k_{\mathrm{L2}})$ and $(k_{\mathrm{P3}}, k_{\mathrm{N3}}, k_{\mathrm{L3}})$ are sensitivity coefficient vectors for three circuits.

Figure 3.5 shows the proposed iterative estimation procedure. First, frequencies of the monitor circuits are predicted by circuit simulation using a transistor model. Measured values are obtained from the chip and then compared with the predicted values. Zero difference refers that

no variation from the values in process parameters defined in the model exists in the chip. If the difference is not zero, then some variations exist in the chip. Linear models of Eq. (3.5) are built by calculating the sensitivity coefficients. Variations of the target parameters are estimated by solving Eq. (3.7). Process parameter values are updated in the transistor model with the estimated amounts of variations and new predictions are made for the frequencies. If the differences between measurements and predictions are not zero, new linear models are built and variations are estimated again. Thus, new set of parameter values will be obtained after each iteration. This way, the whole process is iterated until the differences between measurements and predictions are zero.

Selection methodology of monitor circuits suitable for this technique will be discussed in the next section.

For monitoring of global variations of process parameters, effect of random variation needs to be canceled out. Large number of stages for ROs will average out random effect, but also consumes large area. Thus, a trade-off has to be made between estimation accuracy and monitor circuit area. The equation below shows the probability distribution function of Δf which is the difference between measurement and prediction.

$$\phi(\Delta f) = a \exp \frac{\mu_{\Delta f} - \Delta f}{2\sigma^2_{\Delta f}}. \tag{3.8}$$

Here, $\mu_{\Delta f}$ is the mean value, $\sigma_{\Delta f}$ is the standard deviation of Δf and a is a constant. For estimation of global variation, the mean value $\mu_{\Delta f}$ is the parameter of interest. Because of random variations, measured value of $\mu_{\Delta f}$ may contain some amount of error. The measured value of $\mu_{\Delta f}$ will fall within the range of $\mu_{\Delta f} \pm 3\sigma_{\Delta f}$ with 99.9% probability.

From Eq. (3.7), the estimation value v_i of a particular parameter can be expressed by Eq. (3.9).

$$v_i = z_{i1}\Delta f_1 + z_{i2}\Delta f_2 + z_{i3}\Delta f_3. \tag{3.9}$$

Here, parameter v_i is the variation to be estimated and parameter z_{ij} is the element of the matrix $\mathbf{S}^{-1}$ of Eq. (3.7). Index i refers to the row number of the matrix. In Eq. (3.8), Δf_1, Δf_2 and Δf_3 follow the probability distribution function of Eq. (3.8). Using the method of moment, distribution σ_{v_i} in estimated value can be calculated with the following Gaussian approximation [125].

$$\sigma^2_{v_i} = \sum_{j}^{3} (z_{ij}\sigma_{\Delta f_i})^2. \tag{3.10}$$

Equation (3.10) gives us the trade-off relationship between the estimation accuracy and the number of stages. By calculating the distributions of each RO frequency, sufficient number of stages can be calculated for a tolerable error range value of σ_{v_i} using Eq. (3.10).

3.4.2 WID Random Variation

Equation (3.5) assumes same amount of variation in all the transistors in the chip. However, during random variation modeling, variation in each transistor need to be considered. Equation 3.5 can be extended to express the relationship between RO frequency and each transistor in the RO. Thus, the oscillation frequency variation, Δf, can be expressed by the following equation considering the effect of each transistor in the RO.

$$\begin{aligned}\Delta f &= f_{\mathrm{m}} - f_0 \\ &\approx \sum_i \left(\frac{\partial f}{\partial V_{\mathrm{thp}_i}} \Delta V_{\mathrm{thp}_i} + \frac{\partial f}{\partial V_{\mathrm{thn}_i}} \Delta V_{\mathrm{thn}_i} + \frac{\partial f}{\partial L_i} \Delta L_i \right).\end{aligned} \tag{3.11}$$

Here, f_0 represents the nominal frequency when there is no variation, i is the transistor index in the RO, and $\Delta V_{\mathrm{thp(n)}_i}$ and ΔL_i are the amounts of variations in threshold voltage and gate length for the i-th MOSFET. Sensitivity coefficients are calculated by circuit simulation. Assuming that the transistor variability are random and have no correlation to each other, the frequency variance is therefore expressed as the sum of variances caused by each random component, as follows.

$$\begin{aligned}&\sigma^2_{\Delta f} = \\ &\sum_i \left(\left(\frac{\partial f}{\partial V_{\mathrm{thp}_i}} \right)^2 \sigma^2_{V_{\mathrm{thp}}} + \left(\frac{\partial f}{\partial V_{\mathrm{thn}_i}} \right)^2 \sigma^2_{V_{\mathrm{thn}}} + \left(\frac{\partial f}{\partial L_i} \right)^2 \sigma^2_L \right).\end{aligned} \tag{3.12}$$

Here, $\sigma_{\Delta f}$ is the standard deviation of oscillation frequency, and $\sigma_{V_{\mathrm{thp}}}$, $\sigma_{V_{\mathrm{thn}}}$, and σ_L are standard deviations of V_{thp}, V_{thn}, and L variations, respectively. The validity of the above model will be discussed later in the section.

In Eq. (3.12), unknown parameters that we want to derive are σ^2_{Vthp}, σ^2_{Vthn}, and σ^2_{L}. Measuring statistically meaningful number of ROs, we can obtain the variance of the measured frequencies. Thus, We can measure $\sigma^2_{\Delta f}$ in Eq.(3.12) is obtained from on-chip measurements. Then, we can build a system of linear equations using the three different RO topologies described above. Using the system of linear equation, we can apply a least-square method to estimate the unknown parameters of our interest.

If we put multiple instances of ROs on the chip, standard deviations can be derived for the RO frequencies. In the test chips, we assume that random variation is the most dominant component in the WID variation. Next, we extract the amount of variations in threshold voltages and gate length from the RO frequency variations using frequency sensitivities to each transistor in the RO. The sensitivity coefficients are calculated with circuit simulation. Sensitivity-based method to extract parameter variations from process-sensitive RO frequencies is discussed in Refs. [4, 121]. In Ref. [4], a system of linear equations is built using the sensitivity coefficients. Then, from the measured frequency deviations, the unknown amounts of parameter deviations are estimated with a maximum likelihood method.

Center point of linearization is an important consideration because the sensitivity values change depending on the center point. To determine the center point of linearization, we need to obtain center values for V_{thP}, V_{thN}, and L. Center values for V_{thP}, V_{thN}, and L are obtained by estimating global variations of these parameters with the estimation method described in Sec. 3.4.1. The transistor model is then updated with the estimated values of global variations for V_{thP}, V_{thN}, and L. Sensitivity coefficients are then calculated with the updated transistor model.

3.5 Monitor Circuits for Process Variation Estimation

A suitable set of monitor circuits is needed to realize the proposed estimation technique described in Section 3.4. The monitor circuits should have different sensitivities to the process parameters. In this section, some design options to realize variation-sensitive monitor circuits from simple inverter cell topologies will be demonstrated. Sensitivities are calculated by circuit simulation. Commercial 65 nm process technology is assumed in our simulation. Based on the simulation results, a methodology to choose the best suitable monitor circuits will be described.

In this section, the concept and topology of on-chip monitor circuits are discussed. Monitor structures suitable for the concept are explored, and a suitable structure is presented.

3.5.1 Design of Monitor Circuit Topology

For the estimation technique proposed in Section 3.4, monitor circuits with the following characteristics are needed.

1. Regularity and low design complexity
2. High sensitivity
3. Digital in nature
4. Small area

Regularity of poly pitch should be maintained in the monitor circuit as it affects gate length variation. Design complexity should be low so that monitor circuits can be ported to different process technology. Next, monitor circuits should have high sensitivities to process parameter variations. Digital nature of the monitor circuits is important for on-chip measurement and processing. Finally, area of the monitor circuits should be small enough so that implementing them does not cause large area overhead.

An RO is a good candidate to be used as a monitor circuit. A simple RO fulfills all the requirements mentioned above except the high sensitivity. In this thesis, the inverter topology in the RO is therefore modified to obtain enhanced sensitivities. The following techniques are used to modify the sensitivities of a RO frequency to process parameters.

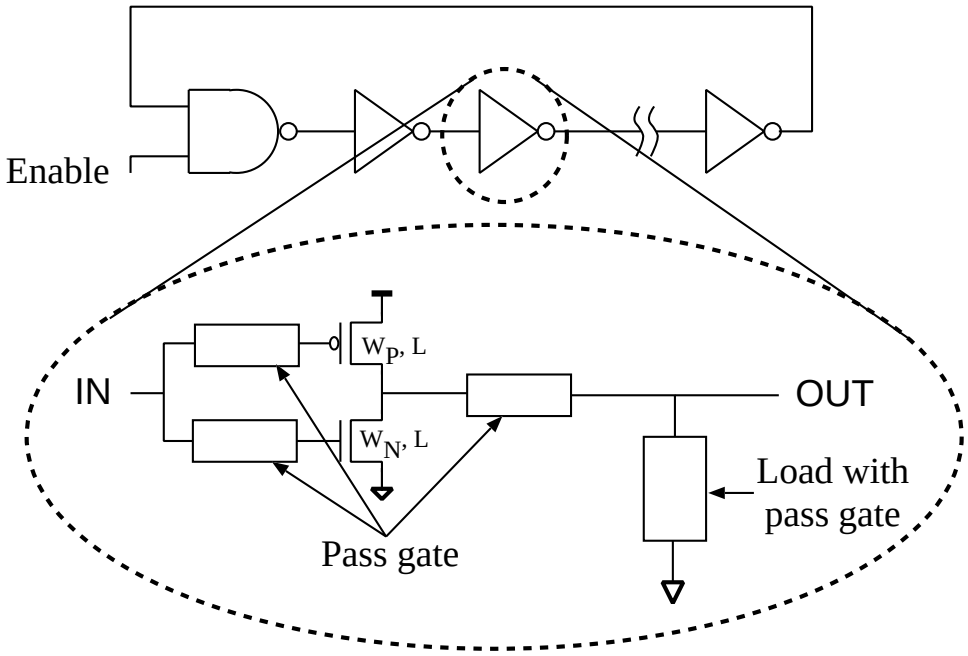


Figure 3.6: RO as monitor circuit. The inverter cell topology can be modified to get enhanced sensitivities. (©2012 IEEE)

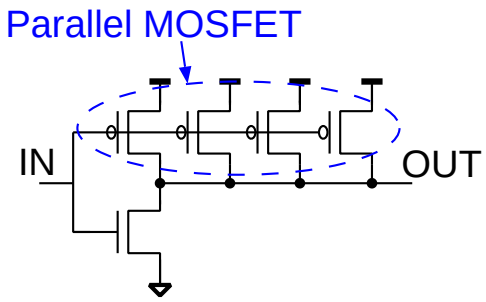


Figure 3.7: An inverter cell with parallel pMOSFETs (“PRICH”).

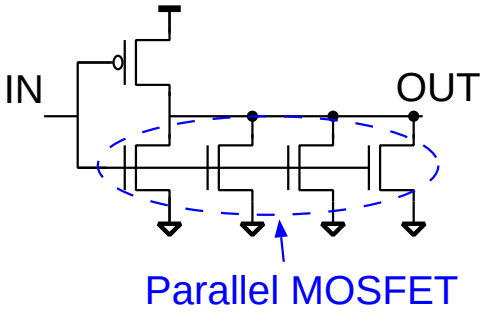


Figure 3.8: An inverter cell with parallel nMOSFETs (“NRICH”).

1. Change gate width and gate length
2. Use pass-gates
3. Use gate capacitance and pass-gate in series

Figure 3.6 shows the modified inverter topology where the inverter topology can be modified to get enhanced sensitivities. Effects on the sensitivities are described below.

RO with Different Gate Length and Gate Width

Increasing the gate width of the pMOSFET or increasing the gate length of the nMOSFET in the inverter will make the RO frequency more sensitive to nMOSFET parameters. We can increase gate width of pMOSFET or we can place multiple pMOSFETs in parallel. In order to

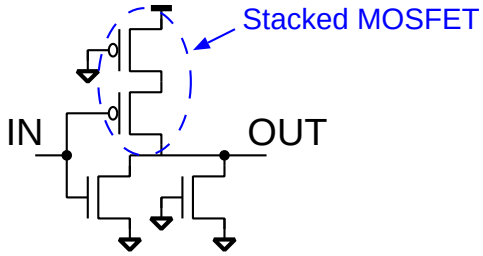


Figure 3.9: An inverter cell with stacked pMOSFETs (NOR2).

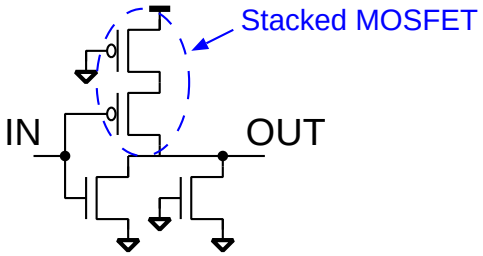


Figure 3.10: An inverter cell with stacked nMOSFETs (NAND2).

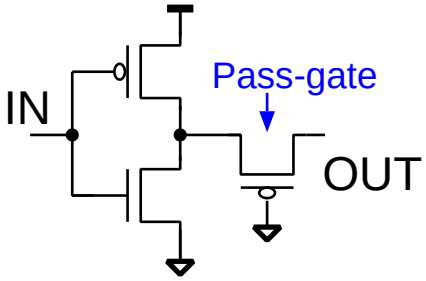


Figure 3.11: An inverter cell with pMOSFET pass-gate at output ("PPASS_O").

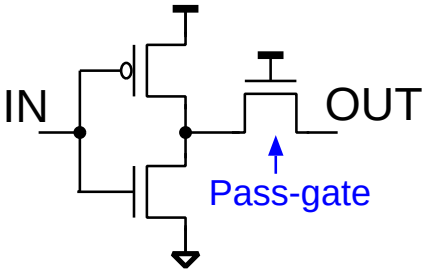
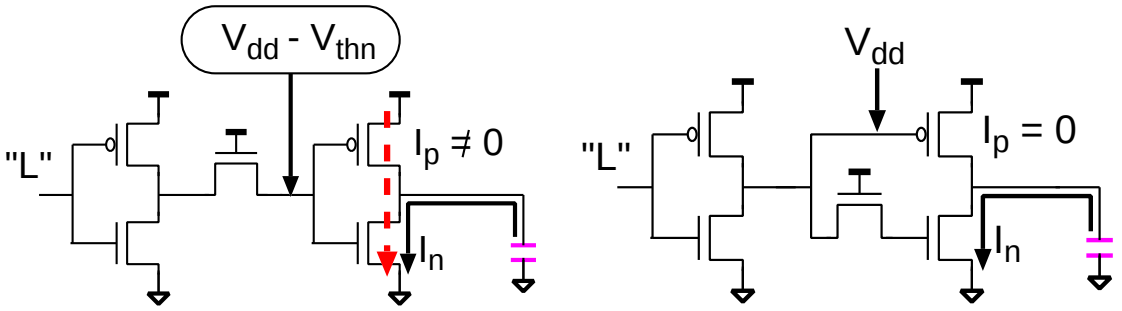


Figure 3.12: An inverter cell with nMOSFET pass-gate at output ("NPASS_O").

maintain regularity, we have designed inverter cells with parallel MOSFETs. Figure 3.7 shows an inverter where pMOSFET is 4 times larger than that of the standard cell. Similarly, inverter topology shown in Fig. 3.8 will be more sensitive to pMOSFET parameters. We call these cells as "PRICH" and "NRICH" respectively, whereas the standard inverter cell is called as "STD". From simulation results for an "PRICH" RO, 21% increase in V_{thn} sensitivity and 20% decrease in V_{thp} sensitivity is calculated compare to that of the "STD" RO. Inverter structure examples to realize larger gate lengths with regular layout is shown in Figs. 3.9 and 3.10. Two stacked transistors are used to realize the effect of larger gate length equivalently.

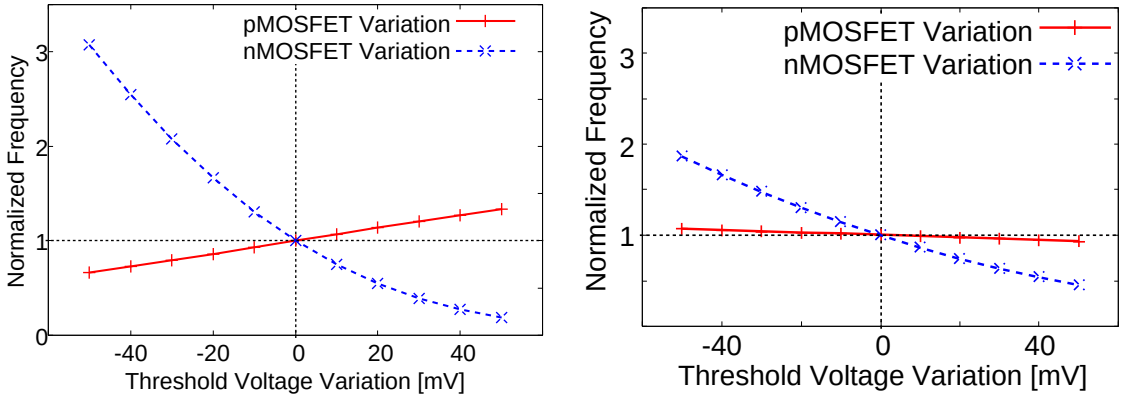
RO with Pass-gate

RO with single pass-gate becomes highly sensitive to threshold voltage variation. The operation of RO with pass-gate is demonstrated in [121]. Figs. 3.11 and 3.12 show inverter cells with an pMOSFET pass-gate and an nMOSFET pass-gate at the output. We call these inverter cells as "PPASS_O" and "NPASS_O" respectively. For nMOSFET pass-gate, voltage drop equal to the nMOSFET threshold voltage occurs during the charging of the output node thus the output node voltage rises only to a value of $V_{\text{dd}} - V_{\text{thn}}$. Similarly, for pMOSFET pass-gate, voltage drop occurs during the discharging of the output node thus the voltage falls only to V_{thp}. Voltage drop across the pass-gate reduces the gate overdrive for the MOSFET of the inverter. Thus, any change in threshold voltage is amplified in the RO frequency. For an "PPASS_O" RO, ΔV_{thp} sensitivity increases by 5 times than that of "STD" RO. For "NPASS_O" RO, ΔV_{thn} sensitivity increases by 7 times than that of "STD" RO.



(a) Pass transistor at output (b) Pass transistor at input

Figure 3.13: Inverter with nMOSFET pass transistor at the input is sensitive to nMOSFET variation only. (©2013 IEICE)



(a) Pass transistor at output (b) Pass transistor at input

Figure 3.14: Sensitivity to MOSFET threshold voltage variation. Topology of Fig. 3.13(b) is sensitive to particular MOSFET variation thus suitable for parameter extraction. (©2013 IEICE)

Same gate sizes are used for pass-gates in the simulations as those in the standard inverter cell MOSFETs. Next, the effect of gate width of pass-gates on the sensitivities are studied. For "NPASS_O" RO, decreasing pass-gate size to half increases the sensitivity to ΔV_{thn} by 8% which is very small compared to the 500% increase in the sensitivity against the "STD" RO. Considering design and layout complexity, pass-gates with same sizes of MOSFETs as in the standard inverter cell are preferable.

Because of the voltage drop across the pass-gate for "PPASS_O" and "NPASS_O" inverter cells, inverter output voltage does not rise to high level during the loading of the next inverter. This voltage drop turns the pMOSFET of the next inverter partially on. This phenomenon is shown in Fig. 3.13(a). For the inverter topology in Fig. 3.13(a), the frequency also changes largely when pMOSFET threshold voltage varies. When pMOSFET threshold voltage lowers meaning pMOSFET becomes faster, the RO frequency decreases instead of increasing. This behavior is the opposite of the behavior found in conventional CMOS digital circuits. This is because when pMOSFET becomes faster, the through current increases resulting in large delay during the pull-down. In order to avoid this, inverter topology of Fig. 3.13(b) is proposed.

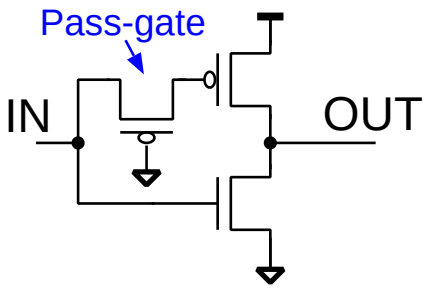


Figure 3.15: An inverter cell with pMOSFET pass-gate at input of pMOSFET gate ("PPASS_I").

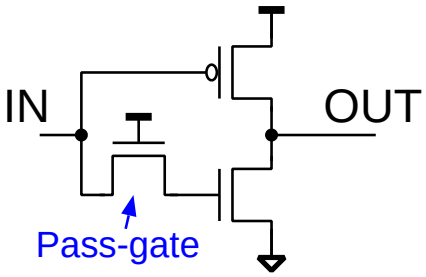


Figure 3.16: An inverter cell with nMOSFET pass-gate at input of nMOSFET gate ("NPASS_I").

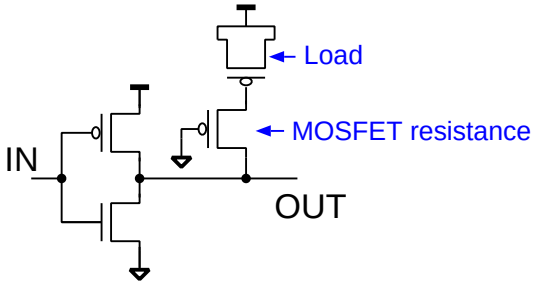


Figure 3.17: An inverter cell with extra load and pMOSFET pass-gate. Time for charging and discharging of the extra load depends on pMOSFET pass-gate threshold voltage. ("PLOAD").

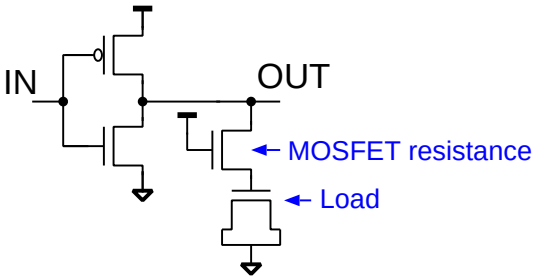


Figure 3.18: An inverter cell with extra load and nMOSFET pass-gate. Time for charging and discharging of the extra load depends on nMOSFET pass-gate threshold voltage. ("NLOAD").

Fig. 3.14 shows the change of RO frequency to MOSFET threshold variations for inverter topologies of Fig. 3.13(a) and Fig. 3.13(b). In Fig. 3.14, both types of pass-gate topologies show high sensitivity to nMOSFET variations. However, for the inverter topology in Fig. 3.13(a), the frequency also changes largely when pMOSFET threshold voltage varies. Thus, inverter topologies shown in Figs. 3.15 and 3.16 are proposed as new inverter topologies having high sensitivity to MOSFET threshold voltage. We call these cells as "PPASS_I" and "NPASS_I" respectively. In Fig. 3.13(b), the RO frequency is not sensitive to pMOSFET variation. For "NPASS_I" inverter, nMOSFET pass-gate contributes to the fall time only, thus the pMOSFET of the next stage is turned off fully. Input and output voltage have full swing during the oscillation similar to the behaviors of the standard cells.

RO with Extra Load

Figs. 3.17 and 3.18 are ROs with an extra load in the output. These cells will be called as "PLOAD" and "NLOAD" respectively where "PLOAD" cell's load is controlled by an pMOSFET pass-gate and "NLOAD" cell's load is controlled by an nMOSFET pass-gate. The loads are realized by MOSFET gate capacitance. For Fig. 3.17, when V_{thp} increases, resistance for the pMOS pass-gate increases. As a result, the inverter sees smaller load and hence delay decreases.

Table 3.1: Sensitivity coefficients of ROs. (©2012 IEEE)

RO Type	k_P	k_N	k_L
STD	−0.038	−0.035	−0.026
PPASS_I	−0.18	−0.033	−0.063
NPASS_I	−0.028	−0.2	−0.034
PPASS_O	−0.24	0.052	−0.085
NPASS_O	0.054	−0.34	−0.029
CPASS	−0.039	−0.036	−0.026
PRICH	−0.031	−0.041	−0.026
NRICH	−0.046	−0.034	−0.027
PLOAD	−0.020	−0.048	−0.023
NLOAD	−0.044	−0.022	−0.027

Thus, the effect of V_{thp} variation gets reduced. Sizing of the load determines the sensitivity for this inverter topology. For "PLOAD" RO where the extra load is equivalent to FO4 of the "STD" cell, sensitivity to V_{thp} decreases by 45% than that of an "STD" cell RO.

Table 3.1 summarizes sensitivity coefficients for these ROs. Sensitivity coefficients are calculated by $k_P = \frac{\Delta f/f0}{\Delta V_{\text{thp}}/\Delta V_{\text{thp0}}}$, $k_N = \frac{\Delta f/f0}{\Delta V_{\text{thn}}/\Delta V_{\text{thn0}}}$ and $k_L = \frac{\Delta f/f0}{\Delta L/\Delta L0}$. Most of the ROs show inversely proportional relationship against the parameter variation meaning RO frequency increases with the decrease of parameter value. However, ROs of "PPASS_O" shows proportional relationship against pMOSFET threshold variation. This characteristic is the result of voltage drop during the discharging of the output node. Similarly, "NPASS_O" shows proportional relationship against nMOSFET threshold variation.

3.5.2 Proposed Set of Monitor Circuits

A set of monitor circuits is needed to extract ΔV_{thp}, ΔV_{thn} and ΔL. The question is how to choose the most suitable set of ROs. The sensitivity matrix plays a major role on defining the robustness of estimation. Angles between the sensitivity vectors are good indicators on how good the sensitivity matrix is for accurate estimation. Figure 3.19 shows the sensitivity vectors for ROs with "STD", "PPASS_O", "NPASS_O", "PRICH" and "NRICH" inverter cells. In Fig. 3.19, "PPASS_O" and "NPASS_O" ROs have large angle between their sensitivity vectors compare to that of "PRICH" and "NRICH" ROs because of their high sensitivities. A quantitative evaluation can be performed by calculating the condition number of the selected ROs.

Figure 3.20 shows the sensitivity vectors of the pass-gate based process monitors. Sensitivity vectors for ROs consisting of inverter, NAND2 and NOR2 gates are also shown. X-axis and Y-axis refer to ΔV_{thn} and ΔV_{vthp} sensitivities respectively. The vector values are normalized with the sensitivity of the standard RO to nMOSFET threshold variation. Standard inverter based RO has similar sensitivities to nMOSFET and pMOSFET variations. NAND2 and NOR2 based ROs do not have enough sensitivity to separate pMOSFET and nMOSFET variation. In the figure, pass-gate based monitors have high sensitivity to nMOSFET or pMOSFET variations.

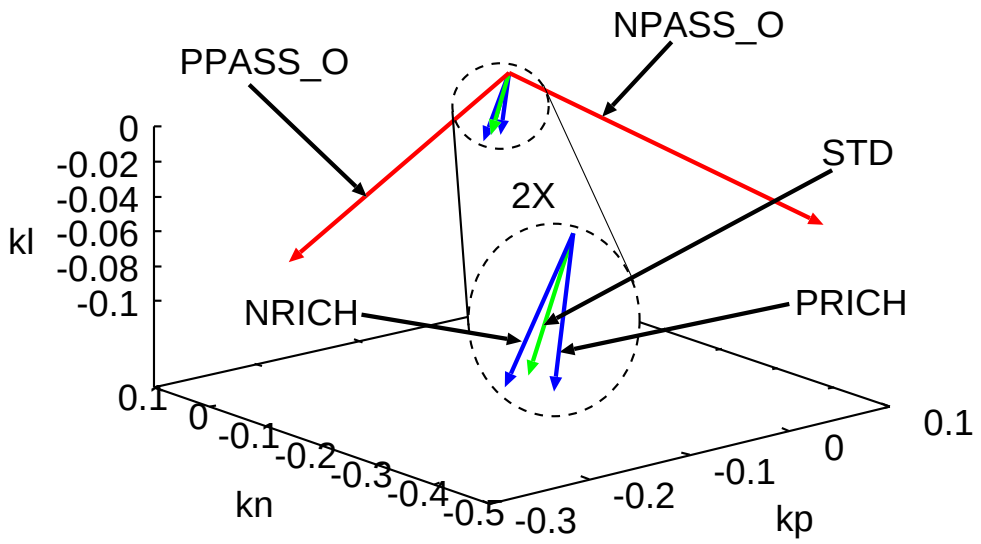


Figure 3.19: Sensitivity vectors of various types of ROs. Sensitivity vectors of "PPASS" and "NPASS" ROs form are near orthogonal referring their robustness in estimation. (©2012 IEEE)

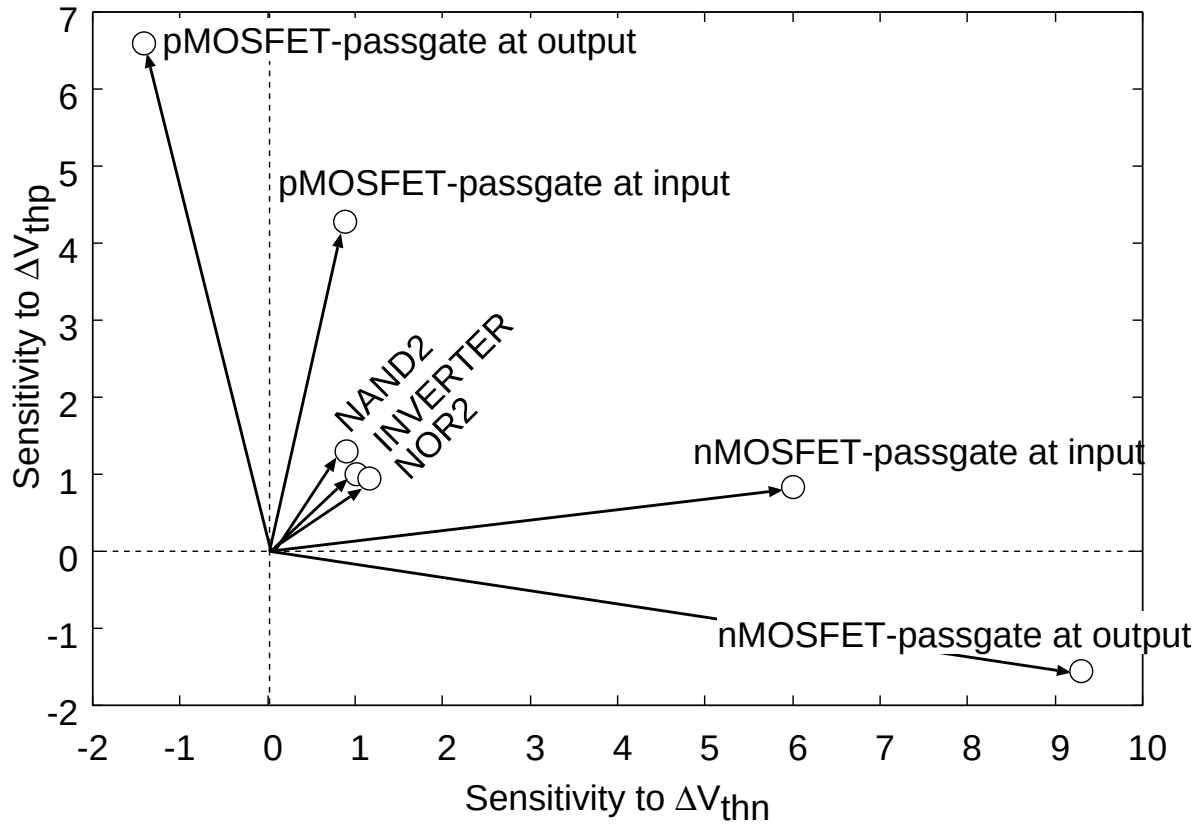


Figure 3.20: Sensitivity vectors of pass-gate based process monitors and RO with standard inverter, NAND2 and NOR2 cells. (©2013 IEICE)

Table 3.2: Condition Numbers of Sensitivity Matrices for different set of ROs. (©2012 IEEE)

No.	RO Set			Condition Number
	RO #1	RO #2	RO #3	
1	STD	PPASS_O	NPASS_O	39
2	CPASS	PPASS_O	NPASS_O	50
3	STD	PPASS_I	NPASS_I	26
4	STD	PLOAD	NLOAD	34
5	STD	PRICH	NRICH	78
6	PRICH	PPASS_I	NPASS_I	28

Condition number is a good indicator on how robust estimation result will be against the uncertainties in sensitivity coefficients or in measurement values. If a matrix has small condition number, the matrix is called a well-conditioned matrix. The condition number of a matrix can be calculated using the infinite norm of the matrix as shown in Eq. (3.13).

$$\text{cond}(\mathbf{A}) = \|\mathbf{A}\|_\infty \cdot \|\mathbf{A}\|_\infty^{-1}. \tag{3.13}$$

Condition number of 1 means that the sensitivity vectors are orthogonal to each other. Bigger the condition number is, smaller the angles between the sensitivity vectors are. The set of ROs having the smallest condition number is most suitable for this estimation technique. Table 3.2 shows condition numbers of the sensitivity matrices for different RO sets. In Table 3.2, "STD", "PPASS_I" and "NPASS_I" ROs have the smallest condition. Thus, "PPASS_I", "NPASS_I" and "STD" ROs are most suitable for the proposed estimation technique.

3.5.3 On-chip Implementation

Figure 3.21 shows an example of the concept of on-chip monitor circuits. Instead of placing single process monitor which is often an RO consisting of standard inverter cells, we propose to distribute the monitor circuits across the chip. Figure 3.22 shows a block diagram where three ROs are used as monitor circuits. The P- and N-monitor circuits are sensitized to P- and N-variations. The proposed "PPASS_I" and "NPASS_I" ROs are suitable to use as P- and N-monitor circuits. RO with "STD" inverter cell can be used as delay-monitor to capture delay deviation. In post-silicon, the ROs are measured and analyzed. The P- and N-monitors give us instant insight on the process condition. The delay monitor is used with the P- and N-monitors for model parameter extractions. These three ROs are then used to estimation global and local variations with the proposed estimation techniques described in Sec. 3.4. For practical purpose, we assume the extraction of these three parameters is enough as they are the dominant factors for performance deviations. Decoders and selectors can be used to select one of the ROs. Dividers can be used to reduce the frequency because the output of each monitor unit is routed globally to a controller. The monitors can also be integrated with on-chip test circuitry such as BIST for

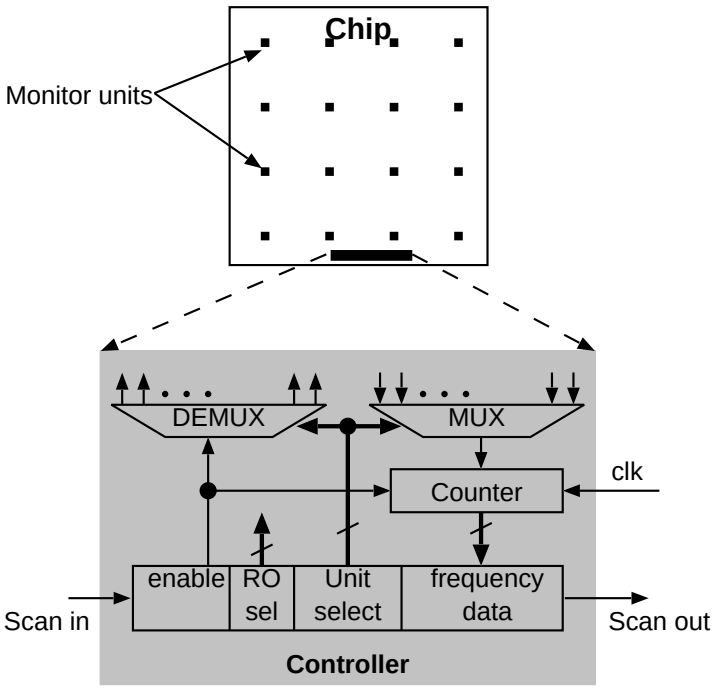


Figure 3.21: One example of on-chip implementation of monitor circuits. Conventional scan-chain based interface is used in this example. (©2013 IEICE)

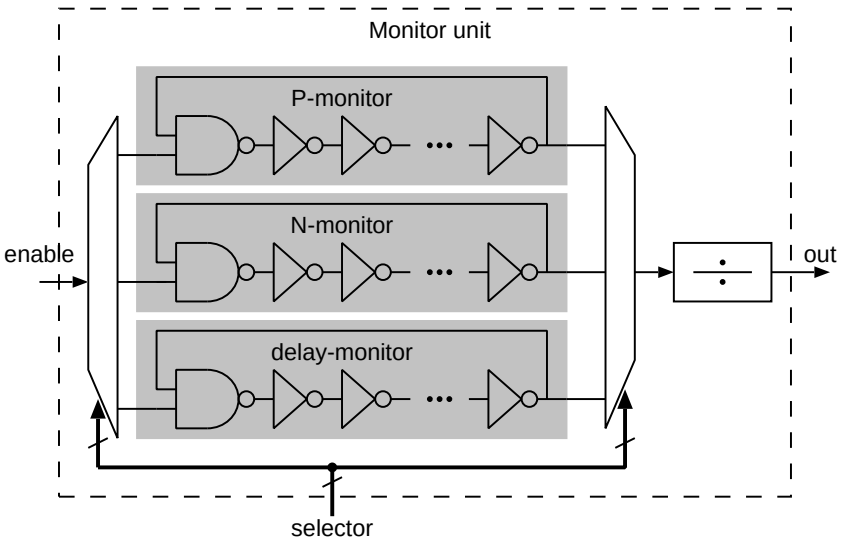


Figure 3.22: One example of monitor unit. The monitor unit consists of three process monitors here to detect process shift and process spread. (©2013 IEICE)

on-chip testing as the output is digital.

3.6 Evaluation of Validity

ROs of "STD", "PPASS_I" and "NPASS_I" inverter cells are proposed as monitor circuits for process parameter estimation. Simulation based experiments have been performed to verify the validity and the robustness of the proposed monitor circuits. Real chip scenario is emulated in

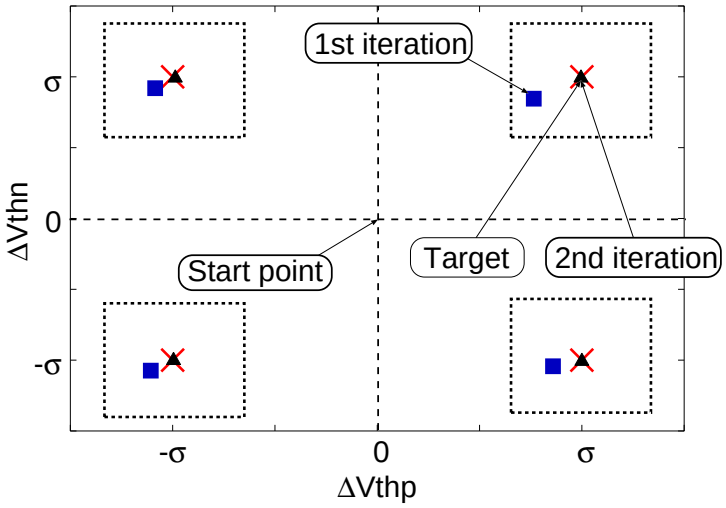


Figure 3.23: Effect of iteration on estimation. Estimation results converge to the target point after several iterations. (©2012 IEEE)

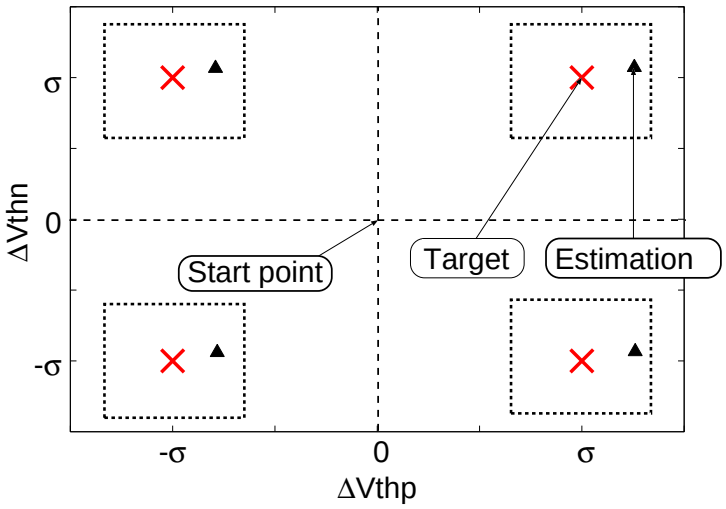


Figure 3.24: Effect of uncertainty such as measurement error in frequency on estimation. Despite of +1% error in each frequency estimation results converge near the target point. (©2012 IEEE)

our simulation.

3.6.1 Experimental Setup

In the experiments, real chip scenario is emulated in the following way. First, we take a transistor model to predict the circuit performances. Values of V_{thp}, V_{thn} and L defined in the model are our reference point or start point. RO frequencies will be predicted from this point. Next, we apply some known amounts of ΔV_{thp}, ΔV_{thn} and ΔL to the transistor model. We call this model as "chip" model because simulation results using this model will be considered as measurement results obtained from the chip. Then, RO frequencies are simulated using the "chip" model.

These frequencies are considered to be the values we can obtain from the chip. Finally, ΔV_{thp}, ΔV_{thn} and ΔL are estimated using the estimation technique described in Section 3.4. ΔV_{thp}, ΔV_{thn} and ΔL are the amount of deviations from the start point. If the estimated ΔV_{thp}, ΔV_{thn} and ΔL match with those in the "chip" model, the estimation becomes correct. In order to emulate measurement errors, some amounts of errors are added to the simulation results obtained from the "chip" model.

3.6.2 Validation of Estimation Technique

The proposed technique has been verified for different values of ΔV_{thp}, ΔV_{thn} and ΔL in the "chip" model. For example, Fig. 3.23 shows the estimation results of ΔV_{thp} and ΔV_{thn} when these values are set to be $\pm\sigma$ in the "chip" model. σ value for ΔL is set in these cases. In Fig. 3.23, X-axis and Y-axis refer to ΔV_{thp} and ΔV_{thn} variation respectively. Cross points refer to the applied ΔV_{thp} and ΔV_{thn} values in the "chip" model. Closed rectangular points refer to the estimation results obtained after the 1st iteration and closed triangular points refer to the results obtained after the 2nd iteration. Regions enclosed by dotted rectangles are used for separating the corresponding estimation results from each other. After the 1st iteration, estimated values of ΔV_{thp} and ΔV_{thn} locate near to the target values but with some amounts of errors. These errors occur from the non-linearity. However, after the 2nd iteration, the estimated values move closer to the target points. Thus, the iterative technique converges and accurate estimations are obtained.

3.6.3 Validation of Robustness

The robustness of the monitor circuits has been verified by applying different error patterns in the frequencies. For example, Fig. 3.24 shows the estimation results when +1% error exists in each of the measured frequencies. Simulation setup and the meanings of the symbols are same as in Fig. 3.23. Closed triangular points are estimation results after the iteration technique has converged. Although some errors are there in the estimations, the important point to note here is that in spite of +1% error in each frequency, estimation results have been converged near the target values. This proves that the proposed circuits are robust for process parameter estimation.

3.7 Test Chip Design

A test chip has been fabricated in a 65-nm process to confirm the validity of our proposed monitor circuits. In this section, the test structure to evaluate the monitor circuits is described. Measurement results and estimation results are discussed next.

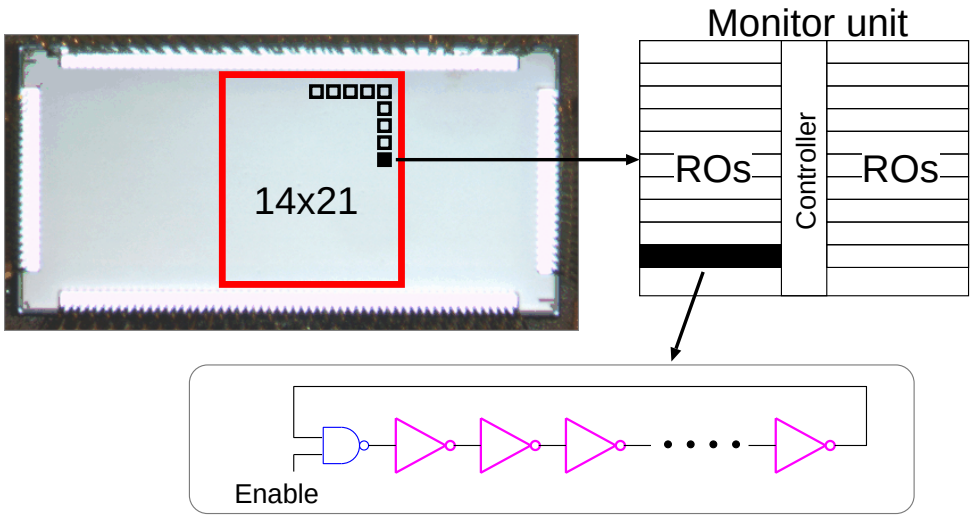


Figure 3.25: Test chip in 65-nm process. (©2013 IEICE)

3.7.1 Chip Design

A test chip in a 65-nm process technology has been fabricated. The process features 1 poly layer, 12 metal layers, copper wiring, and low-κ insulating material techniques. The physical gate oxide thickness is 1.7 nm. ROs of Table 3.1 are implemented in the test chip. In order to evaluate the validity of the monitor circuits, the effect of random variation needs to be evaluated as well. Therefore, an array based test structure methodology proposed in [126] is used to get both D2D and WID variations. Figure 3.25 shows the chip micrograph where 294 sections are integrated into an array of 14×21 onto the chip. Each section contains an instance of a particular type of RO. Therefore, 270 ROs of the same type are integrated in a single die. Figure 3.26 shows the block diagram of our test structure. Selectors and decoders are used to select an RO to oscillate and capture the waveform outside the chip. Local divider and on-chip counter are used to reduce the frequency below 1 MHz so that the waveform does not get distorted outside the chip. Enable signals are generated locally inside the chip to avoid harmonic oscillation [127]. The number of stages for each RO is chosen to a prime number 13 to minimize the probability of harmonic oscillation. We get 294 frequency measurements for a single RO; thus WID variation can be obtained. From this WID variation, we can calculate the number of stages required for a tolerable range of error in the estimation. Global variation is obtained by averaging the 270 measured frequencies. We have 30 chips. So, 30 global values of frequencies are obtained for each RO.

3.7.2 Measurement Procedure

The overall procedure for RO frequency measurement is as follows. First, an RO instance is enabled using the selectors. Then, the total time for a fixed number of oscillation is measured with a resolution of 12.5 ns using an 80 MHz clock signal. Then, the next RO instance is selected

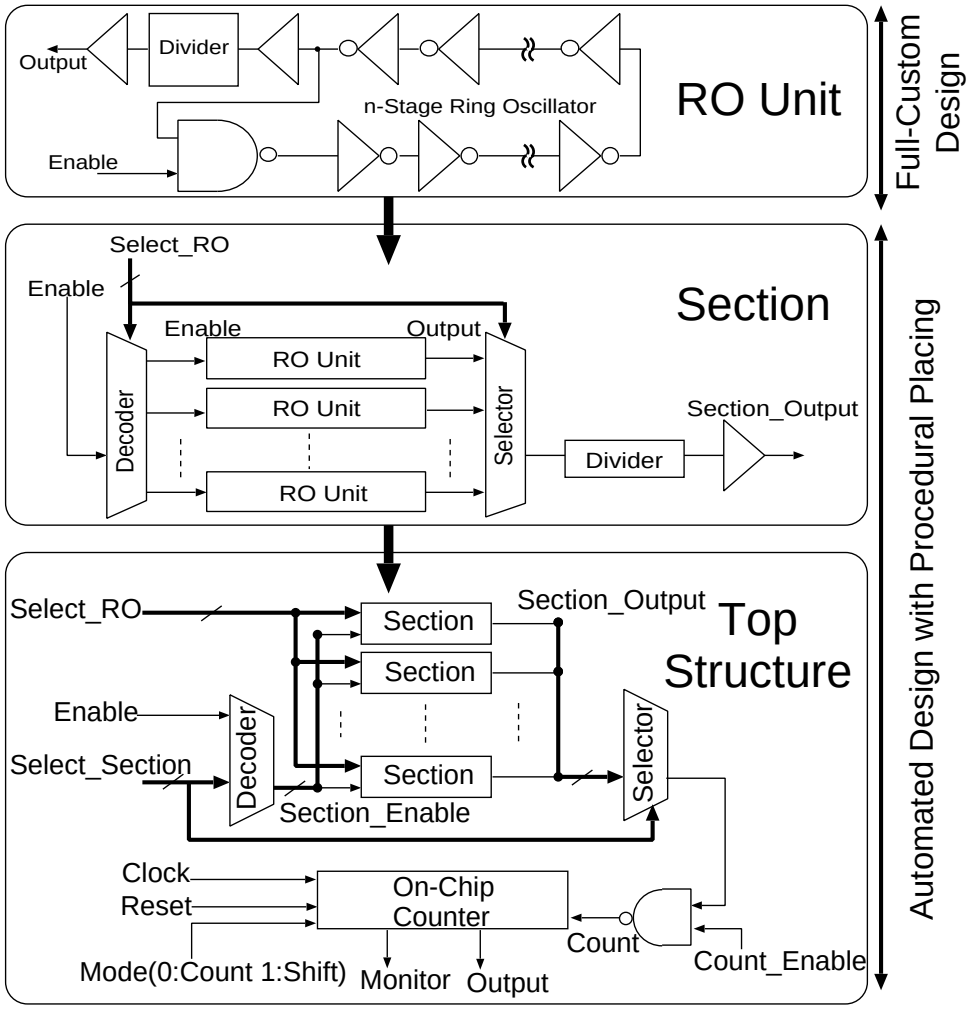


Figure 3.26: Block diagram of test structure. (©2012 IEEE)

and measured. The number of oscillation is set to 1024 in our procedure. The frequency outside the chip is around 1 MHz, maximum error for this procedure is $\pm 1/(1024 * 80) = \pm 0.001\%$ under an ideal condition where there is no fluctuation in environmental parameters such as supply voltage or temperature. However, in real measurement, environmental parameters fluctuate as well as stochastic noises exist. In order to check measurement precision, frequency of the same RO instance is measured 100 times. Standard deviation for the measured 100 frequencies is 0.022%.

3.8 Estimation of Global Variation

In this section, we show that how the measured on-chip frequencies can be used to estimate several process parameter deviations so that correlation between model and silicon can be obtained during the manufacturing process of a product.

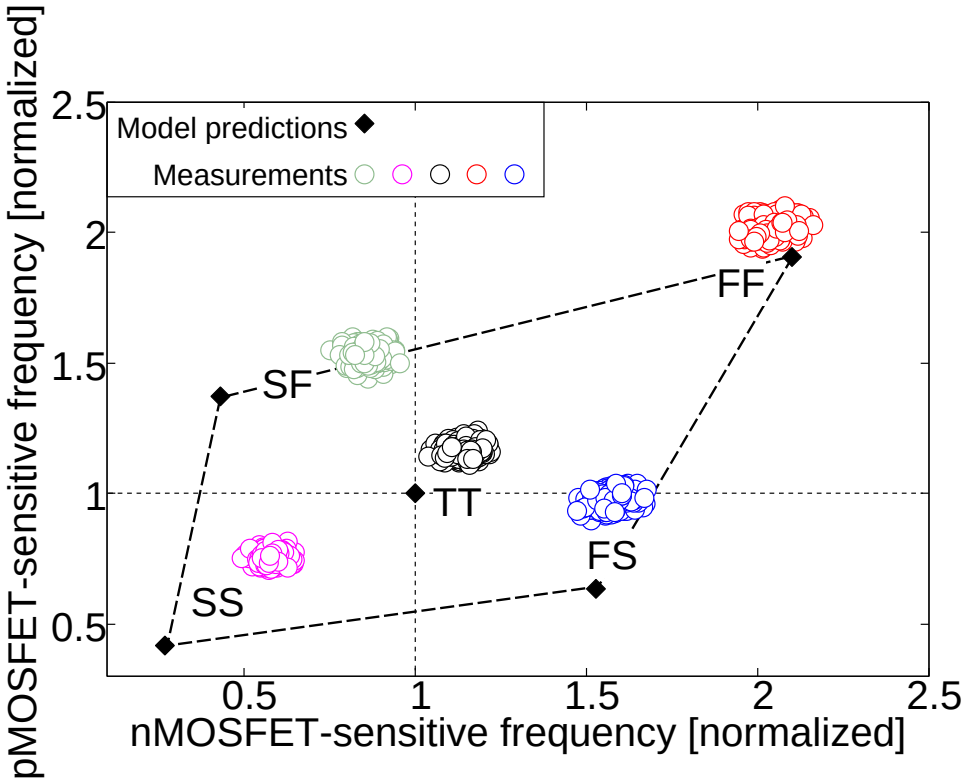


Figure 3.27: Measured monitor frequencies from 5 chips. 5 chips represent 5 process corners. Each chip contains 294 instances of each monitor. (©2013 IEICE)

3.8.1 Measurement Results

Figure 3.27 shows the measured frequencies of "PPASS_I" and "NPASS_I" ROs from 5 chips (open circles). The chips represent five process corners of "TT", "SS", "FF", "FS" and "SF". The values are normalized by the values estimated with the "TT" corner model. Frequency values estimated using the corner models are also plotted in the figure (closed squares). In Fig. 3.27, process shifts from the "TT" model prediction are observed. In Fig. 3.27, deviations are observed between the predicted and measured corner frequencies. Clear deviations are observed for "TT", "SS", "SF" and FS" corners. The silicon values are higher than the model predictions. With comparison with the models, we can have quick understanding of process shift for each chip. This information allow us to take decisions for silicon debug and test pattern generation. By doing further analysis, model-hardware correlation can be performed which allow us to tune the designs. Model-hardware correlation results are presented in the next section.

Table 3.3 shows the measured frequencies from our test chip. Frequencies shown in Table 3.3 are the average values of all frequency measurements from 30 chips. Predictions for RO frequencies using a "TT" (Typical-Typical) transistor model are also shown in the table. In this paper, "TT" model is used as the reference for estimation. Predicted values are compared with the measured values. Positive value of difference refers that measured value is higher than predicted value. Large differences between measurements and predictions are observed for "PPASS_I" and "PPASS_O" ROs. These ROs are highly sensitive to ΔV_{thp}; thus ΔV_{thp} is

Table 3.3: Average values of measured frequencies for 30 chips which are fabricated targeting "TT" corner. Maximum WID variation is shown here. Predicted values for the frequencies and deviation in measurement from the prediction is also shown. (©2012 IEEE)

RO type	Measurement Ave. [MHz]	WID Variation σ/μ [%]	Prediction [MHz]	Deviation [%]
STD	2151	1.42	1907	12.8
PPASS_I	774	2.84	555	39.5
NPASS_I	769	3.78	692	11.1
PPASS_O	443	3.73	244	81.8
NPASS_O	393	6.66	400	−1.78
CPASS	907	1.2	819	10.7
PLOAD	1192	1.47	1114	7.02
NLOAD	1146	1.32	1006	13.9
PRICH	1219	1.29	1141	6.81
NRICH	1199	1.2	1046	14.6

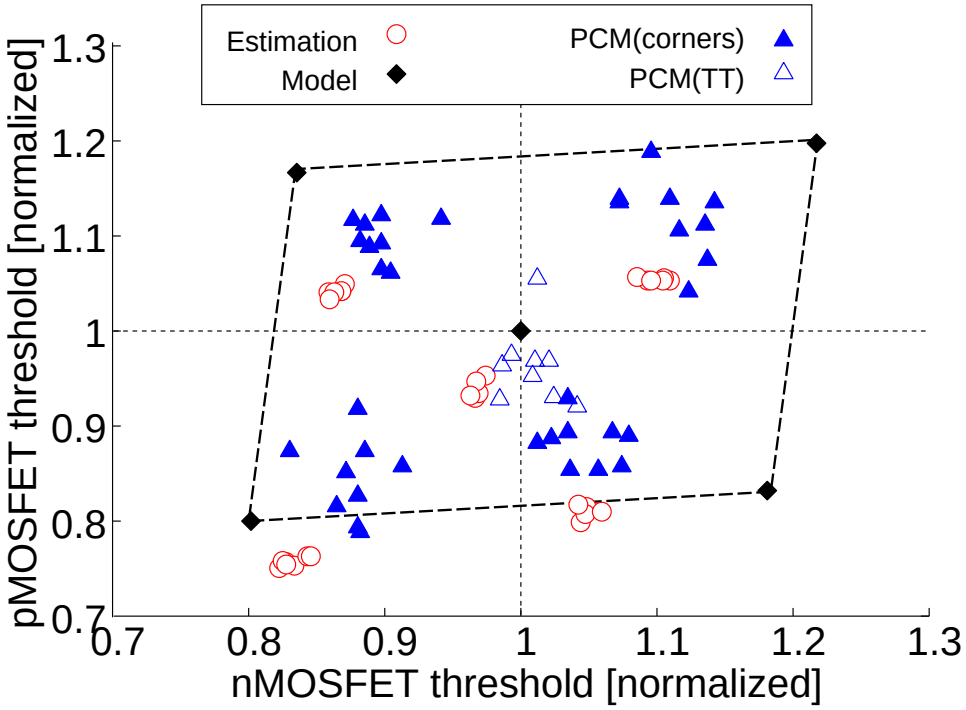


Figure 3.28: Comparison between threshold voltages in corner model and in estimations. PCM data for "TT" corner wafer as well as other corners are also plotted. (©2013 IEICE)

expected to be larger. Maximum amount of WID variation among 30 chips for each RO is also shown in Table 3.3. ROs with higher sensitivities have larger WID variations.

3.8.2 Estimation Results

After detecting the process shift of the product chips, those shifts can be decomposed into the model parameters. Ref. [2] shows a methodology to decompose RO frequencies into several

Table 3.4: Estimation parameter deviation using the monitor circuit measurements against the typical model parameter values.

Corner	delvt0 [mV]		ΔL [nm]
	nMOSFET	pMOSFET	
TT	−17 ~ −12	−32 ~ − 17	−1.8 ~ −0.4
SS	43 ~ 55	7 ~ 34	−0.6 ~ 3.0
FF	−96 ~ −85	−130 ~ −111	−1.4 ~ 1.2
FS	−70 ~ −57	27 ~ 49	−3.5 ~ −1.2
SF	10 ~ 23	−123 ~ −104	1.0 ~ 3.1

key process parameters such as threshold voltages and gate length. An iterative estimation procedure is used in this methodology where the differences between measured frequencies and predicted frequencies are decomposed into model parameters using the sensitivities to each parameter. In this test chip, we have extracted ΔV_{thn}, ΔV_{thp} and ΔL against the values in the "TT" corner model for all the chips using the above method.

WID random variation affects the accuracy of the estimation. As random variation follows Gaussian distribution, the error can be reduced by increasing the number of stages. The relationship between the number of stages and the estimation accuracy is discussed in Sec. 3.8. In our test chip, 294 ROs each having 13 stages are used for global variation monitoring. The number of stages thus becomes $294 \times 13 = 3822$ which we consider to be large enough to cancel the random variation effect. In case of product chips, however, the number of stages is limited. The number need to be chosen based on the amount of random variation and the desired tolerable range of estimation. The adequate number of stages for the monitor circuits are calculated for a fixed tolerable range of estimation error due to WID variation using Eq. (3.10). From Table 3.3, we get the WID variations for each RO. Using these variations, the number of stages are calculated to be 171 when the standard deviation for threshold voltage estimation is set to 2mV.

In this extraction, ΔV_{th} is expressed by the transistor model parameter "delvt0" which is a dedicated parameter for modeling threshold voltage shift in HSPICE [123]. The values are shown in Table 3.4. Here, negative values of threshold voltage deviation refers that the absolute value is lower than that in the "TT" model. One key characteristics derived from the real chip measurement is that the gate length does not vary much compare to the variations in the threshold voltages. Large threshold voltage shifts are observed for the "FF" corner chip.

Figure 3.28 shows the threshold voltage corner in the model and the estimated corner from the RO frequencies. Threshold voltage values from the PCM data of "TT" corner wafer as well as other corners are also plotted. In this lot, some deviations are observed between extracted values and those by the corner models. Some amounts of deviations are also observed between the estimated values and those in the PCM data. These deviations can be caused by a number of factors. One possible reason can be the difference between the layouts in the PCM circuits and our test circuits as differences in poly and diffusion densities can affect MOSFET performances.

Table 3.5: Comparison between measurements and predictions for RO frequencies for a chip. Predictions are made using the estimated ΔV_{thp}, ΔV_{thn} and ΔL. (©2012 IEEE)

RO	Measurement[MHz]	Prediction[MHz]	Difference[%]
STD	2145	2145	0.0
PPASS_I	753	753	0.0
NPASS_I	766	766	0.0
PPASS_O	421	395	−6.0
NPASS_O	398	399	0.3
PLOAD	1189	1224	2.9
NLOAD	1135	1158	2.0
CPASS	901	915	1.6
PRICH	1216	1270	4.4
NRICH	1187	1181	−0.5

3.8.3 Validation

As transistor variation cannot be measured directly for the test chip, the biggest challenge of this work is to show the validity of the estimation results that is showing the estimation accuracy for the real chips. Two methodologies have been used to show the validity of the estimation results. First is to predict circuit performances consisting of various types of logic gates with the estimation parameter variations. Second is to apply external body bias to the chip and emulate threshold voltage shift. Validation results are described in this section.

Predictability of Circuit Performance

Predictions of circuit performances can be made using the estimation results for each chip. Close match between predictions and measurements for all circuits will confirm the validity of the estimation results. Predictions are made for our RO frequencies using estimated ΔV_{thp}, ΔV_{thn} and ΔL for 30 chips. Table 3.5 shows the mismatch between predictions and measurements for a particular chip. For the top three ROs in the table, no mismatch is found because these ROs are used for the estimation. The key point is whether the predictions for circuits other than the ones used for the estimation match closely with the measurements. In Table 3.5, predictions match with the measurements within maximum mismatch of 6% which is small compare to the differences in Table 3.3. Thus, the circuit performances can be predicted with high accuracy using the estimated values. Therefore, the proposed technique can be used for post-silicon tuning.

Different Body Bias Condition

Threshold voltages can be changed by applying body biases to the chip. So, the monitor circuits can be validated by estimating process variations at different bias conditions. If the estimated values correlate to the applied body bias values, then the monitor circuits will be proved to be

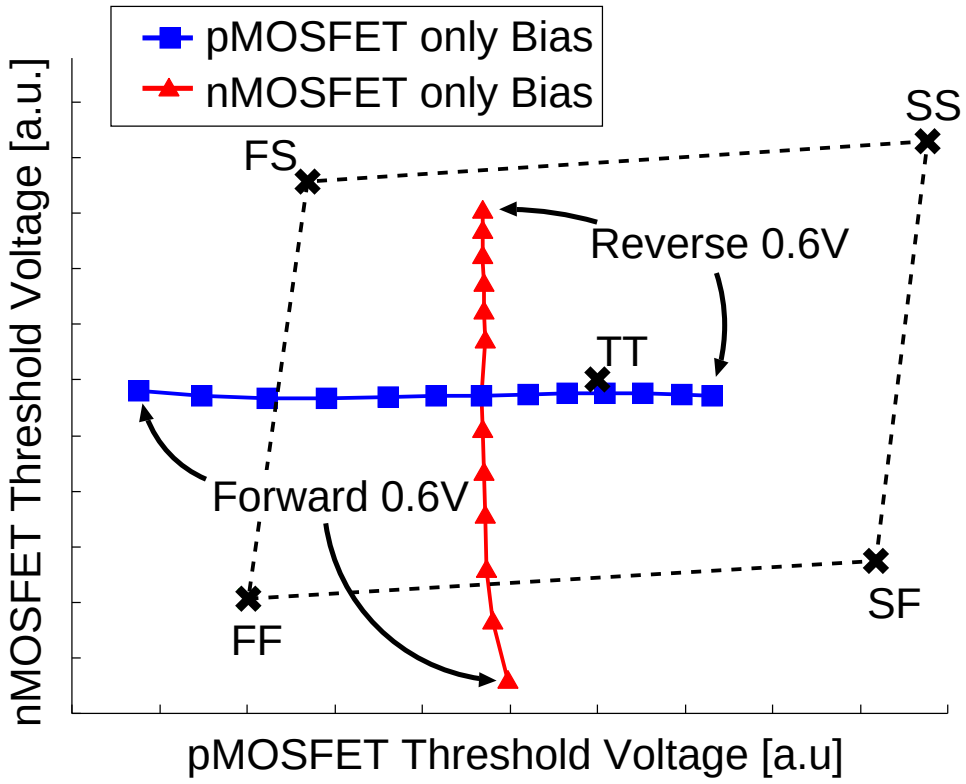


Figure 3.29: Estimation of V_{thp} and V_{thn} at different bias conditions. Threshold change is detected properly with the proposed monitor circuits. (©2012 IEEE)

valid for correct monitoring of process variations.

Figure 3.29 plots the values of ΔV_{thp} and ΔV_{thn} estimated at different body bias conditions for a particular chip. In Fig. 3.29, X-axis refers to ΔV_{thp} estimation and Y-axis refers to ΔV_{thn} estimation. Rectangular points are estimated values of ΔV_{thp} and ΔV_{thn} when only pMOSFET is biased. Triangular points refer to estimated values of ΔV_{thp} and ΔV_{thn} when only nMOSFET is biased. When only pMOSFET is biased, the estimated point moves in horizontal direction referring only ΔV_{thp} is changed in the estimation. When only nMOSFET is biased, the estimated point moves in vertical direction referring only ΔV_{thn} is changed in the estimation. Thus, it is proved that any change in the threshold voltage can be detected correctly by the proposed monitor circuits.

Standard deviation of estimated ΔL values at different body bias conditions are calculated as 0.7% which is small. So, ΔL estimation remains the same at different bias conditions referring to the validness of the monitor circuits.

3.9 Estimation of WID Variation

In order to estimate WID variation of threshold voltage and gate length, frequency distributions are measured for the monitor circuits. Then the measured standard deviations are decomposed into the parameter variations. Measurement and estimation results are described here.

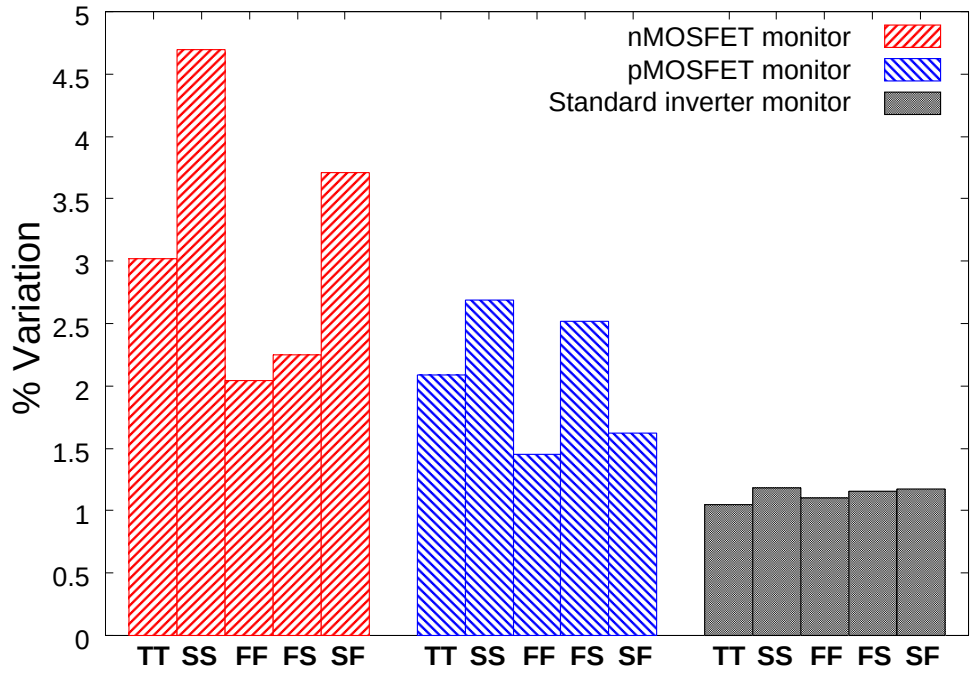


Figure 3.30: WID variation observed in nMOSFET monitor, pMOSFET monitor and standard inverter ROs at the corner chips. (©2013 IEICE)

3.9.1 Measurement Results

In Fig. 3.27, 294 measured values for P- and N-monitors are plotted for each chip. Variation among the frequencies in a chip is observed which represent the process spread for that chip. The amount of spread can be different from chip to chip depending on the location of the chip in the process space. In order to evaluate the amount of spread, we have calculated the standard deviations for the frequencies in a chip. The results are shown in Fig. 3.30. The frequency variation is the function of the sensitivity coefficients and the amount of process variation. As P- and N-monitor ROs have larger sensitivities than the standard inverter RO, they are showing larger variations in Fig. 3.30. The difference in the variations between the "PPASS_I" and "NPASS_I" ROs also reflects the difference in the pMOSFET and nMOSFET variations of the chip. From standard inverter RO measurements, similar WID variation is observed for all the corner chips. However, for the "PPASS_I" and "NPASS_I" ROs, significant differences in the amount of spread between the chips are observed. For "FS" corner chip, nMOSFET monitor's variability becomes smaller and pMOSFET monitor's variability becomes larger than the "TT" corner chip. This indicates that the intrinsic variability in the nMOSFET and pMOSFET performances are different in the unbalanced corners which may cause the yield to decrease drastically [128]. The extend of process spread needs to be accurately monitored and feedback into the design. Our monitors are suitable for distinguishing nMOSFET and pMOSFET variations. In order to model the effects of WID variations, the variation needs to be expressed by transistor model parameters so that designers can use them. In Sec. 3.9.3, obtained WID variations are decomposed and modeled into variations of three transistor model parameters of threshold voltages and gate length.

Table 3.6: Extracted standard deviation of MOSFET threshold voltages and gate length from RO frequency measurements. (©2013 IEICE)

Corner	$\sigma_{V_{thn}}$ [mV]	$\sigma_{V_{thp}}$ [mV]	σ_L [nm]
TT	16.6	11.9	0.89
SS	18.3	14.5	0.53
FF	20.9	16.6	1.14
FS	18.2	13.3	0.99
SF	18.2	13.6	0.99

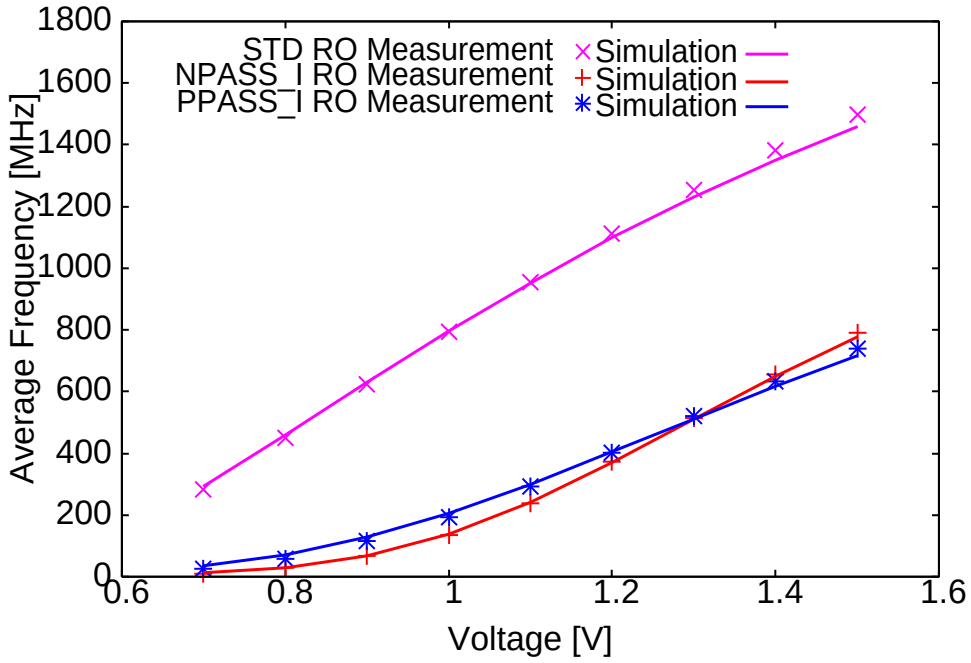


Figure 3.31: Measured average frequency and simulated frequency of homogeneous ROs. Simulation is performed at the estimated center point in the process corner. (©2013 IEEE)

3.9.2 Estimation of Center Point

The center values of V_{thP}, V_{thN}, and L are estimated from the average frequency values for the three ROs of "PPASS_I", "NPASS_I", and "INV-STD". Figure 3.31 shows the simulated frequencies of the three ROs at the estimated center points of V_{thP}, V_{thN}, and L and measured frequencies. Simulated frequencies match with the measured frequencies at the supply voltage range of 0.7 to 1.5 V, confirming the validity of the estimation. The estimated center points are used to build linear models for extraction of WID variations.

3.9.3 Extraction of WID variability

Using the sensitivity coefficients, the standard deviations of V_{thp}, V_{thn} and L are derived from the measured standard deviations in Fig. 3.30. Derived values are shown in Table 3.6. nMOSFET variation is larger than that of pMOSFET. The extracted amounts of variations for nMOSFET

Table 3.7: Delay mismatch between silicon and model prediction for NAND2 and NOR2 delay paths. Delays between silicon and prediction match close when process calibration is done with the estimated process parameter shifts. (©2013 IEICE)

	NAND2 delay mismatch [%]		NOR2 delay mismatch [%]	
Process Corner	w/o calibration	w/ calibration	w/o calibration	w/ calibration
TT	7.6	1.0	8.7	0.6
FS	9.7	0.0	5.8	0.8
SF	7.7	1.3	14	0.9
SS	-13	1.1	-12	0.5
FF	33	1.2	36	1.1

and pMOSFET threshold voltage match closely with the values reported in Ref. [66] where a similar 65-nm process is used. These information are extremely useful for statistical design, yield analysis and post-silicon timing analysis. In product chips, limitations in the numbers of samples may cause some errors in the estimations. These limitations need to be considered during practical applications.

3.10 Application

In this section, several applications are discussed which can be benefited from on-chip monitor circuits.

3.10.1 Performance Prediction

Predicting the performance of the designed circuit in real silicon is difficult under the presence of large process variation. Depending on the location of the chip in the process space, the critical paths change which affects the maximum operating frequency. The effect of process variation is more severe for analog and RF circuits. For analog and RF circuits, the mismatch between pMOSFET and nMOSFET may cause the circuits to fail. In this section, we show that how on-chip monitors that gives us information on the process shift can be useful for predicting performances for both the digital and analog circuits.

Digital Circuit

Post-silicon statistical processing can be used to identify faulty paths and predict the delays of paths [129]. Process calibration is needed for post-processing. Specific information of key process parameters are required for process calibration. For example, in SSTA, the delay, d of a path is modeled using the sensitivity coefficients as follows.

$$d = \mu_d + \sum k_{p_i} \Delta p_i + N(0, \sigma^2_{\text{rnd}}). \tag{3.14}$$

Here, μ_d is the mean value, Δp_i is the i-th parameter that has variation and k_{p_i} is the sensitivity coefficient to parameter p_i. $N(0, \sigma^2_{\text{rnd}})$ is the random component which is the result of random variation. Process parameters can be extracted with on-chip monitor circuits. After the process parameters are known for a chip, the delay can be calculated using Eq. (3.14). The accuracy of the estimation of Δp_i in Eq. (3.14) holds the key for accurate timing analysis. Here, we show that the delays of different paths can be predicted with high precision using the estimated parameters.

Delays of two path types are tested on silicon. The first path consists of NAND2 cells and the second path consists of NOR2 cells. We have 294 instances of identical paths across the chip to evaluate mean delay value and the standard deviation. Table 3.7 shows the amount of mismatch between predicted mean delay values and real silicon mean delay values for 5 corner chips. Mismatches are shown for two cases. The first case is when process calibration is not considered. The second case is when process calibration is performed. When the process calibration is not considered, the prediction is done with the "TT" corner model provided by the foundry. As expected, large mismatches are found for the path delays when there is large variation which is represented by the corner chips here. Next, process calibration is performed by considering the parameter shifts from the "TT" model. In Table 3.7, the predictions with the estimated parameters and silicon values match closely with maximum error of 1.3%. This proves the validity of the monitor circuits and the estimation procedure. Thus, the monitor circuits are extremely helpful for post-silicon timing validation and performance prediction.

Analog Circuit

On-chip monitors can be useful to predict analog circuit performances. For analog circuits mismatch between transistors and P/N-ratio is important for deciding circuit parameters. If the process variation is not controlled very well, large amount of yield loss can happen. In [130], use of on-die monitor circuits are proposed to identify process variations. The information of process variations are then used to guide the test and to allow the estimation of selected performance figures. In their approach, replica of several parts of the DUT is used for monitor circuits. The use of replica of DUT parts gives us direct information on the DUT performance, but it is not useful for debugging the cause of failure. For example, when the DUT fails, the reason can be either unbalanced P/N-ratio or excessive random variation causing neighborhood transistor mismatch or manufacturing defect. P/N-monitors give us information on P/N-ratio and help us debugging the causes.

3.10.2 Model Mismatch Detection

In analog circuit, often wider gate lengths are used to get better linearity. However, using wider gate lengths have risks of large mismatch between transistor model and silicon. Characterizing single transistor is expensive and data processing takes time. Often we are interested on the key parameters such as threshold voltage and ON-current. The estimation method using the

proposed set of monitor circuits can be used for the detection of model mismatch. Monitor circuits need to be designed with the same gate length as used in the DUT. In post-silicon, the monitor circuits will be measured and the measured values will then be used for the estimation of threshold voltages and gate lengths. The estimated parameters will give us the information on model mismatch. This method is particularly useful when the design is implemented on multiple fabs and processes.

3.10.3 Adaptive Testing

In the scaled technology, any path in a design has the potential to be critical depending on the process condition. Testing each path is impossible in today's SoCs where there are millions of paths. Delay test takes time and thus smaller the number of paths to be tested, smaller the test cost is. In order to reduce test cost as well as increase product quality adaptive testing is very attractive [131, 132]. In adaptive testing, instead of using a fixed test strategy and fixed test patterns, they are changed over time based on the information from the products. On-chip monitor circuits come into play a very important role here as they provide information of process parameters for each product chip.

There are various approaches of adaptive testing. One use of the on-chip monitor circuits which tell us the location of the chip in process space is for parametric fault testing. In [132], an adaptive test flow is proposed to detect parametric fault for SSTA based designs. The idea is to cluster the critical paths for several process conditions and generate test patterns for each process condition during the design phase. During the test phase, the monitor circuits are measured first. From the monitor circuit outputs, the process condition of the target chip is identified and the corresponding test patterns are used for path delay testing. This approach can save a large amount of test time which in turn saves test cost. The discussed on-chip monitor circuits are ideal for this kind of application.

3.11 Summary

In this chapter, on-chip monitor circuits for detection of process variation for variability modeling, post-silicon diagnosis, process characterization and model-hardware correlation are proposed. Special inverter topologies are used to make the ring oscillator frequency sensitive to either nMOSFET or pMOSFET variation. The proposed monitors embedded in the product chips, enable quick detection of transistor variation in the transistor model parameters such as threshold voltage and gate length. Extraction techniques for model-hardware correlation are presented. Global and WID variations in key transistor model parameters are successfully derived from chip measurements for several process corners designed in a 65 nm process. Predictions of performances have been made for various types of circuits using our estimated amount of variations. Predicted values match closely with the measured values referring to the validness of the estimation technique. The monitor circuits are also verified at different body bias

conditions; thus the validity of the monitor circuits for process parameter monitoring is confirmed. The proposed monitor circuits can be used for post-silicon diagnosis of parametric fault and performance mismatch of digital and analog circuits. The monitor circuits are also suitable for adaptive testing based on process condition. The proposed monitor circuits can be used in post-silicon compensation techniques and model-to-hardware correlation.

Although the proposed monitor circuit topologies give us accurate variation models, large number of samples for each of the RO need to distributed onto the chip. In order to obtain models for systematic and location correlated variations as well, compact monitor circuit architecture is required. Thus, monitor circuit which can be placed on different locations without large area overhead need to be developed. Chapter 4 of this thesis develops an area-efficient monitor circuit architecture where various types of variations are monitored with the single monitor circuit instance.

Chapter 4

Topology-Reconfigurable Universal On-chip Monitor Circuit

In the previous chapter, monitor circuit topologies suitable for D2D and WID variation extraction are proposed. The conventional way of putting the monitor circuits across the chip consumes lot of area. In this chapter, a topology-reconfigurable monitor circuit structure will be proposed for area-efficient implementation of monitor scheme. By reconfiguring the monitor circuit, various topologies for monitor stage and monitor circuit structure can be realized giving different sensitivities to variation.

4.1 Introduction

Depending on the target variation, the implementation strategy of monitor circuits differ. In order to monitor global variation, the effect of random variation need to be canceled out. This can be achieved by making the number of logic stages in the circuit bigger. In order to monitor random variation, large number of instances of the same circuit needs to be implemented on chip. The circuit for random variation monitoring should have high sensitivity to random variation. This is achieved by making the number of stages small. Random variation is obtained by calculating the standard deviation of measurements from large samples. Thus, monitoring of global and random variation requires two different types of implementation. Although global variation can be obtained by averaging the large samples measured for random variation monitoring, monitor circuits with large number of stages are required for adaptive performance compensation.

In order to predict circuit performance accurately and debug the causes of failure, transistor-by-transistor characterization under switching condition is required. Monitor circuit topology where same inverter topology is used for each of the stages do not have the capability to give an insight on transistor-by-transistor variability. An inhomogeneous structure is proposed in an attempt to characterize dynamic variation such as RTN at the device level [5]. However, the existing techniques have two fundamental problems. One is the large area required for large

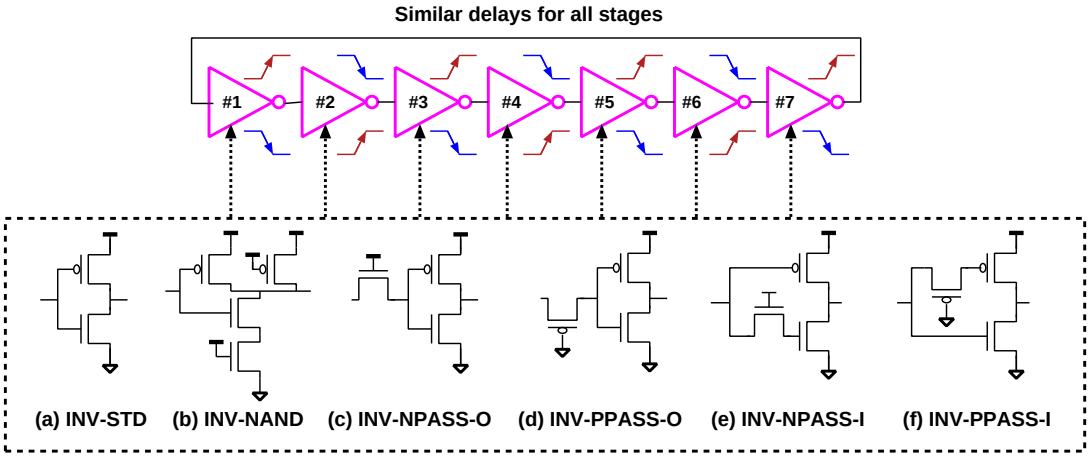


Figure 4.1: A conventional seven-stage RO structure where the same type of inverter structures are used for all stages. As the RO oscillation period is the sum of each inverter stage delay, variation in a particular stage is not directly visible. (©2013 IEEE)

samples. The other is the inability of device identifications for accurate characterization and modeling. In order to overcome these problems, we propose a reconfigurable RO structure where each inverter stage can be configured to several structures. Large measurement samples can be obtained from a single RO by reconfiguring the RO. The proposed structure has the following key advantages over the conventional structure.

1. Runtime monitoring of nMOSFET and pMOSFET performance
2. Statistical evaluation from a single monitor instance
3. Variation of each transistor can be characterized allowing accurate characterization of variation

4.2 Inhomogeneous RO Structure for Variability Enhancement

The topology-reconfigurable monitor circuit uses the concept of inhomogeneous structure. The reconfigurability allows implementing multiple inhomogeneous structure. Before presenting the topology-reconfigurable circuit structure, we first describe the concept of the inhomogeneous structure [5] in this section. Suitable inverter topology for creating the inhomogeneous structure is explained.

4.2.1 Basic Concept

Figure 4.1 shows a conventional RO structure where an identical inverter structure is used at all stages. Inverters of different topologies are used to capture variation and process characteristics.

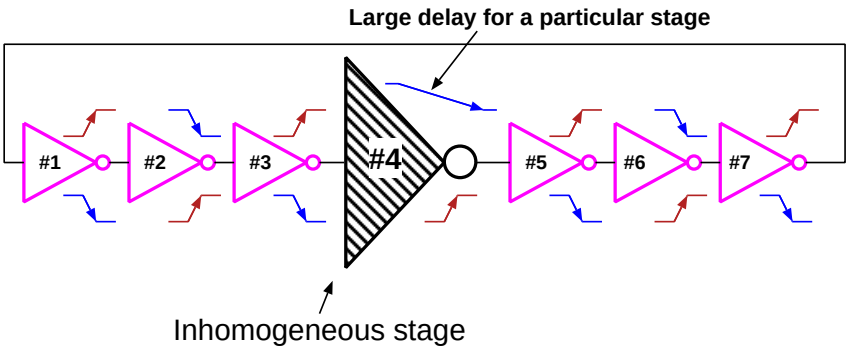


Figure 4.2: Proposed inhomogeneous RO structure. A particular stage is designed to have a large enough delay compared to other stages so that the output frequency is a strong function of that particular stage's delay. Any variation in the inhomogeneous stage becomes directly visible to the output frequency. (©2013 IEEE)

In such cases, the characteristics of all the transistors are averaged out and only one output observable parameter that is the RO frequency is measured. This frequency gives us a general idea of the underlying process characteristics. In order to capture WID random variability, multiple instances of the same RO type are measured. The distribution of the RO frequency gives us the amount of random variation, but fails to give us information on the underlying causes of variation. Several modified inverter structures are proposed in [2, 120] to extract detailed information on the underline variation causes such as threshold voltage and gate length. Even though we can get considerable information from the conventional homogeneous ROs, it is difficult to characterize dynamic variations like RTN, as these variations occur at the transistor level.

In this section, an inhomogeneous RO structure is proposed that enhances the sensitivity of RO frequency to a particular inverter stage delay multiple times more than its homogeneous counterpart. Furthermore, the number of sensitive transistors can be reduced to a small number, meaning transistor-by-transistor characteristics become possible. This enables easy characterization and modeling of complex phenomena like RTN.

The basic concept is as follows. If the delay of a particular inverter stage or a transistor in a particular inverter stage is the dominant delay factor for the overall delay, the RO frequency becomes a strong function of that particular stage. Any variation in the inhomogeneous stage is directly visible to the output frequency. Thus, variability of the transistors in the inhomogeneous stage is enhanced. Figure 4.2 shows an RO circuit with an inhomogeneous element. When a standard inverter, as shown in Figure 4.1(a), is used for the inverter stages, it becomes a conventional homogeneous RO. If a pass-transistor loaded inverter, as in Figure 4.1(c–f), is used for a particular stage and standard inverter for the remaining stages, it becomes an inhomogeneous RO. As will be shown later, the pass-gate loaded inverter stage becomes the dominant factor for the overall variation.

First, the concept of inhomogeneous structure and its effect are explained with the following simple approximations. Let the rise and fall delay of the i-th stage in the conventional RO be

T_{rise_i} and T_{fall_i}. The period of the RO oscillation T_{total} becomes as

$$T_{\text{total}} = \sum_{i}^{N} \left(T_{\text{rise}_i} + T_{\text{fall}_i}\right). \tag{4.1}$$

As shown in Figure 4.2, the inverter cell of a certain stage is replaced by a special inverter cell whose fall delay is much larger than the other inverter cells. The period then can be rewritten as

$$T_{\text{total}} = \sum_{i}^{N-1} \left(T_{\text{rise}_i} + T_{\text{fall}_i}\right) + T_{\text{rise}_s} + T_{\text{fall}_s}. \tag{4.2}$$

$$1 = \frac{T_{\text{others}}}{T_{\text{total}}} + \frac{T_{\text{fall}_s}}{T_{\text{total}}}. \tag{4.3}$$

where, $T_{\text{others}} = \sum_{i}^{N-1} \left(T_{\text{rise}_i} + T_{\text{fall}_i}\right) + T_{\text{rise}_s}$. If the fall time of the inhomogeneous stage is the dominant delay component in the oscillation period, then Equation (4.3) can be approximated as follows.

$$1 \approx \frac{T_{\text{fall}_s}}{T_{\text{total}}}. \tag{4.4}$$

Thus, the period becomes directly proportional to the fall delay of the inhomogeneous inverter. In the case of a homogeneous RO, the effect of delay variation in a particular inverter stage is reduced by the number of stages. In the case of an inhomogeneous structure, the effect of delay variation translates directly to the RO oscillation period; thus, larger sensitivity can be obtained as compared to the homogeneous structure.

4.2.2 Design of Inhomogeneous Element

In order to implement the idea of the inhomogeneous RO, inverter structures capable of creating inhomogeneity are required. In the previous chapter, inverter topologies are discussed to realize ROs with different sensitivity. For example, all the inverter cells in Figure 4.1 are replaced by an inverter loaded with an nMOSFET pass-gate. We name this type of RO homogeneous "INV-NPASS-O" RO. Sensitivity of RO frequency to each transistor's threshold voltage variation is shown in Figure 4.3. As expected, similar sensitivities are observed for the transistors in each inverter stage.

We now show the effect of inhomogeneity on sensitivity. Figure 4.4 shows a seven-stage inhomogeneous "INV-NPASS-O" RO structure. Figure 4.5 shows the sensitivity coefficient of each transistor in the RO for two different supply voltages of 1.2 V and 0.8 V. The pass-gate and the MOSFETs in the inhomogeneous stage have very high sensitivity compared to the others. At a nominal voltage of 1.2 V, the sensitivities of the pass-gate and the nMOSFET are 18 times larger and the sensitivity of the pMOSFET is seven times larger than the other transistors. This is because the voltage drop across the pass-gate increases the sensitivity of the pass-gate and the nMOSFET. The sensitivity of the pMOSFET increases because it remains partially on during

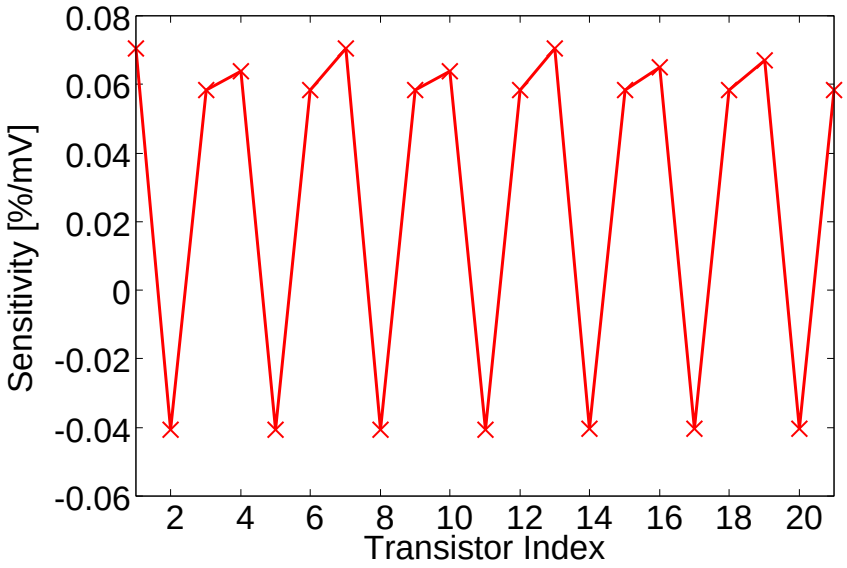


Figure 4.3: Sensitivity of each transistor in a seven-stage homogeneous RO with conventional nMOSFET pass-gate loaded inverter. Similar sensitivities are observed for the MOSFETs in each inverter stage. (©2013 IEEE)

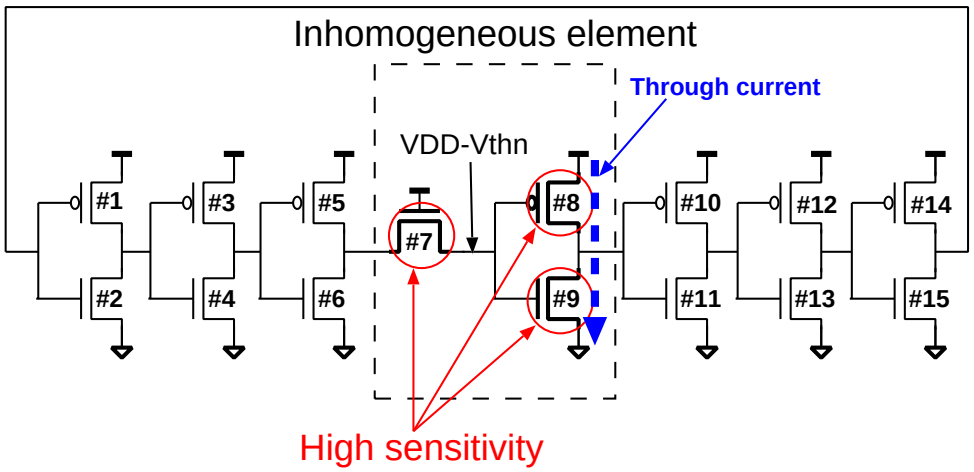


Figure 4.4: Conventional pass-gate loaded inverter structure to create inhomogeneous element. RO frequency is sensitive to the three transistors in the inhomogeneous stage. (©2013 IEEE)

the pull down, where it contributes largely to the delay. Reducing the supply voltage will increase the sensitivities of these MOSFETs drastically, as MOSFET performance becomes more sensitive with reduced gate over-drive. In Figure 4.5, sensitivities of these transistors become more than 100 times as high as the others. With a pMOSFET-inserted inhomogeneous RO, the sensitivities are more than 40 times larger. Performance of inhomogeneous ROs is strongly affected by the variability of dominant transistors. This shows that the local variability of sensitive transistors can be estimated by measuring the inhomogeneous RO's frequency variability.

With a conventional pass-gate inverter structure as shown in Figure 4.1(c) and Figure 4.1(d), we can achieve more than 40 times greater sensitivity for three transistors in the inhomogeneous

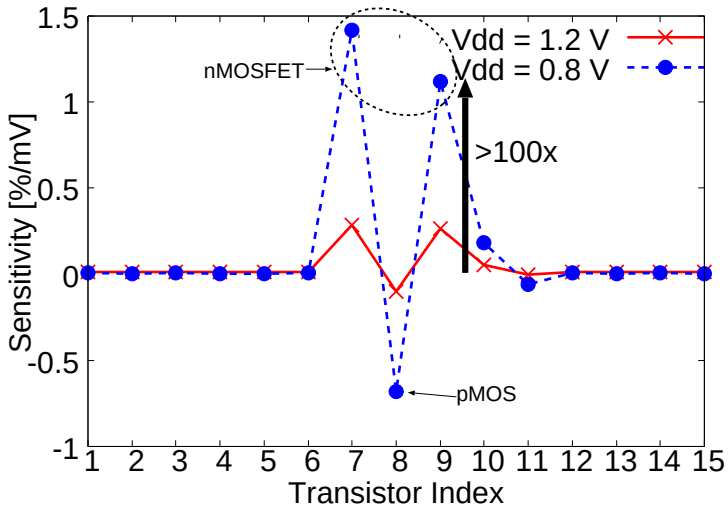


Figure 4.5: Sensitivity of each transistor in a seven-stage inhomogeneous "INV-NPASS-O" RO of 4.4. (©2013 IEEE)

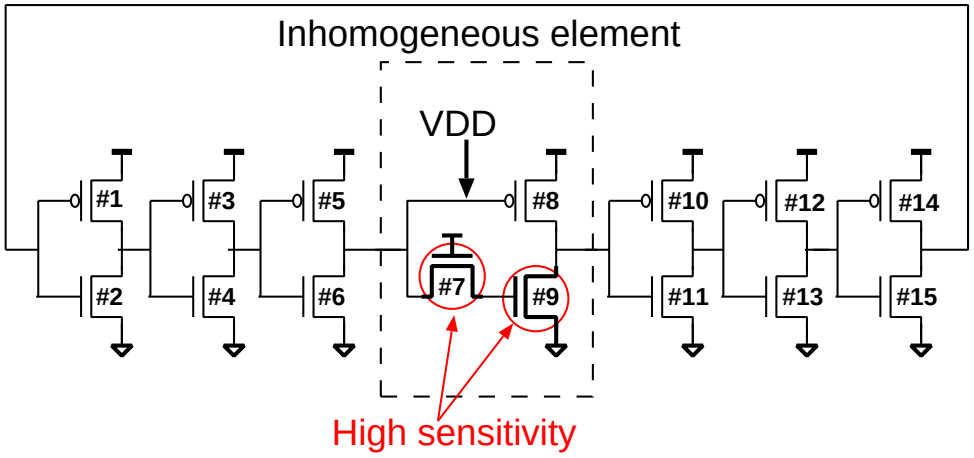


Figure 4.6: Proposed pass-gate-based inverter structure for inhomogeneous element. RO frequency is sensitive to the nMOSFETs of the inhomogeneous stage only. (©2013 IEEE)

stages. These three transistors include both the pMOSFET and nMOSFET, which is not desirable when we want to observe variation on a particular type of device. We, therefore, use the structure shown in Figure 4.6. Figure 4.6 shows a seven-stage inhomogeneous RO with the inverter cell named "INV-NPASS-I". We call this RO as inhomogeneous "INV-NPASS-I" RO. Figure 4.7 shows the sensitivities for each transistor in the RO for supply voltages of 1.2 V and 0.8 V. Only the pass-gate and the other nMOSFET in the inhomogeneous stage have higher sensitivity as compared to the "INV-NPASS-O" structure. More than 50 times sensitivity is achieved for the nMOSFETs only, thus characterization on the nMOSFET becomes possible. The pMOSFET-sensitive counterpart can be designed similarly. We name this RO the "INV-PPASS-I" RO. We omit discussion of "INV-PPASS-I" RO as the operation is just the opposite of the nMOSFET counterpart.

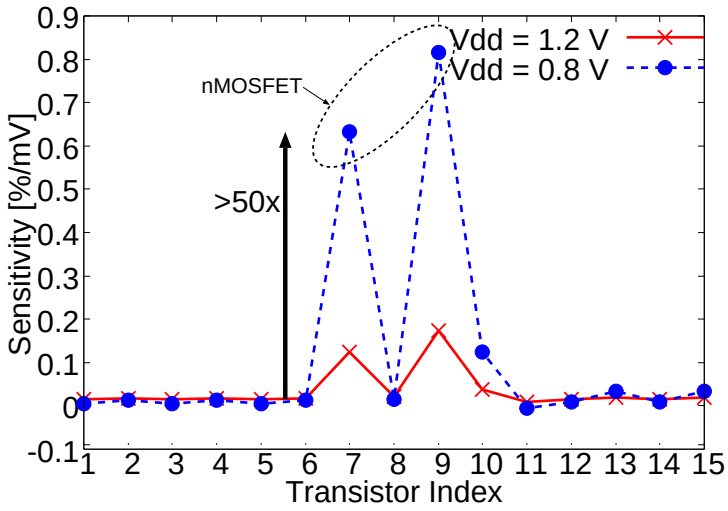


Figure 4.7: Sensitivity of each transistor in a seven-stage inhomogeneous RO of 4.6. (©2013 IEEE)

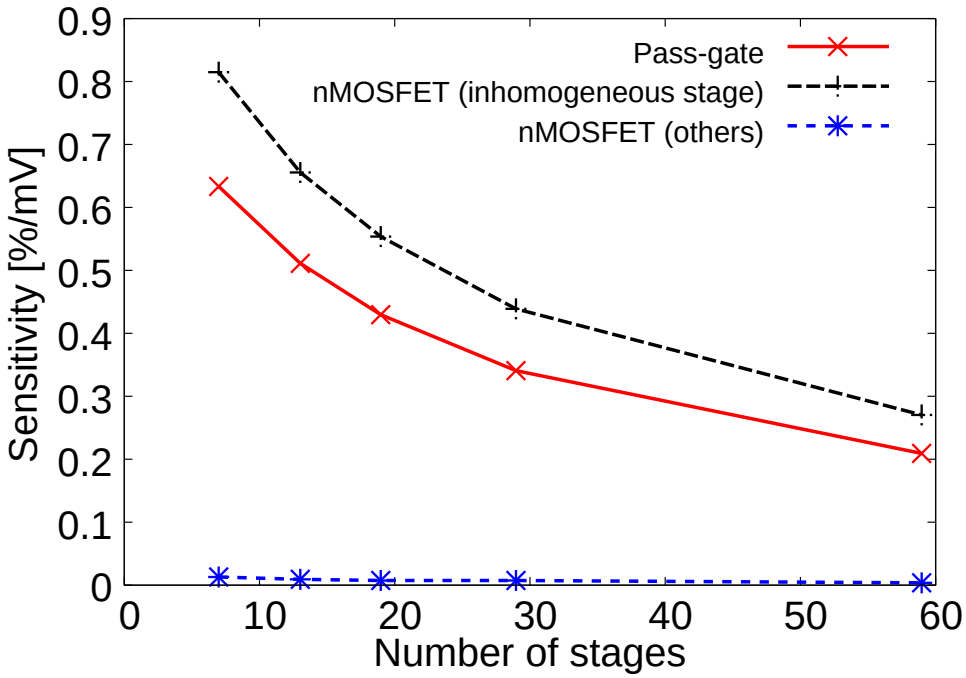


Figure 4.8: Change of frequency sensitivity to individual transistor variation against the number of stages. Increase in the number of stages reduces the sensitivities of the transistors in the same proportion. (©2013 IEEE)

4.2.3 Effect of Number of Stages

Frequency sensitivity to a particular stage is reduced with the increase in number of stages. Figure 4.8 shows the change in frequency sensitivities to individual transistors for an "INV-NPASS-I" inhomogeneous RO against the number of stages. The sensitivities are reduced by the number of stages. However, as the sensitivities for all transistors are reduced in the same proportion, the ratio between the sensitivity of inhomogeneous nMOSFETs and the sensitivity

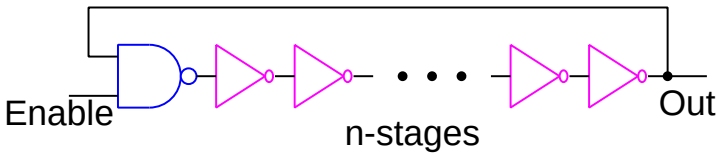


Figure 4.9: Schematic of ring oscillator circuit used for variation characterization. Conventionally, the inverter structure of each stage is fixed. Multiple ROs with different inverter structures are implemented for extracting various variation information.

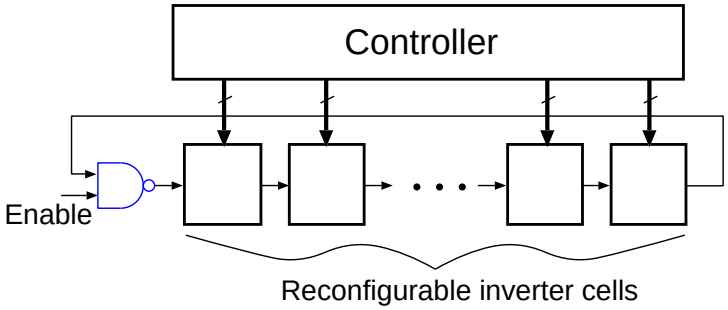


Figure 4.10: Proposed reconfigurable ring oscillator structure. Each inverter stage can be configured to several delay modes.

of other nMOSFETs remains constant at around 50. The number of stages has little impact on the observability of the transistor-by-transistor variability.

4.3 Topology-Reconfigurable Monitor Circuit

In this section, a topology-reconfigurable monitor circuit is proposed by which different kinds of variation can be monitored with the same circuit. Monitor architecture and variation monitoring methodology are discussed here.

4.3.1 Reconfigurable Ring Oscillator Structure

RO based measurement gives us characteristics of digital circuit behavior. Conventional RO structure uses the same inverter structure for each inverter stage. Figure 4.9 shows a conventional RO structure. In order to extract various variation information, inverters of various structures are used [3, 120, 121]. As the inverter structure is fixed, multiple designs and implementation of ROs with various inverter structures are needed. For a conventional RO structure, variation of each transistor is averaged out and thus transistor-by-transistor level characterization is not possible. An inhomogeneous RO structure is proposed where the RO frequency can be made sensitive to a small number of transistors [5]. In this section, we describe a reconfigurable RO structure by which device identification is possible allowing accurate device level variation characterization.

Figure 4.10 shows our proposed reconfigurable RO structure. The inverter structure of each stage is configurable so that various structures can be obtained with a single RO instance. A con-

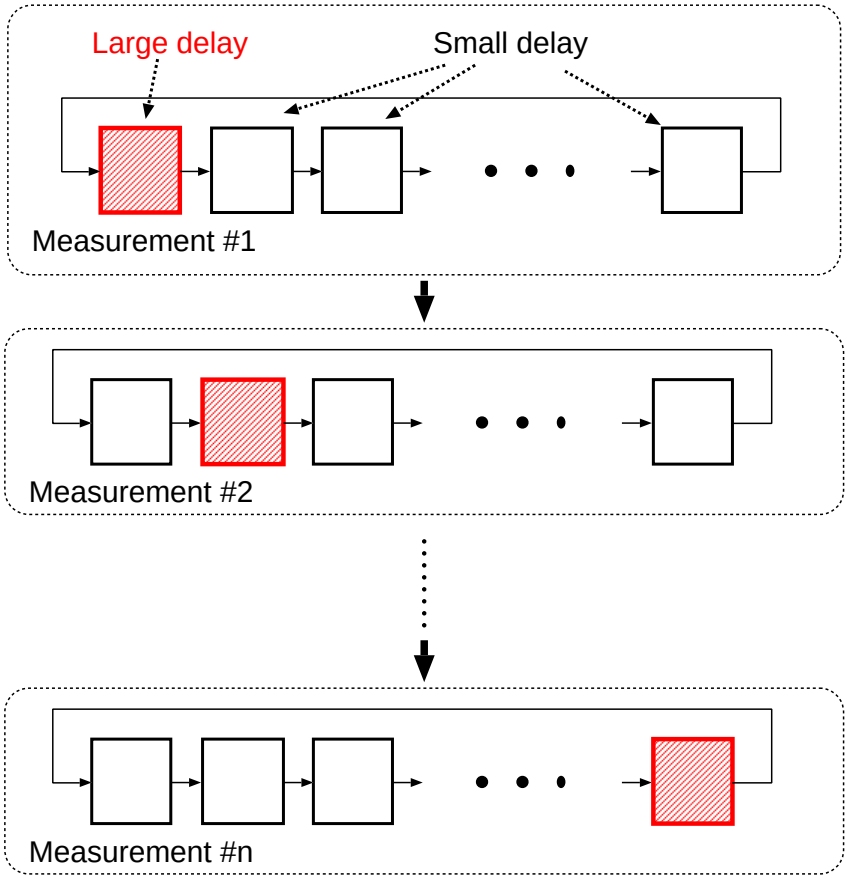


Figure 4.11: Characterization of delay variation using inhomogeneous configuration of our proposed ring oscillator. The delay of a particular stage becomes dominant in the oscillation frequency. By scanning the inhomogeneous stage, delay variation of each stage can be measured. (©2014 JJAP)

troller sets the configuration signals for each inverter stage. Using the inhomogeneous structure, the RO frequency can be made sensitive to the inhomogeneous stage. By scanning the inhomogeneous stage, large measurement samples can be obtained allowing statistical evaluation of variations. We show this measurement procedure to evaluate delay variation of each stage in Figure 4.11. With the same inhomogeneous configuration, by reconfiguring the inhomogeneous stage again, device identification becomes possible.

4.3.2 Variability Monitoring Methodology

Using the measurement method shown in Fig. 4.11, WID random variation is measured. With an N-staged RO, N types of inhomogeneous configurations are achieved. By reconfiguring the inhomogeneous stage further to be either nMOSFET or pMOSFET-sensitive, statistical evaluation of nMOSFET and pMOSFET becomes possible. Reference [5] shows that the sensitivity of transistors in the inhomogeneous stage can become as large as 40 times compared with other transistors at 0.8 V operation. This makes the delay variation in the inhomogeneous stage multiple times larger than the delay variation in other stages. The sensitivity is defined as the

percentage change of frequency due to one mV of threshold voltage change. Measuring the frequencies by scanning the inhomogeneous stages gives us a distribution which is a strong function of the delay variation between the inhomogeneous stages. By making the inhomogeneous stage to be either nMOSFET- or pMOSFET-sensitive, frequency distribution reflecting either nMOSFET or pMOSFET are obtained. The accuracy of the characterization depends on the sensitivity increase of transistors in the inhomogeneous stage compared with the transistors in the other stages. For example, assuming the frequency variation caused by each stage to be random and independent, frequency f_i for an inhomogeneous configuration where i-th stage is the inhomogeneous stage can be expressed as follows.

$$f_i = f_0 + \sum_{j=1}^{i-1} x_j + x_i^s + \sum_{j=i+1}^{N} x_j \tag{4.5}$$

$$= f_0 + x_i^s - x_i + \sum_{j=1}^{N} x_j \tag{4.6}$$

Here, f_0 is the frequency at typical condition and x_j is the amount of frequency fluctuation caused by the j-th stage. x_i^s is the amount of frequency fluctuation caused by the i-th inhomogeneous stage. N is the number of stages. As the measurements of Fig. 4.11 is performed using the same circuit instance, $\sum_{j=1}^{N} x_j$ is constant. Thus, the frequency variation σ_f becomes the following.

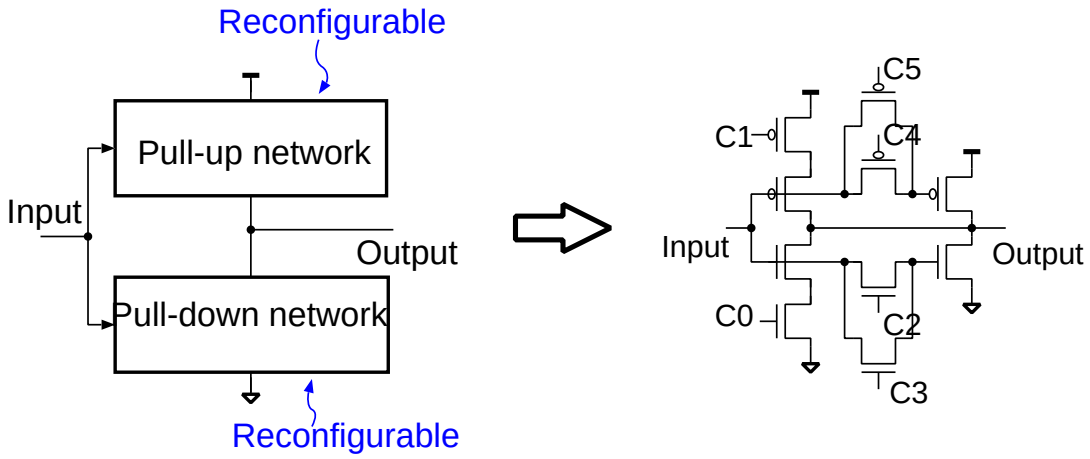
$$\sigma_f^2 = \sigma_{x^s}^2 + \sigma_x^2, \tag{4.7}$$

$$\sigma_f = \sqrt{m^2 + 1} \cdot \sigma_x. \tag{4.8}$$

Here, σ_{x^s} and σ_x are the amounts of frequency variation caused by the inhomogeneous configuration and the default configuration which is used in the stages other than the inhomogeneous one. m is the sensitivity increase for the inhomogeneous stage compared with the other stages. Thus, variation between the inhomogeneous stages can be obtained by exploiting the sensitivity enhancement with the proposed measurement method of Fig. 4.11. From Eq. (4.8), larger the sensitivity increase m is, more enhancement is achieved in the measured frequency variation. In case of an 8 times increase in the sensitivity of an inhomogeneous configuration, frequency variation will be dominated by the variation caused by the inhomogeneous stages by more than 99%. Statistical properties can be obtained by choosing the number of stage N to be sufficiently large.

D2D variation is measured by configuring the RO as homogeneous [3]. As the area of the circuit is very small, placing the circuit at several locations on a chip can capture systematic variations as well.

The proposed monitor scheme is also suitable for monitoring dynamic variations such as RTN because of its high sensitivity to specific transistors. Using the topology-reconfigurable feature, devices having RTN-induced fluctuation can be identified by observing frequency fluc-



(a) Reconfigurable inverter cell (b) Schematic

Figure 4.12: (a) Design of a reconfigurable inverter cell. Pull-up and pull-down networks are reconfigurable separately. (b) One design example of a reconfigurable inverter cell and its transistor level schematic. Various delay modes are realized. (©2014 JJAP)

tuation for each inhomogeneous configuration. For an inhomogeneous configuration, the oscillation frequency is dominated by the delay of the inhomogeneous stage [5]. When a transistor inside an inhomogeneous stage is affected by RTN induced variability, the fluctuation becomes directly visible to the oscillation frequency due to its high sensitivity. By reconfiguring the pull-up or pull-down network, RTN affected devices become distinguishable. Thus, device level characterization becomes possible under switching condition. A homogeneous configuration may also show RTN-induced frequency fluctuation. Devices responsible for frequency fluctuation are distinguished by configuring the RO as inhomogeneous. This helps us to characterize device level variability as well as its effect on circuit behavior.

4.4 Topology-Reconfigurable Universal Monitor Cell

This section proposes a topology-reconfigurable monitor cell for the realization of topology-reconfigurable monitor circuit. Monitor cell structure and its various operation modes are explained here.

4.4.1 Monitor Cell Structure

In order to realize the reconfigurable RO of Figure 4.10, inverter cell with reconfigurable structure is required. In Figure 4.12(a), we show a general concept of a reconfigurable inverter cell. Pull-up and pull-down networks are configurable separately to allow the measurements of pMOSFET and nMOSFET independently. Figure 4.12(b) shows one example of such a reconfigurable inverter cell structure. The inverter cell can be configured to several delay modes which is shown in Figure 4.13. Cell configuration in Figure 4.13(a) has equal rise and fall delay. Cell configuration in Figure 4.13(b) has large fall delay compare to the rise/fall delays of Fig-

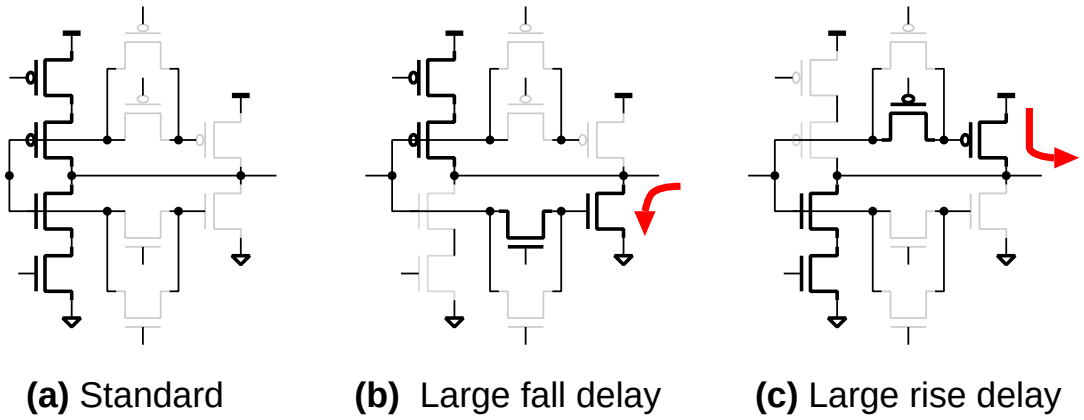


Figure 4.13: Several pull-up and pull-down network configuration to realize various delay modes. (©2014 JJAP)

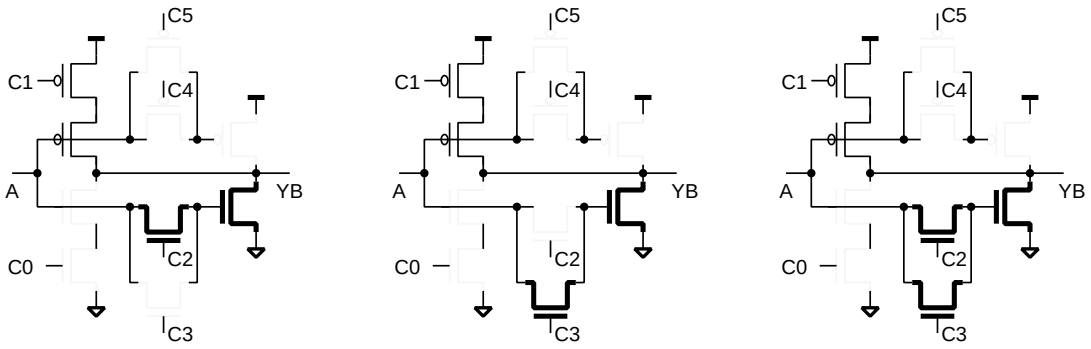
Table 4.1: Delay configurations for the proposed monitor cell.

C5	C4	C3	C2	C1	C0	Delay Mode
1	1	0	0	0	1	Standard
1	1	0	1	0	0	nMOSFET-dominant (pass-gate 1)
1	1	1	0	0	0	nMOSFET-dominant (pass-gate 2)
1	0	0	0	1	1	pMOSFET-dominant (pass-gate 1)
0	1	0	0	1	1	pMOSFET-dominant (pass-gate 2)

ure 4.13(a) structure. Similarly, Figure 4.13(c) has large rise delay compare to Figure 4.13(a) structure. Each of the structures in Figure 4.13(b) and Figure 4.13(c) can be further reconfigured. For example, Figure 4.13(b) has two pass-gates in parallel each of which can be either turned ON or OFF. We get three different configurations for the structure in Figure 4.13(b) excluding both pass-gate OFF configuration.

The three configurations for the structure Figure 4.13(b) is shown in Figure 4.14. Two pass-gates in parallel of the reconfigurable inverter allows us to measure device mismatch directly by taking the difference between the frequencies by turning one of the two parallel pass-gates ON and the other OFF, and vice versa. Device identification with RTN is done by observing frequency fluctuation for each of the configurations of Figure 4.14. If the Figure 4.14(a) configuration shows RTN induced variability and the other configurations does not, it is concluded that the RTN is occurring at the C2 device. Similarly, if Figure 4.14(b) configuration only shows RTN, then the RTN is occurred at the C3 device. This way nMOSFET identification is done by reconfiguring the inhomogeneous stage of an inhomogeneous structure. pMOSFET identification is done in the same way as nMOSFET identification.

Table 4.1 shows several delay mode configurations for the proposed monitor structure. pMOSFET dominant delay path is realized by turning OFF the C1 pMOSFET and the nMOSFET pass-gates, and turning ON the C0 nMOSFET and one of the pMOSFET pair pass-gates. nMOSFET dominant delay path is realized by vice versa. The key idea is to use the pMOSFET pair and nMOSFET pair based pass-gate configuration. Each device in the pair can be



(a) C2 only ON (b) C3 only ON (c) C2 and C3 ON

Figure 4.14: Three different pass-gate configurations for the reconfigurable inverter structure. By swapping the pass-gate configuration, transistor with RTN can be identified. The frequency difference of (a) and (b) configurations gives the mismatch of two adjacent devices. (©2014 JJAP)

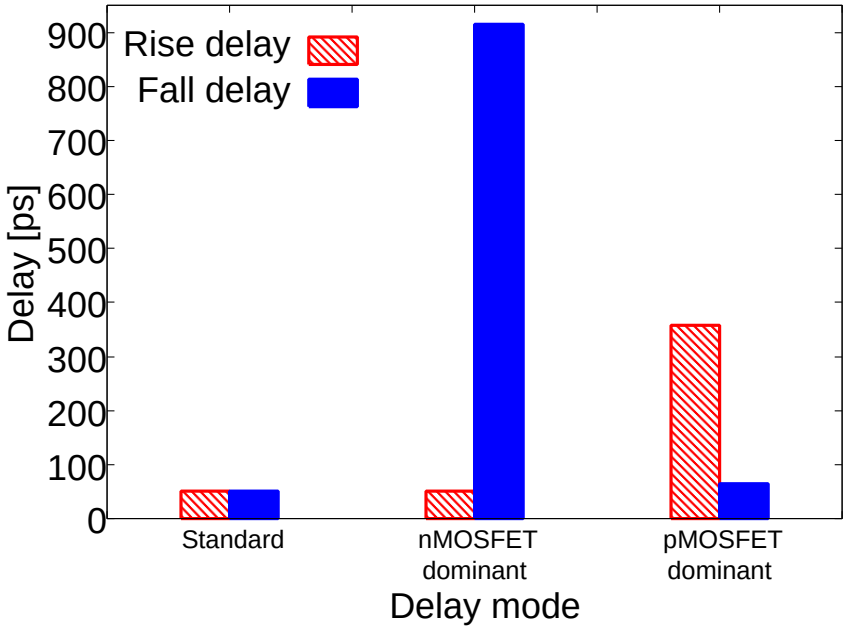


Figure 4.15: Rise and fall delays of the proposed monitor structure for several delay modes at 0.8 V supply. (©2013 IEEE)

either turned OFF or ON. The pair implementation gives us the following advantages. Firstly, by turning a single pass-gate ON and then swapping the pass-gate ON configuration, performance differences between these two devices can be measured. Secondly, while characterizing dynamic variations such as RTN, device level characterization is possible by reconfiguring the pass-gates. Figure 4.15 shows the rise and fall delays of the proposed cell for three different configurations at 0.8 V supply. The fall delay of nMOSFET dominant configuration and the rise delay of pMOSFET dominant configuration are multiple times larger than the standard inverter delays.

In Figure 4.12, the output nodes of pass-gates become floating when both the pass-gates are

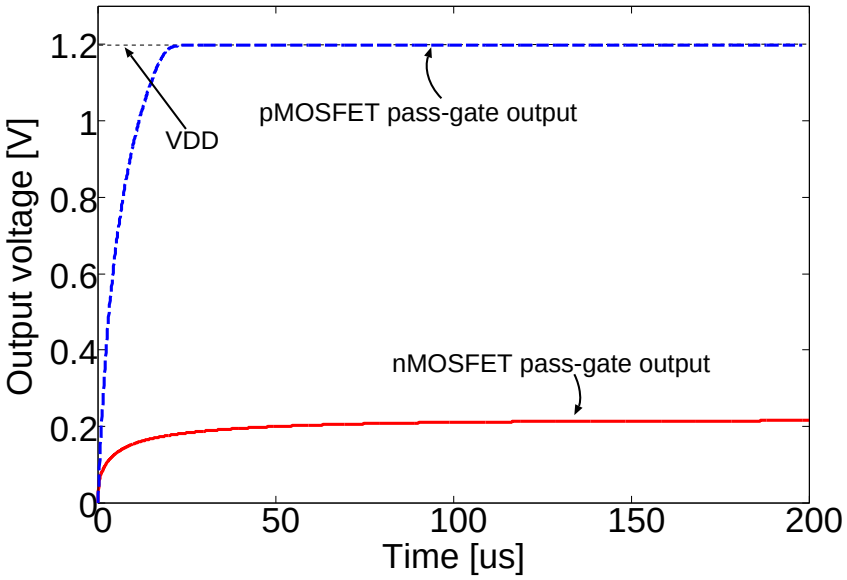


Figure 4.16: DC characteristics of nMOSFET and pMOSFET pass-gate. Output voltages are set to 'L' initially. Then input is raised to 'H'. (©2013 IEEE)

turned OFF. When the input is at 'L', the output of the nMOSFET pass-gate pair becomes 'L' as well (Figure 4.16) cutting the pull-down path completely OFF. Similarly, when the input is at 'H', the pMOSFET pass-gate pair output becomes 'H' cutting the pull-up path OFF. Thus, short-circuit current through partially ON pMOSFET and nMOSFET are avoided.

During the switching operation, when the pass-gates are turned off in Fig. 4.13 configurations, the output voltages of the pass-gates are determined by the pass-gate off-resistances. The output nodes have capacitances associated with the MOSFET gates and interconnect, thus forms an RC low pass filter. For example, in case of Fig. 4.13(c) configuration, the output voltage of nMOSFET pass-gate is determined by the charging and discharging currents during the oscillation. During the charging when the input is "H", pass-gate gate–source voltage becomes negative with the increase of output voltage. The pass-gate is also affected by reverse bias with the voltage increase. These effects lead to an exponential increase in the nMOSFET pass-gate off-resistance as the output voltage increases with a rate similar to the subthreshold slope factor which is typically several tens mV per decade. During the discharging, on the other hand, gate–source voltage remains zero thus the off-resistance does not change much. As the oscillation period is much smaller than the RC time constant, the output voltage is maintained to a value close to "L" which is determined by the off-resistance ratio between charging and discharging. Similarly, for Fig. 4.13(c) configuration, the pMOSFET pass-gate output voltage is maintained to a value close to "H" making the pull-up pMOSFET off.

4.4.2 Layout

Figure 4.17 shows a layout example of our proposed monitor structure. The pair MOSFETs share the same source diffusion to minimize any layout induced variation. Thus, differences of pair MOSFET performances will give pure random variation. The size of the cell is 3.2 μm $\times$

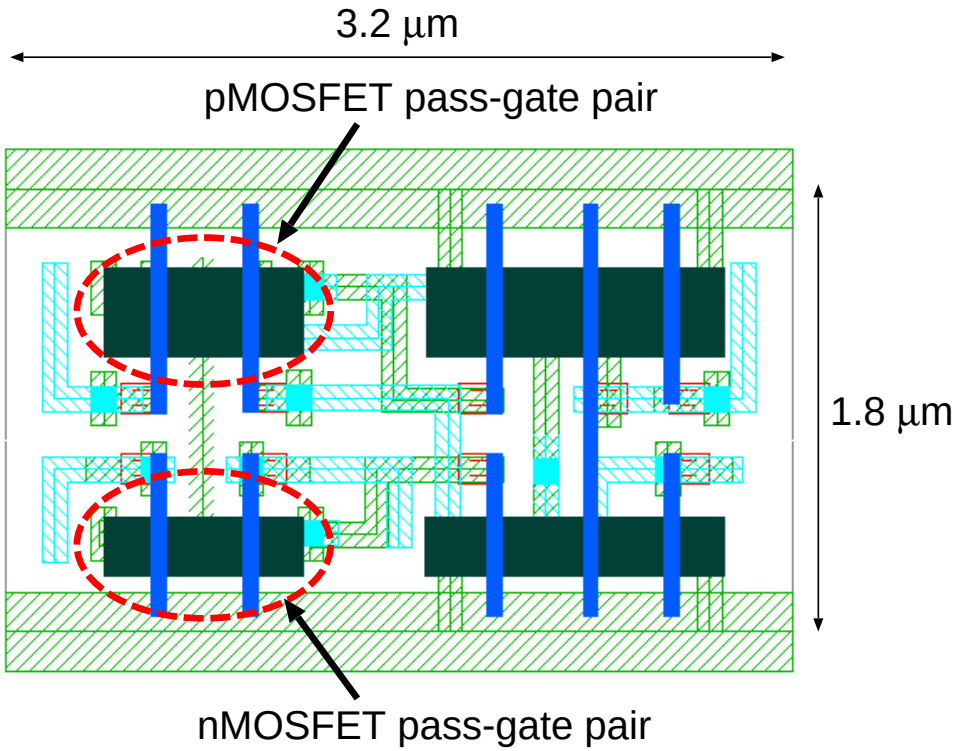


Figure 4.17: Layout example of the proposed monitor cell structure. (©2013 IEEE)

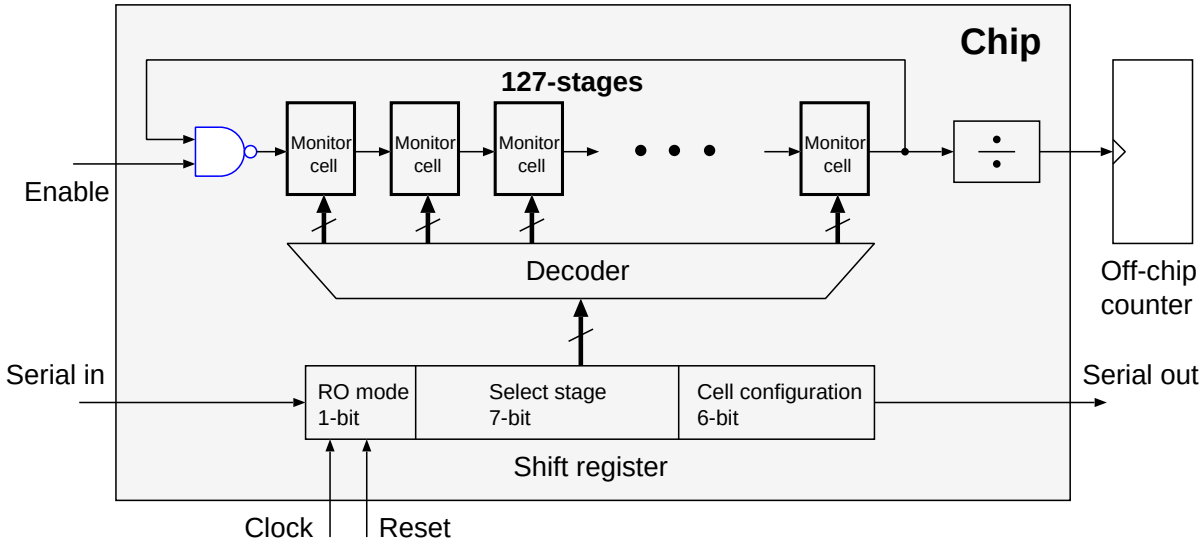


Figure 4.18: Block diagram of the on-chip monitor test structure. (©2013 IEEE)

1.8 μm which is five times larger than the standard inverter cell. The monitor cell has the same height as the standard cells to facilitate conventional place and route design.

4.5 On-chip Monitor Scheme

A monitor scheme is developed to measure several variations including statistical properties of random variation utilizing the flexibility of the proposed monitor cell structure. Figure 4.18

shows the block diagram of the developed monitor scheme. The monitor scheme consists of a single RO of 127 stages. The proposed monitor cell is used as the inverting gate for each stage except the first one. An NAND gate is used to control the oscillation. The scheme has a serial interface to set the values of a 14-bit shift register. A decoder decodes the shift register values and send configuration signals to each inverter stage. Independent configuration of each stage will give lot of flexibility but also increase design complexity. In order to reduce design complexity and area, the following selective configurations are implemented. For monitoring MOSFET performance inside a particular stage, the RO is configured to have inhomogeneous structure as proposed in Sec. 4.2. One bit (RO mode) is used to configure the RO into a homogeneous structure or an inhomogeneous structure. For the homogeneous structure, the inverter cells can be configured to any of the delay modes as shown in Table 4.1. Global variations in pMOSFET and nMOSFET can be measured with this configuration. For the inhomogeneous structure, only the inhomogeneous stage is set to be configurable with the cell configuration signal bits. Inverter stages other than the inhomogeneous stage are set to a default configuration which is the standard inverter cell configuration in this scheme. The inhomogeneous stage can be chosen with select stage signal bits. This way, 126 inhomogeneous configurations are achieved. By scanning all the inhomogeneous configurations and measuring their corresponding frequency values, statistical properties are derived.

4.6 Test Chip Design and Measurement Results

In this section, test chip for the validation of the proposed monitor circuit is explained. Monitoring results for global, local and dynamic variation are presented here.

4.6.1 Test Chip Design

A test chip is fabricated in a 65-nm process to validate our proposed structure. The process features one poly layer, 12 metal layers, copper wiring, and low-κ insulating material technology. The physical gate oxide thickness is 1.7 nm. Figure 4.19 shows the chip micrograph, the layout of the reconfigurable RO and the controller. The inverter cell has the same height as that of the standard cells to allow conventional cell based design. This approach reduces implementation and design cost. The proposed RO thus can be embedded in any digital circuit. The number of stages for the implemented RO is 127. The oscillation signal is divided by 4 inside the chip and then output to the outside of the chip. Output frequency is around 750 kHz at 0.8 V supply for an inhomogeneous configuration. The frequency is counted using an external counter. The sizes of pMOSFET and nMOSFET are 360 nm and 240 nm respectively. The area of our proposed RO with the controller is 0.0085 mm^2. With the proposed reconfigurable RO, 252 nMOSFET-sensitive inhomogeneous configurations and 252 pMOSFET-sensitive inhomogeneous configurations are achieved. In order to obtain the above number of samples, the conventional approach[5, 7] would require 504 ROs in total. Considering 13-stage ROs, the

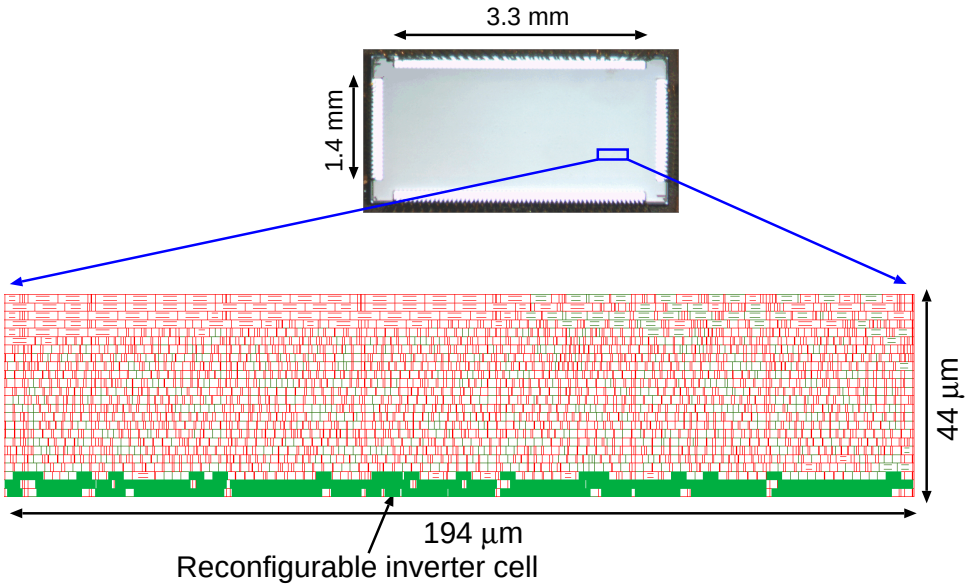


Figure 4.19: Chip micrograph and layout of the proposed reconfigurable ring oscillator. (©2013 IEEE)

proposed RO consumes only 5% area that of the conventional approach. The conventional approach puts the RO instances to several locations of the chip, thus routing and placing the ROs will increase area overhead. Besides, with small stage ROs, the oscillation frequency needs to be divided on-chip requiring additional circuitry.

Sensitivities are calculated for each of the configurations for the proposed RO using post-layout simulation. For an nMOSFET-sensitive inhomogeneous configuration using Fig. 4.13(b) structure, 18 times of sensitivity increase for the nMOSFET pass-gate and 24 times of sensitivity increase for the pull-down nMOSFET are achieved compared with the other nMOSFETs in the RO. Supply voltage of 0.8 V is used in the simulation. For pMOSFET-sensitive inhomogeneous configuration using Fig. 4.13(c) structure, 14 times of sensitivity increase for the pMOSFET pass-gate and 19 times of sensitivity increase for the pull-up pMOSFET are achieved compared with the other pMOSFETs.

Next, pass-gate output voltage characteristics are simulated for the reconfigurable inverter cell using RC extracted netlist. Fig. 4.20 shows the simulation results of pass-gate output for Fig. 4.13(c) configuration under AC condition at 0.8 V operation. AC signal of 5 MHz is applied to the input. The off-resistance ratio between charging and discharging is around 14 times making the output voltage 1/15 of the supply voltage. The voltage is around 0.05 V which is small enough compared with the nMOSFET threshold voltage. Small amount of voltage fluctuation occurs at the output node during the switching due to the coupling capacitances between the input and output. This voltage fluctuation is also small enough compared with the threshold voltage. Similarly, when the pMOSFET pass-gates are turned off as in Fig. 4.13(b), pass-gate output voltage fluctuates around 750 mV.

Below we show several measurement results of WID static variation and RTN induced variation at 0.8 V supply voltage to validate our proposed RO structure. By changing the supply

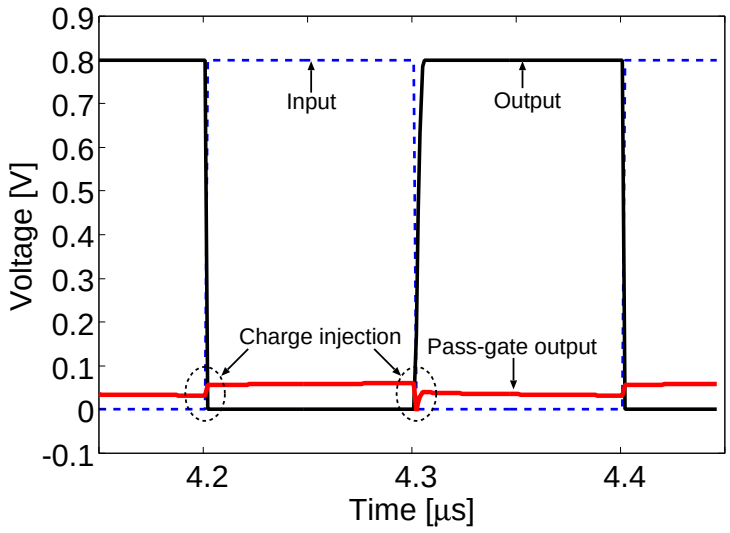


Figure 4.20: Simulated waveform of the nMOSFET pass-gate output for the inverter structure in Fig. 4.13(c). Input and output signals are also plotted. AC signal of 5 MHz is applied to the input. (©2014 JJAP)

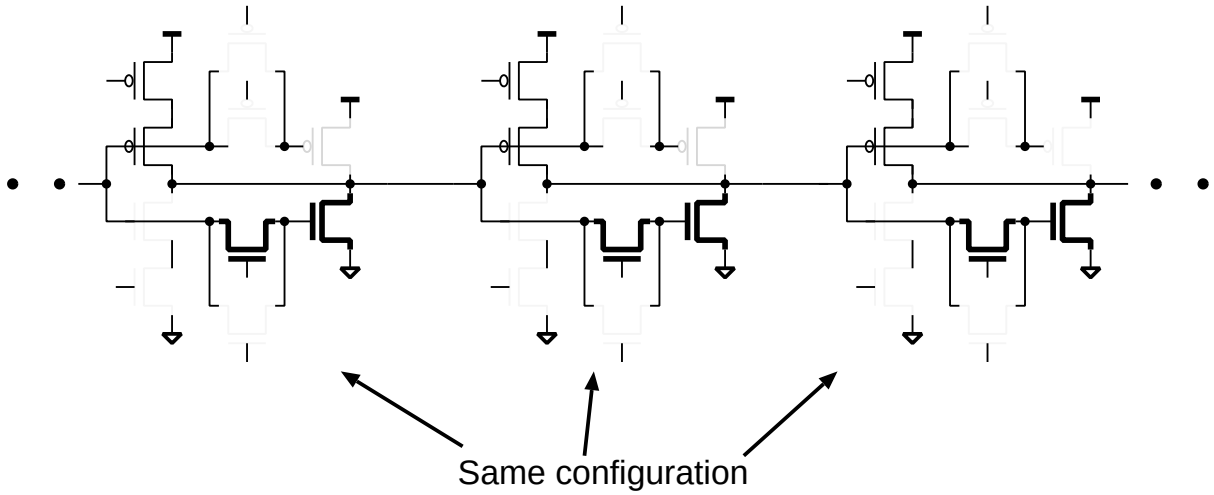


Figure 4.21: Homogeneous Structure for nMOSFET global variation monitoring.

voltage, various properties of these variations can be extracted [101, 120, 121].

4.6.2 Global Variation Monitoring

Figure 4.21 shows a homogeneous configuration for the proposed monitor scheme to monitor nMOSFET global variation. Similarly, pMOSFET global variation is monitored by configuring the homogeneous structure using pMOSFET-sensitive inverter topology. Figure 4.22 shows the measured frequency of pMOSFET-sensitive homogeneous topology against the measured frequency of nMOSFET-sensitive homogeneous topology. Simulation results using post-layout netlist including RC for five process corners are also plotted. Frequencies are measured for 30 chips. For this wafer, inter-chip variation is small. However, deviation from the predictions using "TT" corner model is observed. nMOSFET-sensitive frequency is much slower than the

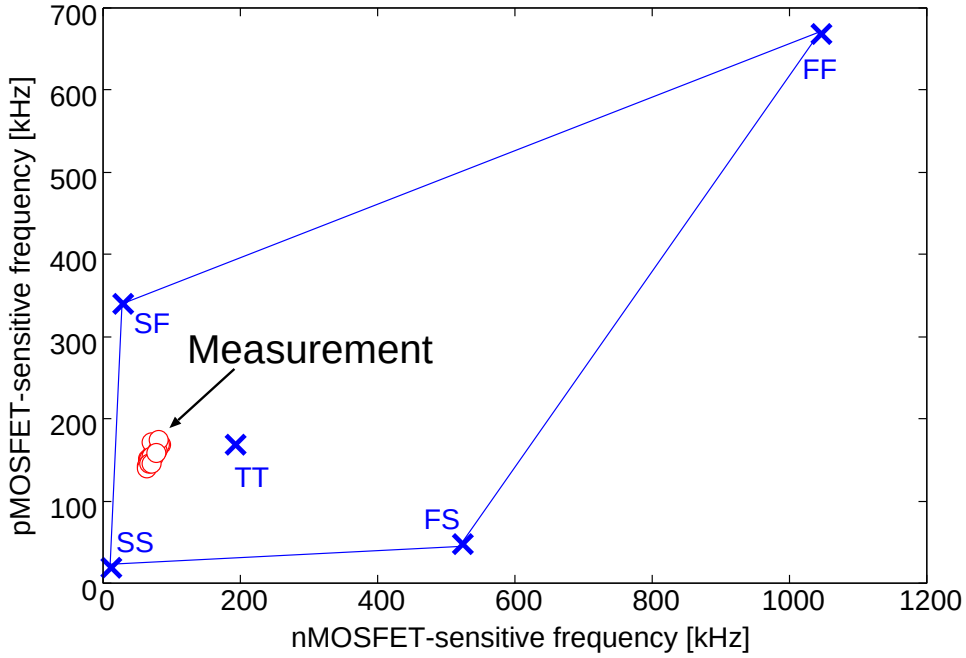


Figure 4.22: Frequency pMOSFET-sensitive homogeneous topology against frequency of nMOSFET-sensitive homogeneous topology for 30 chips.

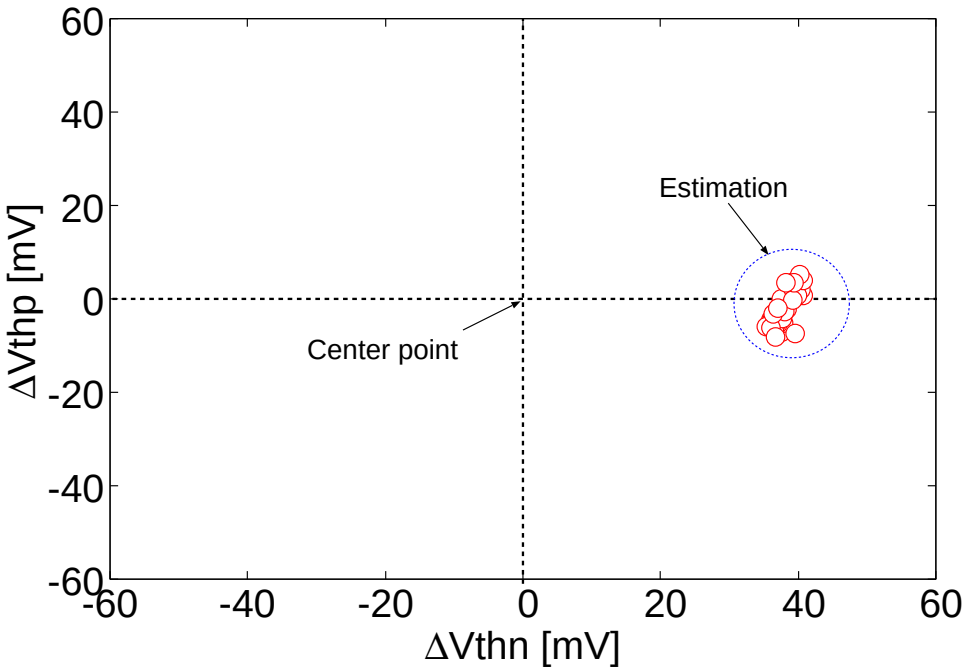


Figure 4.23: Estimation results of nMOSFET threshold voltage and pMOSFET threshold voltage for 30 chips.

"TT" corner model.

Next, the frequency variations are converted into threshold voltage variation using sensitivity analysis. Figure 4.23 shows the estimated threshold voltages of nMOSFET and pMOSFET for 30 chips. nMOSFET is deviated from the "TT" model by around 40 mV.

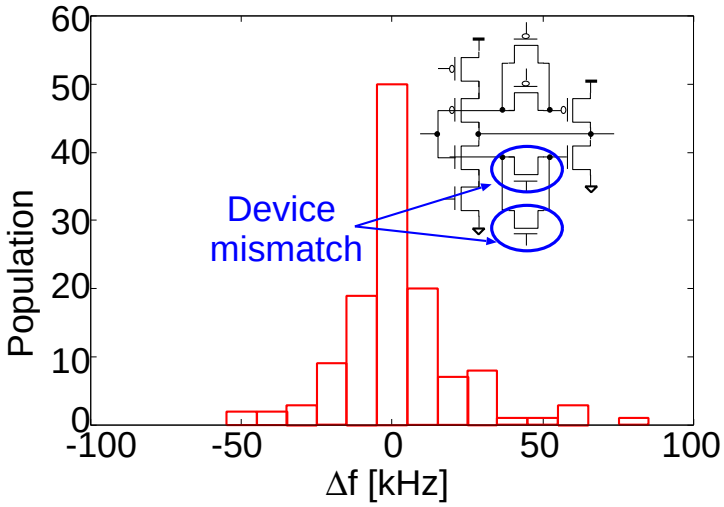


Figure 4.24: Histogram of frequency difference for two nMOSFET pass-gate configurations in the inhomogeneous stage. Frequency difference represents the device mismatch of the two nMOSFET pass-gates. (©2014 JJAP)

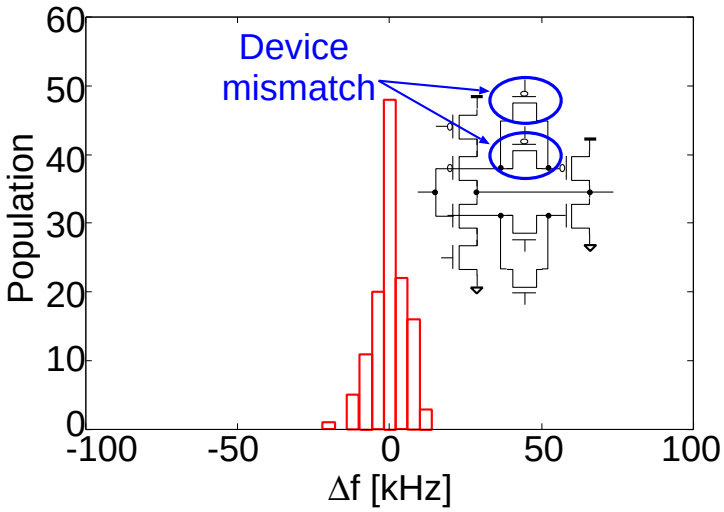


Figure 4.25: Histogram of frequency difference for two pMOSFET pass-gate configurations in the inhomogeneous stage. Frequency difference represents the device mismatch of the two pMOSFET pass-gates. (©2014 JJAP)

4.6.3 WID Random Variation Monitoring

Figure 4.24 shows the histogram of frequency differences when C2 and C3 pass-gates (nMOSFET pass-gates) are ON alternately. The frequency difference is the function of the device mismatch between these two transistors. Thus, device mismatch is directly measured with our proposed structure. Figure 4.25 shows the histogram when C4 and C5 pass-gates (pMOSFET pass-gates) are ON alternately. Both of the distributions are Gaussian suggesting we could measure the random variation. These variations can be further analyzed to extract the underlying variation sources such as threshold voltage by sensitivity based analysis [3, 5].

Assuming threshold voltage to be the main source of static device variability, oscillation

frequencies for an nMOSFET-sensitive inhomogeneous configuration can be expressed with the following linear approximations.

$$f_{n,1} = f_{n,1_0} + k_{n,1}\Delta V_{thn,1} + \alpha, \tag{4.9}$$

$$f_{n,2} = f_{n,2_0} + k_{n,2}\Delta V_{thn,2} + \alpha. \tag{4.10}$$

Here, $f_{n,1}$ is the frequency for the first nMOSFET pass-gate (C2 pass-gate) only being ON and $f_{n,2}$ is the frequency for the second nMOSFET pass-gate (C3 pass-gate) only being ON. $f_{n,1_0}$ and $f_{n,2_0}$ are nominal frequency values. $k_{n,1}$ and $k_{n,2}$ are sensitivity coefficients to the nMOSFET pass-gate threshold voltage variations. α is the sum of delay contributions of stages other than the inhomogeneous stage. As the delay contribution from each stage is much smaller than the inhomogeneous stage, random variation effect of other stages is expected to be canceled out because of large number of stages. Thus, α here is assumed to be constant. If we layout the two nMOSFET pass-gates to be identical, $f_{n,1_0}$ and $f_{n,2_0}$ become equal. Similarly sensitivity coefficients $k_{n,1}$ and $k_{n,2}$ are equal too. From the above two equations, the relationship between frequency mismatch and threshold voltage mismatch for the two different pass-gate inhomogeneous configurations can be obtained as follows.

$$\Delta f_n = k_n \Delta V_{thn}. \tag{4.11}$$

Here, Δf_n is the frequency mismatch for two pass-gate based inhomogeneous configurations, ΔV_{thn} is the threshold voltage difference between the two nMOSFET pass-gates in the inhomogeneous stage, and k_n is the sensitivity coefficient to each of the pass-gate threshold voltage. Next, by scanning the inhomogeneous stage and measure each configurations frequency mismatch as shown in Figure 4.11, threshold voltage variation ($\sigma_{V_{thn}}$) for nMOSFET is calculated as follows.

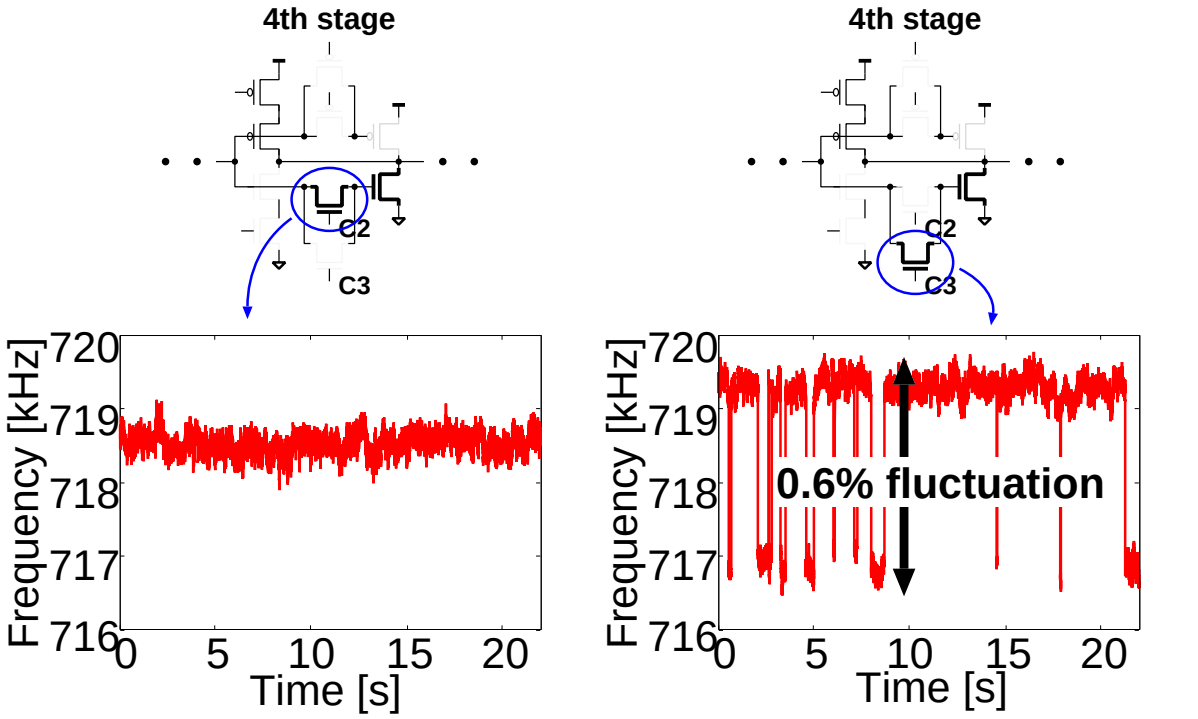
$$\sigma_{\Delta f_n} = k_n \sigma_{\Delta V_{thn}}, \tag{4.12}$$

$$\sigma_{V_{thn}} = \frac{\sigma_{\Delta V_{thn}}}{\sqrt{2}}. \tag{4.13}$$

pMOSFET threshold voltage variation can be calculated the same way. The amounts of WID variability for nMOSFET and pMOSFET threshold voltages are calculated to be 6.9% and 3.2% respectively. nMOSFET variability is measured to be larger than that of pMOSFET which agrees with the device array based measurement results [66, 133]. Thus, area-efficient device level measurement of static variation is performed with our proposed reconfigurable RO.

4.6.4 RTN Monitoring

For scaled technology nodes, effect of dynamic variation such as RTN has become comparable to static variation. Therefore, accurate characterization of dynamic variation is required to build variation models. Device level variation monitoring is required to characterize RTN effect.



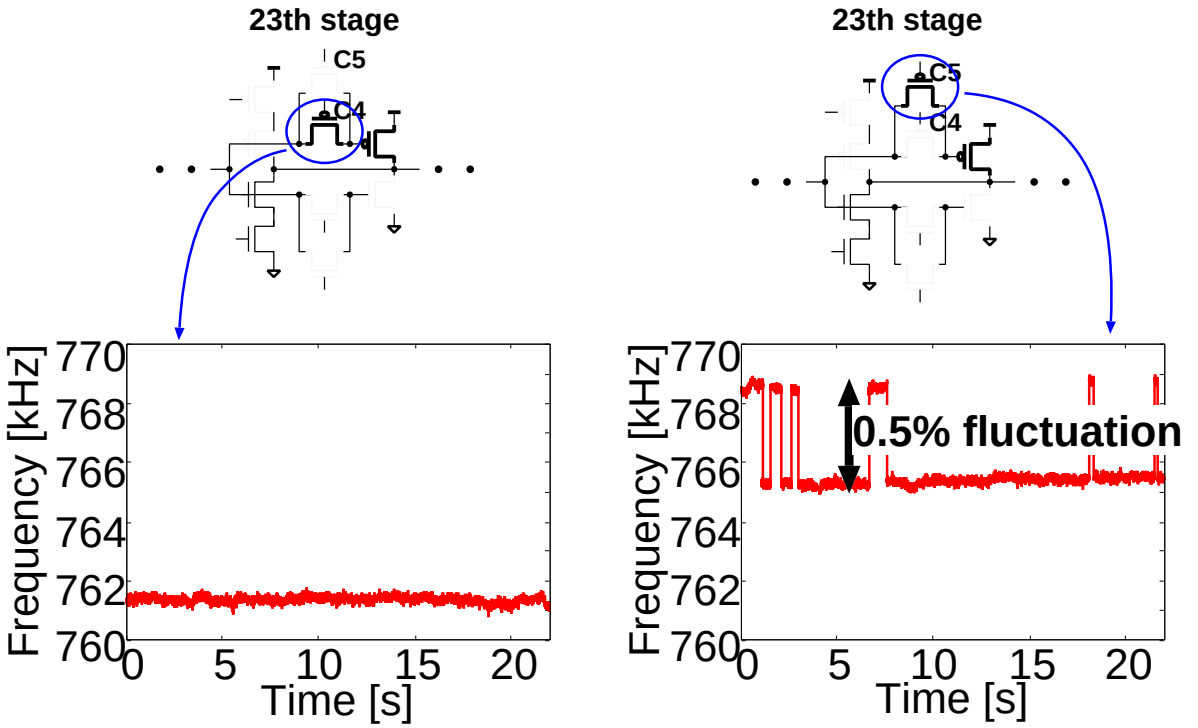
(a) C2 pass-gate ON, C3 pass-gate OFF **(b) C2 pass-gate OFF, C3 pass-gate ON**

Figure 4.26: Example of nMOSFET identification with RTN. C3 nMOSFET pass-gate of the 4th stage is identified with RTN occurring. (©2014 JJAP)

Conventional on-chip monitor circuits are unable to provide device level variation monitoring. With the proposed topology-reconfigurable monitor circuit, device level monitoring becomes possible. This section shows monitoring results of RTN at the device level.

Device Identification

In order to measure RTN effects, RO frequency for each inhomogeneous configuration is measured for 22 s with an integration time of 1 ms. In this test chip, there are 126 stages of reconfigurable inverter cells thus 126 configurations of inhomogeneous RO structure for nMOSFET and 126 configurations of inhomogeneous RO structure for pMOSFET can be measured. For each inhomogeneous configuration measurement (Figure 4.11), the pass-gate configuration in the inhomogeneous stage is swapped to identify devices with RTN. 67 inhomogeneous configurations out of 126 configurations showed RTN induced frequency fluctuation for nMOSFET. For pMOSFET, the number of configurations showing RTN induced variability is 69. Figure 4.26 shows one example of frequency fluctuations observed over time for an nMOSFET-sensitive inhomogeneous configuration. Figure 4.26(a) shows the frequency fluctuation when the C2 pass-gate is turned ON and C3 pass-gate is turned OFF. Figure 4.26(b) shows the frequency fluctuation for an opposite configuration (C2 OFF and C3 ON). C2 pass-gate ON configuration shows no binary fluctuation whereas C3 pass-gate ON configuration shows binary fluctuation which indicates that C3 pass-gate is RTN affected. Figs. 4.27(a) and 4.27(b) show two exam-



(a) C4 pass-gate ON, C5 pass-gate OFF **(b) C4 pass-gate OFF, C5 pass-gate ON**

Figure 4.27: Example of pMOSFET identification with RTN. C5 pMOSFET pass-gate of the 23rd stage is identified with RTN occurring. (©2014 JJAP)

ples of frequency fluctuation for an pMOSFET-sensitive inhomogeneous configuration. RTN is observed when C4 transistor is turned OFF and C5 transistor is turned ON referring that C5 transistor is RTN affected. After device identification, the probability of RTN occurring in an nMOSFET is calculated to be 24% and the probability of RTN occurring in an pMOSFET is calculated to be 26%. Thus, one out of four transistors will be affected by RTN induced variability which is a concern for large scale circuits especially for SRAMs. Thus, by changing the configuration of our reconfigurable RO, we can identify devices with RTN induced variability which will help us characterize the effects accurately. Methods presented in Refs. [101, 134–136] can be used to extract key RTN parameters such as time constant, threshold voltage fluctuation etc.

Multiple RTN

Now that we can identify transistors with RTN occurring on it, we can further explore and investigate devices with multiple RTN, different time constants and frequency fluctuations which will help us to build accurate models. In Figure 4.28, we show one example of an nMOSFET device showing multiple RTN. Several states are being observed. Relatively larger frequency fluctuation of 1.42% is observed in this case. Figure 4.29 shows the histogram of the frequency fluctuation of Figure 4.28. Four states are clearly distinguishable. Device identification allows us accurate characterization of RTN induced variability.

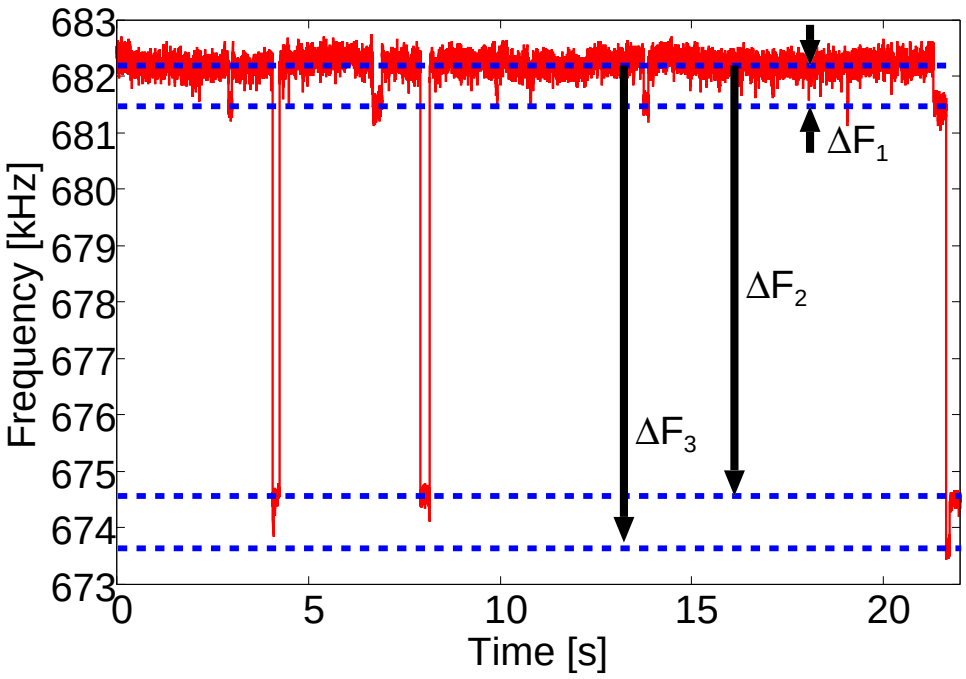


Figure 4.28: Frequency fluctuation over time showing complex RTN occurring on an nMOSFET pass-gate. (©2012 IEEE)

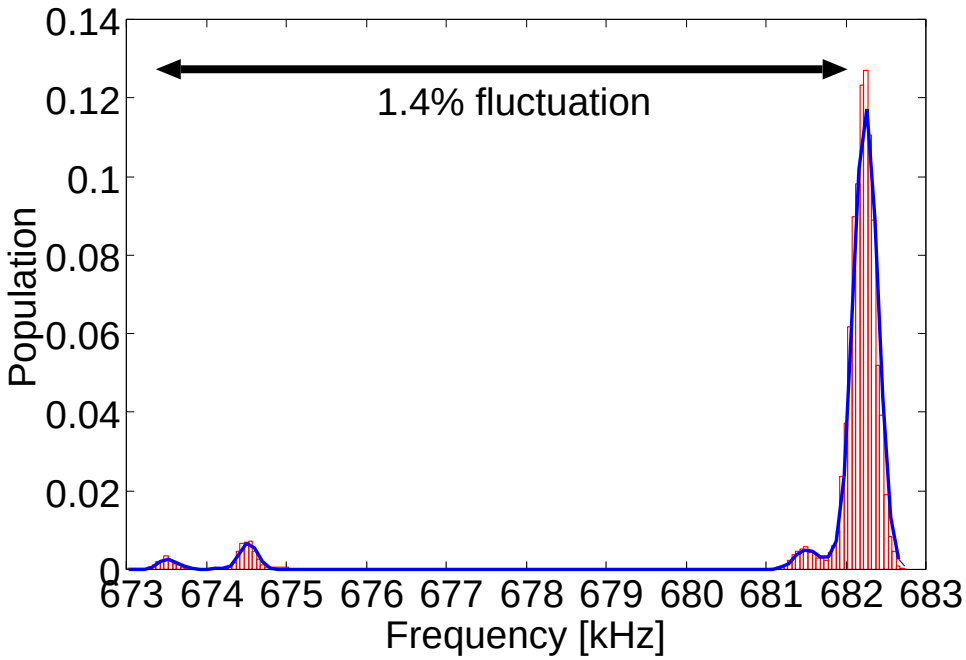


Figure 4.29: Histogram of frequency fluctuation of Figure 4.28. Four states are clearly distinguishable showing possibility of two traps being involved. (©2012 IEEE)

CDF

We have measured frequency over time for 126 inhomogeneous configurations for nMOSFET and pMOSFET characterization. For each configuration, we have two samples of frequency fluctuation by turning either of the pass-gates ON. Thus, 252 samples of frequency fluctua-

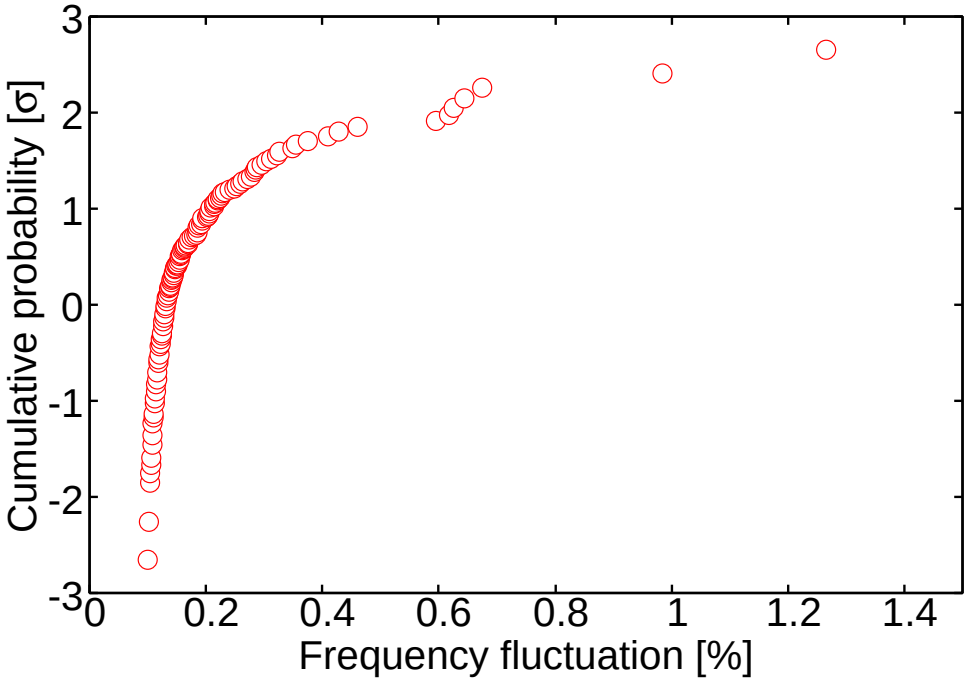


Figure 4.30: CDF of frequency fluctuation for nMOSFET-sensitive inhomogeneous configuration. Long tail exists referring that RTN induced variability is observed.

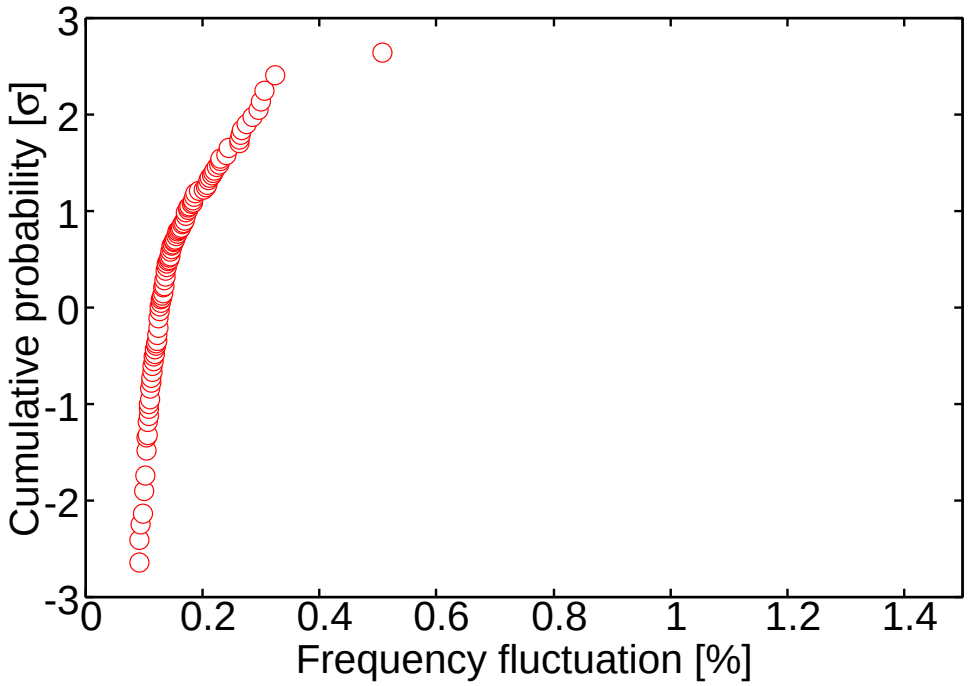


Figure 4.31: CDF of frequency fluctuation for pMOSFET-sensitive inhomogeneous configuration. Long tail refers that RTN induced variability is observed.

tion are obtained. Figs. 4.30 and 4.31 show cumulative distribution function (CDF) of frequency variation for nMOSFET and pMOSFET respectively. Frequency variation is calculated by $\Delta f / f_{\max}$ where Δf is the difference between the maximum and minimum frequencies and $f_{\max}$ is the maximum frequency. As expected, long tails are observed for both the nMOS-

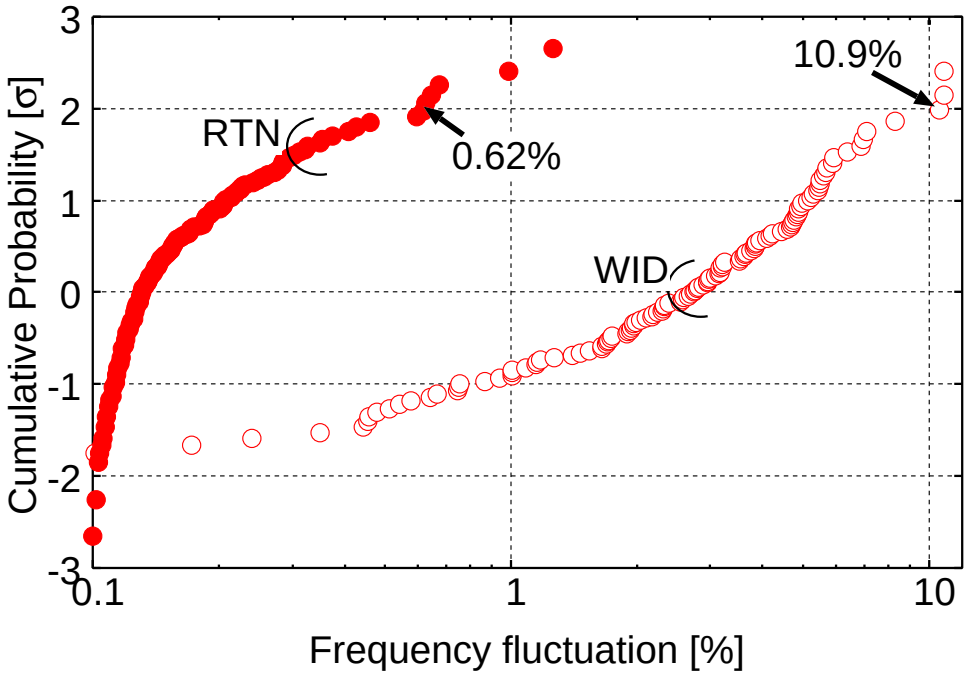


Figure 4.32: Comparison between RTN-induced frequency fluctuation and static process variation induced frequency fluctuation for nMOSFET. (©2014 JJAP)

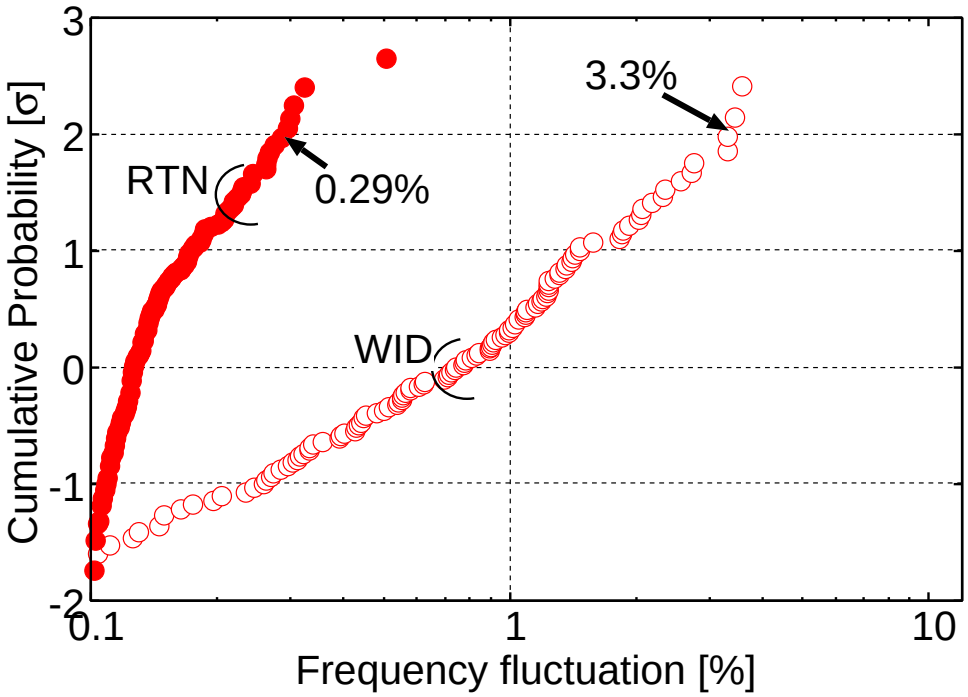


Figure 4.33: Comparison between RTN-induced frequency fluctuation and static process variation induced frequency fluctuation for pMOSFET. (©2014 JJAP)

FET and pMOSFET-sensitive frequency fluctuations. nMOSFET variation is larger than that of pMOSFET. For nMOSFET, maximum of 1.42% frequency fluctuation is observed. Whereas for pMOSFET, maximum of 0.53% frequency fluctuation is observed.

Next, we compare RTN induced frequency fluctuation with static process variation induced

frequency fluctuation. With the proposed RO, frequency distribution for inhomogeneous configuration can be measured for nMOSFET and pMOSFET separately. Figure 4.32 plots the CDF of RTN induced frequency fluctuation and process variation induced frequency fluctuation for nMOSFET. Process variation induced frequency fluctuation is calculated by $(|\mu_f - f|)/\mu_f)$ where μ_f is the average value of frequency measurements. At 2σ level, RTN induced fluctuation is 0.62% whereas process variation induced fluctuation is 10.9%. Thus, RTN induced fluctuation is 5.7% of process variation induced fluctuation at 2σ level. Similarly, Figure 4.33 plots the CDF of RTN and process variation induced frequency fluctuation for pMOSFET. At 2σ level, RTN induced fluctuation is 0.29% and process variation induced fluctuation is 3.3%. RTN induced fluctuation is 8.8% of process variation induced fluctuation. Measuring data from multiple ROs will give 3σ , 4σ level comparison. Implementing the proposed reconfigurable RO with various device sizes will give us more information on the relationship between device size and RTN effect. For finer process, RTN variability is reported to be increasing. Thus, RTN is a big concern for reliable operation of future LSI.

4.7 Summary

An area-efficient topology-reconfigurable monitor circuit structure is proposed for characterizing static and dynamic variations at both device and circuit level. The proposed circuit consists of inverter cells whose pull-up and pull-down networks can be reconfigured during runtime. Device level variation can be monitored using an inhomogeneous structure and by reconfiguring the inhomogeneous stage. Measurement results from a 65-nm test chip confirms the validity of the proposed structure.

With the proposed circuit structure, transistor-by-transistor characteristics of both static and dynamic variations can be characterized. Thus, the proposed circuit enables to investigate and understand on the interactions between these variations under dynamic switching condition. The proposed structure can be used for transistor-by-transistor NBTI characterization too. The proposed circuit can be used to understand the complex relationships between difference device phenomena which will help circuit designers to establish models and tools. Furthermore, the proposed RO can be implemented with very small area thus can be used as on-chip variation sensors. The information from on-chip device characteristic sensors can be used to reduce the design margins, energy-efficient mapping between supply voltage and frequency for ASV and DVFS architectures, post-silicon timing prediction and diagnosis. On-chip monitoring gives confidence for implementing various techniques to increase energy-efficiency of the entire system. The use of sensors will reduce test cost. Self-healing and self-diagnosis can be now implemented for different blocks is a SoC. On-chip sensors will allow to adapt the circuit to environmental changes.

Chapter 5

Runtime Compensation of Performance Variability for Energy-efficient LSI

In Chapters 3 and 4, digital monitor circuits are proposed to estimate device characteristic variations. As the D2D and location-correlated variation components are global for the particular location, this kind of variation affect can be compensated in post-silicon by tuning parameters of supply voltage or body bias. In this chapter, based on the developed monitor structures, a simple and digital post-silicon runtime compensation technique will be proposed. Test chip measurement shows that energy-efficient performance compensation can be achieved based on the monitor circuits. This will help designers to reduce design margins as well as tune the transistor characteristics on the runtime to yield maximum energy-efficiency. In this chapter, first variability effect on LSI performance is evaluated by simulation using a simple model circuit. It is shown that large energy loss occurs for the conventional worst-case design. Variability effect on DVFS architecture is explained. Both the supply voltage and threshold voltage tuning are effective for improving energy-efficiency. Then, a built-in self-adjustment scheme is proposed for runtime compensation of performance variability. Overall architecture of the scheme and details of the implementation is explained. Measurement results from test chip designed in a 65-nm process are presented.

5.1 Introduction

Lowering the supply voltage is the most effective method for energy reduction. However, reducing the supply voltage degrades the circuit speed drastically thus limits their application. In many systems, the workload varies with time. Thus it is possible to reduce the energy consumption of the system by changing the supply voltage and clock frequency depending on the workload. This method is called the dynamic voltage and frequency scaling (DVFS). In systems where high work loads occur rarely, DVFS technique promises large reduction of energy.

Designing LSI for DVFS is a challenge. Aggressive scaling of supply voltage and clock frequency is required in order to reduce energy drastically. Low voltage operation is required for

achieving minimum energy per operation. Wide voltage range DVFS has the following design difficulties. Firstly, finding the optimum mapping between the clock frequency and the supply voltage is difficult. Secondly, the effect of PVT (Process, Voltage, and Temperature) variation complicates the above problem. Furthermore, the effect of variation is amplified at lower supply voltage. For example, frequency spread of 28% at 1.2 V and 62% at 0.8 V between the fastest and slowest cores on a die is reported in Ref.[23]. Conventionally, circuits are designed considering worst-case scenarios. Device aging adds to this uncertainty and results in large amount of design margins to ensure reliable operation. Finally, pMOSFET and nMOSFET performance varies with supply voltage and thus P/N ratio changes with supply voltage. Unbalanced P/N ratio results in performance decrease and energy increase. This chapter addresses this issue and shows that at low voltage operation this P/N unbalance causes significant energy increase. Conventional worst-case design is simply way too inefficient especially at low voltage operation.

In order to encounter the variation problem on wide voltage range DVFS, two methods can be applied. The first method is the variation aware DVFS [23]. In this method, the maximum operation frequency is measured during test time for each core and the mapping is configured based on the measurement. The second method is to apply body bias to reduce the effect of variation. Body bias is an attractive choice to mitigate variation effects as it can also be applied to compensate device aging too. Several implementations of body bias to reduce the impacts of variation are reported [36, 46, 137, 138]. Adaptive body bias can be used to adjust the threshold voltage imbalance and reduce minimum supply voltage [139]. Integrating adaptive body bias with DVFS is reported to be effective [43]. In their approach, the mapping between supply voltage and operating frequency is delayed until test-time based on a simple leakage measurement. Leakage based bias selection can be effective for dies where both the pMOSFET and nMOSFET performances are matched. As will be demonstrated in the rest of the chapter, when pMOSFET and nMOSFET performance vary toward opposite direction, large amount of leakage and performance degradation occurs at low voltage operation. Thus, independent control of pMOSFET and nMOSFET performance is required to improve energy efficiency.

This chapter proposes an P/N-performance self-adjustment scheme. The proposed scheme consists of pMOSFET monitor and nMOSFET monitor. Body bias is generated depending on the monitor circuit performances. As the monitor circuits have large sensitivity to either pMOSFET or nMOSFET, balanced P/N-ratio is achieved which gives us optimal energy performance. As the proposed scheme adjusts transistor performance adaptively and dynamically, design margin required for global process variation and device aging can be eliminated. Simulation results based on simple circuit model show that energy efficiency can be improved by more than 100% for low voltage operation by integrating dynamic performance tuning with the DVFS architecture. More than 50% performance recovery is achieved with small leakage overhead for 0.7 V operation.

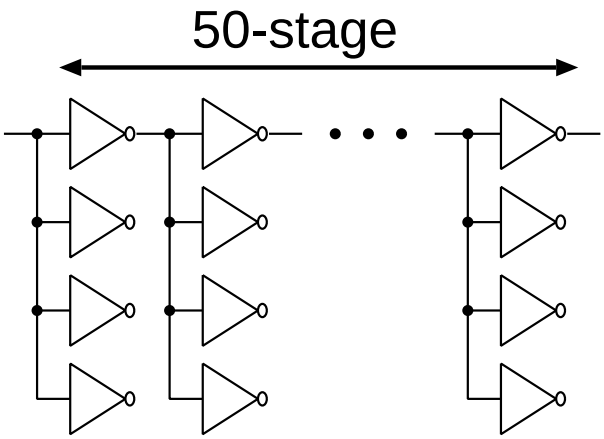


Figure 5.1: Circuit model used for delay and energy calculation. 50-stage fan-out 4 inverter chain as the circuit model.

5.2 LSI Performance Evaluation under Process Variation

In this section, the effect of transistor variation on LSI performance is discussed. Variation effect on DVFS architecture is explored and the need for on-chip monitor based compensation is shown.

5.2.1 Simulation Setup

Analysis based on an inverter chain gives us a general understanding of circuit behavior such as delay and energy. In Ref.[140], an inverter chain with identical stages is used for exploring circuit performance. In this paper, an inverter chain of 50 identical stages with fan-out 4 is used to investigate the effects of supply voltage and process variation on circuit delay and energy. Figure 5.1 shows the schematic of the inverter chain used in the simulation. SPICE [123] simulation is performed using a commercial 65-nm process. A single rise transient signal is applied to the input of the inverter chain and the delay to propagate the signal to the output is measured. Energy is calculated by integrating the current consumed within the transfer time of the input signal. For circuit activity of α, energy per cycle is calculated with the following equations.

$$E_{\mathrm{dyn}} = V_{\mathrm{dd}} \cdot \int_{t_0}^{t_0+t_p} I_{\mathrm{dd}} \cdot dt, \tag{5.1}$$

$$E_{\mathrm{leak}} = V_{\mathrm{dd}} \cdot I_{\mathrm{leak}} \cdot t_p, \tag{5.2}$$

$$E_{\mathrm{total}} = \alpha \cdot E_{\mathrm{dyn}} + (1-\alpha) \cdot E_{\mathrm{leak}}. \tag{5.3}$$

Here, t_0 is the start time of measurement and t_p is the transition time of the input signal which is the minimum clock period the circuit operate correctly. E_{dyn}, E_{leak}, and E_{total} are dynamic, leakage, and total energy per cycle respectively.

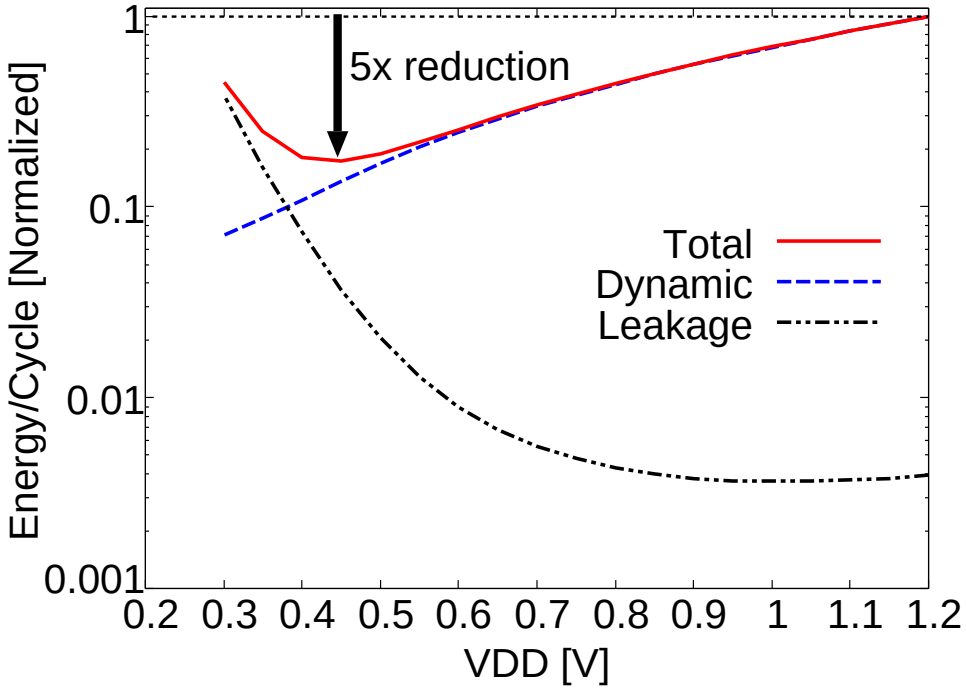


Figure 5.2: Simulated energy per cycle at different supply voltages. Model circuit is used in the simulation. Activity of 0.1 is assumed. Lowering the supply voltage by half improves the energy efficiency by 4.2 times.

5.2.2 Energy Efficiency against Supply Voltage

NTV (Near-Threshold Voltage) operation is considered as one of the key technique for the advancement of LSI as energy-efficiency increases at NTV operation. The effect of supply voltage lowering on energy-efficiency is simulated. Figure 5.2 shows the simulated energy per cycle at various supply voltage for activity rate of 10%. The X-axis is the supply voltage V_{dd} and the Y-axis refers to energy per cycle. Energy per cycle is normalized to the value at nominal V_{dd} (1.2 V). Leakage, dynamic and total energy are shown separately. With the lowering of V_{dd}, circuit operation frequency degrades. Dynamic power has quadratic relationship to V_{dd}, thus dynamic energy decreases with the decrease of supply voltage drastically. However, because of the increase of clock period, circuit idle time increases which causes leakage energy to dominate the total energy consumption. Leakage energy starts to increase drastically at V_{dd} below 0.7 V. Ultimately, total energy reaches to a minimum value which is found at 0.44 V supply. At 0.6 V, the energy efficiency is 4 times larger than operating at the nominal voltage of 1.2 V. Thus, lowering the supply voltage whenever possible is the key to the improvement of energy efficiency. However, leakage current need to be considered carefully. As process variation increases leakage current, careful selection of supply voltage based on circuit topology, activity and operating frequency otherwise energy-efficiency may degrade rather than increasing.

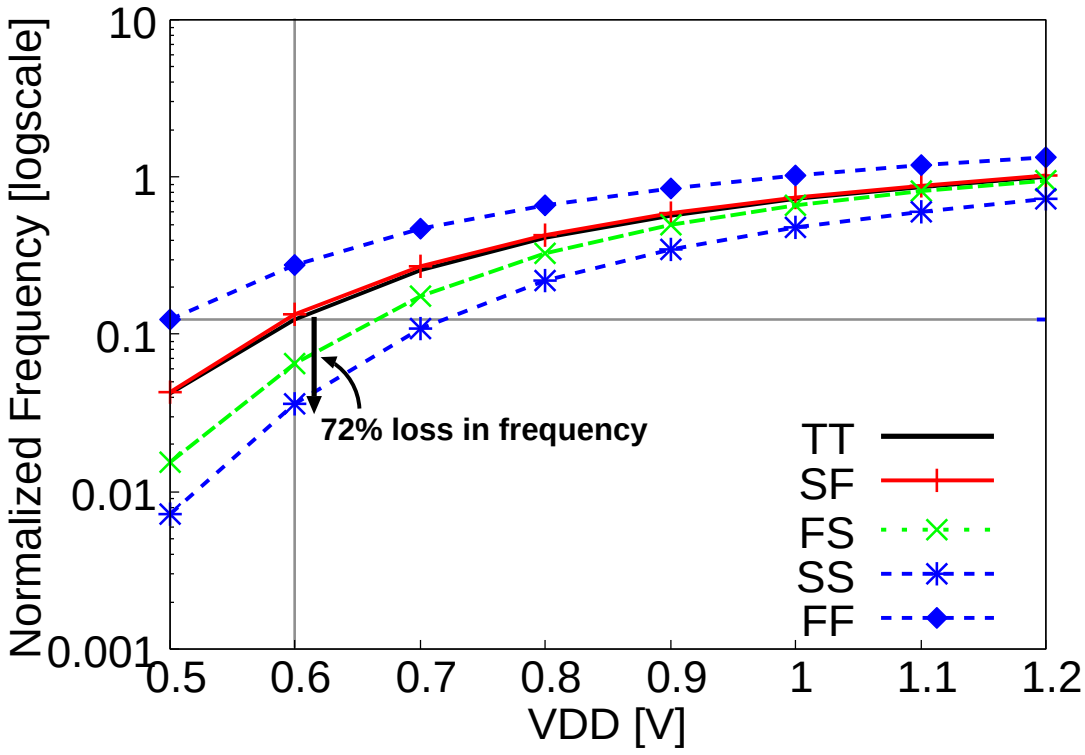


Figure 5.3: Maximum operating frequency at several corners against the supply voltage.

5.2.3 Variability Effect on Circuit Speed

The effect of process variation on circuit operation speed at different supply voltages is illustrated in Figure 5.3. Figure 5.3 shows the maximum operating frequency at several process corners against supply voltage change. Frequency values are all normalized by the value at nominal supply voltage for the "TT" corner. Frequency degrades about 8 times at 0.6 V operation than that at the nominal supply for the "TT" (Typical pMOSFET, Typical nMOSFET) corner. At 0.6 V supply, performance variation between the "TT" and "SS" (Slow pMOSFET, Slow nMOSFET) corners is 72% while the difference is 22% at nominal supply. In order to improve yield, the circuit needs to be designed considering the worst-case scenario. If we select the frequency and the corresponding V_{DD} for worst-case scenario (which is "SS" in this simulation), large amount of energy loss will occur when the process moves near to "FF" (Fast pMOSFET, Fast nMOSFET). Effect of worst-case design will be explained in Sec. 5.2.5.

5.2.4 Effect of P/N Mismatch at Low Supply Voltage

As shown in Figure 5.2, aggressive scaling of supply voltage is required to maximize the energy efficiency of LSI. However, aggressive scaling of supply voltage has two problems. One is the degradation of performance. The other is the effect of process variation. Process variation has severe effects on the circuit performance at low supply voltage. At above-threshold region, gate delay has linear relationship to threshold voltage variation. However, at near-threshold region, gate delay shows non-linear relationship to variation. This behavior is illustrated in Figure 5.4.

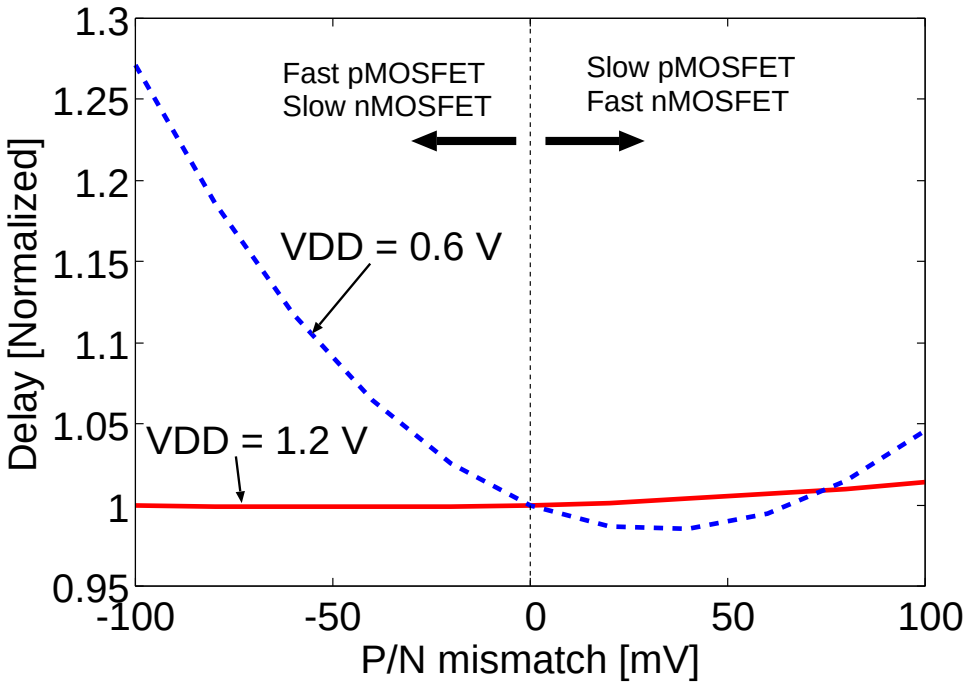


Figure 5.4: Circuit delay against P/N mismatch at two difference supply voltage of 1.2 V and 0.6 V. A 50-stage inverter chain of fan-out 4 is assumed.

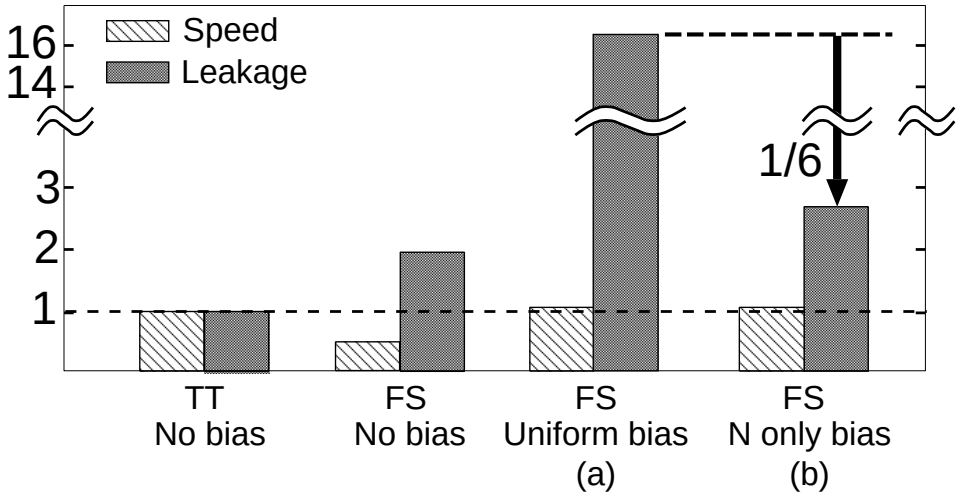


Figure 5.5: Circuit speed and leakage for two corners of "TT" and "FS". Forward body bias is applied for "FS" corner to compensate speed based on (a) critical path monitoring (uniform bias) and (b) P/N-sensitive monitoring (N only bias). (©2012 IEEE)

Variation effects on circuit delay against P/N mismatch for two different voltages of 1.2 V and 0.6 V are shown in the figure. P/N mismatch is realized by deviating pMOSFET and nMOSFET threshold voltages by the same amount in the opposite direction in the transistor model. For example, P/N mismatch of -50 mV refers to −25 mV deviation in pMOSFET threshold voltage and +25 mV deviation in nMOSFET threshold voltage. At 1.2 V operation, circuit delay linearly depends on threshold voltage thus no delay deviation is observed for skewed variation

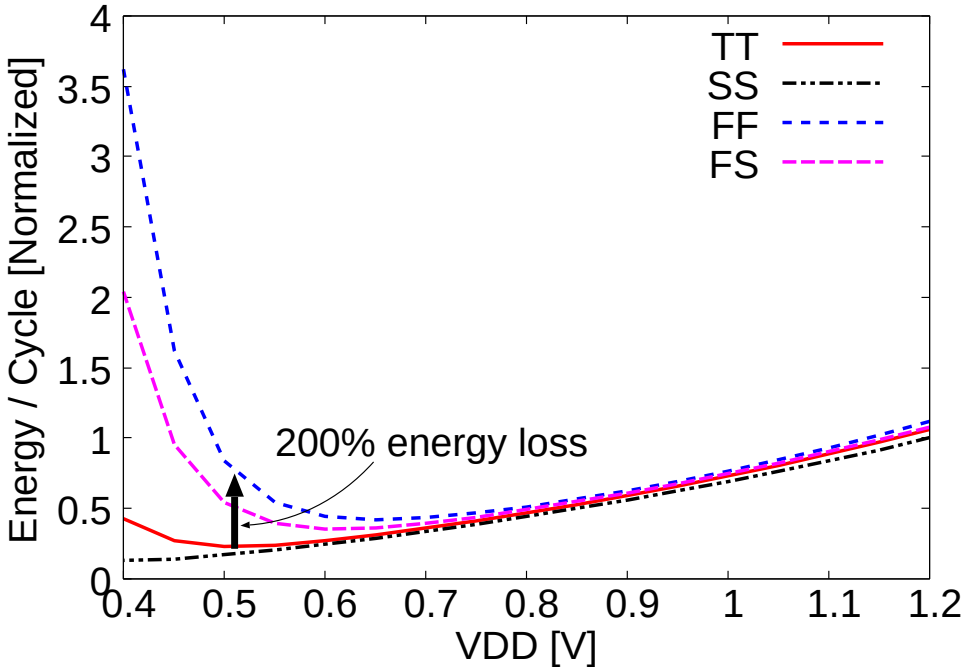


Figure 5.6: Energy efficiency at different supply voltages when operating at worst-case speed.

when pMOSFET and nMOSFET deviate in the opposite direction. However, at 0.6 V operation which is near to MOSFET threshold voltage, P/N mismatch causes large increase in delay. Thus, P/N mismatch needs to be taken care for energy-efficient implementation.

The effect of P/N mismatch is severer for logic gates with stacked transistors as NAND or NOR gates. In order to achieve higher energy-efficiency at low supply voltage operation, transistor performance need to be considered. For adaptive techniques such as adaptive body bias, pMOSFET and nMOSFET need to be tuned independently. Figure 5.5 shows simulated frequency of a 2-input NAND gate ring oscillator (RO) and leakage current of a conventional LSI for an unbalanced P/N corner ("FS") along with a balanced corner ("TT"). We can observe that frequency decreases and leakage increases at "FS" corner. If we apply forward body bias for both the nMOSFET and pMOSFET uniformly to compensate circuit speed to "TT" value, the leakage current increases by 16 times which is unacceptable. However, if we apply forward body bias to the slower device only, which is nMOSFET in this case, the leakage overhead is only 2.6 times, which means 6 times leakage saving can be achieved than the conventional method of uniform biasing. Thus, adaptive biasing based on independent nMOSFET and pMOSFET variations is required for achieving higher energy-efficiency.

5.2.5 Worst-case DVFS

Worst-case based design methodology has severe effect on the energy efficiency which is shown in Figure 5.6. Energy per cycle is shown for four process corners of "SS", "TT" (Typical pMOSFET, Typical nMOSFET), "FF" (Fast pMOSFET, Fast nMOSFET), and "FS" (Fast pMOSFET, Slow nMOSFET). The circuit is operated at the worst-case operating speed. The minimum en-

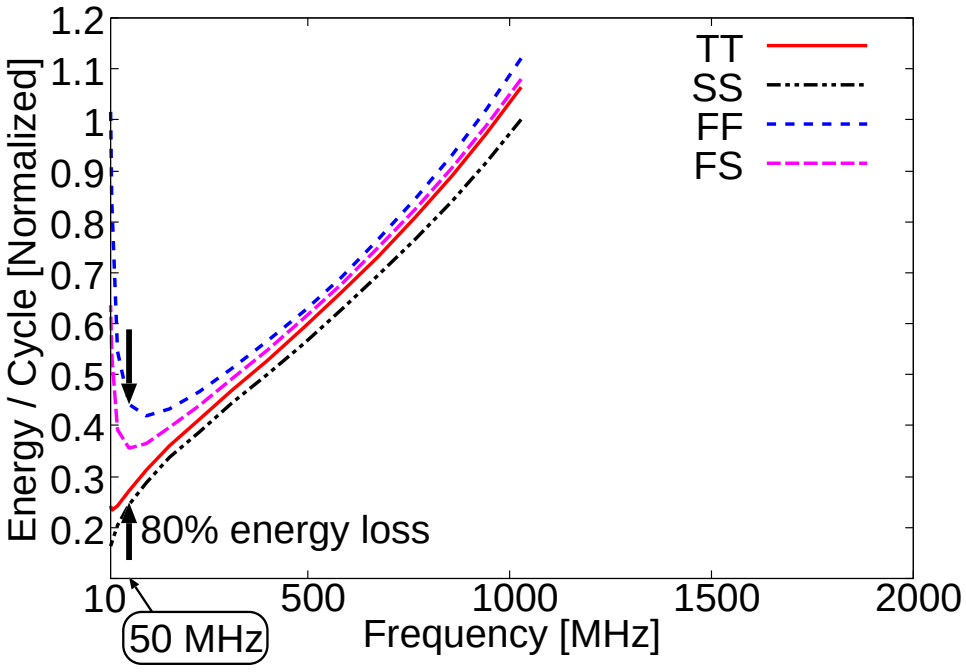


Figure 5.7: Energy per cycle at different operating frequency for several corners. Worst-case operating frequency is chosen for different supply voltages.

ergy point moves to higher supply voltage for "TT" and "FF" corners than the "SS" corner. In Figure 5.6, when the process is in the "SS" corner, V_{dd} can be set to as low as 0.5 V. However, 200% times more energy will be consumed when the process move to the "FF" corner. There are two ways to solve this problem. One is to map the frequency and V_{dd} for each chip based on the process corner. Performance monitors are needed to select the optimum frequency and V_{dd} in this case. The other is to compensate variations to achieve a fixed energy consumption. Performance compensation techniques are required here where body bias and supply voltage will be tuned based on the transistor performances. For the "FS" corner, more than 100% energy is consumed at 0.5 V operation. Thus, unbalanced corners have severe effects on both the circuit speed and energy consumption.

Figure 5.7 shows the comparison of energy consumption between the process corners against the operating frequency. Here, V_{DD} and operating frequency is adjusted to the worst-case scenario which is the "SS" corner model in this case. Large amount of energy loss is observed for corners of "TT", "FS" and "FF". At 50 MHz operating frequency, the energy loss is 90% when the process moves to the "FF" corner and 40% when the process moves to the "FS", a skewed corner. Leakage current increases exponentially with the threshold voltage reduction, thus skewed corners ("FS" and "SF") increases leakage drastically resulting in lower energy efficiency at low operating frequency. Balancing of P/N-performance is thus required for energy efficiency improvement.

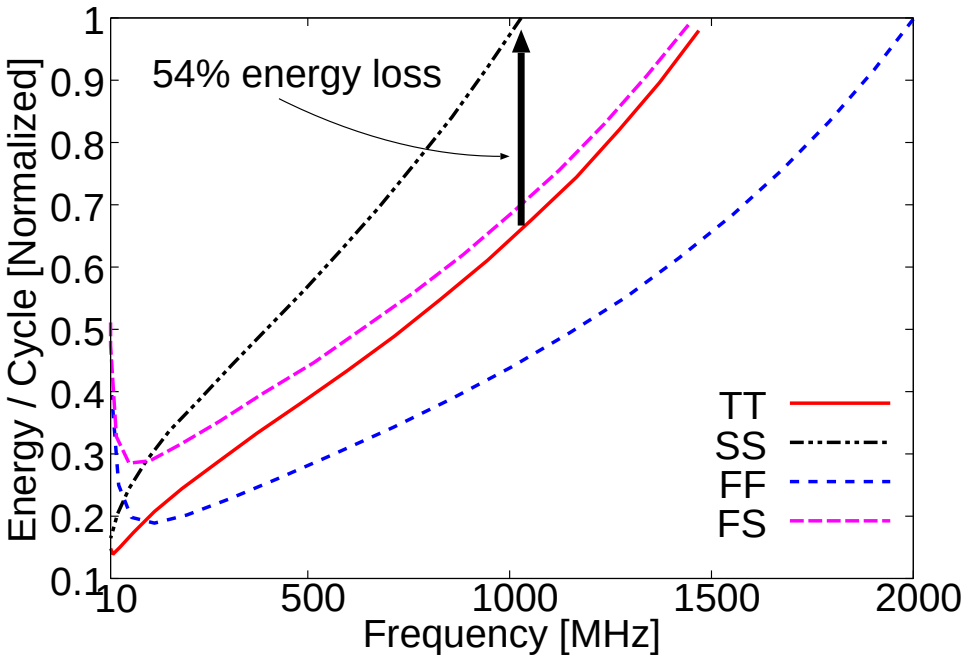


Figure 5.8: Energy per cycle at different operating frequency for different corners when circuit activity rate is 0.2. Operating frequency is chosen according to each corner's potential.

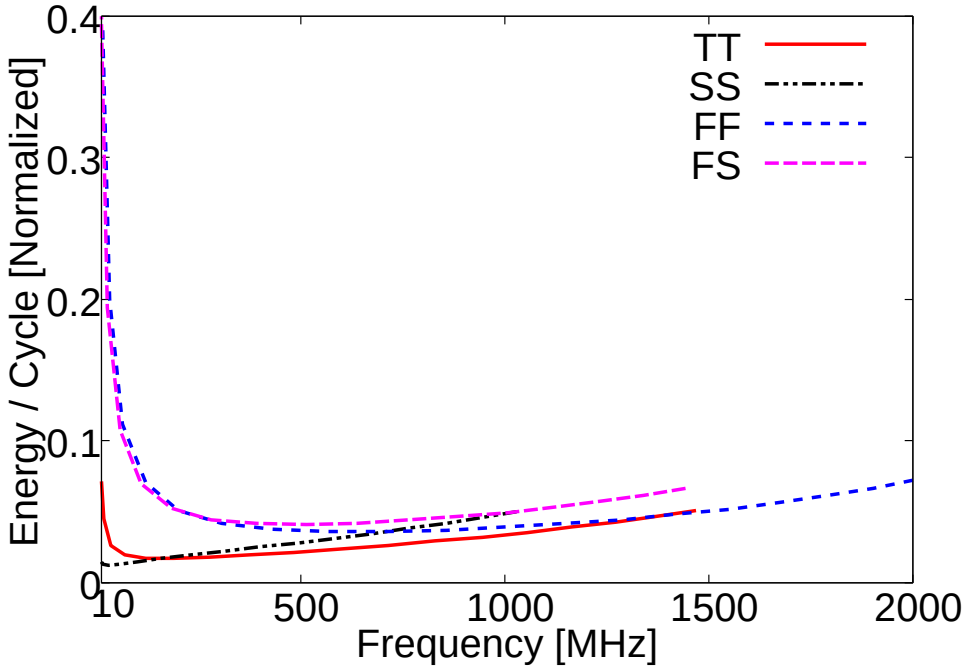


Figure 5.9: Energy per cycle at different operating frequency for different corners when circuit activity rate is 0.01. Operating frequency is chosen according to each corner's potential.

5.2.6 Variation-aware DVFS

Variation aware DVFS refers to the mapping between frequency and supply voltage for each chip by measuring its maximum operating frequency at each supply voltage. Figures 5.8 and 5.9 shows the energy consumption against operating frequency for two activity rates of 20% and

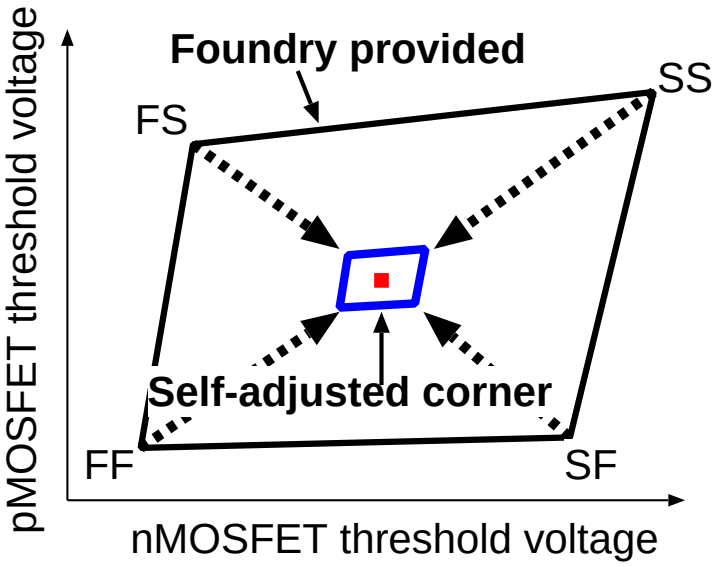


Figure 5.10: Conventional design corners vs. self-adjusted design corners. Corners are adjusted automatically by applying adaptive body bias.

1%. V_{DD} is set accordingly such that the target operating frequency can be achieved for each process corner in this figure. In Figure 5.8 (higher activity rate), the "FF" corner shows the most energy efficiency for frequency above 50 MHz. Dynamic power is dominant at higher activity, thus lowering the threshold voltage becomes more energy efficient. However, when the activity is as low as 1%, the energy profile is different. In Figure 5.9 (lower activity rate), the "TT" corner shows the most energy efficiency. One key point to note here is that the energy efficiency at "FS" corner is less that the "FF" corner. Thus, depending on the activity rate of circuits, optimum supply voltage and threshold voltage setting differs. Therefore, both supply voltage and threshold voltage need to be tuned dynamically to achieve maximum energy efficiency.

5.3 A Built-in Scheme for Runtime Performance Compensation

As discussed in the previous section, balancing of P/N is needed for better energy-efficiency. In order to reduce design margin, runtime compensation of variation is needed. In this section, a simple and digital built-in scheme for LSI performance compensation is proposed.

5.3.1 Process Corner Self-adjustment for Design Margin Reduction

Figure 5.10 shows the conventional process corners for pMOSFET and nMOSFET threshold voltages. As discussed in the previous section, designing for worst-case is inefficient in terms of both frequency and energy efficiency. In this paper, we propose a self-adjustment scheme which adjusts the threshold voltage to some fixed desired values. This self-adjustment allows us to design our circuits focusing on a fixed process corner rather than multiple process corners.

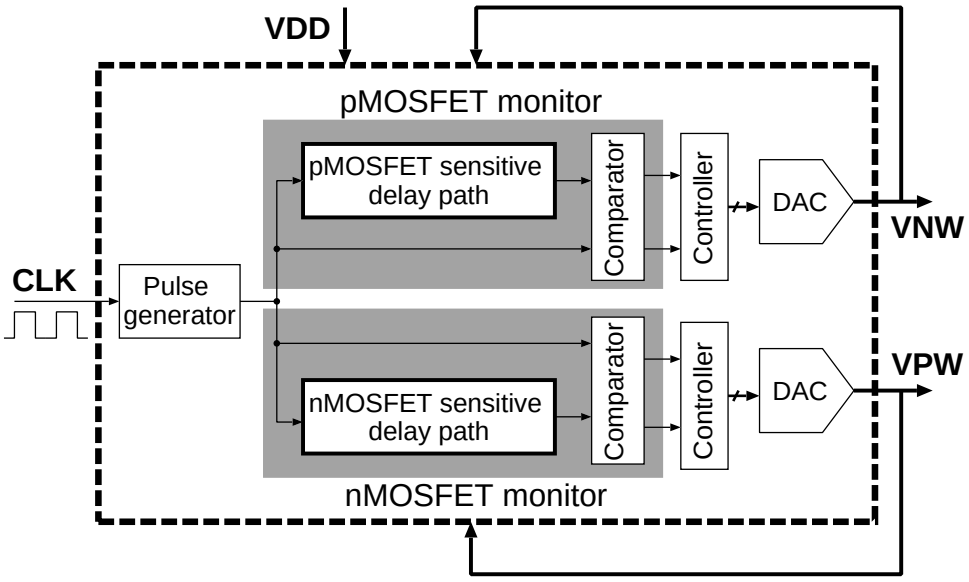


Figure 5.11: Schematic of built-in self-adjustment scheme. P/N variations are detected comparing the clock with the delays of monitor paths. System supply and clock is used to generate body voltages. (©2012 IEEE)

Thus, design margin required for worst-case design can be eliminated. The self-adjustment scheme can be used for P/N balancing which yields energy efficiency improvement.

5.3.2 Overall Architecture

Figure 5.11 shows the proposed self-adjustment scheme with body bias for P/N-performance compensation. The scheme consists of P/N-performance monitors, a controller, and DACs (Digital to Analog Converter) to generate body voltages for pMOSFET and nMOSFET independently. The system supply voltage V_{dd} and clock signal are used only for monitoring P/N-performance and generating body voltages. The proposed scheme does not require additional supply voltage or signals. The P/N-performance monitors are realized by comparing the delays of two different delay paths to the clock signal. Monitor cell structures proposed in Chapters 3 and 4 are used in the delay paths so that the delay becomes particularly sensitive to either pMOSFET or nMOSFET performance. The delay cells are standard cell compatible thus standard place and route based design is possible. The digital approach to monitor P/N-performance has less design complexity than the analog counterpart [141]. The delay path is designed so that the delay is equal to the clock period for the target process. The output signals from the P/N monitors are then fed into the controller. The controller decides the amount of biases (forward or reverse) for pMOSFET and nMOSFET and generate digital signals. Two DACs translate the digital signals to analog values of body voltages for pMOSFET and nMOSFET. The delay paths in the monitors are affected by the body voltages as well. In the next comparison, the monitor outputs will be updated thus a constant feedback loop is created.

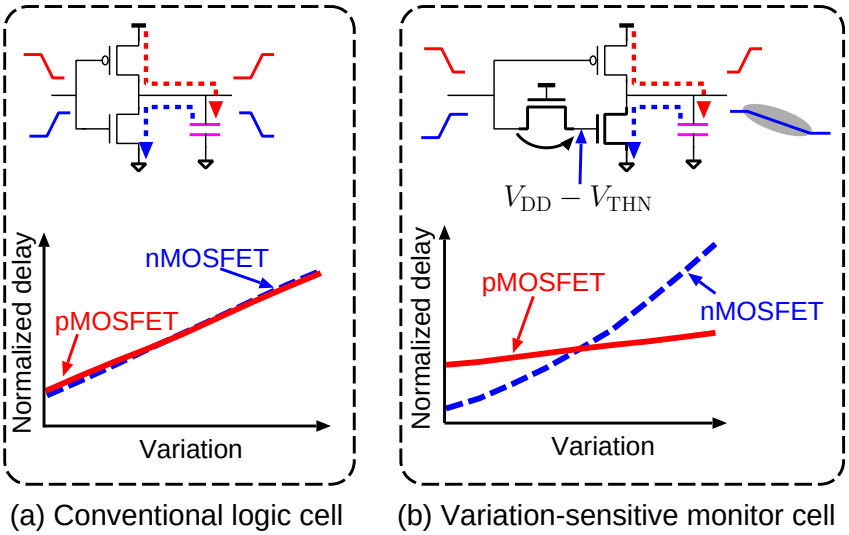


Figure 5.12: Sensitivity of the proposed monitor cell compared to conventional logic cells. The proposed monitor cell is sensitive to a particular type of MOSFET variation only.

5.3.3 Digital P/N-sensitive Monitor Cells

Figure 5.12 shows the proposed nMOSFET-sensitive monitor cell structure along with the conventional inverter structure. The nMOSFET pass-gate inserted in between the input port and the pull-down nMOSFET causes threshold voltage drop. As a result, this structure makes the fall delay 4 times larger than the rise delay at 0.7 V operation. Sensitivities of rise and fall delays to MOSFET threshold voltage variations are shown in the figure. The fall delay of monitor cell in Figure 5.12(b) is highly sensitive to nMOSFET variation whereas the rise delay has very less sensitivity to pMOSFET variation. Similarly, pMOSFET sensitive monitor cell have large sensitivity to pMOSFET performance variation. The proposed monitor cells are very simple to design and can be used in the standard cell-based design flow.

5.4 Test Chip Design

A test chip has been fabricated in a 65 nm process to demonstrate the validity of runtime performance compensation based on on-chip monitor circuits. The implemented scheme and chip layout are described in this section.

5.4.1 Operation Mode

The schematic of the compensation scheme illustrated in Figure 5.11 is implemented in the test chip. During the stable condition, comparison between delay path and clock signal needs not to be performed frequently. This allows us to reduce dynamic power by decreasing the number of comparison cycles. A pulse generator is used instead of the clock signal itself to achieve this. Pulses are generated once in every 1024 cycles. Thus, the activity of the controller circuitry

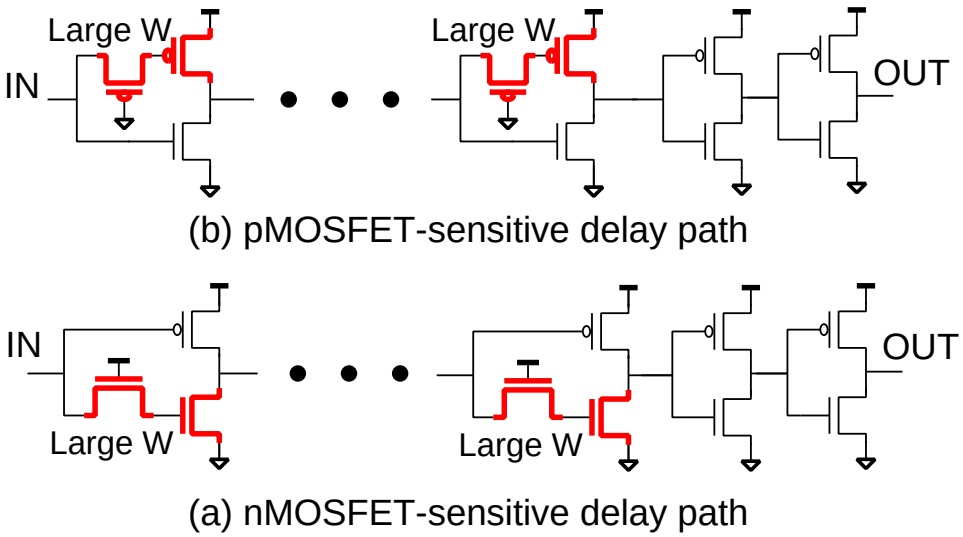


Figure 5.13: Proposed variation-sensitive delay path structure. Pass-gate at input makes the delay highly sensitive to MOSFET variation.

becomes less than 0.1% and therefore dynamic power of the controller circuitry becomes negligible. The delays are compared with the generated pulse width. Pulse width is made longer than the original clock period so that the number of stages for delay paths can be increased to reduce random variation effect. Delay T_{mon} of the monitor is set to be $T_{\mathrm{mon}} = T_{\mathrm{pulse}} + \alpha$ where T_{pulse} is the pulse width and α is for guard band. A phase comparator detects the phase difference between the delayed pulse and the input pulse, and generates up/down signals.

5.4.2 Delay Path Design

Figure 5.13 shows the delay paths to monitor pMOSFET and nMOSFET variations where monitor cells described in Sec. 5.3.3 are used. In order to reduce random variation effect, gate widths for the sensitive MOSFETs are made 4 times larger than those in the standard cells. Standard inverter cells are used at the last few stages to reshape the waveform. The monitor cell delay sensitivity to variation is multiple times larger than that of standard inverter cell. Thus, monitoring capability will not be affected by the standard cells at the output. Figure 5.14 plots the delay changes for pMOSFET and nMOSFET monitors against the threshold voltage changes. The X-axis shows nMOSFET monitor delay and the Y-axis shows pMOSFET monitor delay. When only pMOSFET threshold voltage is varied, pMOSFET monitor delay changes largely whereas nMOSFET delay does not change much. Similarly, when nMOSFET threshold voltage is varied only, nMOSFET monitor delay changes largely. Thus, P/N performances can be detected and compensated independently thanks to the high sensitivity of the monitor cells. Table 5.1 shows the comparison between different monitor circuits.

Target operation for the chip is set to $V_{\mathrm{dd}} = 0.7$ V and clock frequency (f_{clk}) of 40 MHz. The delays of the monitor circuits are set to 27 ns which is slightly larger than the clock period of 25 ns.

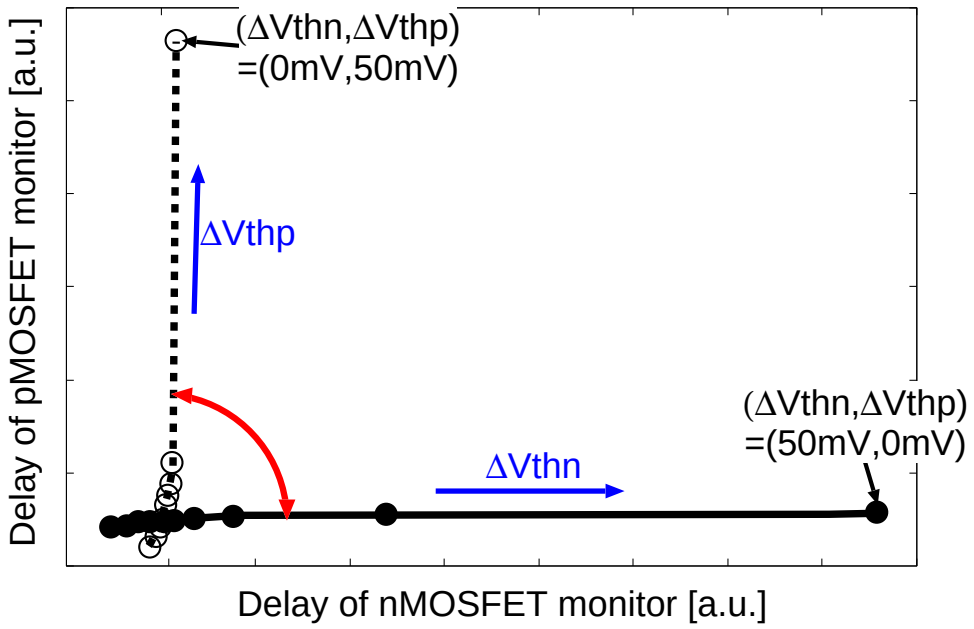


Figure 5.14: Correlation between monitor delays and MOSFET threshold voltage change. pMOSFET monitor has high sensitivity to pMOSFET threshold voltage variation. Similarly, nMOSFET monitor has high sensitivity to nMOSFET threshold voltage variation.

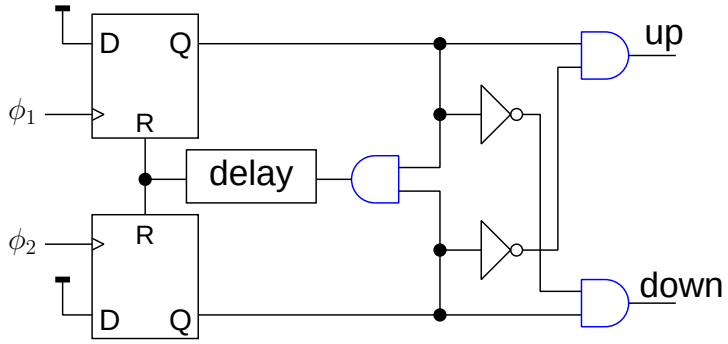


Figure 5.15: PFD used in the implementation of built-in self-adjustment scheme.

5.4.3 Comparator and Controller

Conventional PFD (Phase Frequency Detector) shown in Figure 5.15 is used to detect the phase difference between the outputs of the P/N-sensitive delay paths and the input pulse. The PFD generates up/down signals which is fed to the controller. The controller consists of 6-bit counter which counts the up/down signals. When the up signal is high, the counter value goes up and the down signal is high the counter value goes down. The counter values are then fed to the DACs which generate well voltages.

5.4.4 DACs

DACs convert the outputs of the controller to the desired body bias values. DACs capable of generating both of the forward and reverse bias are required for compensating process variations and realize process corner shrinking as shown in Fig. 5.10. In this thesis, forward body bias is

Table 5.1: Comparison between different monitor circuits.

	Sensitivity	P/N monitoring	Calibration	Design overhead
Leakage[142]	**High**	**Yes**	Analog	High
Critical path[36]	Low	No	**Digital**	**Medium**
Proposed	**High**	**Yes**	**Digital**	**Low**

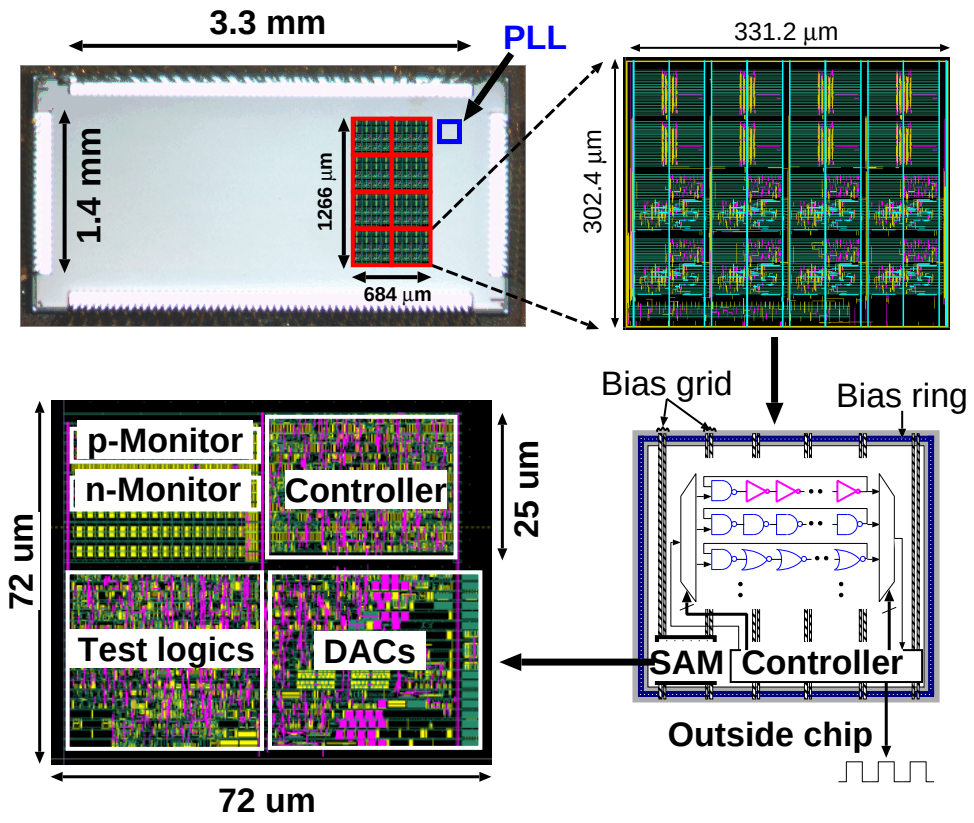


Figure 5.16: Chip photograph and layout of the self-adjustment scheme. ROs of several logic gates are implemented to evaluate critical path delays. (©2012 IEEE)

implemented in the design as demonstration of the concept of runtime variation compensation. However, reverse body bias can be applied with the proposed monitor circuits as well. In the proposed scheme, monitor and controller circuits are all digital and can be designed with the standard cell-based design flow. In order to reduce design cost, cell-based design of DACs is required. A cell-based design of charge-redistribution based DAC is presented in Ref.[110]. The area of the DAC in Ref.[110] is also very small compared with the other DAC designs reported in the literature. Therefore, the DAC proposed in Ref.[110] is adopted in this scheme for area-efficiency. The details of the DAC design can be found in [110].

5.4.5 Chip Layout

Figure 5.16 shows the chip photograph and layout of the proposed runtime compensation scheme. In order to measure gate speeds, ROs are integrated in the target substrate area. ROs consist of standard inverter (INV), 2-input NAND gate and 2-input NOR gate. ROs consisting of the proposed monitor cells are also designed to monitor process variation directly. Additional test circuits are implemented to observe the internal states of the controller, DACs and well voltages. Total area of the scheme excluding the test circuits is 2564 μm^2. In order to compensate variation at a fine-grain level, target circuit area for compensation is set to be 0.1 mm^2 in this demonstration. Area overhead is 2.6%. If the target circuit area is increased, only the design of DACs will need to be adjusted. Monitor delay paths and the controller circuits remain unchanged. Thus, Area overhead will decrease with the increase of the target circuit area. Thus, a trade-off exists on the grain size and area overhead of the compensation scheme.

5.5 Measurement Results

Test chips have been fabricated targeting "TT" condition, as well as four corners of "SS", "FF", "FS" and "SF". In this section, first we show that the proposed monitor circuits can detect P/N variations correctly. Next, we show that the well voltages are generated according to the monitored variations. Finally, speed compensation results for different logic gates and corresponding leakage current will be presented. All the measurements are done at V_{dd} = 0.7 V.

5.5.1 Transient Response

Transient response of the system is measured when self-adjustment is enabled. Figure 5.17 shows measured transient response for "FS" (Fast pMOSFET and slow nMOSFET) corner. After self-adjustment is enabled, body voltage of nMOSFET (VPW) is gradually increased until the delay of nMOSFET monitor is smaller than the target value. pMOSFET body voltage remains constant as pMOSFET is faster than the target performance. Thus, independent control of nMOSFET and pMOSFET is confirmed.

5.5.2 Monitor Outputs and Body Voltage Measurements

Figure 5.18 plots the output frequency of pMOSFET-sensitive RO against the output frequency of nMOSFET-sensitive RO to illustrate the corner conditions. Speeds are normalized by their target values. Open circles show the measured values before self-adjustment and closed circles show the values after self-adjustment. Generated body biases for each corner are also shown. Maximum of 0.34 V body bias was required to adjust speed at "SS" corner. After self-adjustment, the "SS" chip is moved to a point where the speeds of both monitors are above the target values. Similarly, "SF", "TT" and "FS" chips are also moved so that the target speeds are achieved.

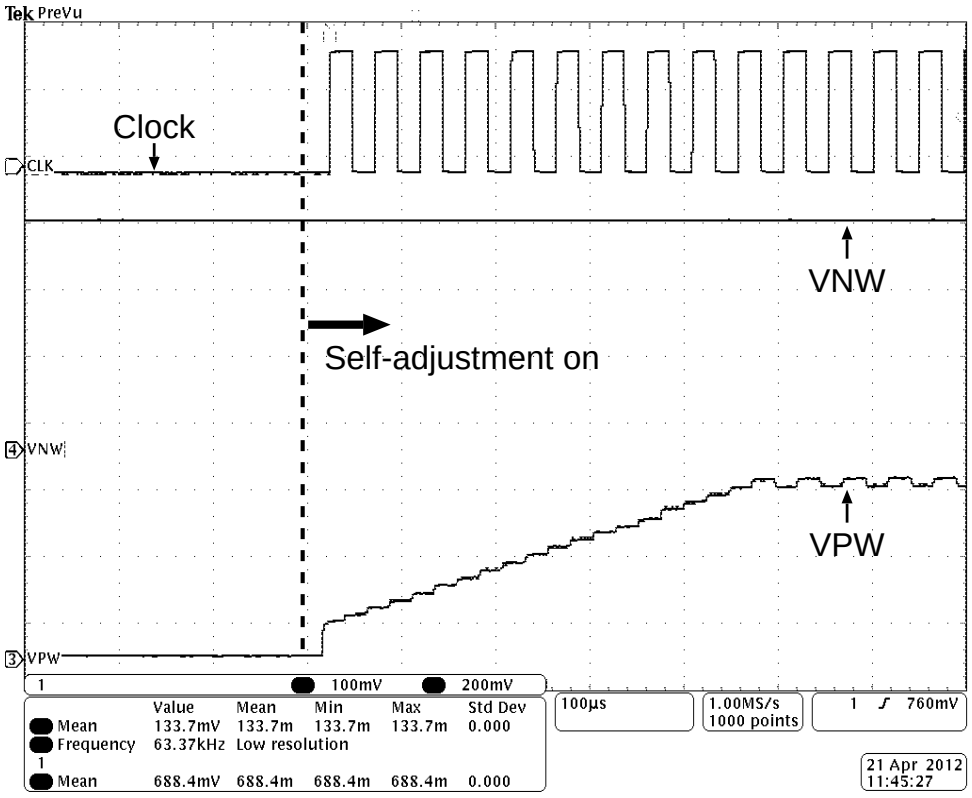


Figure 5.17: Measured transient response of the self-adjusting module when self-adjustment is enabled. System stability and independent control of body bias is confirmed. (©2012 IEEE)

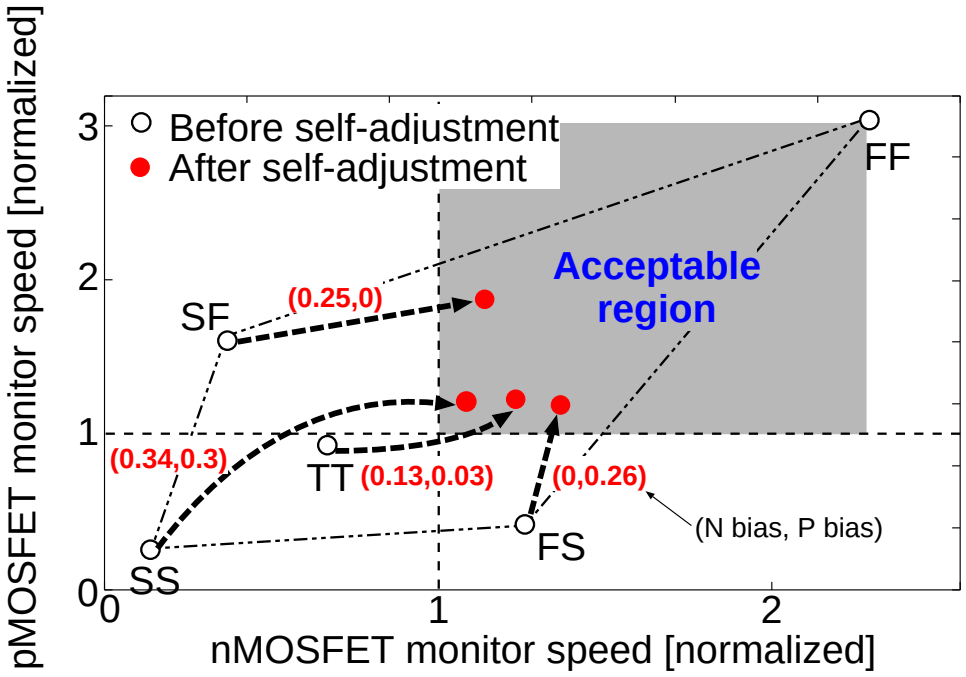


Figure 5.18: Output performances of pMOSFET-sensitive and nMOSFET-sensitive ROs. Performances are measured before and after self-adjustment. Generated body biases for each corner are shown in closed bracket. (©2012 IEEE)

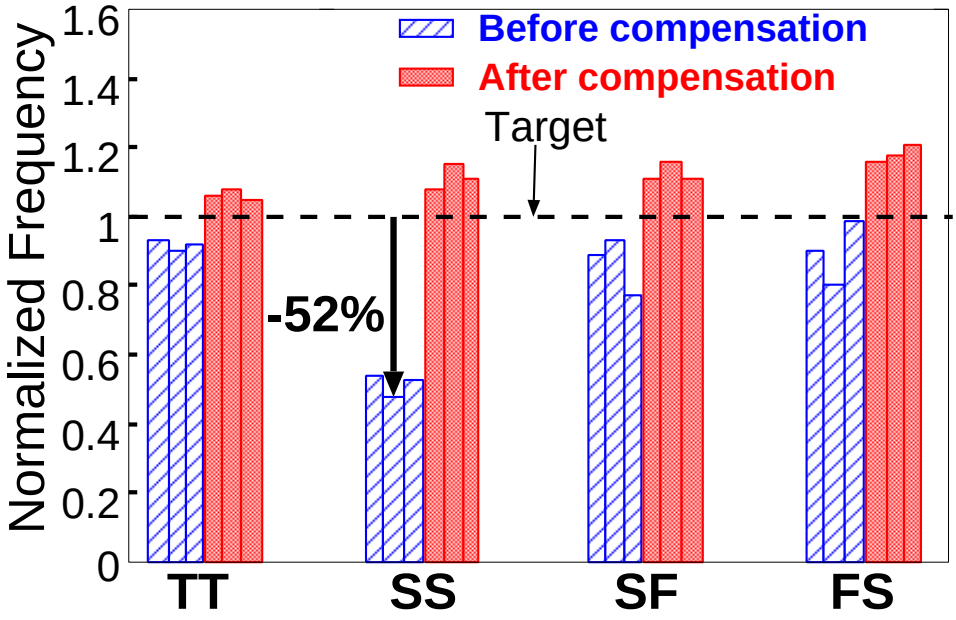


Figure 5.19: Frequencies of ROs consisting of various kinds of gates. INV, NAND and NOR frequencies are plotted for each corner from the left. (©2012 IEEE)

5.5.3 Speed Measurement

Speeds of different gates are measured through RO frequencies as they are the most concern. Figure 5.19 shows the values of INV, NAND and NOR RO frequencies for all the corner chips. Frequencies before and after self-adjustment are shown. The values are normalized by the target values. The worst case speed degradation here is −52% at "SS" corner. After self-adjustment, the frequencies go over the target values. For "FS" and "SF" corner chips, variations among the gate speeds become significant as pMOSFET and nMOSFET move to opposite directions. After self-adjustment, all the gates achieve the target speeds.

5.5.4 Leakage Measurement

Leakage currents are measured before and after self-adjustment. Two types of self-adjustment is compared to demonstrate the need for P/N-sensitive monitors. One is the conventional critical path delay based adjustment using uniform biasing, another is the proposed one. Figure 5.20 shows the leakage currents for "TT", "FS" and "SF" corners. 2.6 times leakage saving is achieved by using the proposed monitor circuits than the conventional critical path based method.

5.5.5 Comparison between Worst-case and Typical-case Designs

The measurement results from the chips targeting all the corners of "TT", "SS", "FF", "FS" and "SF" prove that global variation component can be compensated by applying adequate

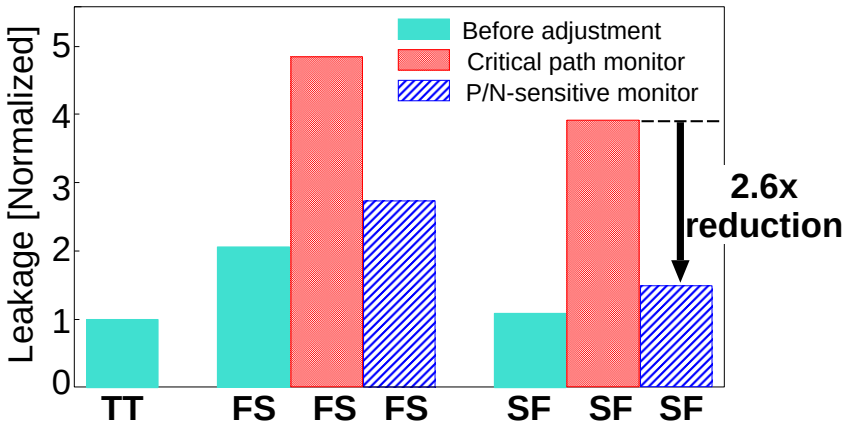


Figure 5.20: Leakage measurement for "TT", "FS" and "SS" chips when (a) both MOSFETs are biased uniformly and (b) proposed scheme is applied. (©2012 IEEE)

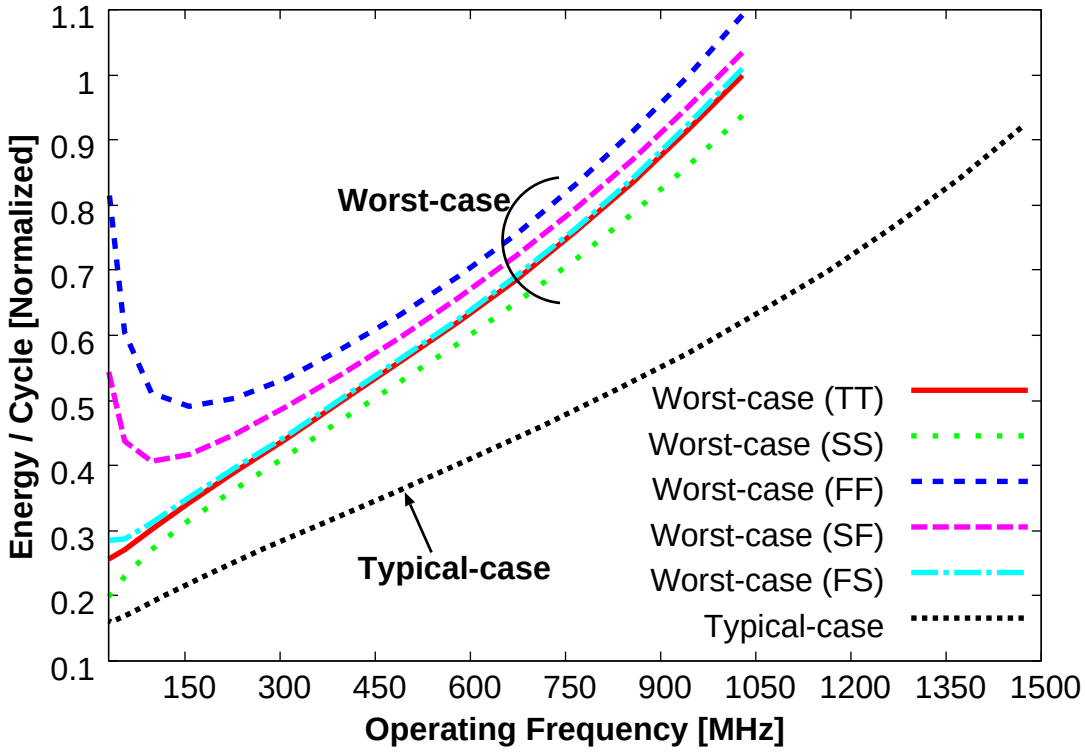


Figure 5.21: Simulated energy consumption per cycle versus operating frequency for worst-case and typical-case designs. Circuit activity of 10% is assumed in the simulation.

body bias based on monitor circuits. Thus, the circuit designer can design the circuit targeting a typical condition rather than the worst-case condition as shown in Fig. 5.10. The design complexity will be largely reduced, therefore target circuit can be operated at higher operating frequency and lower supply voltage. Using the model circuit of Fig. 5.1, the effect of worst-case design and typical-case design on circuit operating frequency and energy consumption are simulated. Figure 5.21 shows simulated energy consumption per cycle against operating frequency for worst-case and typical-case designs. Circuit activity of 10% is assumed in the

simulation. For the worst-case design, the maximum operating frequency is determined by the worst process condition which is the "SS" corner in this case. Typical-case refers to the design where the circuit is designed targeting only the "TT" process corner. Any deviation from the "TT" corner is assumed to be compensated using the on-chip monitor circuits and applying body bias. Maximum of 1000 MHz operating frequency can be achieved under the supply voltage of 1.2 V for the worst-case design, whereas under the same supply voltage, typical-case design can achieve operating frequency of 1450 MHz. When the target operating frequency is 1000 MHz, typical-case design reduces energy consumption per cycle by 33–42% compared with the worst-case design. 33% of energy reduction is achieved when the process moves to the "SS" corner and 42% of energy reduction is achieved when the process moves to the "FF" corner. When the target operating frequency is as low as 50 MHz, typical-case design reduces energy consumption by 31–75%. At lower operating frequency, leakage current becomes dominant, thus larger amount of energy reduction can be achieved by compensating process variation compared to a higher operating frequency. The amount of energy consumption also varies depending on circuit activity and temperature. Adjusting transistor threshold voltages considering circuit activity and temperature along with process variation will further reduce the energy consumption.

5.6 Summary

A simple and digital built-in self-adjustment scheme is developed for runtime variability compensation. Measurements from several corner chips proves that performance compensation based on the developed on-chip monitor circuits is feasible. Self-adjustment of P/N performances is confirmed. Performance measurements of logic gates confirm more than 50% of speed recovery at 0.7 V operation. 2.6 times leakage saving is achieved compared to the conventional critical path based method using uniform biasing. Dynamic power of the propose monitor scheme is made negligible by reducing the activity rate to less than 0.1%. The proposed scheme eliminates the need for large design margins allocated for global variations. Furthermore, the proposed scheme can be applied for adjusting optimum threshold voltages at different supply voltages to achieve maximum energy-efficiency. Experimental results based on simple model circuit show that 42% of energy reduction at 1000 MHz operation and 75% of energy reduction at 50 MHz operation can be achieved considering on-chip monitor circuit based variation compensation during the design phase.

Chapter 6

Conclusion

In this thesis, on-chip monitor circuits suitable for monitoring of device variations have been proposed. An area-efficient universal monitor scheme is developed to monitor different types of variations for different MOSFET types. The monitor circuit thus provides an interface to the designers as well as to the system to fine tune both of the design and the system for energy-efficient operation. In this chapter, key contributions of the thesis is discussed in Sec. 6.1. Future work and limitation of this work are discussed in Sec. 6.2. Finally, Sec. 6.3 puts final remarks on the thesis.

6.1 Key Contributions

The key contributions of this thesis is two-fold. Firstly, digital monitor circuit topologies have been developed to estimate process parameter variations which can be used to build accurate variation models. The variation models are then used to estimate LSI performance accurately. The key feature is that on-chip monitor circuits will allow each chip to have its own variation model as variation models may differ from chip to chip. This will also reduce debugging time as each chip can be tested based on its own model. Chapter 3 discussed the proposed estimation methodology of global and local variations. Monitor circuits suitable for parameter estimation have been proposed. The monitor circuits are designed to have high sensitivity to a particular variability source enabling accurate monitoring of different variations. Measurement and analysis of different types of variations are time consuming and costly. Simple digital circuit based monitor circuit is developed allowing fast measurement of variation. The validity of the proposed monitor circuits has been confirmed by measuring test chips fabricated in a 65-nm process technology. Successful extraction of global and local variations in MOSFET threshold voltage and gate length variation have been performed. The estimated variations are verified by multiple ways. The results match with device level measurements reported in the literature showing the accuracy of the estimation. The proposed on-chip monitor circuits allow the designers to get accurate variation information for their circuits which will eliminate unnecessary margins thus reduce pessimism. The reduced pessimism will relax the design constraints thus circuits with

higher speed and lower energy consumption can be designed. Using the digital nature of the developed monitor circuit topologies, an area-efficient topology-reconfigurable monitor circuit scheme is developed by which both of the static and dynamic variations for nMOSFET and pMOSFET can be monitored. Chapter 4 discussed the monitor circuit architecture and monitor cell design to realize topology-configurability. Measurement of random variation has been made possible for the first time from a single monitor instance. This is achieved by developing a new measurement method which exploits an inhomogeneous configuration for ring oscillator circuit. By making the inhomogeneous stage having sensitivity multiple times larger than the other stages, measurement of gate delay variation becomes possible by scanning and measuring the inhomogeneous configurations. As only a single monitor instance is now required, the area of the monitor circuit has been reduced by 95% than the conventional approach where large number of monitor instances are placed to acquire statistically sufficient sample number. The extremely small area of the topology-reconfigurable monitor circuit will also allow the designers put monitor circuits at various parts of the chip. Amount of variation differs depending on the device sizes, layout, etc. Monitor circuits with different device sizes will provide accurate variation models for the target device size. This thesis establishes all-digital small area on-chip monitor circuit which can be used for energy-efficient LSI design.

Secondly, dynamic compensation scheme of global component of variation ,which is common for all the devices in a chip, has been developed using the proposed monitor circuit topologies. Designing LSI for worst-case scenario is energy-inefficient. Chapter 5 discussed the overall architecture and key design issues for dynamic compensation with adaptive body bias. Effect of variation on LSI performance is addressed. The thesis shows that body bias can be used as a mean to tune MOSFET threshold voltage effectively for energy reduction. The key feature of this approach is that it targets compensating the sources of variations by digitally monitoring delay. The digital and compact nature of the proposed monitor circuits allow area-efficient implementation of the scheme thus fine-grain compensation can be possible. Utilizing the high sensitivity of the proposed monitor topology towards a single MOSFET type, independent control of pMOSFET and nMOSFET has been made possible. The proposed scheme adjusts MOSFET threshold voltages to their target values. Thus, application-specific threshold voltage assignment can be made possible with the proposed technique. Compared to the other approaches where analog components are used, this thesis has successfully demonstrated that independent control of MOSFET variation is possible using digital on-chip monitor circuits. The scheme is validated with test chips targeting all the extreme process corners and successful compensation has been achieved even for the worst possible corner. This thesis reveals that variability effect of LSI energy is severe at low supply voltage operation. More than 50% of speed compensation at a low supply voltage operation of 0.7 V has been confirmed. The implemented scheme shows the validity of the proposed on-chip monitor circuits. Utilizing the dynamic auto compensation mechanism during the design phase, LSI energy consumption can be reduced by maximum of 75%. Design and test cost will reduce exponentially as the circuits can be designed now targeting a single process corner rather than various extreme corners. One

of the key contributions of this thesis is it develops cell-based design techniques for monitoring and controlling units which reduces design cost. Targeting a circuit with area of as small as 0.1 mm^2, the area overhead of the proposed scheme is less than 3%.

6.2 Future Work

In order to utilize the proposed monitor circuits during the design phase, support from the EDA tools is required. At present, the design flow is closed in the design phase. The present design methodology thus need to be enhanced to utilize the on-chip monitor circuits during the design phase. One example of such a design flow can be to first design the circuit with some amount of margin based on the foundry based models. Then after some test runs, the variation information monitored by the on-chip monitor circuits will be fed to the tools. Tools will update the models and then the design will be tuned accordingly. Furthermore, integration and measurement of on-chip monitor circuits need to be made seamless so that users do not have any extra cost in integrating the monitor circuits to their target circuits. The monitor circuits are independent from the target circuit thus they will not affect the performance of the target circuit. The test tools need to adapt the use of variation information from the on-chip monitor circuits. Thus, several challenges exist for the EDA tool developers. The biggest challenge is to make the transition from the present design methodology to the on-chip monitor circuit oriented design methodology as seamless as possible.

As the present SoC contains multiple circuit blocks with different functionality, compensation scheme for each block is required to maximize energy-efficiency as activity of each block may differ largely. Furthermore, each block has different operating speed, thus design strategies will differ. The proposed dynamic compensation technique using on-chip monitor circuits can be applied to each block. Larger blocks can be divided into smaller groups further for fine-grain compensation of variations. The present EDA tools already has the support to divide the design into several voltage islands. The tools need to support to divide the chip into several body bias islands. Placement and routing monitor circuits along with the body bias generation circuits need to be automated. Testing the circuits which is divided into several voltage and body bias islands is a challenge. New test strategies need to be developed to efficiently test each part with the help of on-chip monitor circuits. The monitor circuits can also be used as a part of testing the functionality of the body bias generators by tracing the output.

In a DVFS (Dynamic Voltage and Frequency Scaling) architecture, the operating system (OS) determines the adequate supply voltage and frequency for each task such that no deadline violation occurs. However, many applications do not need high reliability such as video streaming application where alternation of several pixels will not be detected by the eye. Thus, opportunity for further reduction of large amount of energy exists. Real-time variation information of the hardware can be used to estimate the rate of violation for example, which then can be used to estimate the level of reliability for a certain supply voltage and frequency combination. If an application can tolerate 5% of reliability loss, then supply voltage can be reduced further

resulting in further energy-efficiency improvement. For applications requiring high reliability supply voltage will be set accordingly. Therefore, on-chip monitor circuits can play a vital role in reducing the system energy consumption. Variability-aware software need to be developed for effectively using each component in a system on chip [143]. The proposed on-chip monitor circuit provides device characteristics on the runtime, thus runtime hardware information can be made available to the operating system. The information can be used to set the operating point based on the required level of reliability. This will provide the system both high reliability as well as energy-efficiency.

In order to tune system parameters effectively for higher energy-efficiency, temperature and power monitors are also required. The target of this thesis has been to monitor LSI variability resulting from device level variations. Integrating the device level variation monitors along with other temperature and power monitors will consume large area and may limit the usage of these monitor for general purpose circuits. In order to widen the use of on-chip monitors, device variation monitors and temperature monitors need to be merged. New monitor architecture can be developed which will provide device variation as well as temperature variation as well.

6.3 Summary

In conclusion, all digital area-efficient on-chip monitor circuits have been proposed by which monitoring of global and local variations as well as dynamic variations for different MOSFET types become possible. The proposed circuits can be used for building accurate variation models, post-silicon debugging during test as well as dynamic tuning of threshold voltage for high energy-efficiency. Test chip implementation in a 65 nm process validates the proposed circuits.

Bibliography

[1] I. A.K.M Mahfuzul, A. Tsuchiya, K. Kobayashi, and H. Onodera, "Process-sensitive Monitor Circuits for Estimation of Die-to-Die Process Variability," in *ACM International Workshop on Timing Issues in the Specification and Synthesis of Digital Systems*, 2010, pp. 83–88.

[2] I. Mahfuzul, A. Tsuchiya, K. Kobayashi, and H. Onodera, "Variation-sensitive Monitor Circuits for Estimation of Die-to-Die Process Variation," in *IEEE Intl. Conference on Microelectronic Test Structures*, 2011, pp. 153–157.

[3] A. Islam, A. Tsuchiya, K. Kobayashi, and H. Onodera, "Variation-sensitive Monitor Circuits for Estimation of Global Process Parameter Variation," *IEEE Transactions on Semiconductor Manufacturing*, vol. 25, no. 4, pp. 571–580, 2012.

[4] S. Fujimoto, I. A. Mahfuzul, T. Matsumoto, and H. Onodera, "Inhomogeneous Ring Oscillator for WID Variability and RTN Characterization," in *IEEE Intl. Conference on Microelectronic Test Structures*, 2012, pp. 25–30.

[5] S. Fujimoto, A. K. M. M. Islam, T. Matsumoto, and H. Onodera, "Inhomogeneous Ring Oscillator for Within-Die Variability and RTN Characterization," *IEEE Transactions on Semiconductor Manufacturing*, vol. 26, no. 3, pp. 296–305, 2013.

[6] I. A. Mahfuzul and H. Onodera, "On-Chip Detection of Process Shift and Process Spread for Silicon Debugging and Model-Hardware Correlation," in *21st IEEE Asian Test Symposium*, Nov. 2012, pp. 350–354.

[7] A. M. Islam and H. Onodera, "On-Chip Detection of Process Shift and Process Spread for Post-Silicon Diagnosis and Model-Hardware Correlation," *IEICE Transactions on Information and Systems*, vol. E96-D, no. 9, pp. 1971–1979, 2013.

[8] A. K. M. M. Islam and H. Onodera, "Area-efficient Reconfigurable Ring Oscillator for Characterization of Static and Dynamic Variations," in *International Conference on Solid State Devices and Materials*, 2013, pp. 132–133.

[9] A. M. Islam, T. Ishihara, and H. Onodera, "Reconfigurable Delay Cell for Area-efficient Implementation of On-chip MOSFET Monitor Schemes," in *IEEE Asian Solid State Circuits Conference*, 2013, pp. 125–128.

[10] I. Mahfuzul, N. Kamae, T. Ishihara, and H. Onodera, "A Built-in Self-adjustment Scheme with Adaptive Body Bias using P/N-sensitive Digital Monitor Circuits," in *IEEE Asian Solid State Circuits Conference*, 2012, pp. 101–104.

[11] A. M. Islam, N. Kamae, T. Ishihara, and H. Onodera, "Energy-efficient Dynamic Voltage and Frequency Scaling by P/N-performance Self-adjustment using Adaptive Body Bias," in *Proceedings of SASIMI*, 2013.

[12] R. H. Dennard, F. H. Gaensslen, H.-N. Yu, V. L. Rideout, E. Bassous, and A. R. Leblanc, "Design of Ion-Implanted MOSFET 'S with Very Small Physical Dimensions," *IEEE Journal of Solid-State Circuits*, vol. 9, no. 5, pp. 256–268, 1974.

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

[14] ITRS, "International Technology Roadmap for Semiconductors," Tech. Rep. [Online]. Available: http://www.itrs.net

[15] V. Joshi, K. Agarwal, D. Blaauw, and D. Sylvester, "Analysis and Optimization of SRAM Robustness for Double Patterning Lithography," in *IEEE/ACM International Conference on Computer-Aided Design*, Nov. 2010, pp. 25–31.

[16] K. Itoh, "Adaptive Circuits for the 0.5-V Nanoscale CMOS Era," in *IEEE International Solid-State Circuits Conference*, no. 1, 2009, pp. 14–20.

[17] P. Tsui, "Limitation of CMOS Supply-Voltage Scaling by MOSFET Threshold-Voltage Variation," *IEEE Journal of Solid-State Circuits*, vol. 30, no. 8, pp. 947–949, 1995.

[18] H. Esmaeilzadeh, E. Blem, R. St. Amant, K. Sankaralingam, and D. Burger, "Dark Silicon and the End of Multicore Scaling," in *ACM International Symposium on Computer Architecture*, New York, New York, USA, 2011, p. 365.

[19] D. Hisamoto, W.-c. Lee, J. Kedzierski, H. Takeuchi, K. Asano, C. Kuo, E. Anderson, T.-j. King, J. Bokor, and C. Hu, "FinFET — A Self-Aligned Double-Gate MOSFET," *IEEE Transactions on Electron Devices*, vol. 47, no. 12, pp. 2320–2325, 2000.

[20] J.-P. Colinge, "Multiple-gate SOI MOSFETs," *Solid-State Electronics*, vol. 48, no. 6, pp. 897–905, Jun. 2004.

[21] X. Wang, A. R. Brown, and A. Asenov, "Statistical Variability and Reliability in Nanoscale FinFETs," in *IEEE International Electron Devices Meeting*, Dec. 2011, pp. 5.4.1–5.4.4.

[22] R. Dreslinski, M. Wieckowski, D. Blaauw, D. Sylvester, and T. Mudge, "Near-Threshold Computing: Reclaiming Moore's Law Through Energy Efficient Integrated Circuits," *Proceedings of the IEEE*, vol. 98, no. 2, pp. 253–266, Feb. 2010.

[23] S. Dighe, S. Vangal, P. Aseron, S. Kumar, T. Jacob, K. Bowman, J. Howard, J. Tschanz, V. Erraguntla, N. Borkar, and Others, "Within-Die Variation-Aware Dynamic- Voltage-Frequency-Scaling With Optimal Core Allocation and Thread Hopping for the 80-Core TeraFLOPS Processor," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 1, pp. 184–193, 2011.

[24] T. Yasufuku, S. Iida, H. Fuketa, K. Hirairi, M. Nomura, M. Takamiya, and T. Sakurai, "Investigation of Determinant Factors of Minimum Operating Voltage of Logic Gates in 65-nm CMOS," in *IEEE/ACM International Symposium on Low Power Electronics and Design*, vol. 2, Aug. 2011, pp. 21–26.

[25] H. Fuketa, K. Hirairi, T. Yasufuku, M. Takamiya, M. Nomura, H. Shinohara, and T. Sakurai, "Minimizing Energy of Integer Unit by Higher Voltage Flip-Flop: VDDmin-Aware Dual Supply Voltage Technique," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 21, no. 6, pp. 1175–1179, 2013.

[26] S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, and V. De, "Parameter Variations and Impact on Circuits and Microarchitecture," in *Design Automation Conference*, 2003, pp. 338–342.

[27] B. C. Paul, K. Kang, H. Kufluoglu, M. A. Alam, and K. Roy, "Impact of NBTI on the Temporal Performance Degradation of Digital Circuits," *IEEE Electron Device Letters*, vol. 26, no. 8, pp. 560–562, 2005.

[28] S. Kumar, C. Kim, and S. Sapatnekar, "Impact of NBTI on SRAM Read Stability and Design for Reliability," in *International Symposium on Quality Electronic Design*, 2006, pp. 210–218.

[29] M. Bohr, "The New Era of Scaling in an SoC World," in *International Solid State Circuits Conference*, 2009, pp. 23–28.

[30] M. Floyd, M. Allen-ware, B. Brock, A. J. Drake, and J. A. Tierno, "Introducing the Adaptive Energy Management Features of the POWER7 Chip," *IEEE Micro*, vol. 31, no. 2, pp. 60–75, 2011.

[31] C. R. Lefurgy, A. J. Drake, M. S. Floyd, M. S. Allen-ware, B. Brock, J. A. Tierno, J. B. Carter, and B. Rd, "Active Management of Timing Guardband to Save Energy in POWER7," in *IEEE/ACM International Symposium on Microarchitecture*, 2011, pp. 1–11.

[32] L. A. D. Bathen, N. Dutt, A. Nicolau, M. Gottscho, and P. Gupta, "ViPZonE : OS-Level Memory Variability-Driven Physical Address Zoning for Energy Savings," in *IEEE/ACM/IFIP international conference on Hardware/software codesign and system synthesis*, 2012, pp. 33–42.

[33] R. Teodorescu and J. Torrellas, "Variation-Aware Application Scheduling and Power Management for Chip Multiprocessors," in *2008 International Symposium on Computer Architecture*, Jun. 2008, pp. 363–374.

[34] K. J. Nowka, G. D. Carpenter, E. W. Macdonald, H. C. Ngo, B. C. Brock, K. I. Ishii, T. Y. Nguyen, and J. L. Burns, "A 32-bit PowerPC System-on-a-Chip With Support for Dynamic Voltage Scaling and Dynamic Frequency Scaling," *IEEE Journal of Solid-State Circuits*, vol. 37, no. 11, pp. 1441–1447, 2002.

[35] J. Howard, S. Dighe, Y. Hoskote, S. Vangal, D. Finan, G. Ruhl, D. Jenkins, H. Wilson, N. Borkar, G. Schrom, F. Pailet, S. Jain, T. Jacob, S. Yada, S. Marella, P. Salihundam, V. Erraguntla, M. Konow, M. Riepen, G. Droege, J. Lindemann, M. Gries, T. Apel, K. Henriss, T. Lund-larsen, S. Steibl, S. Borkar, V. De, R. V. D. Wijngaart, and T. Mattson, "DVFS in 45nm CMOS A 48-Core IA-32 Message-Passing Processor with DVFS in 45nm CMOS," in *International Solid State Circuits Conference*, vol. 9, no. 2, 2010, pp. 58–59.

[36] J. Tschanz, J. Kao, S. Narendra, R. Nair, D. Antoniadis, A. Chandrakasan, and V. De, "Adaptive Body Bias for Reducing Impacts of Die-to-Die and Within-Die Parameter Variations on Microprocessor Frequency and Leakage," *IEEE Journal of Solid-State Circuits*, vol. 37, no. 11, pp. 1396–1402, Nov. 2002.

[37] J. Tschanz, S. Narendra, R. Nair, and V. De, "Effectiveness of Adaptive Supply Voltage and Body Bias for Reducing Impact of Parameter Variations in Low Power and High Performance Microprocessors," in *Symposium on VLSI Circuits*, 2002, pp. 310–311.

[38] K. Chae, X. Zhao, S. Lim, and S. Mukhopadhyay, "Tier Adaptive Body Biasing: A Post-Silicon Tuning Method to Minimize Clock Skew Variations in 3-D ICs," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 10, no. 3, pp. 1720 – 1730, 2013.

[39] M. Sumita, "High Resolution Body Bias Techniques for Reducing the Impacts of Leakage Current and Parasitic Bipolar," in *International Symposium on Low Power Electronics and Design (ISLPED)*, New York, New York, USA, 2005, p. 203.

[40] H. Mostafa, M. Anis, and M. Elmasry, "A Novel Low Area Overhead Direct Adaptive Body Bias (D-ABB) Circuit for Die-to-Die and Within-Die Variations Compensation," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 19, no. 10, pp. 1848–1860, 2011.

[41] H. Mostafa, M. Anis, and M. Elmasry, "NBTI and Process Variations Compensation Circuits Using Adaptive Body Bias," *IEEE Transactions on Semiconductor Manufacturing*, vol. 25, no. 3, pp. 460–467, 2012.

[42] S. M. Martin, T. Mudge, K. Flautner, and D. Blaauw, "Combined Dynamic Voltage Scaling and Adaptive Body Biasing for Lower Power Microprocessors under Dynamic Workloads," *IEEE International Conference on Computer-Aided Design*, vol. 0, no. 2, pp. 721–725, 2002.

[43] A. Bonnoit, S. Herbert, D. Marculescu, and L. Pileggi, "Integrating Dynamic Voltage/Frequency Scaling and Adaptive Body Biasing using Test-time Voltage Selection," in *ACM/IEEE international symposium on Low power electronics and design*, New York, New York, USA, 2009, p. 207.

[44] S. Herbert and D. Marculescu, "Variation-Aware Dynamic Voltage/Frequency Scaling," in *IEEE International Symposium on High Performance Computer Architecture*, Feb. 2009, pp. 301–312.

[45] M. Elgebaly and M. Sachdev, "Variation-Aware Adaptive Voltage Scaling System," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 15, no. 5, pp. 560–571, May 2007.

[46] M. Basoglu, M. Orshansky, and M. Erez, "NBTI-Aware DVFS: A New Approach to Saving Energy and Increasing Processor Lifetime," in *16th ACM/IEEE international symposium on Low power electronics and design*, 2010, pp. 253–258.

[47] M. Fujii, H. Suzuki, and H. Notani, "On-chip leakage monitor circuit to scan optimal reverse bias voltage for adaptive body-bias circuit under gate induced drain leakage effect," in *European Solid-State Circuits Conference*, vol. 2, 2008, pp. 258–261.

[48] D. Ernst, N. S. Kim, S. Das, S. Pant, R. Rao, T. Pham, C. Ziesler, D. Blaauw, T. Austin, K. Flautner, T. Mudge, B. Ave, and A. Arbor, "Razor: A Low-Power Pipeline Based on Circuit-Level Timing Speculation," in *IEEE/ACM International Symposium on Microarchitecture*, 2003, pp. 7–18.

[49] T. Sato and Y. Kunitake, "A Simple Flip-Flop Circuit for Typical-Case Designs for DFM," in *International Symposium on Quality Electronic Design,*, 2007, pp. 539–544.

[50] D. Blaauw, S. Kalaiselvan, K. Lai, S. Pant, C. Tokunaga, S. Das, and D. Bull, "Razor II: In Situ Error Detection and Correction for PVT and SER Tolerance," in *IEEE International Solid-State Circuits Conference (ISSCC)*, 2011, pp. 400–402.

[51] B. P. Das and H. Onodera, "Warning Prediction Sequential for Transient Error Prevention," in *International Symposium on Defect and Fault Tolerance in VLSI Systems*, 2010, pp. 382–390.

[52] A. Drake, R. Senger, H. Deogun, G. Carpenter, S. Ghiasi, T. Nguyen, N. James, M. Floyd, and V. Pokala, "A Distributed Critical-path Timing Monitor for a 65nm High-Performance Microprocessor," in *IEEE International Solid-State Circuits Conference*, 2007, pp. 398–399.

[53] X. Wang, M. Tehranipoor, and R. Datta, "Path-RO: A Novel On-Chip Critical Path Delay Measurement Under Process Variations," in *IEEE/ACM International Conference on Computer-Aided Design*, no. 1455, Nov. 2008, pp. 640–646.

[54] Q. Liu and S. S. Sapatnekar, "Synthesizing a Representative Critical Path for Post-Silicon Delay Prediction," in *International symposium on Physical design*, New York, New York, USA, 2009, p. 183.

[55] J. Park and J. a. Abraham, "A Fast, Accurate and Simple Critical Path Monitor for Improving Energy-Delay Product in DVS Systems," in *IEEE/ACM International Symposium on Low Power Electronics and Design*, Aug. 2011, pp. 391–396.

[56] S. Wang, J. Chen, and M. Tehranipoor, "Representative Critical Reliability Paths for Low-Cost and Accurate On-Chip Aging Evaluation," in *International Conference on Computer-Aided Design*, 2012, pp. 736–741.

[57] W. Huang, C. Lefurgy, W. Kuk, A. Buyuktosunoglu, M. Floyd, K. Rajamani, M. Allen-Ware, and B. Brock, "Accurate Fine-Grained Processor Power Proxies," in *IEEE/ACM International Symposium on Microarchitecture*, Dec. 2012, pp. 224–234.

[58] A. Bartolini, M. Cacciari, A. Tilli, and L. Benini, "A Distributed and Self-Calibrating Model-Predictive Controller for Energy and Thermal management of High-Performance Multicores," in *Design, Automation & Test in Europe*, Mar. 2011, pp. 1–6.

[59] I. A. M. Elfadel, R. Marculescu, and D. Atienza, "Closed-Loop Control for Power and Thermal Management in Multi-core Processors: Formal Methods and Industrial Practice," in *Design, Automation & Test in Europe Conference & Exhibition (DATE), 2013*, New Jersey, 2013, pp. 1879–1881.

[60] P. Chen, C.-c. Chen, C.-c. Tsai, W.-f. Lu, and A. A. Cmos, "A Time-to-Digital-Converter-Based CMOS Smart Temperature Sensor," *IEEE Journal of Solid-State Circuits*, vol. 40, no. 8, pp. 1642–1648, Aug. 2005.

[61] C.-K. Kim, J.-G. Lee, Y.-H. Jun, C.-G. Lee, and B.-S. Kong, "CMOS temperature sensor with ring oscillator for mobile DRAM self-refresh control," *Microelectronics Journal*, vol. 38, no. 10-11, pp. 1042–1049, Oct. 2007.

[62] P. Ituero, J. L. Ayala, and M. Lopez-Vallejo, "A Nanowatt Smart Temperature Sensor for Dynamic Thermal Management," *IEEE Sensors Journal*, vol. 8, no. 12, pp. 2036–2043, Dec. 2008.

[63] P. Chen, S. Chen, Y. Shen, and Y. Peng, "All-Digital Time-Domain Smart Temperature Sensor After One-Point Calibration," *IEEE Transaction on Circuits and Systems*, vol. 58, no. 5, pp. 913–920, 2011.

[64] S. Ohkawa, M. Aoki, and H. Masuda, "Analysis and Characterization of Device Variations in an LSI Chip Using an Integrated Device Matrix Array," *IEEE Transactions on Semiconductor Manufacturing*, vol. 17, no. 2, pp. 155–165, May 2004.

[65] K. Agarwal, F. Liu, C. Mcdowell, S. Nassif, K. Nowka, M. Palmer, D. Acharyya, and J. Plusquellic, "A Test Structure for Characterizing Local Device Mismatches," in *Symposium on VLSI Circuits and Technology*, vol. 00, no. c, 2006, pp. 67–68.

[66] T. Tsunomura, A. Nishida, F. Yano, A. T. Putra, K. Takeuchi, S. Inaba, S. Kamohara, K. Terada, T. Hiramoto, and T. Mogami, "Analyses of 5σ Vth Fluctuation in 65nm-MOSFETs using Takeuchi Plot," in *Symposium on VLSI Technology*, 2008, pp. 156–157.

[67] T. Sato, H. Ueyama, N. Nakayama, and K. Masu, "Accurate Array-Based Measurement for Subthreshold-Current of MOS Transistors," *IEEE Journal of Solid-State Circuits*, vol. 44, no. 11, pp. 2977–2986, Nov. 2009.

[68] M. Mani, A. Singh, and M. Orshansky, "Joint Design-Time and Post-Silicon Minimization of Parametric Yield Loss using Adjustable Robust Optimization," *2006 IEEE/ACM International Conference on Computer Aided Design*, pp. 19–26, Nov. 2006.

[69] M. Ketchen, M. B. and Bhushan, "Product-representative ‘‘ at speed ’’ test structures for CMOS characterization," *IBM Journal of Research and Development*, vol. 50, no. 4, pp. 451–468, 2006.

[70] C. Zhuo, D. Blaauw, and D. Sylvester, "Post-Fabrication Measurement-Driven Oxide Breakdown Reliability Prediction and Management," in *International Conference on Computer-Aided Design*, New York, New York, USA, 2009, p. 441.

[71] M. Ketchen, M. Bhushan, D. Pearson, and Y. Heights, "High Speed Test Structures for In-line Process Monitoring and Model Calibration," in *IEEE ICMTS*, vol. 18, no. April, 2005, pp. 33–38.

[72] M. Ketchen, M. Bhushan, and R. Bolam, "Ring Oscillator Based Test Structure for NBTI Analysis," in *IEEE International Conference on Microelectronic Test Structures*, 2007, pp. 42–47.

[73] K. Agarwal, J. Hayes, and S. Nassif, "Fast Characterization of Threshold Voltage Fluctuation in MOS Devices," *IEEE Transactions on Semiconductor Manufacturing*, vol. 21, no. 4, pp. 526–533, Nov. 2008.

[74] A. Chang, D. Boning, and Others, "A Test Structure for the Measurement and Characterization of Layout-Induced Transistor Variation," Ph.D. dissertation, Massachusetts Institute of Technology, 2009.

[75] J. Keane, T. Kim, and C. H. Kim, "An On-Chip NBTI Sensor for Measuring pMOS Threshold Voltage Degradation," *IEEE Transaction on Very Large Scale Integration (VLSI) Systems*, vol. 18, no. 6, pp. 947–956, 2010.

[76] H. K. Alidash, A. Calimera, A. Macii, E. Macii, and M. Poncino, "On-Chip NBTI and PBTI Tracking through an All-Digital Aging Monitor Architecture," in *Power And Timing Modeling, Optimization and Simulation (PATMOS)*, 2012, pp. 155–165.

[77] T. Kim, R. Persaud, and C. H. Kim, "Silicon Odometer : An On-Chip Reliability Monitor for Measuring Frequency Degradation of Digital Circuits," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 4, pp. 874–880, 2008.

[78] E. Saneyoshi, K. Nose, and M. Mizuno, "A Precise-Tracking NBTI-Degradation Monitor Independent of NBTI Recovery Effect," in *IEEE International Solid-State Circuits Conference (ISSCC)*, Feb. 2010, pp. 192–193.

[79] S. Mukhopadhyay and K. Kim, "An On-Chip Test Structure and Digital Measurement Method for Statistical Characterization of Local Random Variability in a Process," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 9, pp. 1951–1963, 2008.

[80] S. Mukhopadhyay and K. Kim, "Statistical Characterization and On-chip Measurement Methods for Local Random Variability of a Process Using Sense-Amplifier-Based Test Structure," in *IEEE International Solid-State Circuits Conference*, 2007, pp. 20–22.

[81] B. Ji, D. Pearson, I. Lauer, F. Stellari, D. Frank, L. Chang, and M. Ketchen, "Operational Amplifier Based Test Structure For Transistor Threshold Voltage Variation," in *IEEE International Conference on Microelectronic Test Structures*, 2008, pp. 3–7.

[82] A. Bassi, A. Veggetti, L. Croce, and A. Bogliolo, "Measuring the effects of process variations on circuit performance by means of digitally-controllable ring oscillators," in *IEEE International Conference on Microelectronic Test Structures*, 2003, pp. 214–217.

[83] T. Yamagishi, T. Shiozawa, K. Horisaki, H. Hara, and Y. Unekawa, "A Standard-Cell Based On-Chip NMOS and PMOS Performance Monitor for Process Variability Compensation," *IEICE Transactions on Electronics*, vol. E96.C, no. 6, pp. 894–902, 2013.

[84] S. Wang and M. Tehranipoor, "TSUNAMI : A Light-Weight On-Chip Structure for Measuring Timing Uncertainty Induced by Noise During Functional and Test Operations," in *Great Lakes Symposium on VLSI*, 2012, pp. 183–188.

[85] E. J. Jangl, A. Gattiker, S. Nassif, and J. A. Abrahaml, "An Oscillation-based Test Structure for Timing Information Extraction," in *IEEE VLSI Test Symposium*, 2012, pp. 74–79.

[86] K. Bernstein, D. J. Frank, A. E. Gattiker, W. Haensch, B. L. Ji, S. R. Nassif, E. J. Nowak, D. J. Pearson, and N. J. Hohrer, "High-performance CMOS variability in the 65-nm

regime and beyond," *IBM Journal of Research and Development*, vol. 50, no. 4/5, pp. 433–449, 2006.

[87] I. Ahsan and N. Zamdmer, "RTA-Driven Intra-Die Variations in Stage Delay, and Parametric Sensitivities for 65nm Technology," in *Symposium on VLSI Technology Technical Digest*, vol. 00, no. April 2005, 2006, pp. 2005–2006.

[88] H. Onodera, "Variability: Modeling and Its Impact on Design," *IEICE Transactions on Electronics*, vol. E89-C, no. 3, pp. 701–704, 2006.

[89] M. Pelgrom, A. Duinmaijer, and A. Welbers, "Matching Properties of MOS Transistors," *IEEE Journal of Solid-State Circuits*, vol. 24, no. 5, pp. 1433–1439, Oct. 1989.

[90] S. Nishizawa and H. Onodera, "A Ring Oscillator With Calibration Circuit for On-Chip Measurement of Static IR-drop," *IEEE Transactions on Semiconductor Manufacturing*, vol. 26, no. 3, pp. 306–313, 2013.

[91] K. Bowman, J. Tschanz, S. Lu, P. Aseron, M. Khellah, A. Raychowdhury, B. Geuskens, C. Tokunaga, C. Wilkerson, T. Karnik, and Others, "A 45 nm Resilient Microprocessor Core for Dynamic Variation Tolerance," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 1, pp. 194–208, 2011.

[92] M. Kirton and M. Uren, "Noise in solid-state microstructures: a new perspective on individual defects, interface states, and low-frequency noise," *Advances in Physics*, vol. 38, no. 4, pp. 367–468, 1989.

[93] C. H. Liu, M. T. Lee, C. Lin, J. Chen, K. Schruefer, J. Brighten, N. Rovedo, T. B. Hook, V. Mukesh, S. Hung, W. Clement, T. Chen, and T. H. Ning, "Mechanism and Process Dependence of Negative Bias Temperature Instability (NBTI) for pMOSFETs with Ultrathin Gate Dielectrics," in *International Electron Devices Meeting*, 2001, pp. 39.2.1–39.2.4.

[94] S. Borkar, "Designing reliable systems from unreliable components: the challenges of transistor variability and degradation," *IEEE Micro*, vol. 25, no. 6, pp. 10–16, Nov. 2005.

[95] N. Aghaee, Z. Peng, and P. Eles, "Process-Variation and Temperature Aware SoC test Scheduling Using Particle Swarm Optimization," in *IEEE International Design and Test Workshop (IDT)*, Dec. 2011, pp. 1–6.

[96] S. Mitra, E. Volkerink, E. McCluskey, and S. Eichenberger, "Delay defect screening using process monitor structures," in *IEEE VLSI Test Symposium*, no. Vts, 2004, pp. 43–48.

[97] N. Tega, H. Miki, and F. Pagette, "Increasing Threshold Voltage Variation due to Random Telegraph Noise in FETs as Gate Lengths Scale to 20 nm," in *Symposium on VLSI Technology*, 2009, pp. 50–51.

[98] K. Takeuchi, T. Nagumo, K. Takeda, S. Asayama, S. Yokogawa, K. Imai, and Y. Hayashi, "Direct Observation of RTN-induced SRAM Failure by Accelerated Testing and Its Application to Product Reliability Assessment," in *Symposium on VLSI Technology*, Jun. 2010, pp. 189–190.

[99] H. Kurata, K. Otsuga, A. Kotabe, S. Kajiyama, T. Osabe, Y. Sasago, S. Narumi, K. Tokami, S. Kamohara, and O. Tsuchiya, "Random Telegraph Signal in Flash Memory: Its Impact on Scaling of Multilevel Flash Memory Beyond the 90-nm Node," *IEEE Journal of Solid-State Circuits*, vol. 42, no. 6, pp. 1362–1369, Jun. 2007.

[100] X. Wang, P. R. Rao, A. Mierop, and A. J. Theuwissen, "Random Telegraph Signal in CMOS Image Sensor Pixels," in *2006 International Electron Devices Meeting*, 2006, pp. 1–4.

[101] T. Matsumoto, K. Kobayashi, and H. Onodera, "Impact of Random Telegraph Noise on CMOS Logic Delay Uncertainty under Low Voltage Operation," in *International Electron Devices Meeting*, Dec. 2012, pp. 25.6.1–25.6.4.

[102] C. Visweswariah, "First-Order Incremental Block-Based Statistical Timing Analysis," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 25, no. 10, pp. 2170–2180, 2006.

[103] H. Chang, V. Zolotov, S. Narayan, and C. Visweswariah, "Parameterized Block-Based Statistical Timing Analysis with Non-Gaussian Parameters, Nonlinear Delay Functions," in *Design Automation Conference*, New York, New York, USA, 2005, p. 71.

[104] S. Tam, R. D. Limaye, and U. N. Desai, "Clock Generation and Distribution for the 130-nm," *IEEE Journal of Solid-State Circuits*, vol. 39, no. 4, pp. 636–642, 2004.

[105] B. Taylor and L. T. Pileggi, "Adaptive Post-Silicon Tuning for Analog Circuits: Concept, Analysis and Optimization," in *IEEE/ACM International Conference on Computer-Aided Design*, Nov. 2007, pp. 450–457.

[106] Y. Hashizume, Y. Takashima, and Y. Nakamura, "Post-Silicon Clock-Timing Tuning Based on Statistical Estimation," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E91-A, no. 9, pp. 2322–2327, 2008.

[107] K. Mishra, A. Faraz, a. D. Singh, and A. Chatterjee, "Path Delay Tuning for Performance Gain in the Face of Random Manufacturing Variations," *2011 24th Internatioal Conference on VLSI Design*, pp. 382–388, Jan. 2011.

[108] S. H. Kulkarni, D. M. Sylvester, and D. T. Blaauw, "Design-Time Optimization of Post-Silicon Tuned Circuits Using Adaptive Body Bias," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 27, no. 3, pp. 481–494, Mar. 2008.

[109] N. Kamae and A. Tsuchiya, "An Area Effective Forward/Reverse Body Bias Generator for Within-Die Variability Compensation," in *IEEE Asian Solid State Circuits Conference*, 2011, pp. 217–220.

[110] N. Kamae, A. Tsuchiya, and H. Onodera, "A Body Bias Generator Compatible with Cell-based Design Flow for Within-die Variability Compensation," in *IEEE Asian Solid State Circuits Conference*, 2012, pp. 389–392.

[111] T. Chen and S. Naffziger, "Comparison of Adaptive Body Bias (ABB) and Adaptive Supply Voltage (ASV) for Improving Delay and Leakage Under the Presence of Process Variation," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 11, no. 5, pp. 888–899, Oct. 2003.

[112] B. Rahbaran and A. Steininger, "Is Asynchronous Logic More Robust Than Synchronous Logic?" *IEEE Transactions on Dependable and Secure Computing*, vol. 6, no. 4, pp. 282–294, Oct. 2009.

[113] B. Devlin, M. Ikeda, and K. Asada, "A 65 nm Gate-Level Pipelined Self-Synchronous Robust Operation," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 11, pp. 2500–2513, 2011.

[114] B. Devlin, M. Ikeda, and K. Asada, "Gate-level Process Variation Offset Technique by using Dual Voltage Supplies to Achieve Near-threshold Energy Efficient Operation," in *IEEE COOL Chips XV*, Apr. 2012, pp. 1–3.

[115] K. Agarwal, S. Nassif, F. Liu, J. Hayes, and K. Nowka, "Rapid Characterization of Threshold Voltage Fluctuation in MOS Devices," in *IEEE International Conference on Microelectronic Test Structures*, Mar. 2007, pp. 74–77.

[116] K. Kuhn, "Reducing Variation in Advanced Logic Technologies: Approaches to Process and Design for Manufacturability of Nanoscale CMOS," in *IEEE Electron Devices Meeting*, 2007, pp. 471–474.

[117] A. Ghosh, R. M. Rao, J. Kim, C.-T. Chuang, and R. B. Brown, "On-Chip Process Variation Detection Using Slew-Rate Monitoring Circuit," in *21st International Conference on VLSI Design*. Ieee, 2008, pp. 143–149.

[118] T. Iizuka, T. Nakura, and K. Asada, "Buffer-Ring-Based All-Digital On-Chip Monitor for PMOS and NMOS Process Variability and AgingEffect," in *13th IEEE Symposium on Design and Diagnostics of Electronic Circuits and Systems*, 2010, pp. 167–172.

[119] H. Notani, M. Fujii, H. Suzuki, H. Makino, and H. Shinohara, "On-chip Digital Idn and Idp Measurement by 65 nm CMOS Speed Monitor Circuit," in *IEEE Asian Solid-State Circuits Conference*, 2008, pp. 405–408.

[120] M. Bhushan, A. Gattiker, M. B. Ketchen, and K. K. Das, "Ring Oscillators for CMOS Process Tuning and Variability Control," *IEEE Transactions on Semiconductor Manufacturing*, vol. 19, no. 1, pp. 10–18, 2006.

[121] M. Bhushan, M. B. Ketchen, S. Polonsky, and A. Gattiker, "Ring Oscillator Based Technique for Measuring Variability Statistics Manjul," in *Microelectronic Test Structures, 2006 IEEE International Conference on*, Mar. 2006, pp. 87–92.

[122] T. Takahashi, T. Uezono, M. Shintani, K. Masu, and T. Sato, "On-die parameter extraction from path-delay measurements," in *IEEE Asian Solid-State Circuits Conference*, no. 4, Nov. 2009, pp. 101–104.

[123] I. Synopsys, *HSPICE User's Manual: Simulation and Analysis*, 2010, vol. 1.

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

[125] R. Spence and R. S. Soin, *Tolerance Design of Electronic Circuits*, 1988.

[126] H. Onodera and H. Terada, "Characterization of WID Delay Variability Using RO-array Test Structures," in *IEEE 8th International Conference on ASIC*, Oct. 2009, pp. 658–661.

[127] M. Bhushan and M. B. Ketchen, "Generation, Elimination and Utilization of Harmonics in Ring Oscillators," in *IEEE Intl. Conference on Microelectronic Test Structures*, vol. 1, 2010, pp. 108–113.

[128] M. Yamaoka and H. Onodera, "A Detailed Vth-Variation Analysis for Sub-100-nm Embedded SRAM Design," in *IEEE International SOC Conference*, Sep. 2006, pp. 315–318.

[129] Q. Liu and S. S. Sapatnekar, "Confidence Scalable Post-Silicon Statistical Delay Prediction under Process Variations," in *44th ACM/IEEE Design Automation Conference*, Jun. 2007, pp. 497–502.

[130] A. Zjajo, M. Barragan Asian, and J. de Gyvez, "BIST Method for Die-Level Process Parameter Variation Monitoring in Analog/Mixed-Signal Integrated Circuits," in *Design, Automation & Test in Europe*, no. Figure 2, 2007, pp. 1–6.

[131] P. Maxwell, "Adaptive Testing: Dealing with Process Variability," *IEEE Design & Test of Computers*, vol. 28, no. 6, pp. 41–49, 2011.

[132] M. Shintani, T. Uezono, T. Takahashi, H. Ueyama, T. Sato, K. Hatayama, T. Aikyo, and K. Masu, "An Adaptive Test for Parametric Faults Based on Statistical Timing Information," in *IEEE Asian Test Symposium*, Nov. 2009, pp. 151–156.

[133] T. Tsunomura, A. Nishida, and T. Hiramoto, "Investigation of Threshold Voltage Variability at High Temperature Using Takeuchi Plot," *Japanese Journal of Applied Physics*, vol. 49, no. 5, p. 054101, May 2010.

[134] K. Ito, T. Matsumoto, S. Nishizawa, H. Sunagawa, K. Kobayashi, and H. Onodera, "The Impact of RTN on Performance Fluctuation in CMOS Logic Circuits," in *International Reliability Physics Symposium*, Apr. 2011, pp. 710–713.

[135] K. Ito and T. Matsumoto, "Modeling of Random Telegraph Noise under circuit operation —Simulation and measurement of RTN-induced delay fluctuation," in *12th International Symposium on Quality Electronics Design*, 2011, pp. 22–27.

[136] T. Matsumoto, K. Kobayashi, and H. Onodera, "Impact of Body-Biasing Technique on Random Telegraph Noise Induced Delay Fluctuation," *Japanese Journal of Applied Physics*, vol. 52, p. 04CE05, Mar. 2013.

[137] J. Tschanz, S. Narendra, and A. Keshavarzi, "Adaptive Circuit Techniques to Minimize Variation Impacts on Microprocessor Performance and Power," *2005 IEEE International Symposium on Circuits and Systems*, pp. 9–12, 2005.

[138] M. Gupta, J. Rivers, and P. Bose, "Tribeca: Design for PVT Variations with Local Recovery and Fine-grained Adaptation," in *IEEE/ACM International Symposium on Microarchitecture*, 2009, pp. 435–446.

[139] G. Ono and M. Miyazaki, "Threshold-Voltage Balance for Minimum Supply Operation," *IEEE Journal of Solid-State Circuits*, vol. 38, no. 5, pp. 830–833, May 2003.

[140] B. Zhai, D. Blaauw, D. Sylvester, and K. Flautner, "The Limit of Dynamic Voltage Scaling and Insomniac Dynamic Voltage Scaling," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 13, no. 11, pp. 1239–1252, 2005.

[141] M. Miyazaki, G. Ono, and K. Ishibashi, "A 1.2-GIPS / W Microprocessor Using Speed-Adaptive Threshold-Voltage CMOS with Forward Bias," *IEEE J. of Solid-State Circuits*, vol. 37, no. 2, pp. 210–217, 2002.

[142] T. Kuroda, T. Fujita, S. Mita, T. Nagamatsu, S. Yoshioka, K. Suzuki, F. Sano, M. Norishima, M. Murota, M. Kako, M. Kinugawa, M. Kakumu, and T. Sakurai, "A 0.9-V, ISO-MHz, IO-mW, 4 mm2, 2-D Discrete Cosine Transform Core Processor with Variable Threshold-Voltage (VT) Scheme," *IEEE Journal of Solid-State Circuits*, vol. 31, no. 11, pp. 1770–1779, 1996.

[143] P. Gupta, Y. Agarwal, L. Dolecek, N. Dutt, R. K. Gupta, R. Kumar, S. Mitra, A. Nicolau, T. S. Rosing, M. B. Srivastava, S. Swanson, and D. Sylvester, "Underdesigned and Opportunistic Computing in Presence of Hardware Variability," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 32, no. 1, pp. 8–23, Jan. 2013.

Publication List

Patent

1. A.K.M. Mahfuzul Islam, and Hidetoshi Onodera, "Delay Circuits for Delay Monitor Circuits and Delay Monitor Circuits for signal-propagation-time measurements in integrated circuits," Patent Application Number 2013-169956.

Journal

1. A.K.M. Mahfuzul Islam, and Hidetoshi Onodera, "Area-efficient Reconfigurable Ring Oscillator for Characterization of Static and Dynamic Variations," *Japanese Journal of Applied Physics*, vol. 53, no. 4S, pp. 04EE08, 2014.

2. A.K.M. Mahfuzul Islam, and Hidetoshi Onodera, "On-Chip Detection of Process Shift and Process Spread for Post-Silicon Diagnosis and Model-Hardware Correlation," *IEICE Transactions on Information and Systems*, vol. E96-D, no. 9, pp. 1971–1979, 2013.

3. Shuuichi Fujimoto, A.K.M. Mahfuzul Islam, Takashi Matsumoto, and Hidetoshi Onodera, "Inhomogeneous Ring Oscillator for Within-Die Variability and RTN Characterization," *IEEE Transactions on Semiconductor Manufacturing*, vol. 26, no. 3, pp. 296–305, 2013.

4. A.K.M. Mahfuzul Islam, Akira Tsuchiya, Kazutoshi Kobayashi, and Hidetoshi Onodera, "Variation-sensitive Monitor Circuits for Estimation of Global Process Parameter, " *IEEE Transactions on Semiconductor Manufacturing*, vol. 25, no. 4, pp. 571–580, 2012.

International Conference

1. A.K.M. Mahfuzul Islam, and Hidetoshi Onodera, "In-Situ Variability Characterization of Individual Transistors using Topology-Reconfigurable Ring Oscillators ," *in IEEE International Conference on Microelectronic Test Structures*, Mar. 2014, pp. 121–126.

2. A.K.M. Mahfuzul Islam, Tohru Ishihara, and Hidetoshi Onodera, "Reconfigurable Delay Cell for Area-efficient Implementation of On-chip MOSFET Monitor Schemes," *in IEEE*

Asian Solid State Circuits Conference, Nov. 2013, pp. 125–128.

3. A.K.M. Mahfuzul Islam, Norihiro Kamae, Tohru Ishihara, and Hidetoshi Onodera, "Energy-efficient Dynamic Voltage and Frequency Scaling by P/N-performance Self-adjustment using Adaptive Body Bias," *Proceedings of SASIMI*, Oct. 2013.

4. A. K. M. Mahfuzul Islam, and Hidetoshi Onodera, "Area-efficient Reconfigurable Ring Oscillator for Characterization of Static and Dynamic Variations," *in International Conference on Solid State Devices and Materials*, Sep. 2013, pp. 132–133.

5. A.K.M. Mahfuzul Islam, and Hidetoshi Onodera, "On-Chip Detection of Process Shift and Process Spread for Silicon Debugging and Model-Hardware Correlation," *in IEEE Asian Test Symposium*, 2012, pp. 350–354.

6. A.K.M Mahfuzul Islam, Norihiro Kamae, Tohru Ishihara, and Hidetoshi Onodera, "A Built-in Self-adjustment Scheme with Adaptive Body Bias using P/N-sensitive Digital Monitor Circuits," *in IEEE Asian Solid State Circuits Conference*, Nov. 2012, pp. 101–104.

7. Shuuichi Fujimoto, A.K.M. Mahfuzul Islam, Takashi Matsumoto, and Hidetoshi Onodera, "Inhomogeneous Ring Oscillator for WID Variability and RTN Characterization," *in IEEE International Conference on Microelectronic Test Structures*, Mar. 2012, pp. 25–30.

8. A.K.M. Mahfuzul Islam, Akira Tsuchiya, Kazutoshi Kobayashi, and Hidetoshi Onodera, "Variation-sensitive Monitor Circuits for Estimation of Die-to-Die Process Variation," *in IEEE International Conference on Microelectronic Test Structures*, Apr. 2011, pp. 153–157.

9. Shuuichi Fujimoto, A.K.M. Mahfuzul Islam, Shinichi Nishizawa, and Hidetoshi Onodera, "Component Analysis of WID Variation," *in IEEE International Workshop on Design for Manufacturability & Yield*, May 2010, pp.

10. A.K.M. Mahfuzul Islam, Akira Tsuchiya, Kazutoshi Kobayashi, and Hidetoshi Onodera, "Process-sensitive Monitor Circuits for Estimation of Die-to-Die Process Variability," *in ACM International Workshop on Timing Issues in the Specification and Synthesis of Digital Systems*, Mar. 2010, pp. 83–88.

On-site Demonstration

1. A.K.M. Mahfuzul Islam, Tohru Ishihara, and Hidetoshi Onodera, "Reconfigurable Delay Cell for Area-efficient Implementation of On-chip MOSFET Monitor Schemes," *in Student Design Contest of IEEE Asian Solid State Circuits Conference*, Nov. 2013.

2. Norihiro Kamae, A.K.M. Mahfuzul Islam, Tohru Ishihara, and Hidetoshi Onodera, "Post-silicon P/N-performance Compensation Scheme Compatible with Cell-based Design," *in University Booth of IEEE Design, Automation and Test in Europope*, Mar. 2013.

3. A.K.M. Mahfuzul Islam, Norihiro Kamae, Tohru Ishihara, and Hidetoshi Onodera, "A Built-in Self-adjustment Scheme with Adaptive Body Bias using P/N-sensitive Digital Monitor Circuits," *in Student Design Contest of IEEE Asian Solid State circuits Conference*, Nov. 2012.

Awards

1. Student Design Contest Award, "Reconfigurable Delay Cell for Area-efficient Implementation of On-chip MOSFET Monitor Schemes" *in IEEE Asian Solid State Circuits Conference*, Nov. 2013.

2. 情報処理学会 システム LSI 設計技術研究会 優秀論文賞, "完全ディジタル型の P/N ばらつきの自律補償回路", Aug. 2013.

3. IEEE SSCS Japan Chapter VDEC Design Award, "P/N ばらつきモニタを用いたチップ間およびチップ内ばらつきの自律補償回路", Aug. 2012.

4. IEEE Kansai Section Award, "Variation-sensitive Monitor Circuits for Estimation of Die-to-Die Process Variation" *in IEEE Int. Conf. on Microelectronic Test Structures (ICMTS)*, Dec. 2011.