Design and Characterization Techniques for Reliable and Secure Integrated Circuits

A DISSERTATION

SUBMITTED TO THE FACULTY OF

UNIVERSITY OF MINNESOTA

BY

Qianying Tang

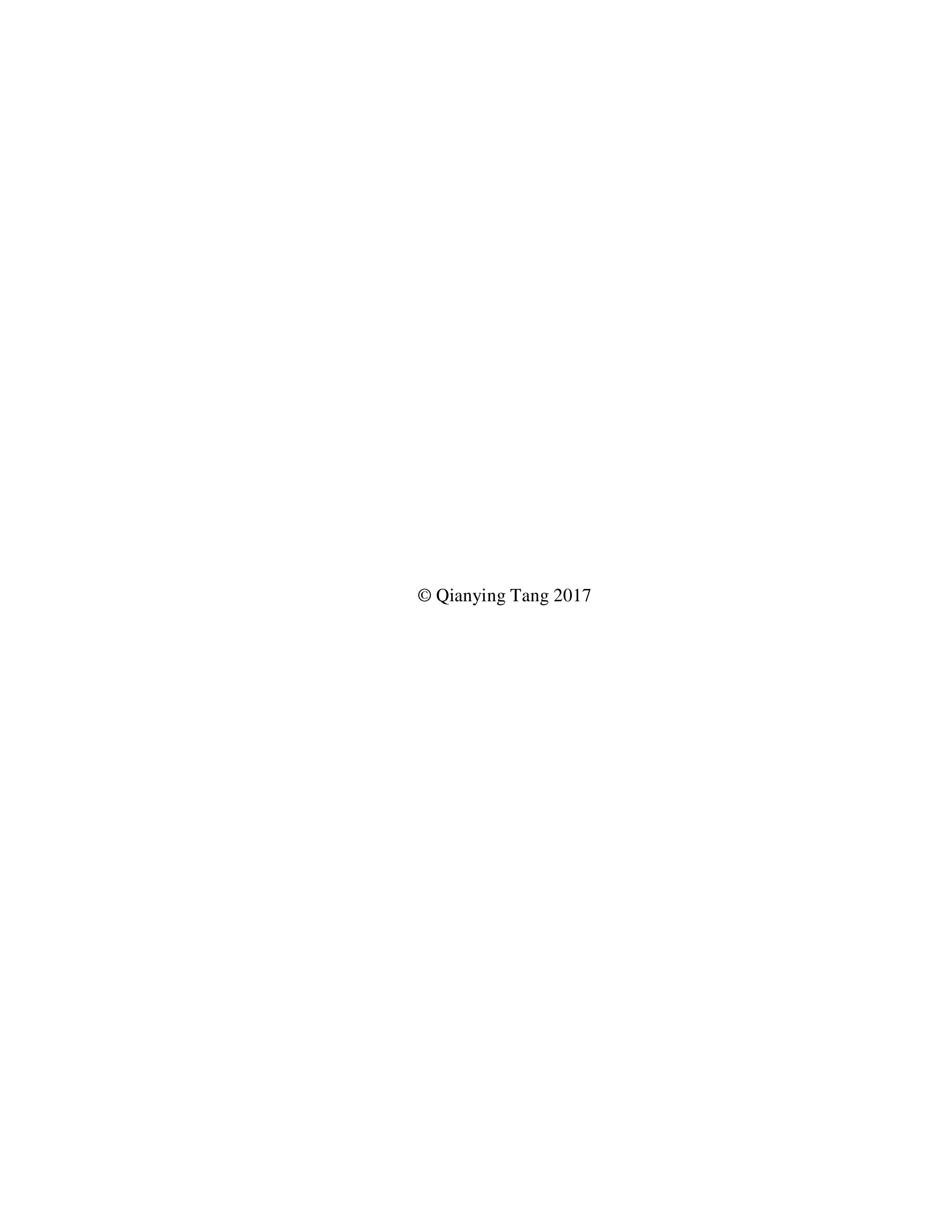
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Advisor: Chris H. Kim

Februray 2017



© Qianying Tang 2017

Acknowledgements

Firstly, I would like to express my sincere gratitude to my advisor, Prof. Chris H. Kim, for his patience, motivation, and constant support throughout my PhD study. For the past four and half years that he has been my advisor, I have learnt and benefitted a lot from his knowledge and expertise in circuit design. His enthusiasm and creativities in research has inspired me to explore all the possibilities and never give up trying new things. I could not have imagined having a better advisor for my Ph.D. study.

Next, I would like to thank Prof. Keshab K. Parhi for his guidance and constructive advising throughout my research. I also would like to than Prof. Kia Bazargan and Prof. Hubert H. Lim, for their insightful comments and guidance in writing this thesis. I appreciate their contribution of time to serve on my committee.

I am also thankful to China Scholarship Council and SRC for funding my study and research.

My sincere thanks also goes to Dr. Sungjae Lee for providing me an opportunity to join his team as intern in IBM. I also would like to Dr. Hongmei Lee, who mentored and helped me fulfill my project during the internship.

I'm deeply grateful to my labmates, for the encouraging discussions and plenty of joys we have had in the lab. I especially thank Xiaofei Wang, Bongjin Kim, Won Ho Choi, Yingjie Lao (Prof. Parhi's group), Chen Zhou and Saurabh Kumar for their devotional collaboration and valuable suggestions. My thanks also goes to the group senior alumni,

Ki Chul Chun, Pulkit Jain, John Keane, for their valuable research legacy that I have learnt a lot from. Also I thank Seung-hwan Song, Xiaofei Wang, Bongjin Kim, Ayan Paul, Won Ho Choi, Hoonki Kim, Jongyeon Kim, Weichao Xu, Somnath Kundu, Paul Mazanec, Muqing Liu, Luke Everson, Po-Wei Chiu, Gyusung Park, Nakul Pande, Minsu Kim and Jeehwan Song.

Last but not the least, I would like to thank my parents Jun Tang and Fan Yang for supporting my pursuits and everything else in my life in general. Without them, I may never have gotten to where I am today.

This dissertation is dedicated to the memories of my grandfather Zesheng Tang who led me to the wonderland of science and research.

Abstract

For the past decades of years, device feature size has continued to shrink for achieving better performance at faster speed, lower power and higher circuit density. However, going to a smaller feature sizes also brings in reliability issues such as greater process variations and more aggressive performance degradation. To address these issues, circuits are designed with certain guard-band to avoid probable failures. In order to determine an appropriate guard-band, it is imperative to develop accurate and efficient methods for characterizing and collecting these reliability metrics. This dissertation considers two important circuit reliability issues: Random Telegraph Noise (RTN) and Radiation induced Soft Error. For characterizing the realistic impact of RTN on logic circuit, we proposed two on chip monitors using a 65nm and a 32nm process respectively based on a Beat Frequency Detection (BFD) technique. The impact of RTN on logic and SRAM performance was analyzed based on the measured data. In the chapter 3, a compact 2 Transistor (2T) radiation sensor with tunable measurement sensitivity implemented in a 65nm LP bulk process is presented. The 2T sensor array exhibits a 117X higher sensitivity as compared to a 6T SRAM cell under an alpha particle radiation test.

Meanwhile, with the electronic devices become increasingly ubiquitous and interconnected, demand for secure system design has also increased. In particular, hardware-oriented security has emerged as a new solution to provide another dimension of security in additional to the conventional software-oriented security. Many of the

hardware security primitives seek to leverage the process variation, in contrast to suppress it for the sake of performance, to against post-silicon attacks. For example, hardware security building blocks such as true random number generators (TRNGs) and physical unclonable functions (PUFs) employ the CMOS devices inherent variation to extract entropy: the former one takes advantage of the time-variant random noise and latter one is based on the manufacturing induced random variation. In this dissertation, one TRNG and two lightweight PUFs are presented. The TRNG measures the frequency difference between two free-running ring oscillators to extract random frequency jitter. Benefitted from the differential structure, the proposed circuit fabricated in 65nm TRNG test chips passed all 15 NIST tests without the use of any feedback or tracking scheme in a supply voltage range from 0.8V to 1.2V. The final part of the dissertation presents two lightweight PUFs that are based on existing Dynamic Random-Access Memory (DRAM) and Successive Approximation Register (SAR) Analog-to-Digital Converter (ADC) blocks respectively.

Table of Contents

Abstract iv

Table of Contents vi

List of Figures..ix

Chapter 1. Backgrounds 1

1.1. Circuit Reliability 1

1.1.1. Random Telegraph Noise..1

1.1.2. Soft Error..4

1.2. Hardware Security..5

1.2.1. True Random Number Generator (TRNG)..8

1.2.2. Physical Unclonable Function (PUF)..9

Chapter 2. On chip RTN Monitors 13

2.1. Introduction..13

2.2. Beat Frequency Technique for RTN Monitoring 15

2.3. Characterizing RTN with a Single Array Structure on a 65nm Process......17

2.3.1. RTN Monitor Design 17

2.3.2. RTN Induced Frequency Shift Measurements on 65nm Test Chip 19

2.4. Characterizing RTN with a Dual Array Structure on a 32nm HKMG Process 25

2.4.1. Dual Ring Oscillator Array Technique 27

2.4.2. RTN Induced Frequency Shift Measurements on 32nm Test Chip 30

2.4.3. RTN Impact on Logic Timing 36

2.4.4. RTN Impact on SRAM Stability and Timing 41

2.5. Conclusion 44

Chapter 3. Compact High-Sensitivity Radiation Sensor Array 46

3.1. Introduction..46

3.2. 2T Sensor Array 47

3.3. Alpha Particle Experiment 50

3.4. Single Event Upset Simulation 57

3.5. Conclusion 60

Chapter 4. True Random Number Generator 61

4.1. Introduction..61

4.2. Beat Frequency Detector based TRNG 63

4.2.1. Test Chip Implementation...65

4.2.2. Measurement Data..66

4.3. Simulation and Modeling 74

4.4. Multi-phase TRNG for Enhancing the TRNG Generation Efficiency 75

4.4.1. Circuit Implementation 75

4.4.2. Measurement Data..77

4.5. Conclusion 79

Chapter 5. Physical Unclonable Function 80

5.1. Introduction 80

5.2. Conventional PUFs 81

5.3. DRAM based PUF for Chip Authentication 83

5.3.1. Proposed DRAM PUF Design 86

5.3.2. Improving DRAM PUF Reliability 90

5.3.3. Test Chip Measurement 93

5.4. Charge-redistribution based PUF for Chip Authentication using a SAR ADC Circuit 99

5.4.1. Background 99

5.4.2. Charge-redistribution PUF Design 104

5.4.3. Test Chip Measurement Result 108

5.5. Conclusion 114

Chapter 6. Conclusion 116

References 119

List of Figures

Figure 1.1: Random trapping and de-trapping of carriers causes fluctuation in Vt, resembling a random telegraph signal. 3

Figure 1.2: Distribution of emission times at 95K and Vgs = 1.15V, showing that the time is Poisson distributed [1]. 3

Figure 1.3: Frequency-domain representation of the RTN signal [2]. 4

Figure 1.4: Charge generation and collection phases in a reverse-biased junction and the resultant current pulse caused by the passage of a high-energy ion [19]. 5

Figure 1.5: A typical authentication protocol involves utilization of TRNG and PUF circuit. 7

Figure 1.6: (a) Conventional authentication scheme stores keys in NVM. (b) Using strong PUF for direct authentication [53]. 10

Figure 1.7: Intra-chip and inter-chip HD under (a) secure condition and (b) unsecure condition [40]. 12

Figure 2.1: Comparison of different RTN characterization techniques. 14

Figure 2.2: Beat frequency detection circuit adopted in this work for measuring RTN induced delay shifts at sub-0.5V supply voltages with high resolution. The output count N represents the number of f_B clock cycles that can fit within a single beat frequency (i.e. f_A-f_B) clock period. 16

Figure 2.3: ROSC array test chip for RTN measurements comprising an on-chip beat frequency detection 17

Figure 2.4: Comparison of three ROSC based RTN measurement techniques. 18

Figure 2.5: Single trap RTN waveforms measured from different ROSCs. 19

Figure 2.6: (b) Single and multi-trap RTN waveform from two different ROSCs. (b) Time Lag Plot (TLP) of the two traces. 20

Figure 2.7: RTN traces from same ROSC at different supply voltages. 21

Figure 2.8: Capture and emission time distributions and exponential fit results for supply voltage from 0.8V to 1.0V. 22

Figure 2.9: Time constant versus supply voltage. 23

Figure 2.10: Power Spectrum Density (PSD) of the frequency shift data. 23

Figure 2.11: Histogram of the number of traps per ROSC. 24

Figure 2.12: Frequency to V_{th} mapping. 24

Figure 2.13: RTN induced V_{th} shift with different supply voltage. 25

Figure 2.14: Limitation of prior art. Due to the wide frequency spread, not all ROSCs under test can achieve high measurement resolution at sub-0.5V supply voltages. 27

Figure 2.15: Measurement resolution comparison when pairing a 64 ROSCs with 3 reference ROSCs (left figure) and 64 reference ROSCs (right figure). A more precise waveform can be reconstructed using 64 reference ROSCs which is critical for collecting high quality RTN statistics at low supply voltages such as 0.5V. 28

Figure 2.16: Block diagram of the proposed dual ROSC array based RTN characterization circuit. By pairing ROSCs from two arrays, the beat frequency detection circuit can achieve a frequency measurement resolution less than 0.01%. The number of inverter stages can be configured from 9 to 15 using scan bits. 29

Figure 2.17: (a) RTN induced frequency shift traces measured at different voltages. (b) Magnitude of frequency shift of 6 RTN traps measured at different voltages. 32

Figure 2.18: (a) RTN induced frequency shift due to the same trap measured at different temperatures. (b) Capture and emission time constants both decrease at higher temperatures. 32

Figure 2.19: RTN induced frequency shift versus the number of ROSC stages. The frequency shift caused by the same RTN trap is reduced as the number of stages increases. 33

Figure 2.20: RTN trap location map measured at different supply voltages. Each cell represents a single ROSC. 34

Figure 2.21: RTN trap location map measured after 0, 2, 6 and 14 hours of 1.8V stress. 35

Figure 2.22: RTN occurrences measured from 6 different chips. 35

Figure 2.23: Logic timing errors for RTN traps located in clock tree. 37

Figure 2.24: Logic timing errors for RTN traps located in combinational logic. 37

Figure 2.25: Logic timing errors for RTN traps located in flip-flop.................................37

Figure 2.26: (a) RTN trap location on DFF signal path for worst case setup time (hold time is opposite location). No traps assumed on clock path. (b) RTN impact on D-flip-flop setup and hold times. 38

Figure 2.27: RTN impact on logic path delay assuming a clock period of 20 FO4 inverter delays and one RTN trap in each block (i.e. logic path, clock tree, input DFF, and output DFF). 39

Figure 2.28: Probability of setup time violation versus timing guard band. 40

Figure 2.29: RTN impact on SRAM (a) read SNM and (b) write SNM. 42

Figure 2.30: Monte Carlo simulations of SRAM (a) read SNM and (b) write SNM, with and without RTN. 43

Figure 2.31: RTN impact on SRAM read timing..43

Figure 2.32: RTN impact on sense amplifier resolving time. 44

Figure 2.33: RTN impact on SRAM read path delay. 44

Figure 3.1: Particle strike induced soft errors are rare in (a) logic gates and (b) SRAM cells because of the strong restore current. (c) The proposed 2T sensor can detect SER strikes with a higher sensitivity by removing the restore current and minimizing the node capacitance...49

Figure 3.2: Proposed 2T sensor array for detecting SEU with high sensitivity. Voltage stored inside the cell (V_{cell}) varies with time and leakage current. 50

Figure 3.3: Overall test sequence for the 2T sensor array. The array pattern is compared with the initial checkerboard pattern to identify particle strike induced SEUs.................51

Figure 3.4: 65nm test chip including a 2T sensor array and SRAM cells. 52

Figure 3.5: Ion beam facility with particle accelerator used for radiation testing (Source: University of Minnesota Characterization Facility)...53

Figure 3.6: (a) Upset probability bit map and (b) the number of upsets per 1Kbit array. .53

Figure 3.7: Measured cross-section increases with lower supply voltage. 55

Figure 3.8: (a) Measured cross-section increases with longer t_{READ} due to the reduced data read margin as shown in (b). 56

Figure 3.9： Measured SER is proportional to alpha particle flux. 56

Figure 3.10: Measured cross-section of 2T cell and SRAM at different supply voltages. 57

Figure 3.11: (a) Simulated current pulse using double exponential model and (b) 2T sensor transient response with Q_{total} equal to 0.25fC and 0.15fC respectively. 58

Figure 3.12: Q_{crit} decrease with a longer t_{READ}. 59

Figure 3.13: Simulated Qcrit for inverter, 6T SRAM, and 2T sensor. 59

Figure 4.1: A conventional meta-stability based TRNG [38] 63

Figure 4.2: (a) ROSC based TRNG employed in IBM POWER7+. (b) Probability of two consecutive bits being different as a function of sampling period (=wait time) for IBM's ring oscillator TRNG. Jitter accumulation time must be >200 ROSC cycles for the sampled bit to be considered random [40]. 63

Figure 4.3: Basic principle of the proposed beat frequency based TRNG circuit. 65

Figure 4.4: TRNG circuit with trimming caps and power saving mode. 66

Figure 4.5: Measured beat frequency count for different trim capacitor settings. 68

Figure 4.6: Percentage of '1's and '0's for each bit of the beat frequency count output. The lower significant bits (e.g. bits 1, 2, 3) have better randomness compared to the higher bits. 68

Figure 4.7: NIST test verifies the randomness of 1st ~3rd LSBs. 69

Figure 4.8: Concatenating LSBs to generate the final TRNG output bit stream. 4th LSB can be used after von Neumann correction. 70

Figure 4.9: Concatenated 1st~4th LSBs passes all NIST test after applying Von Neumann correction on the 4th LSB. 71

Figure 4.10: One-time calibration of average count during start up. 72

Figure 4.11: Stability under continuous operation. 73

Figure 4.12: Measured count under different voltages. 73

Figure 4.13: (Upper) Individual ROSC frequency distributions estimated using statistical model and measured data. (Lower) Measured count distribution shows good agreement with simulated data. 74

Figure 4.14: Multi-phase TRNG implementation (3 phase example). 76

Figure 4.15: The number of LSBs with good randomness increases under the same sampling time as compared to the single-phase version...77

Figure 4.16: Measured count output from single-phase and multi-phase TRNGs. 78

Figure 4.17: The number of random bits per output that passes all NIST test as well as the TRNG generation efficiency improves using the proposed multi-phase structure. 78

Figure 4.18: Multi-phase TRNG utilizing fewer ROSC stages shows improved bit rate and efficiency. 78

Figure 4.19: Single-phase and multi-phase TRNG chips in 65nm. 79

Figure 5.1: Schematic of an arbiter PUF. 82

Figure 5.2: SRAM based PUF. 83

Figure 5.3: Qualitative comparison between SRAM PUF and DRAM PUF. 85

Figure 5.4: (a) 2T DRAM cell schematic and leakage components in hold mode. (b) A DRAM cell generates a different response depending on the write data and retention time. 86

Figure 5.5: The proposed authentication scheme consists of four steps: (1) write random 128 bit challenge to DRAM upper array, (2) allow 10% of bits to flip due to retention failure, (3) transfer data to lower array according to random mapping info from server, and finally (4) repeat step (2). The inherent DRAM retention failure rate is utilized for generating a unique and secure response. For the chip demonstration, we chose a 128-bit random input pattern, a 128 x 10 bit random address mapping info (=128+128x10=1,408 total challenge bits) and a 128-bit response. 88

Figure 5.6: Overall enrollment and authentication flow of the proposed DRAM PUF. New techniques proposed in this work are highlighted in red. 90

Figure 5.7: DRAM PUF soft response distribution for data '1' and data '0'. Soft response is defined as the average response value over 500 trials. For example, if the output for a

particular memory cell is 1 for 90% of the time and 0 for 10% of the time, the soft response for this memory cell is 0.9. 91

Figure 5.8: Repetitive write back scheme for improving DRAM PUF stability. (a) Waveforms of DRAM cell storage voltage with repetitive write back. (b) Marginally stable bits can be stabilized to the opposite value with repetitive write-back...................92

Figure 5.9: Percentage of unstable cells decreases with more write-backs. Cells that remain unstable after 10 write-backs will be flagged and masked by the server during chip authentication. 93

Figure 5.10: 65nm DRAM PUF chip micrograph and summary table.............................94

Figure 5.11: Measured inter-chip (15 chips and 10,000 different challenges) and intra-chip Hamming Distance with and without bit masking...95

Figure 5.12: DRAM retention failure map measured under different supply voltages and temperatures. The failure probability can be kept within the desired range of 9.5%<P<11% before each authentication test using the calibration scheme described in Figure 5.11..97

Figure 5.13: A pre-authentication calibration scheme to mitigate the V, T drifts induced intra-chip variation...97

Figure 5.14: Distributions of Hamming distance measured at different supply voltages before and after calibration...98

Figure 5.15: Distributions of Hamming distance measured at different temperatures before and after calibration...99

Figure 5.16: Comparison between conventional arbiter, memory and the proposed charge sharing PUFs. 101

Figure 5.17: 3-bit SAR ADC architecture. 101

Figure 5.18: (a) SAR algorithm; (b) DAC switching procedures [74]. 103

Figure 5.19: Transfer curve for a 3-bit SAR ADC with and without capacitor mismatch. 104

Figure 5.20: Capacitor array layout for a 10-bit SAR ADC [75]....................................105

Figure 5.21: (a) Schematic and (b) timing diagram of the proposed charge redistribution PUF. 107

Figure 5.22: Measured soft response distribution for the charge redistribution PUF. 109

Figure 5.23: (a) Intra-chip HD and (b) percentage of discard CRPs for different enrollment threshold. 110

Figure 5.24: Percentage of discard CRPs for different # of enabled unit capacitors. 111

Figure 5.25: (a) Measured mean, standard deviation and maximum value of intra-chip Hamming Distance with a V_{ref} from 0.8~1.2V. A better intra-chip HD is obtained when the enrollment is performed at lower supply voltage. (b) More CRPs are discard at a lower V_{ref} during chip enrollment. 112

Figure 5.26: Measured inter-chip and intra-chip Hamming Distance distributions and evaluation conditions. 113

Figure 5.27: 65nm charge-redistribution based PUF chip micrograph. 114

Chapter 1. Backgrounds

1.1. Circuit Reliability

1.1.1. Random Telegraph Noise

Parametric shifts caused by temporal random trapping and de-trapping of carriers in the channel, also known as Random Telegraph Noise (RTN), have become a growing concern in extremely scaled CMOS. RTN coupled with Random Dopant Fluctuation (RDF) is predicted to have a detrimental effect on SRAM cell stability beyond 15nm. These traps are believed either be defects created during the fabrication process or generated by voltage stress during normal operation. This situation has spurred a number of studies focusing on the characterization and mitigation of RTN effects. Conventional defect models assume that the defect can exist in two states, one is charged and the other is discharged. Defects capture and emit carriers in a random pattern as shown in Figure 1.1(a). As a consequence, the threshold voltage would switch between two discrete voltage levels, consistent with a two-state Markov process, as shown in Figure 1.1(b).

A typical single trap RTN impact on CMOS devices can be characterized by three parameters: the capture time (t_e), the emission time (t_c) and the amplitude ΔV_{th}. The capture and emission times are randomly distributed while the amplitude is usually fixed. By collecting the RTN data for a sufficiently long period of time, it can be observed that

the probability of t_e and t_c appear to be Poisson distributed as shown in Figure 1.2. The distribution is described as follow [1]:

$$\Pr(t) = \frac{1}{\tau}\exp\left(-\frac{t}{\tau}\right)$$

where τ is the RTN time constant defining the average time a trap site stays in the captured state or in the emission state. The capture and emission time constants (τ_c and τ_e) then can be extracted separately from the collected data. It is reported that RTN time constants range from microseconds to seconds which depends on both the operation conditions, i.e. supply voltage, temperature, and defect locations. For circuits operated at a frequency higher than the RTN time constant (e.g. >1MHz), a capturing event induced threshold voltage increase can be treated as a constant parametric shift which may induce errors that are un-recoverable within one clock period. By applying a Fourier transform of the stationary two-level signal, the corresponding power spectrum is a Lorentzian as shown in Figure 1.3 [2]. The corner frequency of such noise spectrum is determined only by τ_c and τ_e defining as:

$$f_{0RTN} = \frac{1}{2\pi}\left(\frac{1}{\tau_c} + \frac{1}{\tau_e}\right)$$

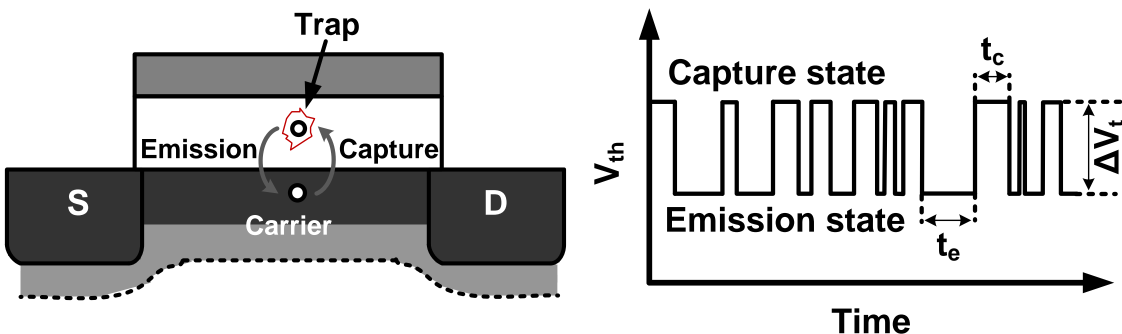


Figure 1.1: Random trapping and de-trapping of carriers causes fluctuation in Vt, resembling a random telegraph signal.

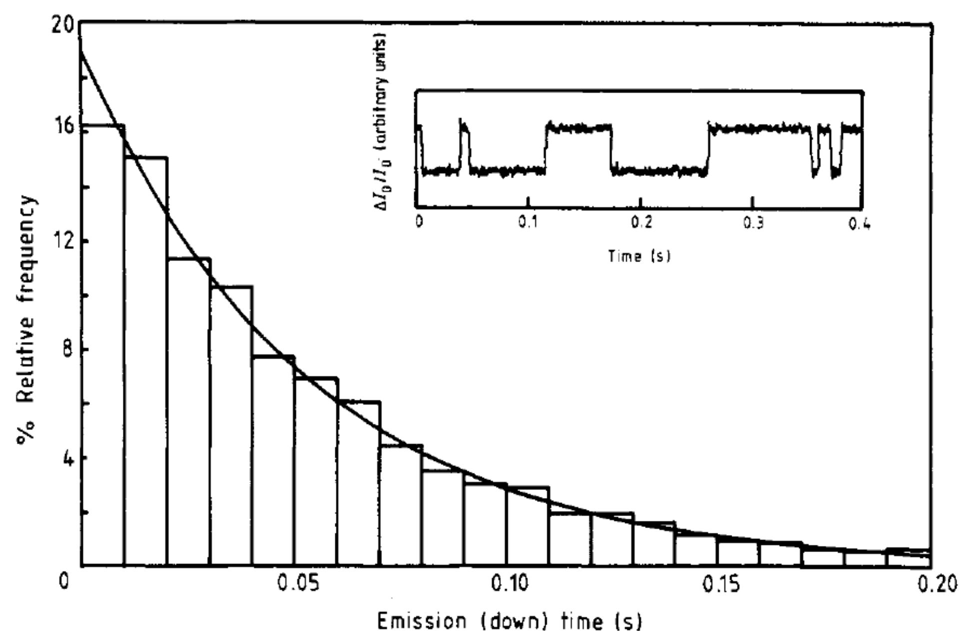


Figure 1.2: Distribution of emission times at 95K and Vgs = 1.15V, showing that the time is Poisson distributed [1].

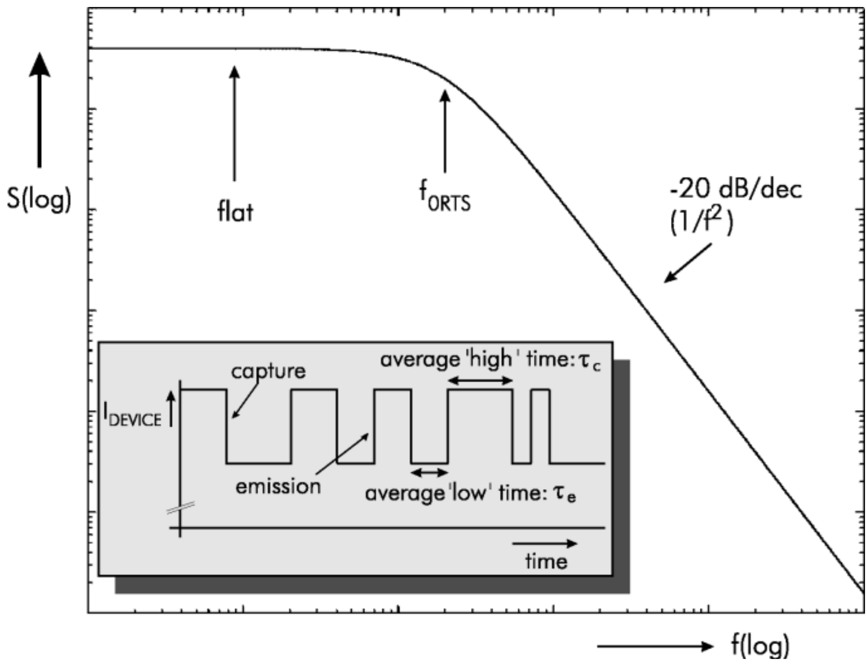


Figure 1.3: Frequency-domain representation of the RTN signal [2].

1.1.2. Soft Error

Another concern in the field of circuit reliability is soft error, which is referred to the type of errors that are uncontrollable, random and usually not catastrophic. These errors are induced by particle strikes, either from the radioactive atoms released by the package material or interaction between cosmic rays and the atmosphere. One of the most commonly seen particles is the alpha particle, which is essentially a helium nucleus that carriers a +2 charge, written as He^{2+}. Compared to other radiation particles, alpha particle is heavy (two protons and two neutrons) and highly ionized therefore has the lowest penetration depth. The interaction between particles and silicon device is shown in Figure 1.4. When a charged particle strikes at the reverse-biased p-n junction, it creates an ionized path through the penetration track. Under the effect of electric field, carriers

created by the strike are collected by the diffusion region resulting in a large transient current pulse. If the current pulse is sufficiently large, a soft error may occur. Nondestructive soft errors can be categorized into two types: (1) Single Event Transient (SET) which generates a voltage glitch propagating through a combinational logic path and (2) Single Event Upset (SEU) which causes a data flip of a memory cell or register.

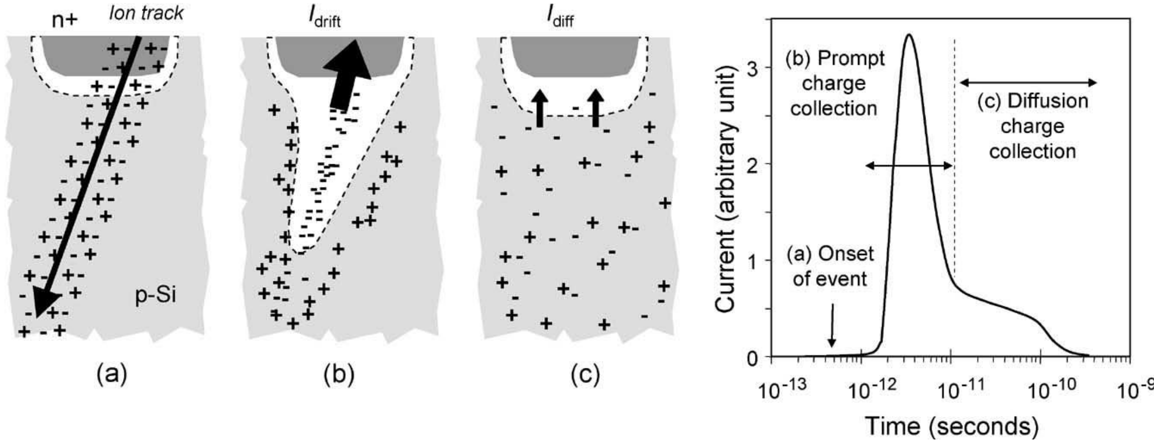


Figure 1.4: Charge generation and collection phases in a reverse-biased junction and the resultant current pulse caused by the passage of a high-energy ion [1].

1.2. Hardware Security

With the development of the Internet of Things (IoTs), achieving secure and trustworthy communication and computation is becoming increasingly challenging. Protections from software level alone is proven to be insufficient, especially against physical attacks such as fault injection, micro-probing and cloning [58]-[62]. Hardware security, aims at providing a silicon solution towards those physical attacks, has emerged and gained attractions in both academia and industry. Traditional hardware security modules are based on cryptographic primitives such as secret key storage,

cipher/decipher blocks (e.g. AES, RSA) and random number generators (RNG). These modules are proven effective and suitable for mainframe computers, especially for those with high performance and require a high level of security. However, due to the computationally intensive algorithm, conventional hardware security primitive are generally power consuming and costly thus are undesired for most of the portable platform. In addition, the fact that mobile devices are distributed, unsupervised and resource-limited further aggravates the situation. As a result, hardware security design are expanded from protecting mainframe servers to distributed lightweight devices. Different lightweight oriented cryptographic protocols has been proposed and discussed [53]-[54] for applications in the source constrained ICs such as sensors, smart cards, and health care gears. Although refining the cryptographic algorithm or protocol will improve the level of security, of equal importance is to provide a more secure entropy source through underlying ICs. The primary focus of our work is to implement hardware security building blocks that are reliable, difficult to break and cost efficient. To begin with, we can consider an authentication scenario between two parties: a resource-abundant server and a resource-limited token.

One of the simplest lightweight authentication schemes is shown in Figure 1.5 [56]. Before deploying the device, a one-time enrollment is required. The server randomly generates some challenges ($C_i \in C_n$) with an RNG. The token device produces the corresponding responses thus forming the Challenge and Response Pairs (CRPs: $<C_i, R_i>$). The server collects and safely stores the CRPs in the database. Before distributing

the devices to clients, the enrollment interface will be permanently destructed in order to prevent any possible micro-probing attacks. In the authentication phase, the server again randomly retrieves a challenge from its database. A genuine token then should return the response that matches the corresponding response stored in the database ($\boldsymbol{R}_i = \tilde{\boldsymbol{R}}_i$) [54].

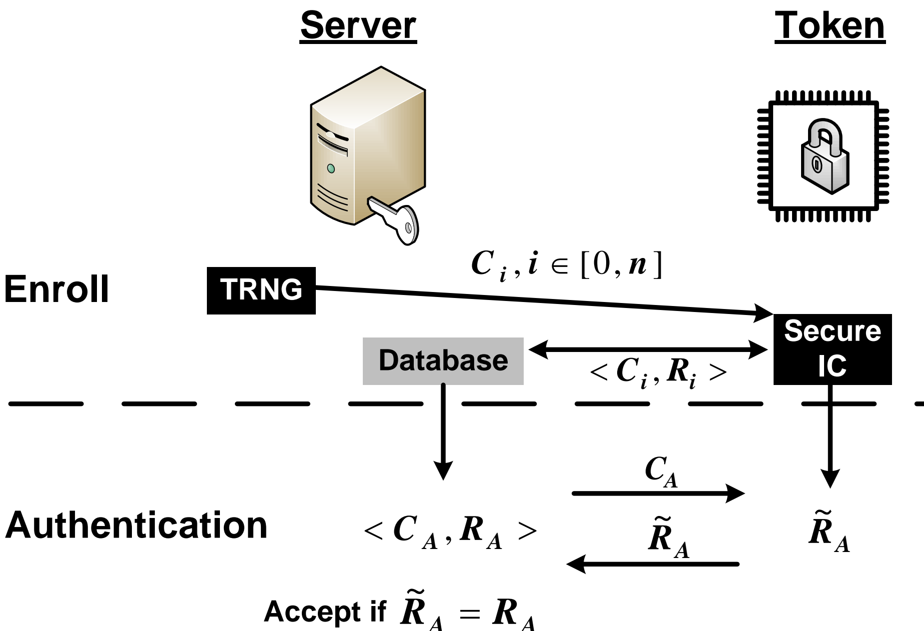


Figure 1.5: A typical authentication protocol involves utilization of TRNG and PUF circuit.

In general, any authentication is breakable given enough time, motivation and resources. Improving the security is essentially to increase the cost for an adversary [53]. In addition to employing a more complex authentication protocol, utilizing novel authentication building blocks can be another solution to improve the security. From the server side, utilizing a True Random Number Generator (TRNG) helps enhance the security by making the random bit-streams (e.g., secret key, challenges for a PUF) more

difficult to guess. From the token side, Physical Unclonable Function (PUF) provides a low cost solution as compared to conventional Non-Volatile Memory (NVM) based authentications.

1.2.1. True Random Number Generator (TRNG)

The security of a network, in a large extent, relies on how unpredictable the secret data is appeared to unauthorized parties. Random numbers therefore are vital to many crypto protocols. Besides generating random challenges for authentications, another important application of random numbers is to provide the private and public key pairs for encrypting messages.

Depending on the approach to generate it, random data can be categorized into pseudo random numbers and true random numbers. Pseudo random numbers, as the name indicated, are essentially generated with algorithms that use mathematical formulae or simply pre-calculated tables. In most of the applications, a Pseudo Random Numbers Generator (PRNG) will be sufficient as the repeating period of the generated random sequence is so long that determinacy can be ignored. However for operation which requires high level security, e.g. the transaction between bank terminals, a truly random data is required. TRNGs are referred to those extracting randomness from physical phenomena, e.g. the circuit noise. The major advantage of TRNG over PRNG is its high unpredictability which increases the difficulty for attackers to guess the random sequence.

1.2.2. Physical Unclonable Function (PUF)

In a cryptographic system, the lightweight devices are generally used in the token side, namely the parties being authorized. Security tokens conventionally store the secret information, e.g. digital signature or biometric data, in NVMs such as EEPROM or fuses. During an authentication, cryptographic blocks import the secret information as a reference key to perform the designated protocols. NVMs are able to provide some basic protections however are becoming less effective on protecting mobile devices due to the following reasons. (1) As the attacking techniques improve significantly in the past years, NVMs are becoming more easily breakable under offline attacks such as cloning, reverse engineering and fault attacks. (2) NVMs are not logic-compatible therefore are oppose to the low-cost requirement for portable devices. (3) Conventional cryptographic hardware is still required to achieve the secure operations.

To address the problems faced by the NVM based authentication techniques, an alternative approach, Physical Unclonable Function (PUF) was proposed. Instead of burning in the secret information in a device, a PUF extracts the random features from the manufacturing induced random variation. PUFs are feasible for a lightweight authentication due to the following reasons: (1) Compatible with the standard logic process therefore less expensive than NVMs; (2) Some PUFs can be directly used for authentication without using additional cipher blocks; (3) Irreversible as the entropy source comes from the physical characteristic of the chip. (4) Immune to offline attacks. PUFs can be categorized in two classes, depending on the number of available CRPs they

can offer. Weak PUFs offers limited number of CRPs therefore is generally employed as a substitute to Non-Volatile Memories (NVMs) to store the secret key. Strong PUFs, in contrast, offers an enormous number of CRPs which are exponentially proportional to the block area. The strong PUF thus can be authenticated directly without using any cryptographic hardware.

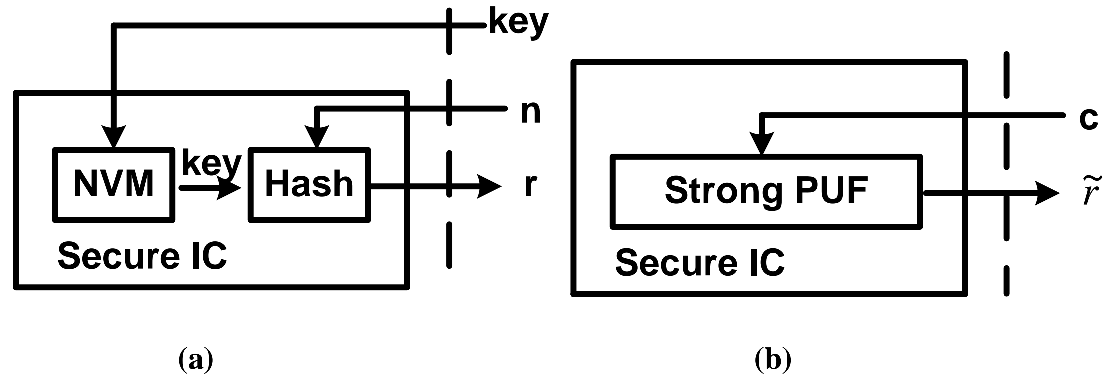


Figure 1.6: (a) Conventional authentication scheme stores keys in NVM. (b) Using strong PUF for direct authentication [54].

Figure 1.6 compares the conventional NVM based and strong PUF based authentication methods. In Figure 1.6(a), a secret key is programmed in NVM during enrollment. An additional cryptographic hardware (e.g. a Hash function) performs the encryption to provide a secure protection. The server sends a random nonce n and receives a response r from the token. The token will be authorized if r = Hash (k,n). Strong PUF based authentication, as shown in Figure 1.6(b), simplify the scheme by incorporating both the secret key and hash function in a single one way function.

To evaluate the PUF performance, we introduce two concepts: intra- and inter- chip Hamming Distance (HD) [73]. The former represents the reproducibility of the responses

while the latter represents the uniqueness of the responses. The intra-chip HD calculates the differences between the expected responses and the actual PUF outputs for the same challenges from a single chip. Inter-chip HD, on the other hand, evaluates how uniquely a PUF is distinguished from other PUFs. This is generally estimated over a group of chips by applying a large set of CRPs. In general, the HD is calculated as follow:

$$HD(x, y) = \sum_{i=1}^{n} (x \oplus y)$$

where x and y represent two n-bit responses. Based on this equation, an ideal token requires 0 inter-chip HD and 0.5 inter-chip HD. However, the PUF behavior is generally very sensitive to environmental variables such as thermal noise, supply voltage and temperature shifts. Therefore, to authentication a token, only an approximated match is required between the actual response and the CRP in the database. As shown in Figure 1.7, to guarantee a safe operation, the intra- and inter- chip HD distribution are required to be sufficiently separated with a certain margin. Otherwise, a post-processing procedure is required [55][74].

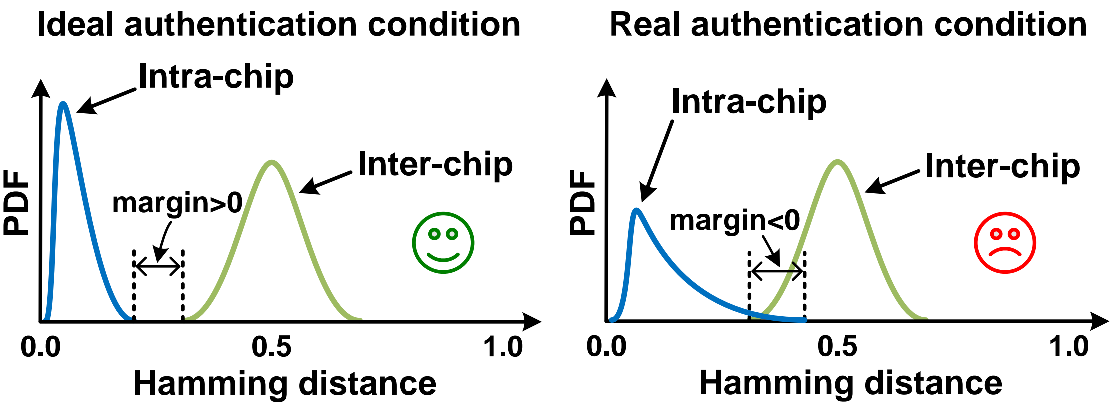


Figure 1.7: Intra-chip and inter-chip HD under (a) secure condition and (b) unsecure condition [55].

Chapter 2. On chip RTN Monitors

2.1. Introduction

One direct impact of RTN on CMOS transistor is the V_{th} fluctuation between capture and emission states which resembles a random telegraph signal. Recent studies on RTN aided by new characterization methods have helped establish a better understanding of the underlying physics. Figure 2.1 compares different RTN characterization techniques. Traditionally, characterization of RTN involved continuously monitoring the transistor drain current for a large population of devices using individual probing [3]-[6]. This method, however, is time-consuming, cumbersome, and provides little insight into the circuit level implications of RTN. Inferring circuit level parameters based on device I-V data is prone to error due to the fast signal switching and complex circuit topology Furthermore, due to the limited sensitivity of prior circuit based approaches, accelerated stress had to be applied in many experiments to amplify the RTN signal. This practice significantly undermines the confidence and applicability of the test results. Circuit based RTN characterization methods are therefore developed. The RTN impact on SRAM is generally monitored through observing the V_{min} shift. However, RTN induced V_{min} shift is so small that very few SRAM cells are within that failure margin making it difficult to be detected. As for logic circuits, very few attempts have been made to assess the true RTN impact. This, we believe, is primarily due to the difficulty of taking high precision

measurements in a short measurement time from realistic circuits such as Ring OSCillators (ROSCs).

	Probing	Circuit based methods	
	Individual transistor	SRAM	Logic circuit or ring oscillator
Schematic		VDD WL WL BL BLB	D Q
Parameter of interest	DC current	SRAM V_{min}	Frequency shift
Pros	1. Simple 2. High resolution	1. Realistic RTN impact on SRAM V_{min} 2. Short test time 3. Small test area	1. Realistic RTN impact on circuit frequency 2. Short test time 3. Small test area
Cons	1. Long test time 2. Large test area 3. Limited insight on circuit level	1. Limited resolution 2. Rare occurrences 3. Averaging effect	1. Limited resolution 2. Long meas. time 3. Averaging effect

Figure 2.1: Comparison of different RTN characterization techniques.

For efficient collection of large RTN statistics, several logic circuit based approaches have been demonstrated for RTN measurements. The metastable behavior of a counter circuit was used in [7]-[8] to extract RTN signatures. On the modeling side, there has been a large body of work analyzing the impact of RTN on circuit parameters such as logic gate delay and SRAM noise margins [11]. For instance, a statistical timing estimation algorithm was proposed in [12] to calculate RTN induced logic delay shift for a large circuit block. However, the lack of experimental data to verify the estimation results undermines the confidence of such work. Therefore an odometer circuit with

capability of high resolution and high sample rate measurement for RTN detection is implemented for accurately characterizing the RTN induced frequency shift in logic circuit.

2.2. Beat Frequency Technique for RTN Monitoring

The basic concept of the beat frequency detection (BFD) technique for measuring the frequency difference between two ROSCs is illustrated in Figure 2.2 [13]. ROSCs are widely used evaluating process variation and reliability issues in logic circuits since they are able to provide circuit designers a straightforward estimation on circuit performance. The beat frequency detector samples the output of one oscillator using a D flip-flop at intervals set by the output of the other. The faster signal A catches up and then overtakes the slower signal B, and as this process repeats, the time between the overlapping points is the period of the beat frequency. This time is measured by counting the number of reference ROSC periods during a single beat period (i.e. $N=floor(f_B/(f_A-f_B))$). This information is then read out through a scan-based interface. The advantage of this technique is that the measurement resolution can be made very high by bringing the two frequencies f_A and f_B closer to each other. For example, when the initial frequency difference is calibrated to be 1%, the initial count output is 100. With a RTN trapping event on ROSC B which increase the frequency difference to 1.5%, the count output changes from 100 to 67. Therefore, the minimum frequency measurement resolution, corresponding to a count change from 100 to 99, is 0.01%. Further details on how to calculate the actual % frequency shift based on the scanned data can be found in the

previous publication [13]. The beat frequency approach enables us to measure changes in transistor switching times as small as one part in 10,000 in less than a microsecond, making it ideally suited for characterizing RTN effects in logic paths.

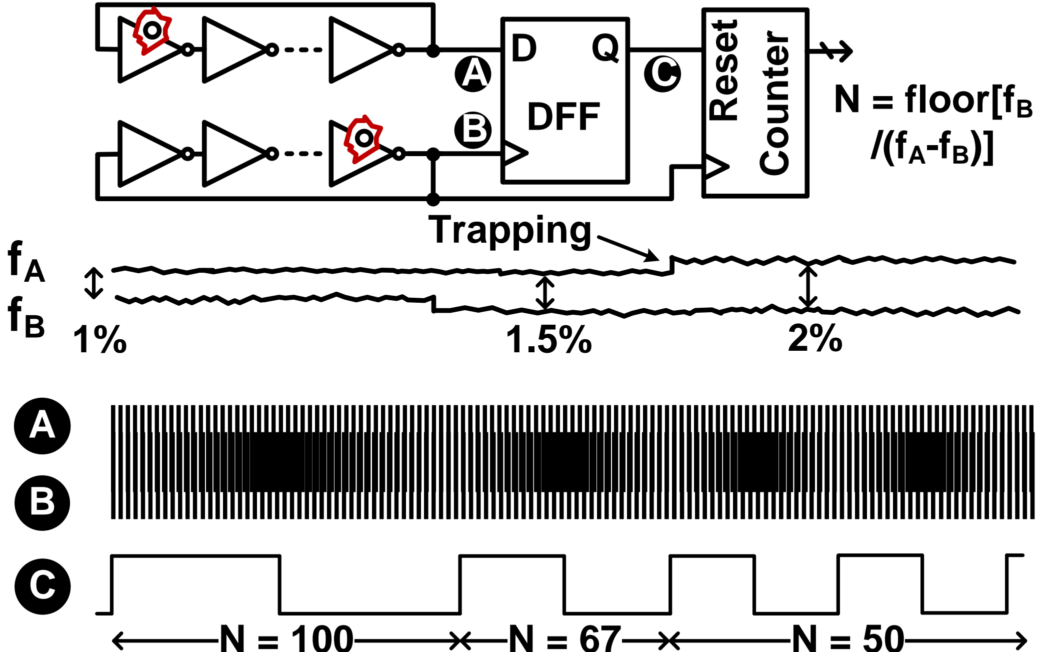


Figure 2.2: Beat frequency detection circuit adopted in this work for measuring RTN induced delay shifts at sub-0.5V supply voltages with high resolution. The output count N represents the number of f_B clock cycles that can fit within a single beat frequency (i.e. f_A-f_B) clock period.

2.3. Characterizing RTN with a Single Array Structure on a 65nm Process

2.3.1. RTN Monitor Design

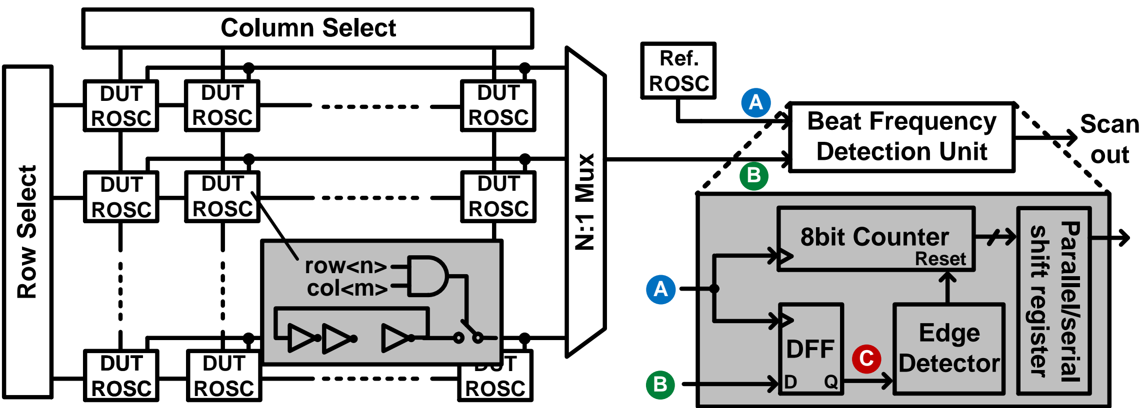


Figure 2.3: ROSC array test chip for RTN measurements comprising an on-chip beat frequency detection

The top level diagram of the 65nm test vehicle is shown in Figure 2.3. It consists of a 10x8 ROSC array, a reference ROSC, and a beat frequency detection unit. Each ROSC has only 11 inverter stages to ensure high RTN sensitivity by minimizing the averaging effect. Prior to the testing, the initial frequency difference between the DUT and reference ROSCs is set to be around 1% (i.e. output count = 100) using on-chip trimming capacitors. This initial setting was found to provide a sufficiently high measurement resolution (= 0.01%) with minimal noise effects. One DUT ROSC is selected at a time for the frequency measurements using column and row select signals. Output signal of

the selected DUT ROSC drives the bitline signal which is multiplexed out and fed to the D flip-flop inside the beat frequency detection unit.

	Single ROSC	*DFF Sensor [8]*	*This Work*
Circuit Schematic	Counter	D Q Counter	D Q reset Counter
Description	**Simple ROSC with frequency divider**	**D flip-flop driven by a simple ROSC to metastable state**	**Differential ROSC structure, beat frequency detection**
Freq. Meas. Resolution	**Low resolution, sensitive to V, T drifts**	**Pass/fail information only**	**High resolution, immune to V, T drifts**
Sampling Time*	**~ 100μs**	**N/A**	**> 1μs****
Capable of Measuring	**N/A (no reported RTN data using this method)**	**Single trap only**	**Multiple traps**
Source(s) of RTN		**ROSC and DFF**	**ROSCs**
Output	**Actual ROSC freq. shift**	**Input frequency divided by 2 or 3**	**Actual ROSC freq. shift**

** For a 0.01% frequency shift resolution and a ROSC period of 10ns*
*** Emission and capture times down to several microseconds can be measured*

Figure 2.4: Comparison of three ROSC based RTN measurement techniques.

For a better understanding, Figure 2.4 compares three ROSC based RTN measurement structures including the proposed technique. A ROSC with a frequency divider can be considered, however, this single-ended configuration suffers from large measurement noise in the presence of voltage fluctuations and temperature drifts. To the best of our knowledge, there haven't been any reports showing RTN data based on this simple approach. Another method for measuring RTN is illustrated in Figure 2.4 where a frequency divider in a metastable state is used [8]. The divider's output frequency switches between f/2 and f/3 according to the RTN capture and emission times. Here, f is

the input frequency. The main drawback of this approach is that it is hard to infer physical parameters such as frequency or V_{th} based on the erratic divider output. The beat frequency scheme, on the other hand, can measure the exact frequency shift due to single or multiple RTN traps with high precision and short measurement time, making it an effective characterization method for RTN effects in logic circuits.

2.3.2. RTN Induced Frequency Shift Measurements on 65nm Test Chip

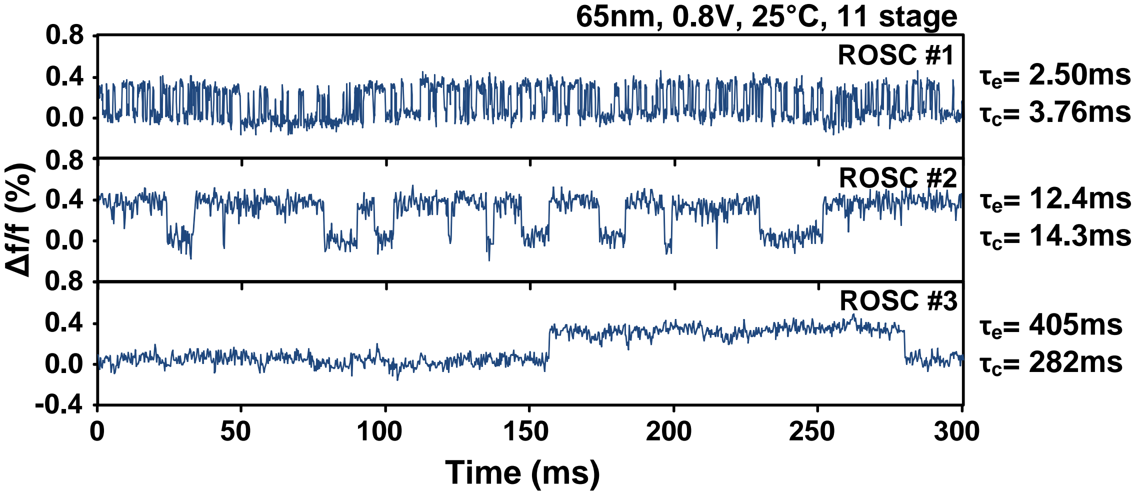


Figure 2.5: Single trap RTN waveforms measured from different ROSCs.

The frequency shift waveforms in Figure 2.5 are from three different ROSCs in the test array. RTN's signature trapping/de-trapping behavior can be clearly observed. Independent of the time constant values, the frequency shift caused by a single RTN trap was approximately 0.4% for the 11 stage ROSC operating at 0.8V and 25℃. Our repeated experiments revealed no RTN in the reference ROSC, although this is purely a statistical occurrence. Due to the identical layout and physical proximity, there is no reason to believe that the chances of having an RTN trap in the reference and DUT

ROSCs will different. The emission and capture time constants (i.e. τ_e and τ_c) range from 2.50ms to 405ms depending on the supply voltage and ROSC instance.

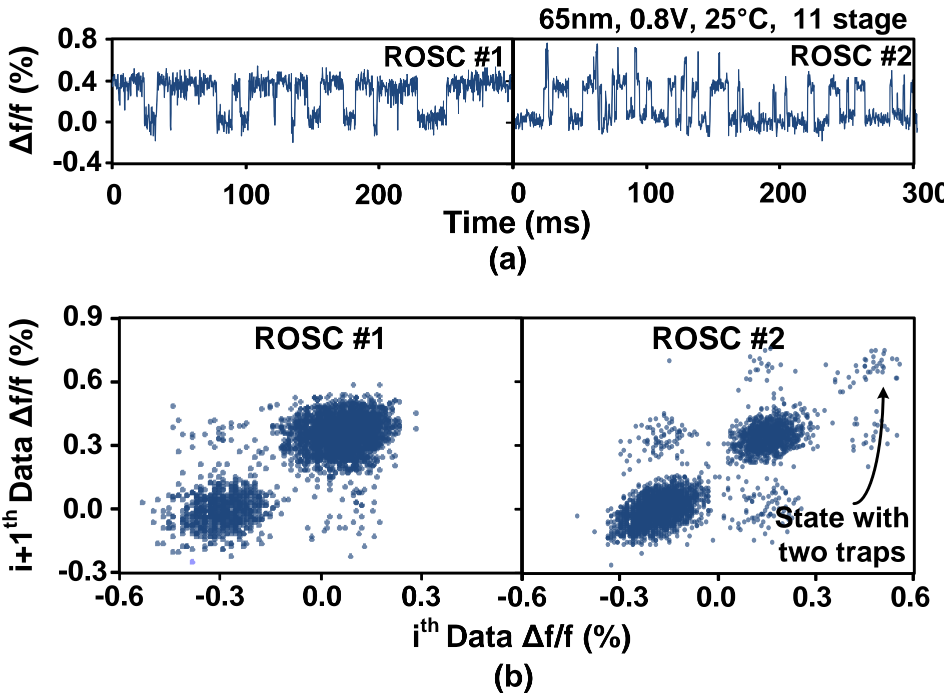


Figure 2.6: (a) Single and multi-trap RTN waveform from two different ROSCs. (b) Time Lag Plot (TLP) of the two traces.

Figure 2.6(a) shows frequency traces of two ROSCs having a single trap and two traps, respectively. The maximum frequency shift induced by two traps is 0.8%, which is almost twice as large as the single-trap one indicating a state in which both traps are in capturing event. But as indicated from the waveform, the time constant at 0.8% freq. shift is relatively smaller than the other two states which makes it difficult to be recognized. A Time Lag Plot (TLP) therefore is proposed [14] for a better visualization of the capturing and emission transitions as shown in Figure 2.6(b). This method plots each sampled points on a plane indicating its current state and the relationship to the next point's state.

As shown in these two TLPs, a single trap has only one capture and one emission state while two trap RTN has three clusters on the diagonal.

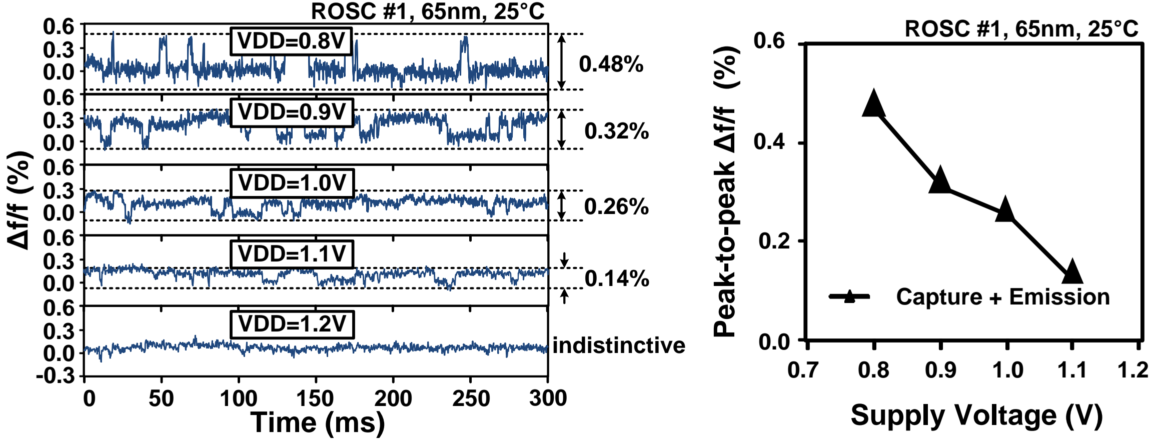


Figure 2.7: RTN traces from same ROSC at different supply voltages.

RTN parameters are shown to have a strong voltage dependence which was experimentally verified from the test chip (Figure 2.7). As the supply voltage is increased from 0.8V to 1.1V, the RTN induced frequency shift decreases from 0.48% to 0.14%. Eventually, RTN becomes indistinctive at 1.2V, the nominal operating voltage of this process. These results are in line with previous studies that have reported a stronger RTN signal at lower supply voltages [14]. RTN time constants are defined as the average time a trap site stays in the occupied state or in the unoccupied state. The capture (τ_c) and emission (τ_e) time constants can be extracted using an exponential model which agrees well with the measured data as shown in Figure 2.8. The distribution of the capture and emission times distribution spreads out as the supply voltage is increased from 0.8V to 1.0V indicating an increase of capture time constant. And the emission time distributions show opposite dependence. For a better understanding we extrapolated the capture and

emission time constants in term of supply voltage as shown in Figure 2.9. The capture time constant shown in the Figure 2.9(a) decreases as the supply voltage is increased indicating an RTN trap on a NMOS. Whereas for RTN trap located on the PMOS, as shown in Figure 2.9(b), the capture time increase with the voltage. The Power Spectrum Density (PSD) of a DUT ROSC with a single trap is shown in Figure 2.10. The spectrum follows a Lorenzian with a corner frequency at ~10ms which is in accordance with the RTN capturing and emitting time constants. Histogram of the numbers of RTN traps per ROSC is shown in Figure 2.11. 66% of the ROSCs did not show any signs of RTN while no ROSC had more than 2 traps. The relatively low number of traps in each ROSC can be attributed to the mature process technology used for the test chip fabrication.

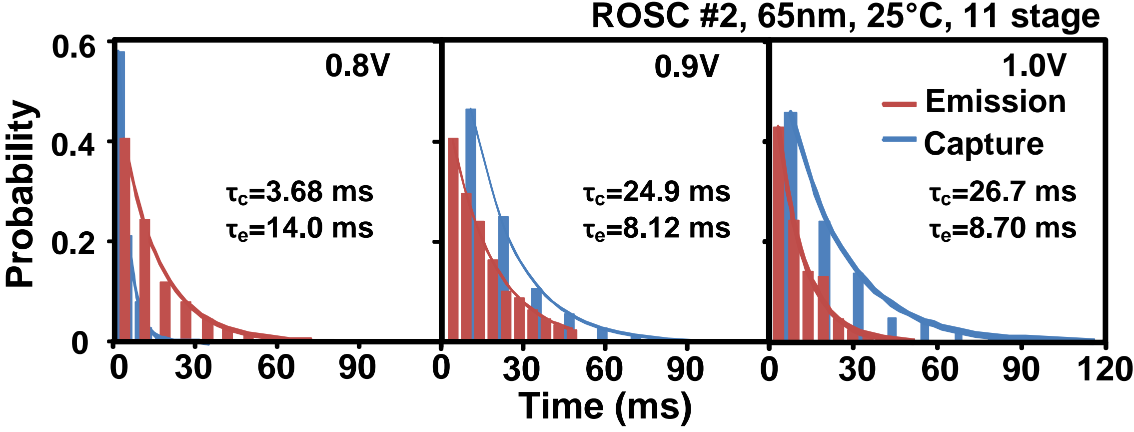


Figure 2.8: Capture and emission time distributions and exponential fit results for supply voltage from 0.8V to 1.0V.

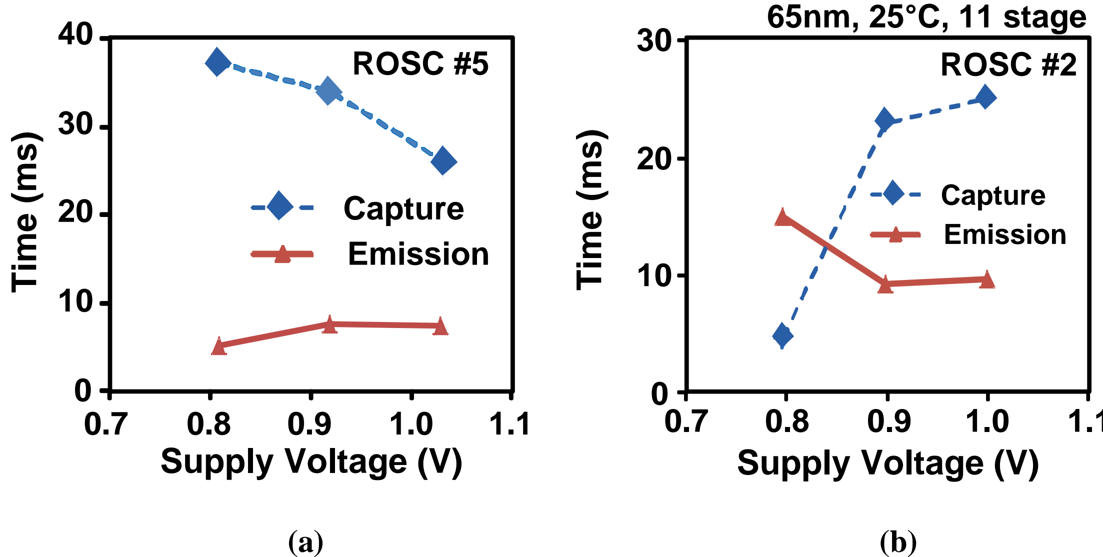


Figure 2.9: Time constants versus supply voltages.

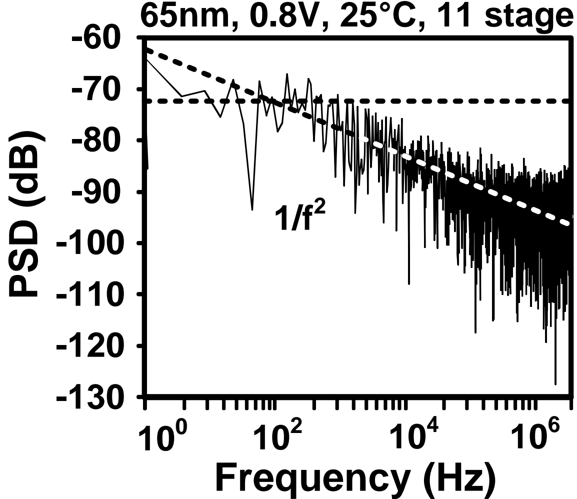


Figure 2.10: Power Spectrum Density (PSD) of the frequency shift data.

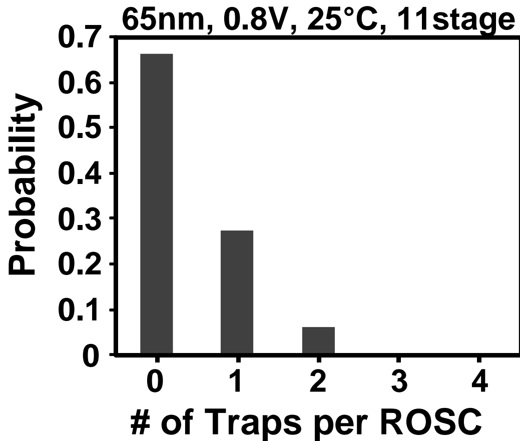


Figure 2.11: Histogram of the number of traps per ROSC.

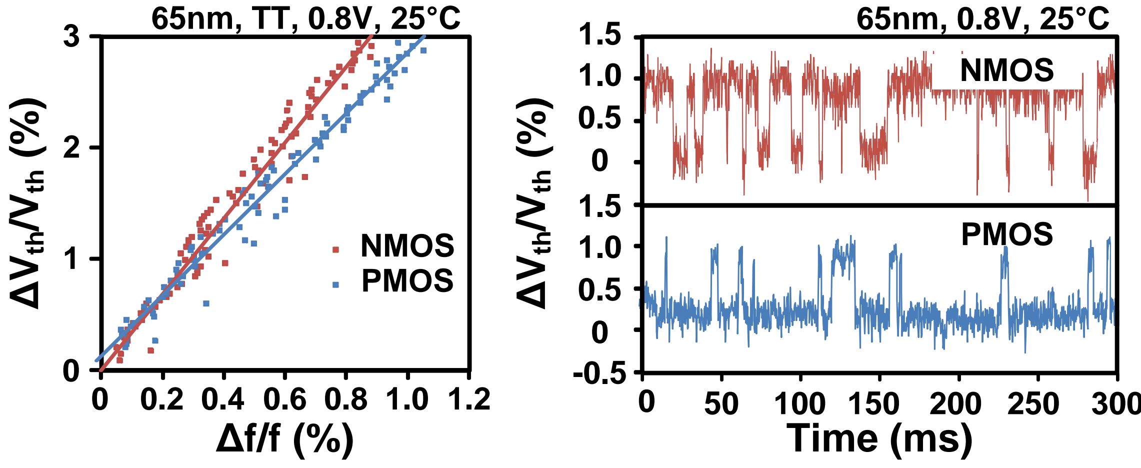


Figure 2.12: Frequency to V_{th} mapping.

For a better understanding, we simulated an 11 stage ROSC with HSPICE and translated the measured frequency data back to the device V_{th} using the mapping curve as shown in Figure 2.12. RTN induced V_{th} shifts can be estimated from the measured frequency shift data with linear fitting. The transient waveforms in terms of V_{th} shift with respect to NMOS and PMOS are shown on the right. It indicates that single trap induced V_{th} shift in NMOS is 1.9% compare to 1.6 % in PMOS. The dependence of frequency

shifts on V_{th} shifts with different supply voltage is simulated and shown in Figure 2.13(a). RTN induced V_{th} shifts with different supply voltages is translated from the measurement data. The transient waveform in Figure 2.13(b) indicates that the same trap induced V_{th} shifts reaches maximum at 0.9V. But for frequency shift, as shown earlier, the maximum value is achieved at 0.8V. This discrepancy implies that both the absolute V_{th} shift value and the parametric dependences should be considered when evaluating the RTN impact on logic circuit frequency shift.

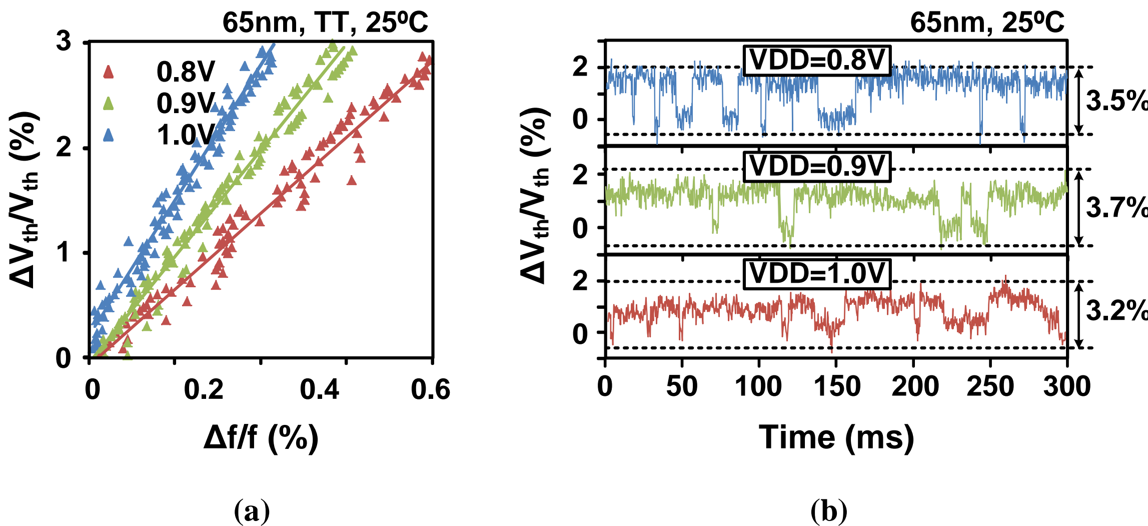


Figure 2.13: RTN induced V_{th} shift with different supply voltage.

2.4. Characterizing RTN with a Dual Array Structure on a 32nm HKMG Process

With scaling of VLSI technology, the transistor feature size is halved every generation however the supply voltage is scaled at a slower rate. The power density therefore is expected to grow in future technology nodes. On the other hand, the chip cooling

capability remains the same which limits the number of simultaneously switching transistor per area [17]. One solution is to utilize the Near Threshold Voltage (NTV) technique which improves the energy efficiency by decreasing the supply voltage to a near threshold region [18]. A major challenge for NTV operation is the higher sensitivity to parametric variations as compared to a nominal operation condition. RTN induced frequency shift in logic circuit at lower supply voltage, as we proved in the previous section, is expected to be more severe than that at a higher voltage. For a better understanding of the RTN impact on circuit performance at a NTV operation, we proposed a dual array structure RTN monitor which is capable of characterizing RTN induced frequency shift in logic circuit in a sub-0.5V region.

The main contribution of this work is that we present detailed RTN induced frequency fluctuation data collected from a 32nm test chip operating at supply voltages as low as 0.45V. One of the main shortcomings of the previous design is that the resolution degrades sharply at low supply voltages due to increased variation between ROSC frequencies, which makes the BFD technique less accurate. Note that RTN effects become more severe at low supply voltages due to the Fermi level change and higher circuit sensitivity. To overcome this limitation, this work proposes a dual ROSC array based test structure which achieves a frequency measurement resolution less than 0.01% for every single ROSC in the array for supply voltages down to 0.45V.

2.4.1. Dual Ring Oscillator Array Technique

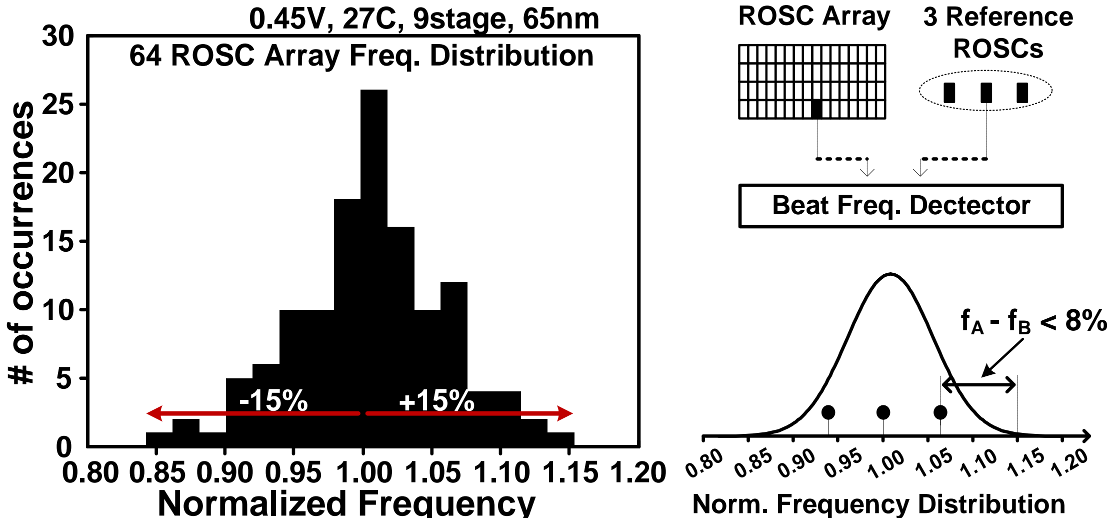


Figure 2.14: Limitation of prior art. Due to the wide frequency spread, not all ROSCs under test can achieve high measurement resolution at sub-0.5V supply voltages.

When a large number of ROSCs need to be measured at low supply voltages however, due to process variation between ROSCs, a small frequency difference (e.g. <1%) between the two ROSCs cannot always be guaranteed. This can be seen in Figure 2.14 where the frequency variation of 64 ROSCs can be as high as +/-15% at 0.45V. In the previous design the ROSC test array is paired with three reference ROSCs, the frequency difference can be as high as 8% which limits the frequency measurement resolution to >0.6% which is not sufficient for precise RTN measurements. Tuning the frequency of individual ring oscillators using dedicated hardware is not desirable since the tuning circuit itself may introduce RTN noise. Furthermore, adding tuning circuits will make the

ROSC less representative and increase the sensitivity to common-mode effects such as temperature and voltage drifts.

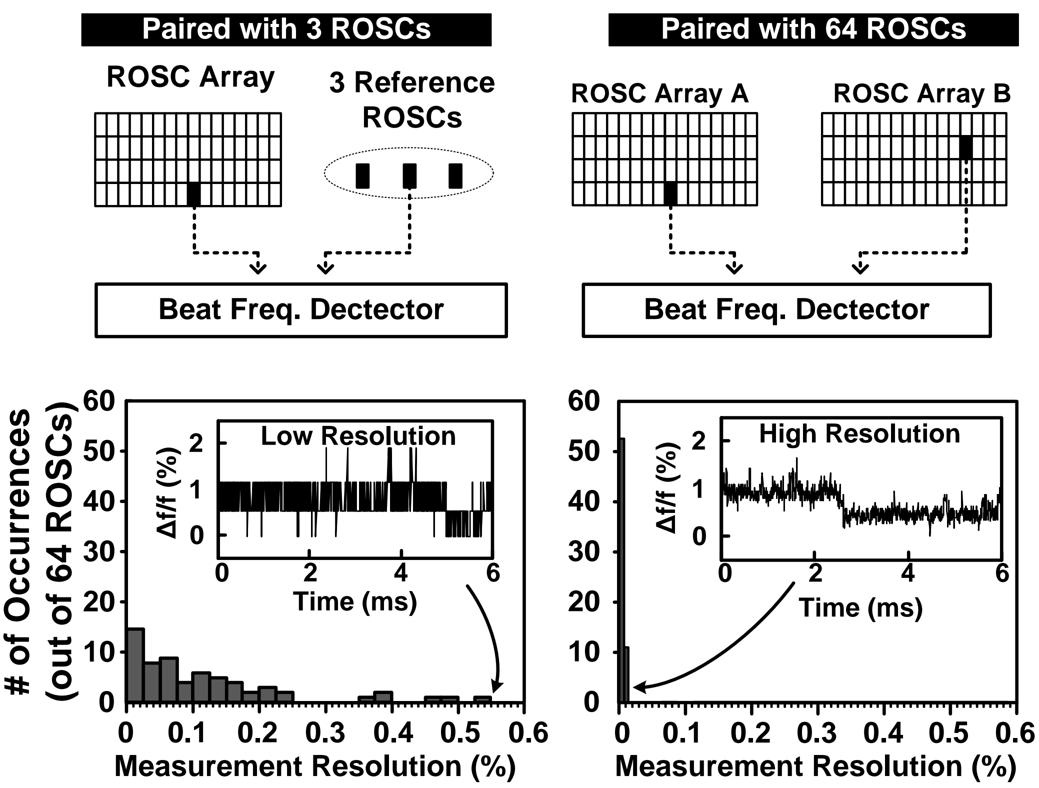


Figure 2.15: Measurement resolution comparison when pairing a 64 ROSCs with 3 reference ROSCs (left figure) and 64 reference ROSCs (right figure). A more precise waveform can be reconstructed using 64 reference ROSCs which is critical for collecting high quality RTN statistics at low supply voltages such as 0.5V.

To overcome this limitation, in this work, we propose a dual-array test structure, which guarantees that a ROSC from the main array can be paired with a ROSC from another array with a frequency difference less than 1%. This ensures a frequency measurement resolution of less than 0.01% even in the worst case. As shown in Figure 2.15, as the number of reference ROSCs increases from 3 to 64, the worst-case

measurement resolution is improved from 0.5% to 0.01% for the proposed dual ring oscillator array configuration. Test chip results in Section 2.4.2 indicate that a frequency resolution of 0.05% is attainable which is significantly less than the frequency shift induced by a single RTN trap.

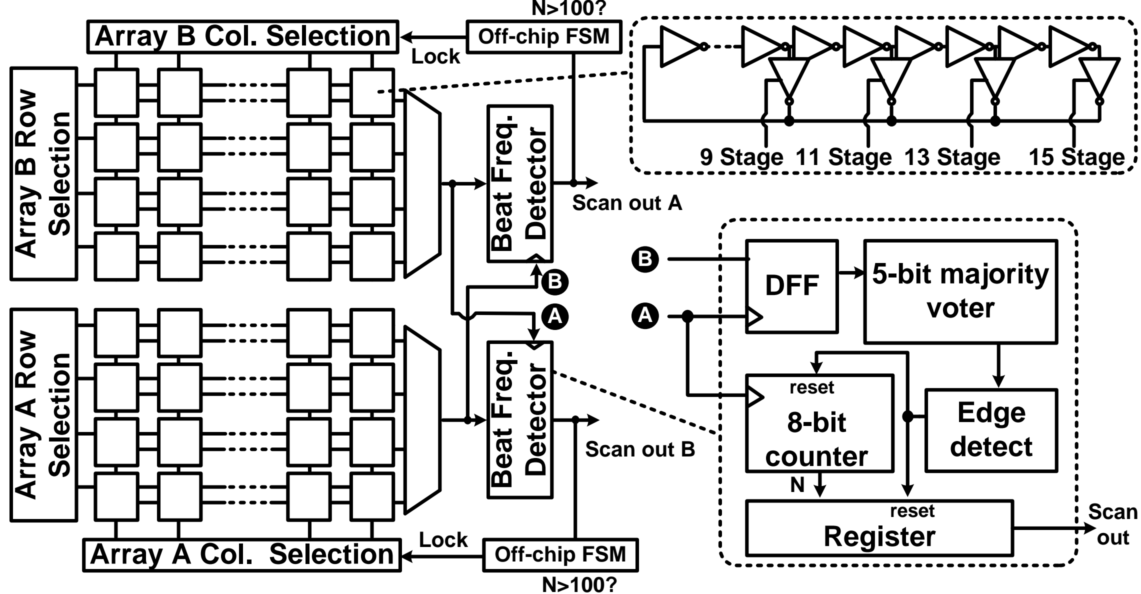


Figure 2.16: Block diagram of the proposed dual ROSC array based RTN characterization circuit. By pairing ROSCs from two arrays, the beat frequency detection circuit can achieve a frequency measurement resolution less than 0.01%. The number of inverter stages can be configured from 9 to 15 using scan bits.

Figure 2.16 shows further details of the 32nm test chip. It consists of two identical ROSC arrays, each comprising 64 ROSCs, along with two separate beat frequency detectors to determine which of the two input frequencies is higher. A 5 bit majority voter circuit is used to prevent functional errors caused by logic bubbles (e.g. lone 0 in a string of 1's and vice versa) which are likely to occur when the two ROSC edges are very close

to each other. A ROSC in one array is sequentially paired with a ROSC in the other array until the BFD count falls within the desired range (e.g. >100). A finite state machine sends out a 'lock' signal to freeze the column and row selection signals, and then the frequency difference is measured and scanned out. The pairing process takes no more than 100 μs using our automated test setup. ROSCs are designed with programmable number of stages (i.e. 9, 11, 13, and 15) to study the impact of the number of inverter stages on the amount of RTN induced frequency shift. PMOS and NMOS transistors used in the ROSC circuit have a width of 624nm and a length of 56nm. The new test structure is well suited for Bias Temperature Instability (BTI) stress experiments since the ROSC can be configured as an open-loop inverter chain using tri-gate stages.

2.4.2. RTN Induced Frequency Shift Measurements on 32nm Test Chip

The proposed dual-array based RTN monitor was fabricated in a 32nm high-k metal-gate process. The nominal supply voltage of this technology is 0.9V. Figure 2.17(a) shows frequency shift traces of a 9 stage ROSC from 0.45V to 0.6V. Measurements show the signature RTN behavior caused by trapping and de-trapping events. The measured RTN induced frequency shift decreases from 0.38% to 0.15% as the supply voltage is increased from 0.45V to 0.6V. The telegraph-shaped RTN waveform was not appreciable at supply voltages higher 0.6V. This suggests that RTN is not a major issue at the nominal supply voltage, but will become more significant when the supply is lowered to near-threshold voltages. The magnitude of frequency shift due to RTN measured from 6 different ROSCs is shown in Figure 2.17(b). Variation in RTN induced frequency shift

can be attributed to the different trap locations in the gate oxide [15]. The frequency shift monotonically decreases at higher supply voltages. One possible reason for this is that ROSC frequency is more sensitive to the same Vth change at lower supply voltages due to the lower gate overdrive voltage. Figure 2.18(a) shows the frequency shift waveforms at 27 ℃, 55 ℃, and 85 ℃. The magnitude of the frequency shift shows little dependence on temperature, however trapping and de-trapping occur more frequently at higher temperatures which is in line with previous studies. RTN time constants are defined as the average time a trap site stays in the occupied state or in the unoccupied state. The capture (τ_c) and emission (τ_e) time constants can be extracted using an exponential model fitted to the measured distribution, as shown Figure 2.18(b) (upper). Theoretically, the proposed BFD can measure time constants shorter than a microsecond. However, due to the slow data scan out, the minimum time constant measureable by our design is a few microseconds. The maximum time constant we can measure is limited only by the measurement time. To study the impact of logic depth on frequency shift, we first selected a ROSC with an RTN trap and then varied the number of stages using scan signals. Experimental data in Figure 2.19 shows that as the number of stages increases from 9 to 15, the frequency fluctuation reduces from 0.38% to 0.24% for the same RTN trap due to the lower sensitivity.

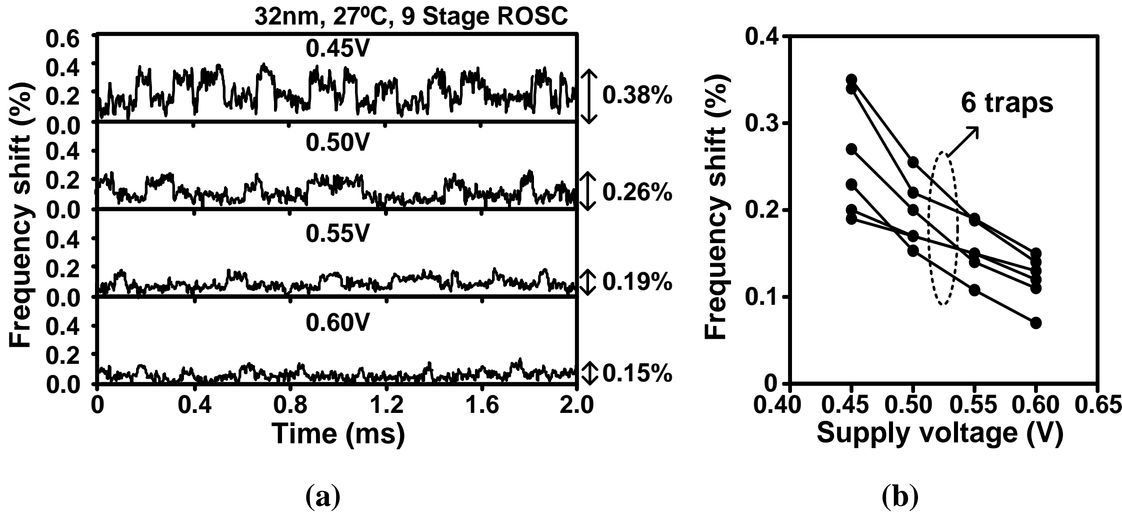


Figure 2.17: (a) RTN induced frequency shift traces measured at different voltages. (b) Magnitude of frequency shift of 6 RTN traps measured at different voltages.

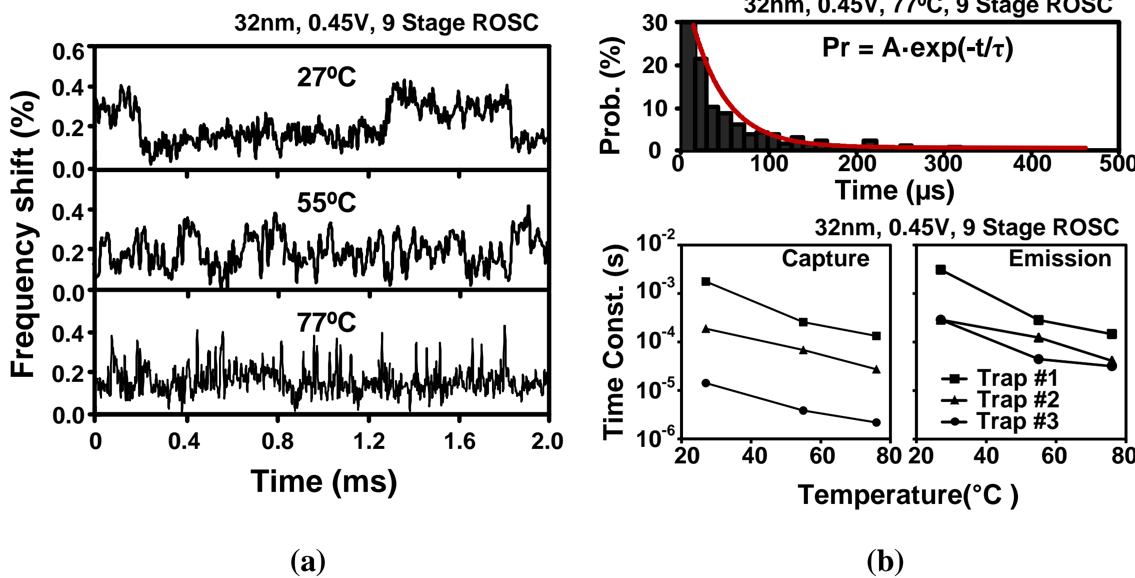


Figure 2.18: (a) RTN induced frequency shift due to the same trap measured at different temperatures. (b) Capture and emission time constants both decrease at higher temperatures.

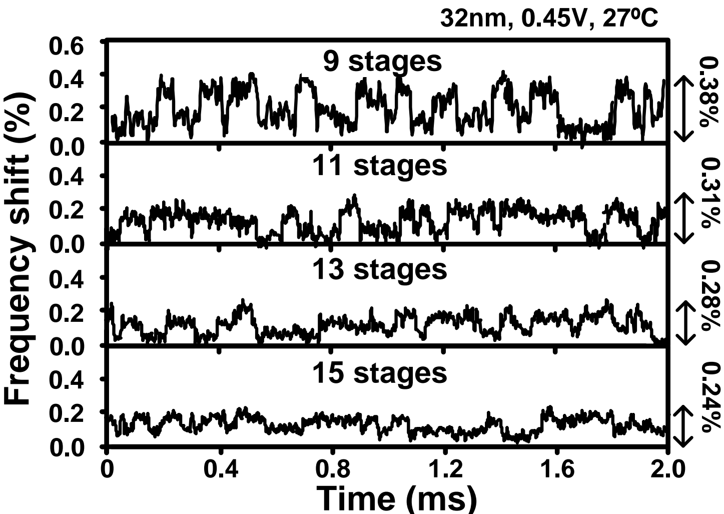


Figure 2.19: RTN induced frequency shift versus the number of ROSC stages. The frequency shift caused by the same RTN trap is reduced as the number of stages increases.

Figure 2.20 shows the occurrence and location of RTN traps across a single test chip from 0.45V to 0.6V. RTN traps may appear or disappear as the supply voltage is varied which we suspect is due to the Fermi level shift [15]. That is, the RTN trap is more likely to be detected if the trap energy level and the Fermi level are closely aligned. The number of ROSCs affected by RTN remained relatively constant under different supply voltages.

Both RTN and BTI have been reported to originate from the same defect sources [16]. To understand the interplay between RTN and BTI better, we measured the location and occurrence of RTN while applying a voltage stress to the ROSC array. The ROSC frequencies were sampled periodically at 0.45V while the test chip was subject to a 1.8V voltage stress (=2x the nominal V_{DD}) for 14 hours. Stress results in Figure 2.21 reveal several newly generated RTN traps as well as few annealed traps. The former can be

attributed to defects created during BTI stress while the latter may be related to the BTI recovery phenomenon [16]. The higher occurrence rate with longer stress time implies that RTN along with BTI further degrades the circuit long-term performance. The percentage of ROSCs affected by RTN measured from 6 different chips is shown in Figure 2.22.

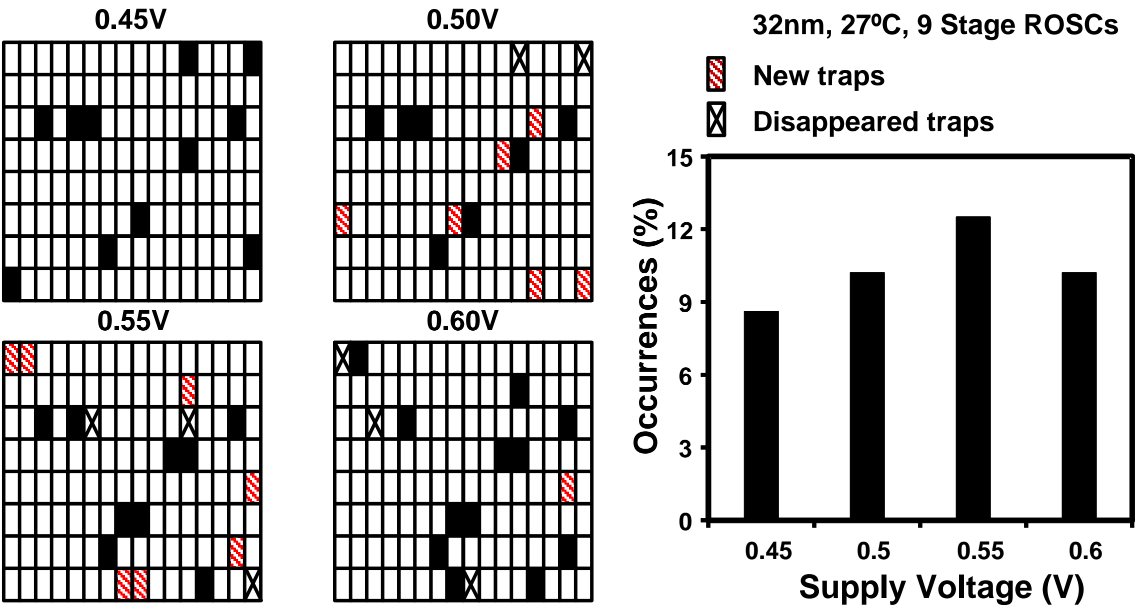


Figure 2.20: RTN trap location map measured at different supply voltages. Each cell represents a single ROSC.

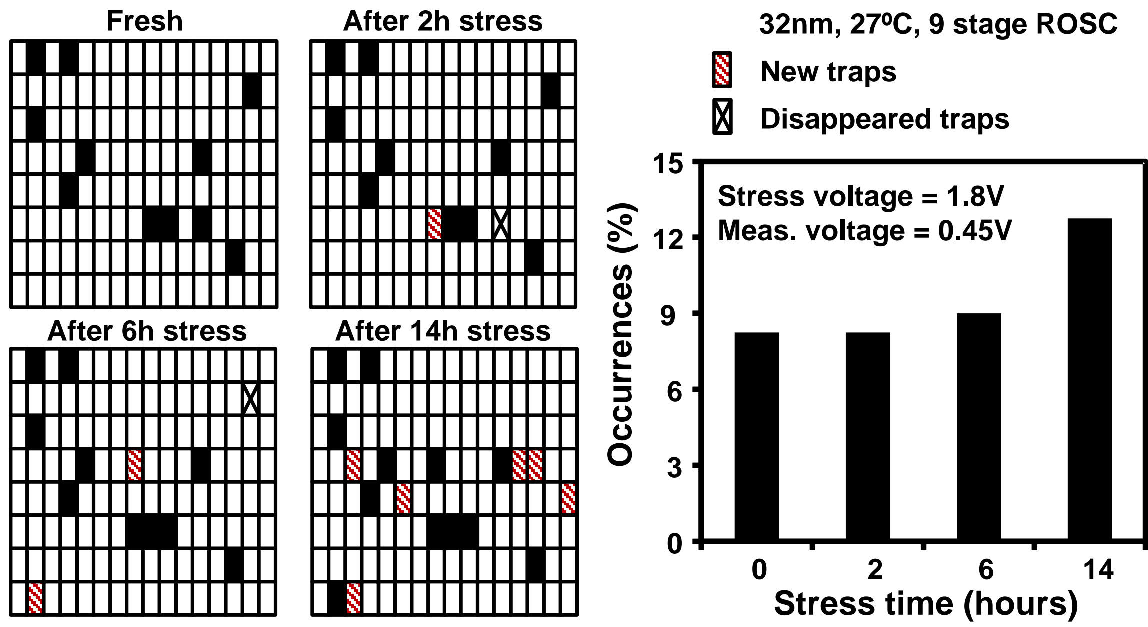


Figure 2.21: RTN trap location map measured after 0, 2, 6 and 14 hours of 1.8V stress.

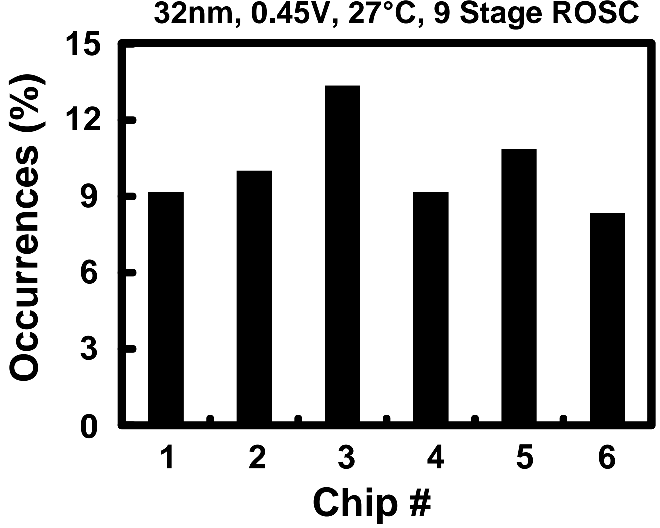


Figure 2.22: RTN occurrences measured from 6 different chips.

2.4.3. RTN Impact on Logic Timing

To estimate RTN induced delay shift in circuits other than simple inverters, we first translated the frequency shift measured from the 32nm test chip to V_{th} shift using the frequency versus Vth relationship simulated in SPICE. Then we apply the Vth shift to various logic gates and D flip-flops (DFFs). We have evaluated three possible RTN induced timing violations in a typical pipeline circuit. Setup time violation is illustrated in Figure 2.23 in the presence of RTN traps in the clock tree. In the worst case, the launching clock CLK1 arrives late and the sampling clock CLK2 arrives early due to RTN. This introduces a skew between CLK1 and CLK2 which reduces the available time for logic computation. The second scenario is shown in Figure 2.24 where the combinational logic delay increases due to RTN. Finally, as shown in Figure 2.25, RTN in the DFF can affect setup and hold times. For better understanding, Figure 2.26(a) shows that the worst case DFF setup time occurs when traps appear in alternating PMOS and NMOS devices on the signal path from D to Q. Figure 2.26 (b) displays the D-to-CLK and CLK-to-Q delays with and without RTN. Since RTN becomes more appreciable at low supply voltages, our simulations are performed at 0.5V. It can be seen that in the presence of RTN, the setup time and hold time curves shift either to the right or left depending on the location of the RTN trap. The setup time varies by -0.08 to 0.18 FO4 inverter delays with RTN.

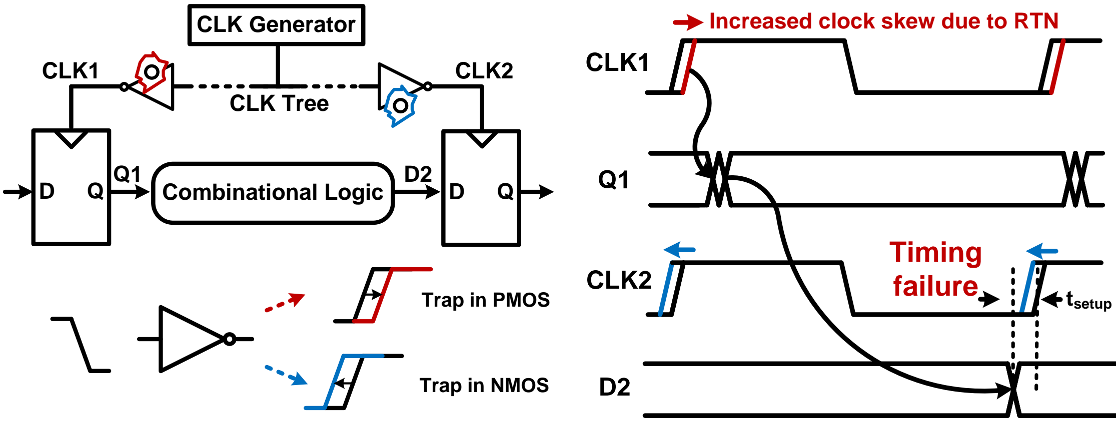


Figure 2.23: Logic timing errors for RTN traps located in clock tree.

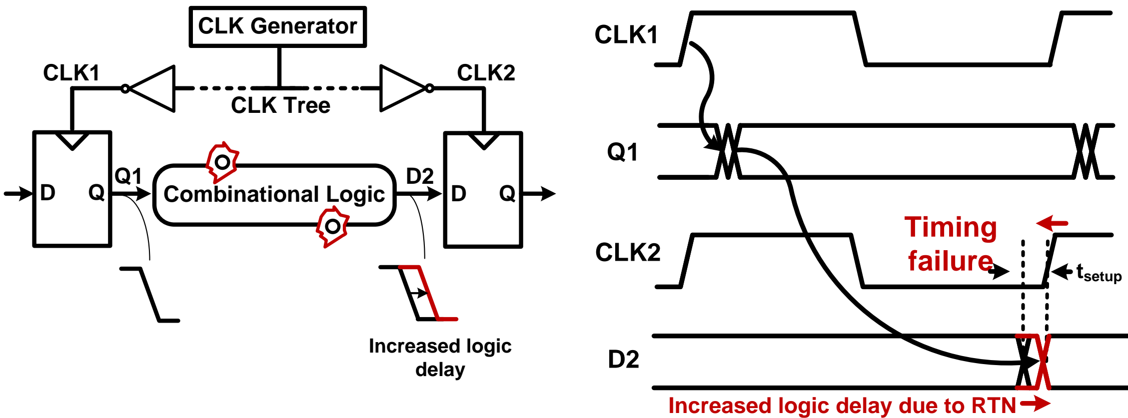


Figure 2.24: Logic timing errors for RTN traps located in combinational logic.

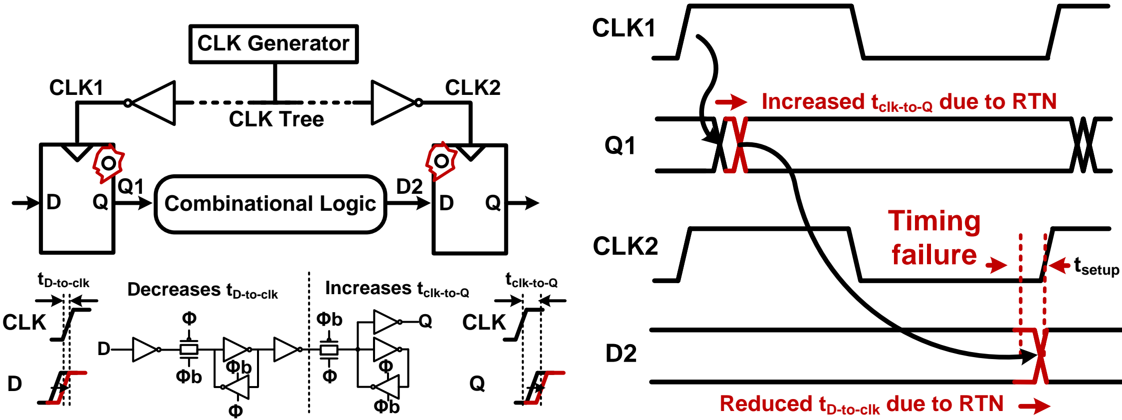
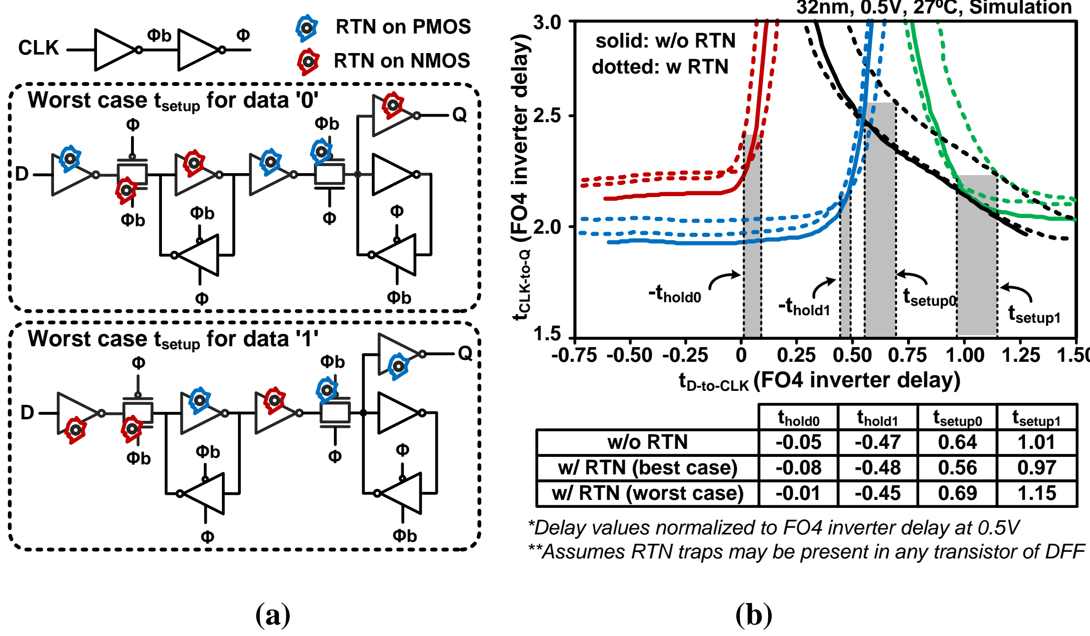


Figure 2.25: Logic timing errors for RTN traps located in flip-flop.



	t_{hold0}	t_{hold1}	t_{setup0}	t_{setup1}
w/o RTN	-0.05	-0.47	0.64	1.01
w/ RTN (best case)	-0.08	-0.48	0.56	0.97
w/ RTN (worst case)	-0.01	-0.45	0.69	1.15

Figure 2.26: (a) RTN trap location on DFF signal path for worst case setup time (hold time is opposite location). No traps assumed on clock path. (b) RTN impact on D-flip-flop setup and hold times.

The following discussion will focus on setup time violation. A similar analysis can be performed for hold time violation. As shown in Figure 2.27(a), to operate without any logic errors, the clock period T_{clk} must be greater than $t_{clk\text{-}to\text{-}q}$ + t_{logic} + t_{setup} + t_{clk_skew}. Figure 2.27(b) compares the max-delay time under different RTN scenarios. In the absolute worst case, traps may be present in the input and output DFFs as well as the clock tree and logic path. The max-delay time allowed for correct operation is reduced by 0.21 FO4 inverter delays under this worst-case condition.

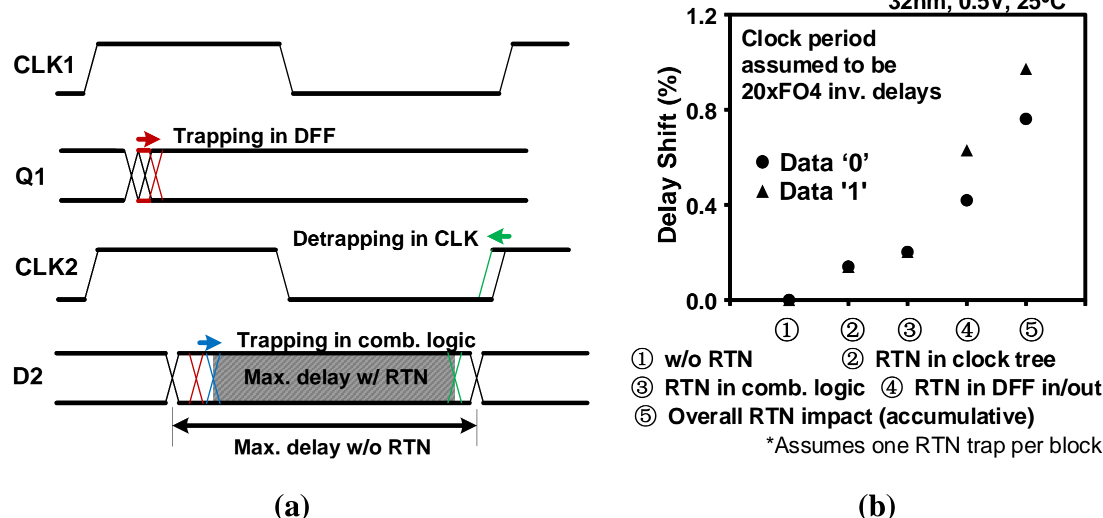


Figure 2.27: RTN impact on logic path delay assuming a clock period of 20 FO4 inverter delays and one RTN trap in each block (i.e. logic path, clock tree, input DFF, and output DFF).

The following two factors have been incorporated for estimating RTN induced timing errors of a large circuit: (1) the frequency shift magnitude of an individual trap, and (2) the spatial distribution of traps. The probability of RTN induced timing errors for a given timing guard band x can be expressed as:

$$\Pr(\text{Timing error} \mid \text{guard band} = \text{x})$$
$$= \frac{1}{2}\left[1 - \sum_{i,j,k} \Pr(N_{clk} = i)\cdot\Pr(N_{data} = j)\cdot\Pr(N_{DFF} = k)\right]$$
$$\forall i,j,k : \Delta t_{skew}(N_{clk} = i) + \Delta t_{data}(N_{data} = j) + \Delta t_{DFF}(N_{DFF} = k) \le x$$

Here, Nclk, Ndata and NDFF represent the number of traps in the clock tree, combinational logic, and DFF, respectively. To prevent RTN induced timing errors, it is required that the guard band x should be greater than the total delay shift (Δt_{skew}+ Δt_{data}+

Δt_{DFF}) of the critical path. Here, we assume the probability of a trap being present in a transistor is independent and identically distributed, and follows the spatial probability measured from the 32nm test chip. The magnitude of RTN induced Vth shift can be modeled using a log-normal distribution [19]. However, for simplicity, Eq. (1) assumes that all RTN traps have the same Vth shift that is equal to the measured average value. We also assume that at any given moment, half the traps are in capture state and half are in emission state. A separate in-depth study will be needed to fully capture Vth shift variation and spatial distribution effects. Based on Eq. (1) and the above-mentioned simplifications, the estimated probability of timing errors for circuit before and after BTI stressed is shown in Figure 2.28. For a fresh circuit, the probability of timing errors due to RTN will be reduced to less than 10^{-12} with a guard band of 1.2 FO4 delay. The number of traps increases with longer BTI stress and therefore the number of RTN induced timing errors increases accordingly.

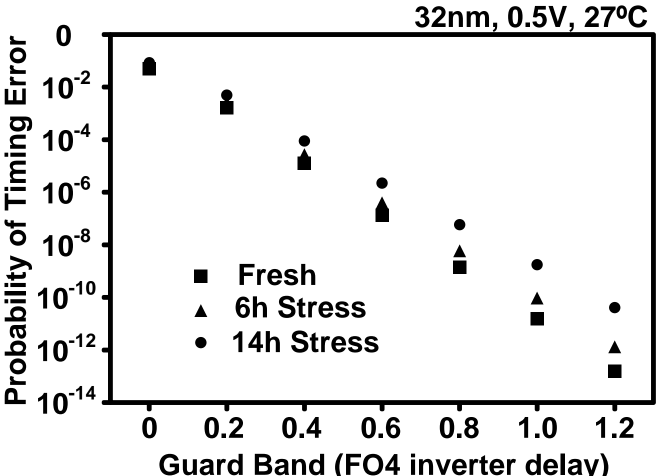


Figure 2.28: Probability of setup time violation versus timing guard band.

2.4.4. RTN Impact on SRAM Stability and Timing

A 6T SRAM cell is shown in Figure 2.29. RTN either improves or worsens the read margin depending on the trap location inside the SRAM cell. The read margin is determined primarily by the relative strength between the pull down NMOS transistor (PD) and the pass gate (PG). In the worst case, the diagonal PD and PU transistor pair becomes weaker while PG becomes stronger due to multiple RTN traps. Figure 2.29(b) shows the RTN impact on SRAM write SNM. Read SNM and write SNM move in opposite directions for the same RTN trap. The worst case for write happens when RTN trapping occurs in PG. For better illustration, we ran Monte Carlo simulations on SRAM read and read SNM under a 0.6V supply voltage assuming random trap locations. As shown in Figure 2.30, with RTN, the 99.9 percentile read SNM and write SNM are reduced by 12% and 3.9%, respectively. Next, we analyze how the SRAM read path delay, namely the CLK to DOUT delay, is affected by RTN. Figure 2.31 shows the schematic and timing diagram of a 128Kbit SRAM subarray used in this discussion. Firstly, when RTN traps are present in the column decoder, the CLK to WL delay increases causing the read delay to increase. Similarly, read delay may increase due to RTN traps in the sense amplifier enable signal (SAE) generation path. The worst case read delay occurs when the trap is located in the PG transistors because SRAM read speed is determined by the read current. Finally, RTN in the sense amplifier may degrade the resolving time. Figure 2.32 shows a typical latch based sense amplifier. When BL is discharged, traps on transistors 2 and 3 increase the SAE to DOUT delay while traps on transistors 4 and 5 decrease the delay. RTN has a stronger impact on sense amplifier delay for smaller bitline voltage differences.

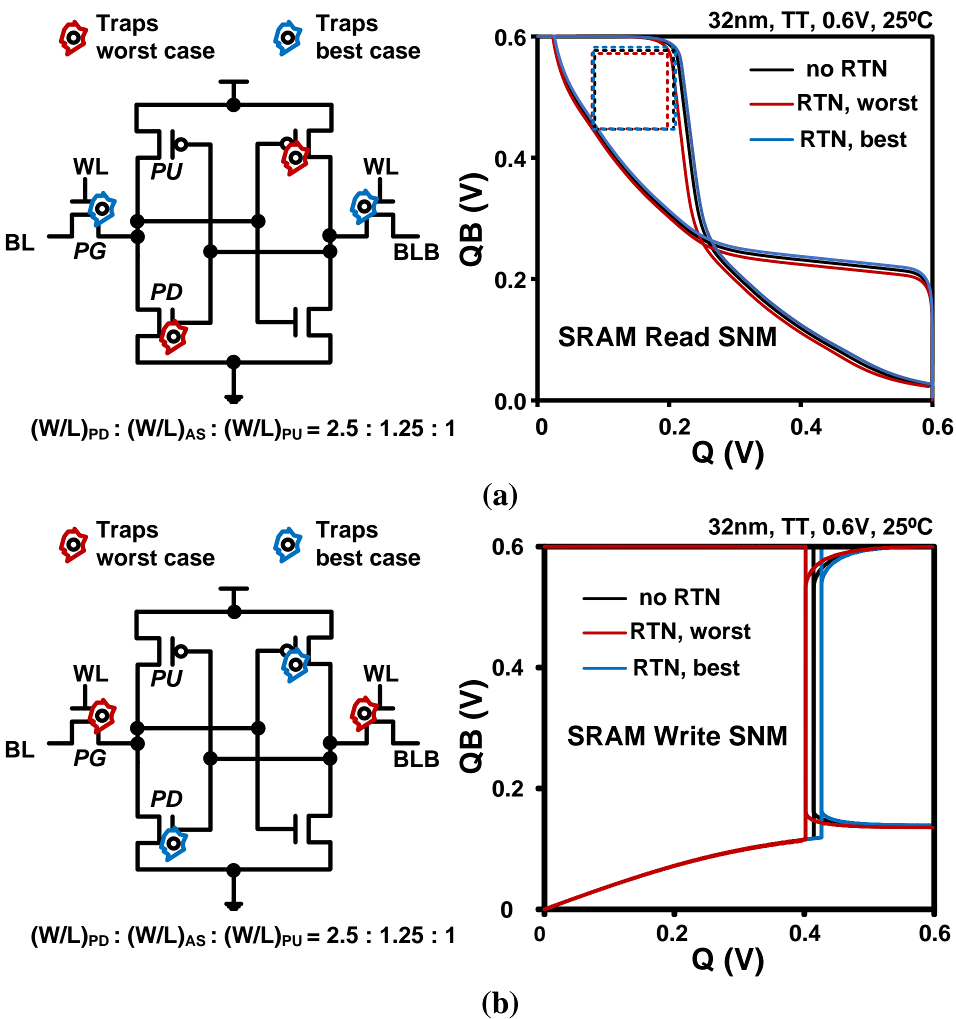


Figure 2.29: RTN impact on SRAM (a) read SNM and (b) write SNM.

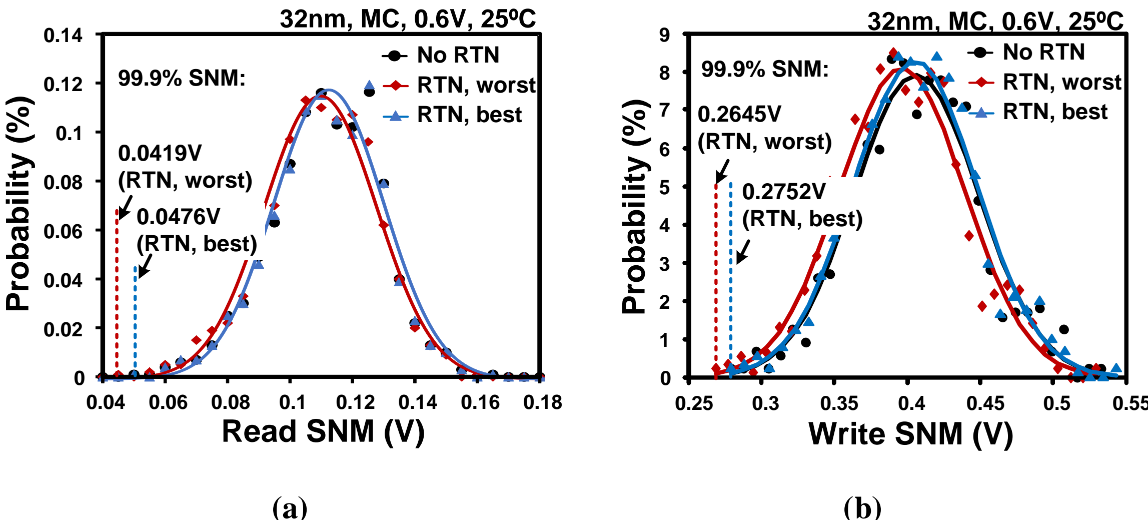


Figure 2.30: Monte Carlo simulations of SRAM (a) read SNM and (b) write SNM, with and without RTN.

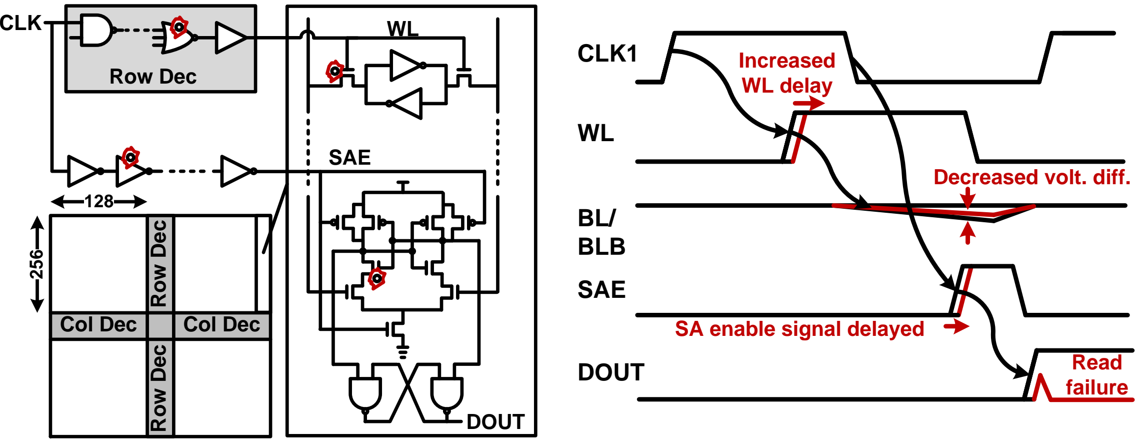


Figure 2.31: RTN impact on SRAM read timing.

To capture the above discussion, we simulated the CLK to DOUT delay of a realistic 128Kbit SRAM sub-array in 32nm technology assuming traps in different locations. As shown in Figure 2.33, traps located in the column decoder show negligible impact on the overall read delay while traps in the sense amplifier have a greater impact. Assuming one

trap each in the column decoder, SRAM access transistor, sense amplifier, and SR latch, the read delay increases by 0.96% at 0.6V and by 9.3% at 0.55V.

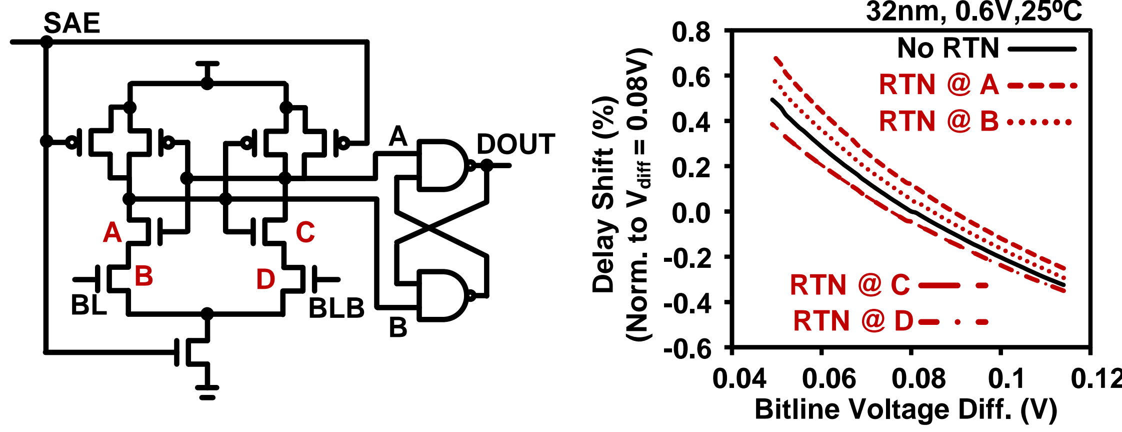


Figure 2.32: RTN impact on sense amplifier resolving time.

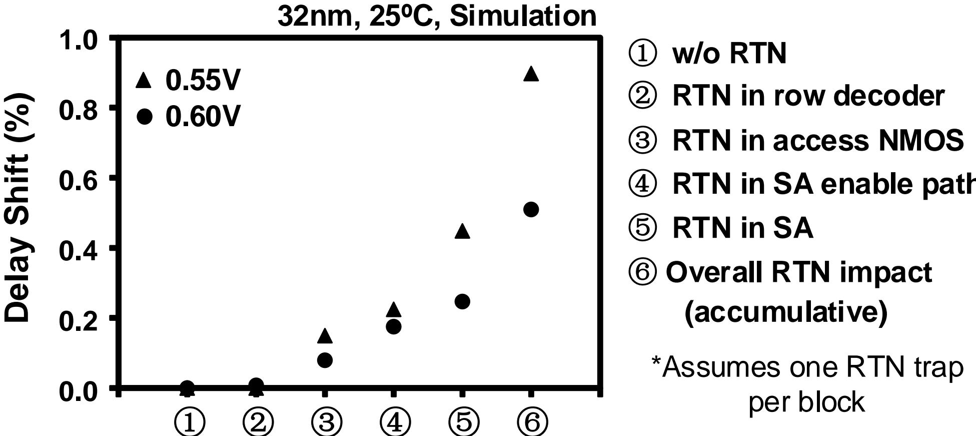


Figure 2.33: RTN impact on SRAM read path delay.

2.5. Conclusion

The impact of random telegraph noise on ring oscillator (ROSC) frequency was measured for the first time using an on-chip beat frequency detection system. The proposed differential sensing scheme achieves high measurement resolution and short

measurement time was first demonstrated in a 65nm LP process. Experimental data from the test chip displays both single trap and multi-trap RTN behavior. The voltage dependencies of the frequency shift and capture/emission times were measured and analyzed. For a more accurate characterization of the impact of random telegraph noise (RTN) on logic timing margin under sub-0.5V supply voltages, a novel method utilized dual ring oscillator array test structure was fabricated in a 32nm HKMG technology. The new test structure improves the frequency measurement resolutions of the tested-and-proven beat frequency detection (BFD) technique by pairing a ROSC from one array with a ROSC from a second array having a similar frequency. It enables fully-automated collection of RTN statistics with high measurement accuracy at supply voltages as low as 0.45V. The magnitude and occurrences of RTN induced ROSC frequency shift were measured across different supply voltages, temperatures, and voltage stress conditions. Based on the measured frequency shift data, we estimated the RTN impact on logic timing margins and SRAM performance.

Chapter 3. Compact High-Sensitivity Radiation Sensor Array

3.1. Introduction

Recent studies have shown that Soft Error Rate (SER) per memory bit is steadily decreasing with technology scaling due to the commensurate reduction in junction area [22]-[26]. Furthermore, SER in FinFET technologies is expected to be 5-10x less than that in planar technologies due to the smaller geometrical contact area between the diffusion and substrate regions [25]. As a result, collecting statistically significant amount of SER data in advanced technologies such as FinFET has become a challenging task, requiring massive number of test circuits exposed to a radiation source for long periods of time. Extracting model parameters from limited number of SEU and SET events results in model inaccuracies. Extrapolating SER based on data collected at low supply voltages can lead to unrealistic SER projections.

To overcome these limitations, we present a compact 2 Transistor (2T) radiation sensor array with a high sensitivity to radiation strikes. The proposed sensor circuit has a critical charge (Q_{crit}) that is more than 10 times smaller compared to a 6T SRAM. This is achieved by eliminating the restore current and minimizing the node capacitance. Alpha particle experiments show that the proposed 2T sensor can detect many strikes that would have otherwise gone undetected in a standard 6T SRAM test structure. The higher

sensitivity to particle strikes and dense area makes the proposed 2T sensor an effective tool for performing detailed radiation studies.

3.2. 2T Sensor Array

In terms of CMOS transistor, soft errors typically occur when the particle strikes the device diffusion area [26][27]. The most sensitive area is generally the drain of an 'off' state transistor. As shown in Figure 3.1(a) and (b), the susceptible nodes are the drains of the alternating PMOS and NMOS along an inverter chain, or the storage nodes in a 6T SRAM cell. SET occurs in combinational logic when the particle induced a sufficiently large current or voltage pulse, e.g. greater than an inverter trip point, to propagate through the logic chain. SET may gradually diminish while propagating along the logic chain, however if the pulse is captured by a latch or DFF then an error occurs. SEU, on the other hand, occurs in a single memory cell which directly lead to a data flip. Not all particle strikes will result in an SEU or SET, it largely depends on the particles energy, strike location and angles, circuit topology and so on. Two parameters are widely used to characterize the single event effects [1].

1) Collected charge (Q_{coll}) is defined as the total charge created by a radiation event at the vicinity of the junction. The amount of Q_{coll} largely depends on the characteristic of the particles and its interaction with the materials. Typical value of Q_{coll} ranges from less than 1fC to hundreds of fC.

2) Critical charge (Q_{crit}) is defined as the amount of charges that is required to trigger a change of the data state which primarily relies on intrinsic circuit parameters such as node capacitance, supply voltage, and restore current.

A soft error occurs when Q_{coll} is greater than the Q_{crit} of a certain circuit node. For an isolated nodes, the critical charge can be approximated as $Q_{crit} = Cs \times V_{DD}$ where Cs is the storage node capacitance and V_{DD} is the supply voltage. In actual circuit implementation, most nodes are interconnected resulting in a charge sharing that greatly affects the effective Q_{crit}. In particular, restore current prevents the cell voltage from being disturbed by replenishing the charge loss. This can be seen in Figure 3.1(a), where the SET pulse shape is determined by the amount of charge generated by a strike as well as the magnitude of the restore current. Similarly, in the 6T SRAM cell shown in Fig. 1(b), the restore current along with the cross-coupled feedback loop reinforces the data and thereby increases the Q_{crit}. Based on this observation, we implemented a two-transistor (2T) structure shown in Figure 3.1(c) to increase the chances of collecting the single events data. The basic idea is to remove the current restore path thereby reduce the effective Q_{crit}. As compared to an inverter chain or an SRAM cell, the 2T sensor has a significantly lower Q_{crit} since (1) the restore current path is removed and (2) the node capacitance is minimized. The 2T sensor cell consists of a PMOS write transistor and an NMOS read transistor. During irradiation mode, the sensitive node voltage V_{cell} is initialized and then left floating by turning off the write PMOS transistor. This way, a particle strike can easily disrupt V_{cell}.

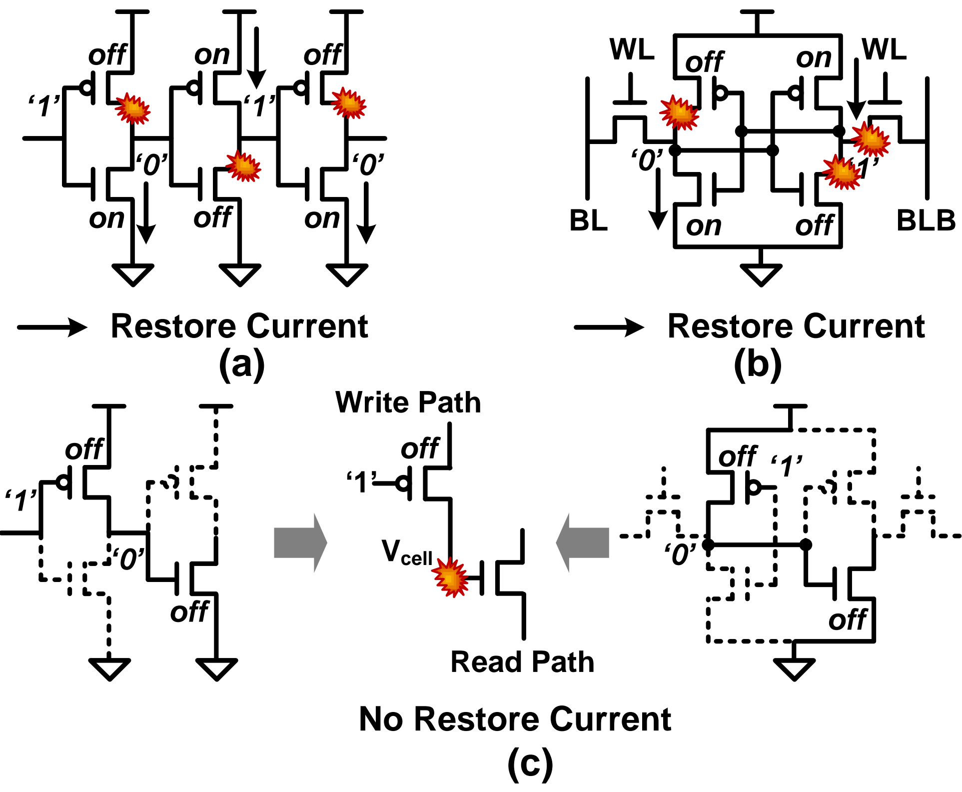


Figure 3.1: Particle strike induced soft errors are rare in (a) logic gates and (b) SRAM cells because of the strong restore current. (c) The proposed 2T sensor can detect SER strikes with a higher sensitivity by removing the restore current and minimizing the node capacitance.

The circuit diagram of a 32x32 sensor array is shown in Figure 3.2. Column decoders are used for decoupled write and read operations. Output voltage of the selected 2T cell is compared with an external reference voltage V_{REF} to determine the cell status. Similar to a DRAM cell, leakage currents surrounding the sensitive node causes V_{cell} to gradually discharge or charge depending on the initial voltage written to the cell. In order to separate retention time induced errors from radiation induced errors, the cell data must be

read out before V_{cell} rises above or falls below the threshold. Also V_{REF} is carefully chosen to make sure an equivalence of the data '1' and '0' margins.

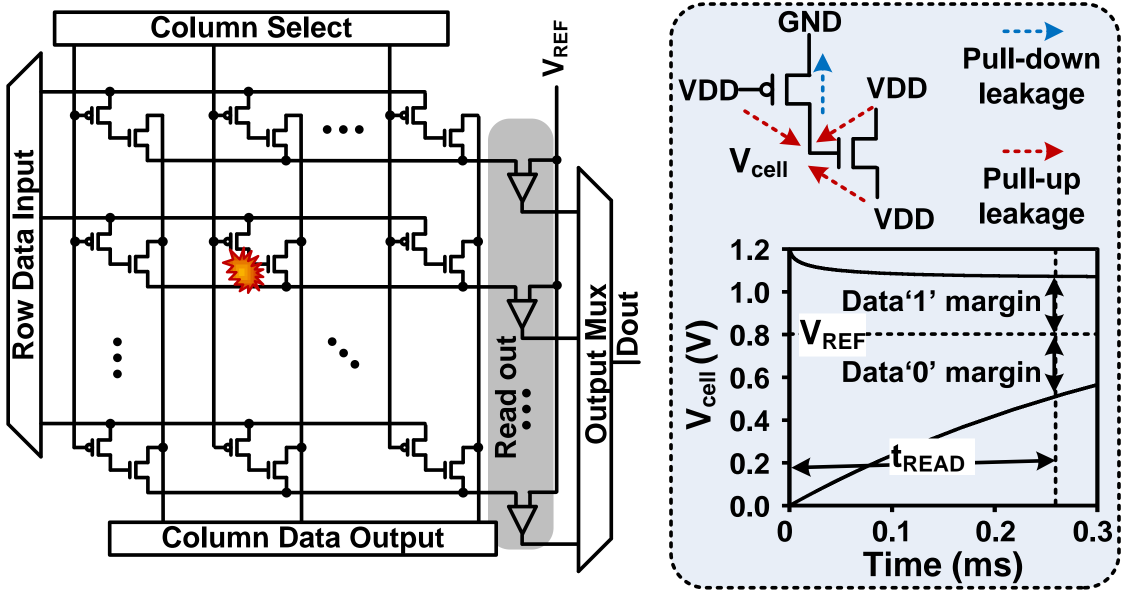


Figure 3.2: Proposed 2T sensor array for detecting SEU with high sensitivity. Voltage stored inside the cell (V_{cell}) varies with time and leakage current.

3.3. Alpha Particle Experiment

The overall experiment flow is shown in Figure 3.3. Although it is known that the storage node is more sensitive to radiation effect when the junction is reverse-biased (written with a data '0'), for test completeness, a checkerboard pattern is written to the sensor array to investigate all possible failures. The tests are conducted with two phases. During pre-calibration phase, the checkerboard pattern is written to the sensor array and subsequently, the cell data is read out after t_{READ} to check for any retention errors. t_{READ} is gradually decreased until retention errors are no longer present. This ensures that no

intrinsic failures occur within a t_{READ} period. During irradiation phase, the array data is initialized and read out repeatedly using the specific t_{READ} interval found in the pre-calibration phase. SER is calculated by comparing the array pattern containing errors with the initial checkerboard pattern.

The 2T sensor array chip was fabricated in a 1.2V 65nm LP bulk process. The die microphotograph and chip specifications are shown in Figure 3.4. Although the total array size is 64Kbit, only 1Kbit cells are utilized as a compromise between t_{READ} and IO speed. A small number of reference SRAM cells were implemented in the same chip for comparison purposes.

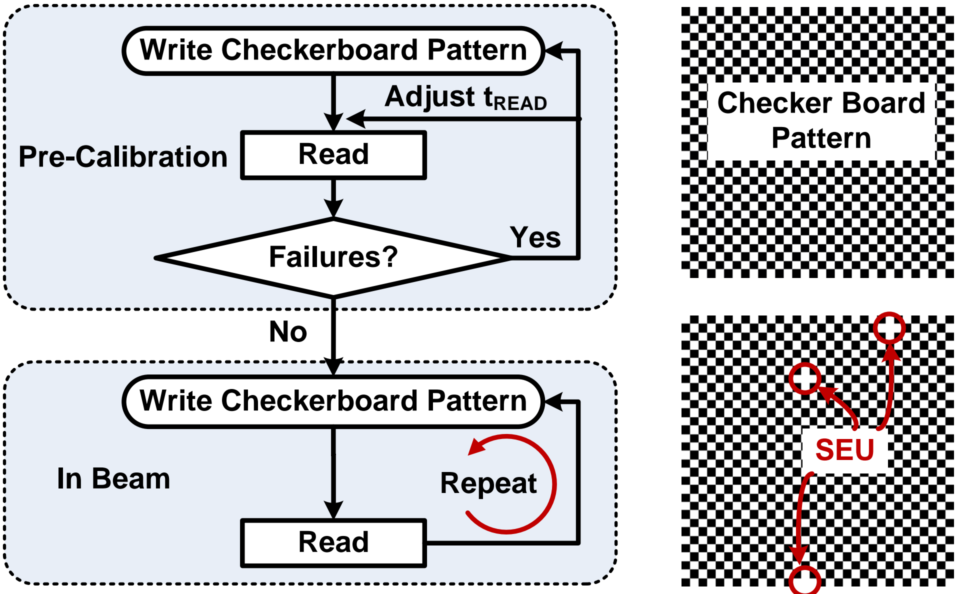
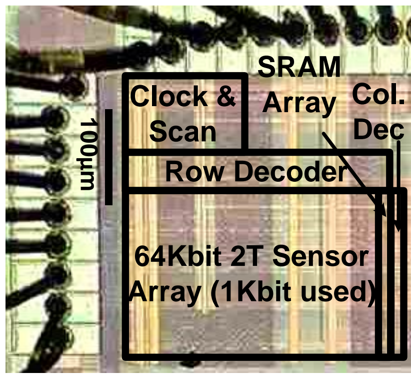


Figure 3.3: Overall test sequence for the 2T sensor array. The array pattern is compared with the initial checkerboard pattern to identify particle strike induced SEUs.



	2T Sensor	SRAM (reference)
Cell area (μm^2)	1.00×1.37	1.48×1.61
Array size	64 Kbit	128 bit
Sensitive area (μm^2)	0.245×0.15	3·(0.18×0.15)
Supply Voltage (V)	1.2~1.5	0.75~1.25
Retention Time (ms)	0.176~0.264	N/A

Figure 3.4: 65nm test chip including a 2T sensor array and SRAM cells.

Although the 2T sensor provides a higher sensitivity to single event effects, characterizing the SER in a natural terrestrial environment is timing consuming. It is reported that, at sea-level, SER of a 24MB cache (SRAM) in the Intel Xeon Processor is only 0.2 to 2 errors/year [21]. An accelerated test is therefore necessary to collect statistically meaningful data within a limited time. An alpha particle accelerator generates alpha particles that have a similar energy but significantly higher particle densities as compared to the real terrestrial environment. Our irradiation testing was performed with an alpha particle accelerator housed in the Characterization Facility at the University of Minnesota as shown in Figure 3.5. Alpha particles were generated from the MAS 1700 pelletron tandem ion accelerator (5SDH). Test chips were placed in an endstation chamber with a controllable angle rotation. The particle energy used in our experiment was 3.8 MeV. Another commonly used term is Linear Energy Transfer (LET) which describes the energy deposited per unit length along the ion penetration track. The LET used in this experimental is estimated to be 0.736MeV·cm^2/mg using the NIST ASTAR calculator [28].

Figure 3.5: Ion beam facility with particle accelerator used for radiation testing (Source: University of Minnesota Characterization Facility).

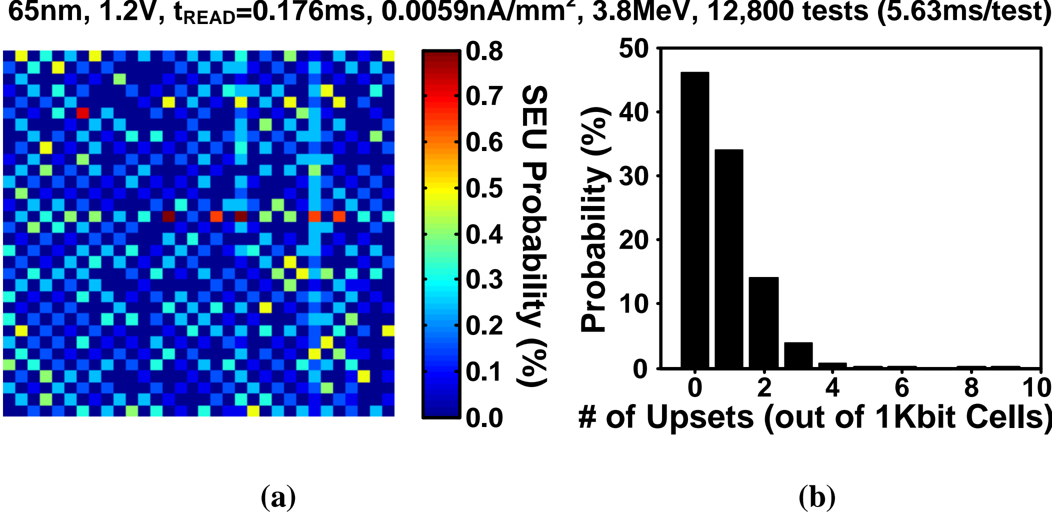


Figure 3.6: (a) Upset probability bit map and (b) the number of upsets per 1Kbit array.

SER map for a 1Kbit array measured from 12,800 consecutive read out cycles is shown in Figure 3.6(a). The per-cell upset probability ranges from 0% to 0.8% under a nominal supply voltage of 1.2V. The ‘0’-to-‘1’ SER is significantly higher than ‘1’-to-‘0’

flips which verifies the susceptibility of reverse-biased p-n junction to radiation events. Figure 3.6(b) shows that the probability of cell with no upset is less than 50% indicating that in most locations the 0'-to-'1' upset happened at least once out of 12,800 tests.

SER is generally characterized based on cross-section, which quantifies the likelihood of a single event effect as the effective area. SEU cross-section is calculated as follow [23]:

$$\text{Cross} - \text{Section} = \frac{\text{total}\,\#\,\text{of errors}}{\text{array size} \times \text{fluence}}$$

Fluence here is defined as the total number of particles received per unit area. The fluence in this experiment is obtained by multiplying the flux by the test time. Figure 3.7 shows that the cross-section increases from 1.86E-11 cm^2/cell to 1.22E-9 cm^2/cell as the supply voltage is lowered from 1.5V to 1.2V with a t_{READ} of 0.264ms. Measurement result with t_{READ} of 0.176ms shows a similar trend. Figure 3.8(a) illustrates the relationship between sensor cell cross-section and t_{READ}. The cross-section increases from 5.32E-11 cm^2/cell to 1.25E-9 cm^2/cell as the read interval t_{READ} is increased from 0.176ms to 0.264ms. As shown in Figure 3.8(b), with a larger read interval, the data '0' margin decrease due to leakage current which continuously pull up the node voltage. Also, the extended exposure time further degrades the data margin resulting in an increased cross-section. To prove that the 2T cell is feasible for a radiation sensor application, we conducted the experiments under three different fluxes. The measured data in Figure 3.9 confirms that cross-sections of the proposed 2T structure is proportional to radiation flux. For a comparison, we measured the SER of a 128 bit 6T

SRAM array from the same chip. Due to rare occurrence of SRAM SEU at nominal supply voltage with the given experimental parameters, the SER is measured with a supply voltage from 0.75V to 1.15V. As shown in Figure 3.10, the measured cross-section of the 2T sensor cell is 1.22E-9 cm^2/cell at 1.2V while a 6T SRAM has a cross-section of 1.04E-11 cm^2/cell at 1.15V for the same beam energy and radiation flux. This corresponds to a 117x higher cross-section area per cell for the 2T sensor as compared to a 6T SRAM.

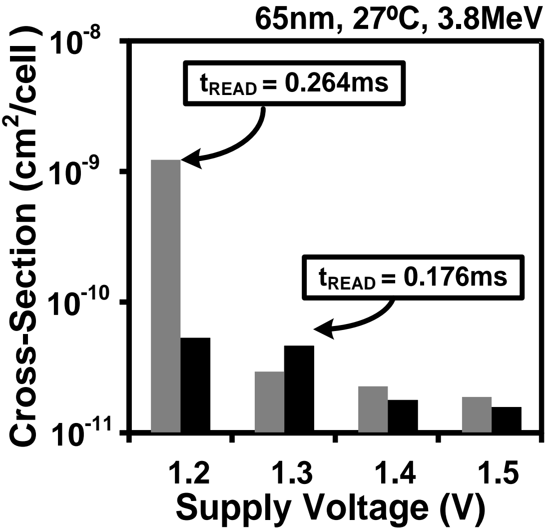
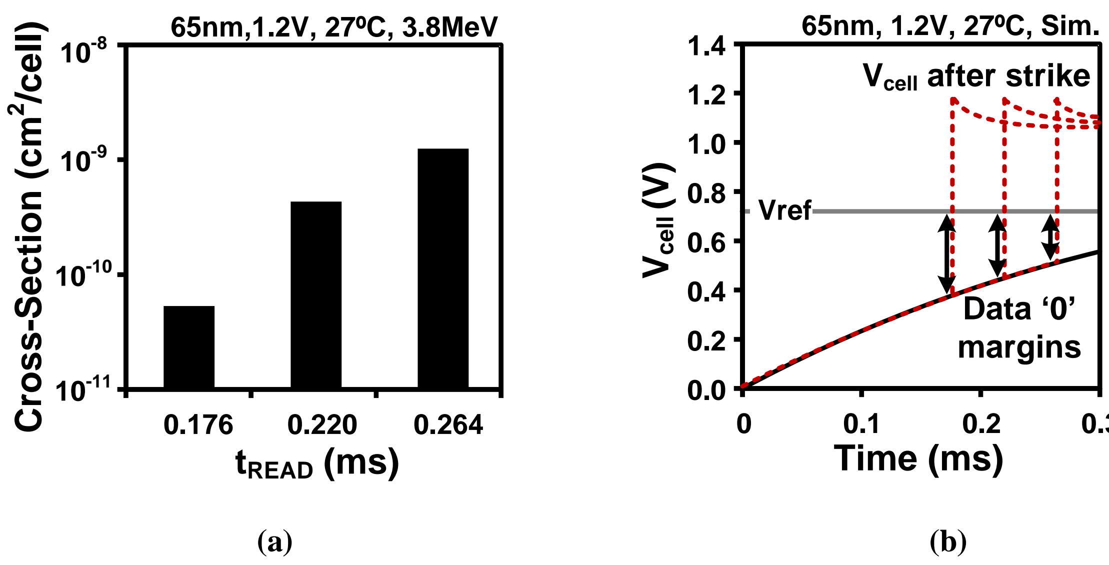


Figure 3.7: Measured cross-section increases with lower supply voltage.



(a) (b)

Figure 3.8: (a) Measured cross-section increases with longer t_{READ} due to the reduced data read margin as shown in (b).

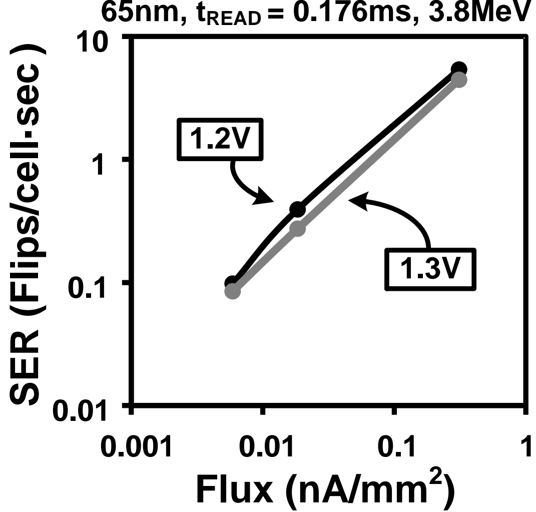


Figure 3.9: Measured SER is proportional to alpha particle flux.

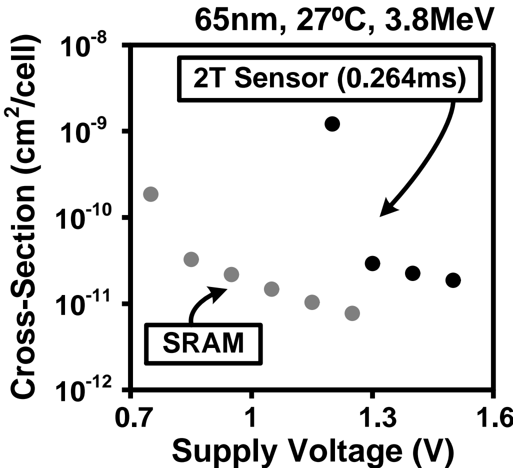


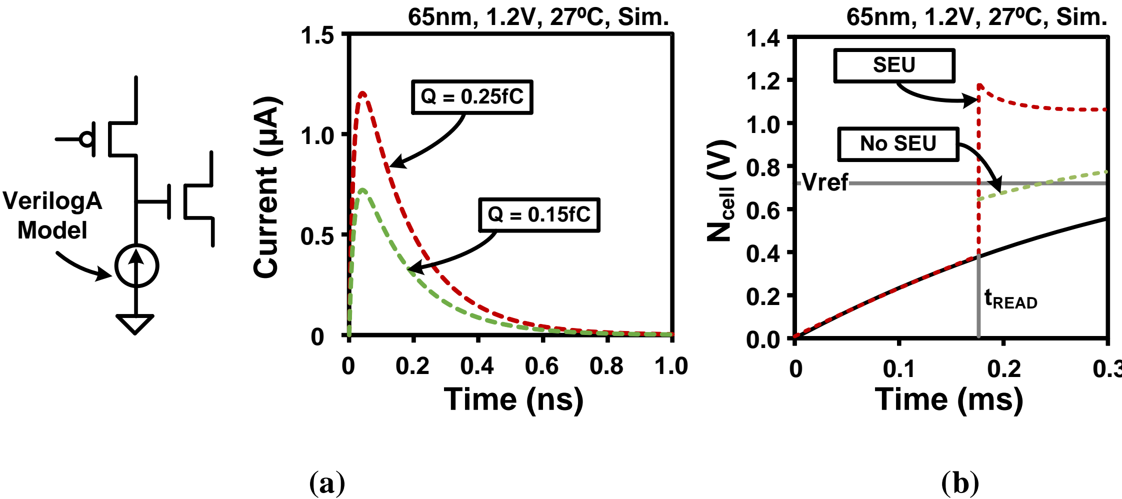
Figure 3.10: Measured cross-section of 2T cell and SRAM at different supply voltages.

3.4. Single Event Upset Simulation

In addition to the irradiation test, the SEU simulation was performed on the 2T structure, 6T SRAM cell and inverter chains for a better understanding of the sensor sensitivity to single event effects [29]-[33]. One of the most commonly used model to characterize Q_{crit} is the double exponential model expressed as follow [33]-[36]:

$$I_{rad}(t) = \frac{Q_{total}}{\tau_f - \tau_r}\left[\exp\left(-\frac{t}{\tau_f}\right) - \exp\left(-\frac{t}{\tau_r}\right)\right]$$

Here, Q_{total} is the amount of charge collected by the diffusion during a radiation event, τ_r and τ_f corresponds to the pulse rise and fall time constant. $I_{rad}(t)$ defines the shape of the injected current pulse which is a dependent of time. In this equation, t = 0 is defined as the moment when the particle strikes the material.



(a) **(b)**

Figure 3.11: (a) Simulated current pulse using double exponential model and (b) 2T sensor transient response with Q_{total} equal to 0.25fC and 0.15fC respectively.

To simulate the radiation effect, a current source is firstly modeled with VerilogA based on the double exponential equation. This current source then is instantiated in Spectre and connected to the sensitive nodes to perform the SEU simulation. Figure 3.11(a) shows the simulated current pulses with 16ps τ_r and 160ps τ_f. Current pulse corresponding to a 0.25fC Q_{total} leads to a '0'- to-'1' upset whereas the pulse with 0.15 Q_{total} shows no upset. By sweeping Q_{total}, the Q_{crit} can be found which is defined as the minimum Q_{total} to flip a cell.

The aforementioned 2T sensor SER dependence on t_{READ} is also captured through the simulation. As shown in Figure 3.12, Q_{crit} decreases monotonically with a long t_{READ}. Figure 3.13 shows the simulated Q_{crit} for an inverter chain, 6T SRAM and 2T sensor cell. Q_{crit} for the inverter chain is defined as the minimum charge that allows the SET voltage pulse to propagate through two inverter stages. It can be seen that Q_{crit} of the proposed 2T

sensor cell is 17x-60x and lower than that of an inverter chain and SRAM cell for supply voltages ranging from 0.75V to 1.5V.

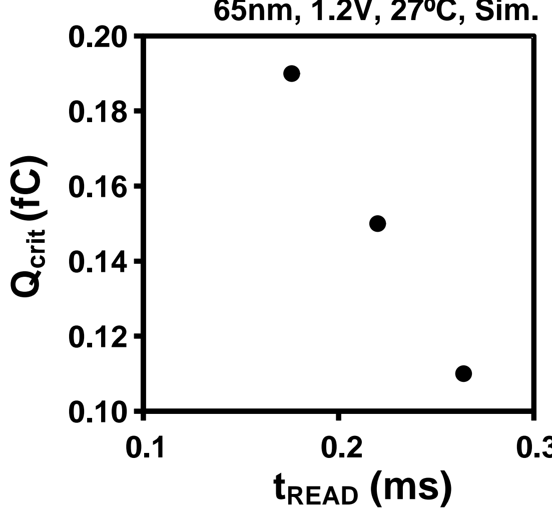


Figure 3.12: Q_{crit} decrease with a longer t_{READ}.

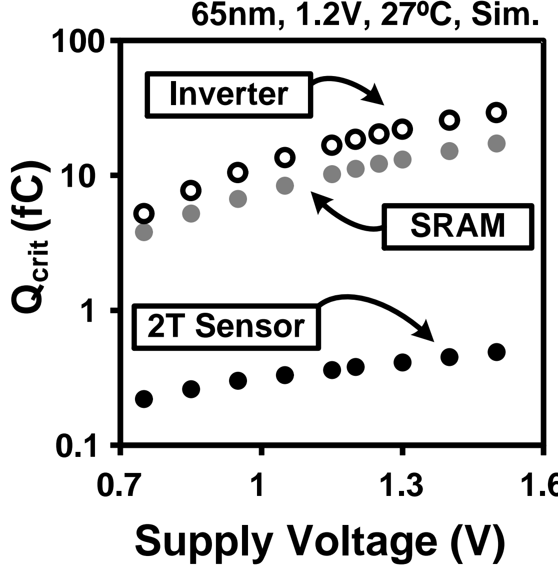


Figure 3.13: Simulated Qcrit for inverter, 6T SRAM, and 2T sensor.

3.5. Conclusion

Ionizing particles, such as alpha particle, can induce soft errors such as Single Event Upset (SEU) in memory cells and Single Event Transient (SET) in logic circuits. As CMOS technology node scales down, the critical charge (Q_{crit}) also decreases due to the smaller circuit parasitic and lower supply voltage. However, the per bit Soft Error Rate (SER) declines with the scaling trend since the shrink of feature size and reduction of corresponding sensitive area has been dominated. In this work, we propose an on-chip compact 2T sensor structure for the efficient statistical collection of the single event data. Test chip implemented on a 65nm bulk process demonstrates a 117X higher measurement sensitivities as compared to 6T SRAM cell under an accelerated alpha particle irradiation. Simulation results shows that Qcrit of the proposed 2T sensor cell is 17x-60x smaller than that of an inverter chain and SRAM cell.

Chapter 4. True Random Number Generator

4.1. Introduction

True Random Number Generators (TRNGs) are used for encrypting purposes in systems or networks that require a high level of security. On-chip TRNGs typically harvest randomness from a circuit that converts transistor level noise such as thermal noise, flicker noise, or RTN [37]-[40] into a voltage or delay signal. For example, a popular type of TRNG utilizes an inverter pair that is initialized to a metastable state to amplify the thermal noise [38][39] as shown in Figure 4.1. In this design, the two output nodes of the inverter pairs are initialized to the same voltage. The switches are then turned off causing the outputs to resolve to either '1' or '0' depending on the random device noise. The main drawback of this circuit is that the meta-stable point is extremely sensitive to the inverter pair's voltage offset. More specifically, it has been shown that randomness can only be guaranteed when the voltage offset is below $\pm 0.24\sigma_{noise}$. Here, σ_{noise} is the standard deviation of the noise [39]. Therefore, in order to guarantee the random number quality in a long-term operation, a continuous calibration loop including a Finite State Machine (FSM), shift registers and tuning circuits is required. Another popular type of TRNG is the delay based TRNGs which employs oscillator circuits. One conventional oscillator based TRNG implementation is the one shown in Figure 4.2 [41]. A D-Flip-Flop (DFF) samples a free-running ROSC at a frequency much lower than the ROSC intrinsic frequency. The ROSC jitter accumulates over time and will become

greater than one ROSC cycle after a sufficient wait period. The DFF therefore samples an uncertain phase and outputs a random bit stream. The actual implementation in IBM's POWER+7 processor consists of 64 parallel ROSCs operating at different frequencies to prevent the ROSCs from locking into each other. An independent low frequency clock samples the ROSC outputs at a rate which is a non-integer multiple of all 64 ROSCs. Because each ROSC operates at different frequencies, the accumulated jitter will be different and independent from each other. Figure 4.2(b) shows the probability of two consecutive bits being different as a function of the sampling period. For an ideal TRNG, this probability should be very close to 50%. The probability traces measured from the parallel ROSCs show that a sampling period greater than ~200 ROSC cycles is required for the bit to be considered random. However, the 200 ROSC cycles criteria only holds when circuit is operated at nominal or higher supply voltage, attacks such as lowering the supply voltage would reduce the output randomness.

In this work, we demonstrate a fully-digital TRNG circuit utilizing the beat frequency detection technique that employs the subtle frequency difference between two identical free-running ROSCs. Instead of directly sampling the ROSC output, the entropy is extracted from the BFD output count value. As compared to the 64-parallel ROSC IBM TRNG, the BFD-TRNG has approximately 3 times power advantage and 2 times area efficiency given the same generation speed [42].

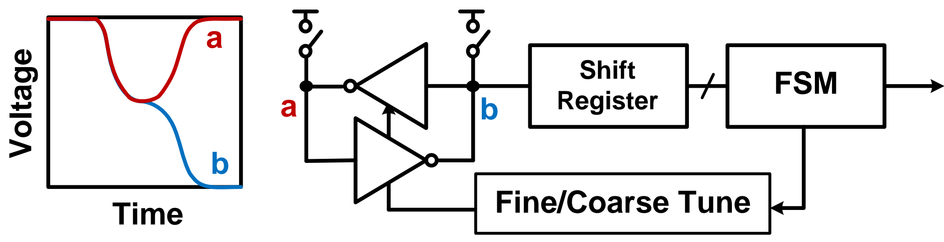
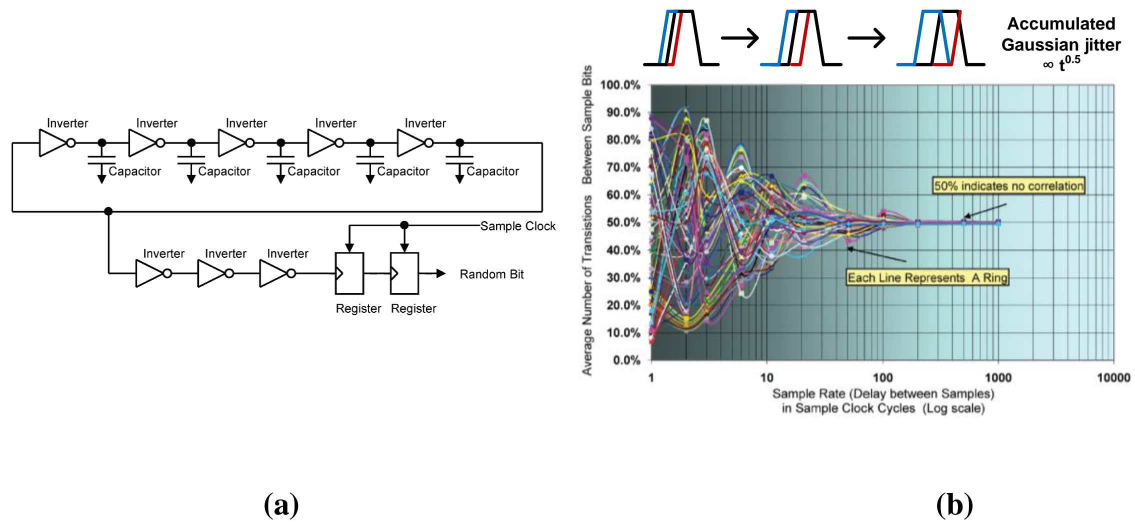


Figure 4.1: A conventional meta-stability based TRNG [39]



(a) **(b)**

Figure 4.2: (a) ROSC based TRNG employed in IBM POWER7+. (b) Probability of two consecutive bits being different as a function of sampling period (=wait time) for IBM's ring oscillator TRNG. Jitter accumulation time must be >200 ROSC cycles for the sampled bit to be considered random[41].

4.2. Beat Frequency Detector based TRNG

The basic concept for capturing the frequency difference between two ROSCs is illustrated in Figure 4.3 [44]. The faster signal A passes, catches up and overtakes the slower signal B repeatedly at intervals determined by the frequency differences of the two ROSCs, namely the beat frequency or Δf. This pattern is recorded by a standard D-

flip-flop where the output of ROSC A is continuously sampled by that of ROSC B. The counter output (N in Figure 4.3) increments every ROSC period until it reaches the beat frequency interval after which the count is sampled and reset. For better illustration, let's consider an example in which the average frequency difference between the ROSC pair is 1% and the maximum frequency difference due to random jitter is 0.01%. Under this condition, the average counter output is 100 while the maximum and minimum counts are 101 and 99, respectively. In this scenario, we can take the least significant bit (LSB) of the output count as the TRNG output. Now suppose the average frequency difference is reduced to 0.5% by adjusting the frequency difference, while the random jitter remains the same at 0.01%. Then, the output count will fluctuate between 196 and 204, thereby providing up to three random bits (1st, 2nd, and 3rd LSBs) per output count and at the same time increasing the randomness of the lower bits. To adaptively change the ROSC frequency difference, we implemented the binary weighted trimming capacitors on each ROSC stage. For instance, to bring the frequencies closer, one can either enable additional capacitors on the faster ROSC or disable the capacitors on the slower ROSC. One should be noted that by making the frequencies even closer, we could generate more random bits from a larger count however at the expense of a longer sampling time. Depending on the application, one can determine the optimal BFD count value range in terms of generation speed and power efficiency.

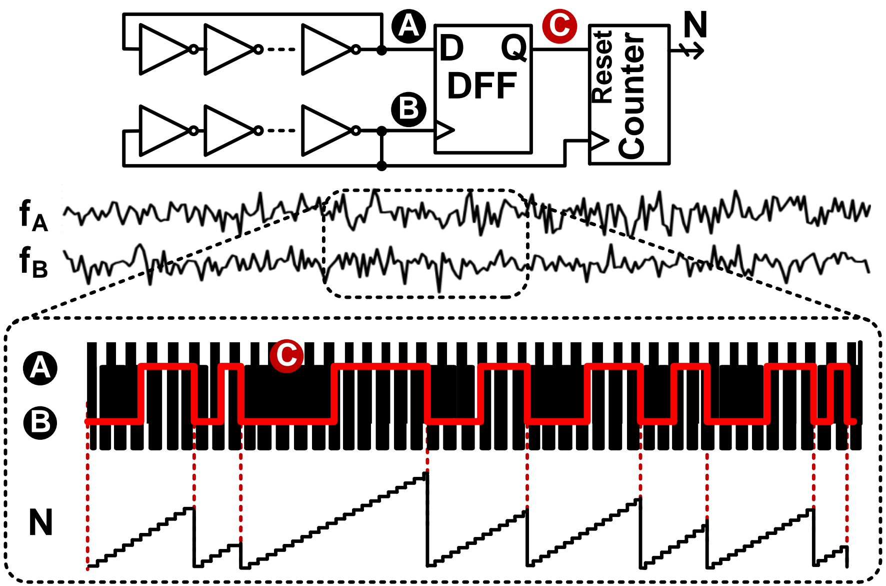


Figure 4.3: Basic principle of the proposed beat frequency based TRNG circuit.

4.2.1. Test Chip Implementation

A 65nm test chip was designed to experimentally verify the proposed TRNG circuit. A block diagram of the main circuit blocks is shown in Figure 4.4. A 6-bit trimming capacitor bank is connected to each ROSC stage for tuning the initial frequencies of the two ROSCS to ensure beat frequency counts are in the desired range. Based on our simulation, each tuning step can provide a frequency resolution of 0.4%. A 5 bit majority voter circuit is used in the beat frequency detector to prevent functional errors due to logic bubbles (e.g. lone 0 in a string of 1) which may occur when the two ROSC signal edges are about to cross each other. In order to synchronize with the post processing circuits, an external sample CLK is applied to reset the counter. This sampling clock period is set to be longer than the one beat frequency cycle to avoid count value aliasing.

To reduce unnecessary switching power consumption, a start/end control logic is implemented to automatically shut down the ROSCs once the beat frequency operation is completed and turn on the ROSCs at the rising edge of the next sample clock. Finally, the lower LSBs are post processed with Von Neumann logic and serialized to produce the final TRNG output bit stream.

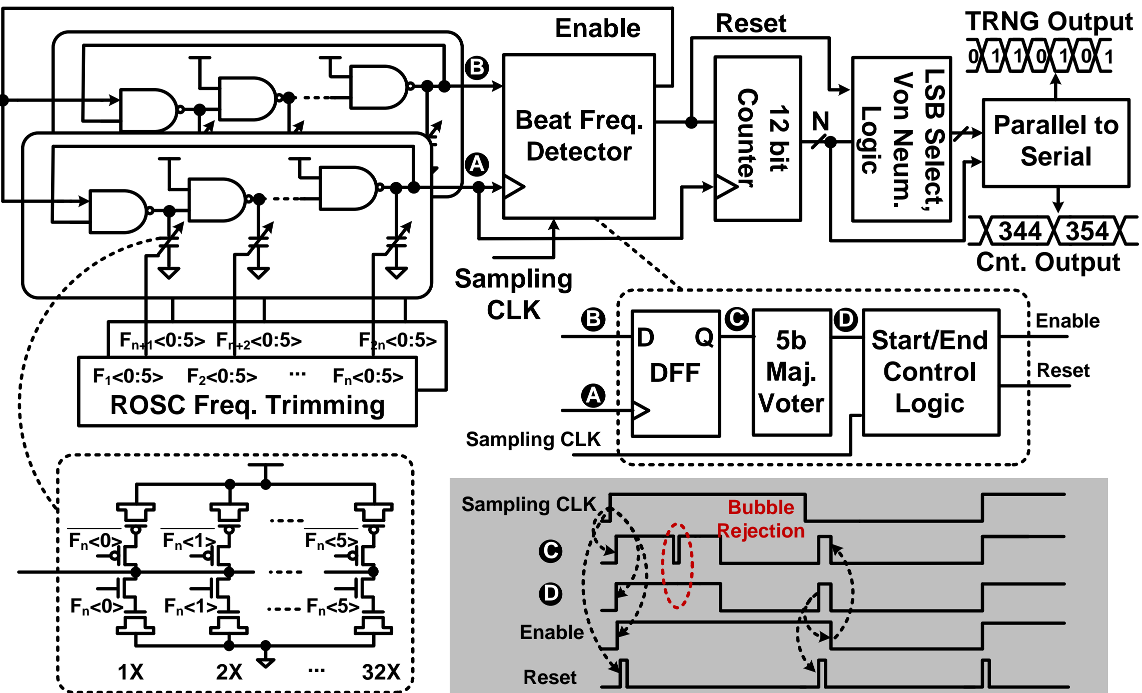


Figure 4.4: TRNG circuit with trimming caps and power saving mode.

4.2.2. Measurement Data

Figure 4.5 shows the measured beat frequency count under different trimming capacitor settings. The average count increases from 191 to 3040 as the ROSC frequencies are brought closer together. Injection locking between the two ROSCs was not observed in any of the tests owing to the good isolation between the two ROSCs. This

can be attributed to the modest amount of decoupling capacitors placed between the ROSCs as can be seen in the layout diagram in Figure 4.19. One of the most basic evaluation of randomness is the Shannon bit entropy which measures the percentage of 1's and 0's in a bit stream. Figure 4.6 shows the Shannon bit entropy of a 1M samples with an average count of 352. In our test, the Shannon bit entropy is assess on each count digit from a 1M continuously measured samples. As shown in Figure 4.6, the proportion of '1' and '0' are equal for the first 4 LSBs which implies that the first 4 LSBs has a high potential to be directly utilized as the random numbers.

However, Shannon bit entropy does not validate all possible weakness, more comprehensive statistical tests are required to assess the bit stream randomness. Standards such as National Institute of Standards and Technology's (NIST) randomness statistical test suite (STS) and DIEHARD test are developed for measuring the quality of random number generators. For example, the NIST STS performs an exhaustive analysis based on 15 different types of non-randomness that could exist in a sequence such as Shannon Entropy, FFT [46]. Figure 4.7 shows the NIST STS results obtained from 1st to 4th LSBs. For a beat frequency output with an average count of 352, the first 3 LSBs passed all 15 NIST tests (P-value $\chi 2 > 0.01$, Proportion>0.949751) without requiring any post-processing steps. Although the 4th LSB appears to be unbiased from a visual check, NIST results reveals that further post processing is necessary.

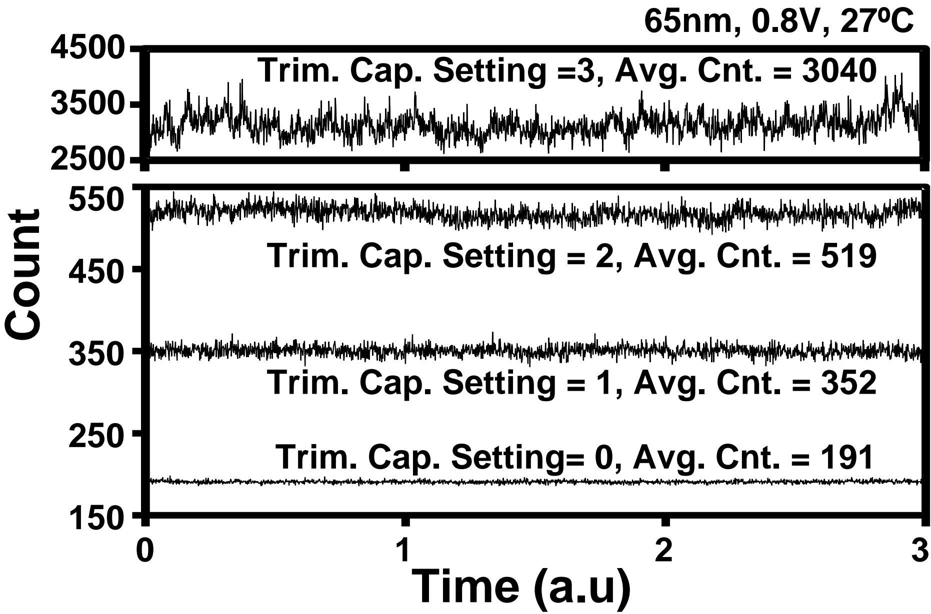


Figure 4.5: Measured beat frequency count for different trim capacitor settings.

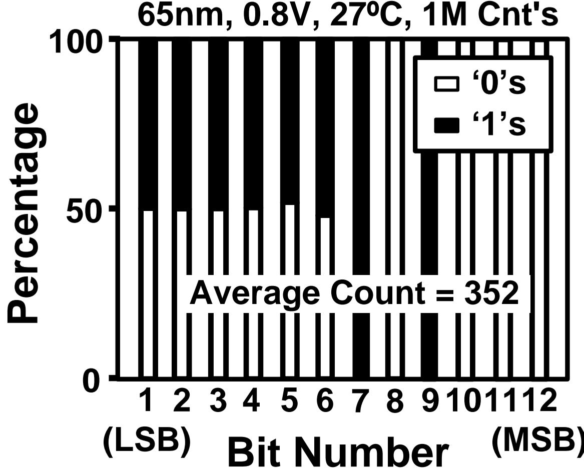


Figure 4.6: Percentage of '1's and '0's for each bit of the beat frequency count output. The lower significant bits (e.g. bits 1, 2, 3) have better randomness compared to the higher bits.

Avg. count = 352	1st LSB		2nd LSB		3rd LSB		4th LSB	
P-Val / Proportion	P-Val	Prop.	P-Val	Prop.	P-Val	Prop.	P-Val	Prop.
Frequency	0.679	0.9818	0.091	0.9818	0.514	1.0000	Fail	Fail
Block Frequency	0.130	1.0000	0.760	1.0000	0.063	0.9818	Fail	Fail
*Cumulative Sums	0.554	0.9818	0.367	0.9818	0.475	1.0000	Fail	Fail
Runs	0.596	1.0000	0.182	1.0000	0.225	0.9818	Fail	Fail
Longest Run	0.049	0.9818	0.760	0.9818	0.596	1.0000	Fail	Fail
Rank	0.305	0.9818	0.335	0.9636	0.305	1.0000	0.596	0.9636
FFT	0.305	1.0000	0.010	1.0000	0.514	0.9818	Fail	Fail
*Nonoverlapping Temp.	0.437	1.0000	0.596	1.0000	0.679	1.0000	Fail	Fail
Overlapping Template	0.249	0.9818	0.249	1.0000	0.798	0.9818	Fail	Fail
Universal	0.868	1.0000	0.475	0.9818	0.071	1.0000	Fail	Fail
Approximate Entropy	0.554	1.0000	0.043	1.0000	0.103	1.0000	Fail	Fail
*Random Excursions	0.672	1.0000	0.740	1.0000	0.500	1.0000	Fail	Fail
*Rand. Excursions Var.	0.740	1.0000	0.602	1.0000	0.637	0.9679	Fail	Fail
Serial	0.637	1.0000	0.401	1.0000	0.637	0.9818	Fail	Fail
Linear Compelxity	0.401	1.0000	0.072	0.9818	0.063	0.9818	0.163	0.9818

** Tests with 2 or more subtest, P-val and Prop shown here are the smaller or median values*
*** Concatenate 1st~3rd LSBs and 4th LSB after von Neumann correction*

Figure 4.7: NIST test verifies the randomness of 1st ~3rd LSBs.

von Neumann	Raw Data	00	01	10	11
	Corr. Data	Drop	0	1	Drop

Figure 4.8: Concatenating LSBs to generate the final TRNG output bit stream. 4th LSB can be used after von Neumann correction.

Figure 4.8 illustrates the procedure for generating the final random bit sequence from the measured beat frequency counts. The first 3 LSBs can be directly concatenated and streamed out without any post-processing. To improve the throughput, von Neumann correction is applied on 4th LSB as shown in Figure 4.8. This popular algorithm takes two consecutive bits from the original data and encodes '01' into an output '0' and '10' into an output '1'. The two other combinations (i.e. '00' and '11') are simply dropped. Using Von Neumann corrector will produce perfect correction with 0 bias but throughput is reduced to less than 25% of its original. In our implementation, a group of four 4th LSBs from the original sequence generates a single corrected output bit that is buffered and inserted into the final bit sequence. The efficiency is increased from 3 bits per count to ~3.25 bits per count using the von Neumann method with almost no additional power

consumption. NIST test results performed using measured data from the proposed TRNG are summarized in Figure 4.9. The final sequence with the insertion of the 4th LSB after von Neumann correct also passed all NIST tests.

Final Output NIST result		
P-Val / Proportion	P-Val	Prop.
Frequency	0.514	1.0000
Block Frequency	0.946	0.9818
Cumulative Sums	0.437	1.0000
Runs	0.720	1.0000
Longest Run	0.475	1.0000
Rank	0.055	0.9636
FFT	0.103	1.0000
Nonoverlapping Temp.	0.679	1.0000
Overlapping Template	0.182	0.9818
Universal	0.063	1.0000
Approximate Entropy	0.600	1.0000
Random Excursions	0.637	1.0000
Rand. Excursions Var.	0.876	1.0000
Serial	0.304	0.9818
Linear Compelxity	0.868	0.9636

Figure 4.9: Concatenated 1st~4th LSBs passes all NIST test after applying Von Neumann correction on the 4th LSB.

The initial count measured from different chips ranges from 200 to 1000 when using the same trimming capacitor setting. Through extensive testing, we found that a count range of 200 to 500 provides a reasonable trade-off between speed and bit efficiency. This count range corresponds to a ROSC frequency difference of 0.5% to 0.2%, respectively. A simple one-time calibration step shown in Figure 4.10 can be used to guarantee that the initial count is in the desired range (200 to 500) across the different

TRNG chips. At the beginning of the random number generation process, an FSM reads the count values and determines either to increase or decrease the two ROSC frequency difference. This can be readily achieved within a few beat frequency periods using minimal hardware overhead during the initial startup. After the start-up calibration, the count value can remain relatively steady for a long period of operation. Figure 4.11 verifies that the average count value has a good stability through a continuous 15 hour operation test. The measured count values are relatively consistent for a supply voltage range of 0.8V to 1.2V as shown in Figure 4.12 suggesting a wide operation range and good tolerance against environmental effects such as supply voltage and temperature variation. The slightly wider count range at 0.8V can be attributed to the larger device noise and higher delay sensitivity.

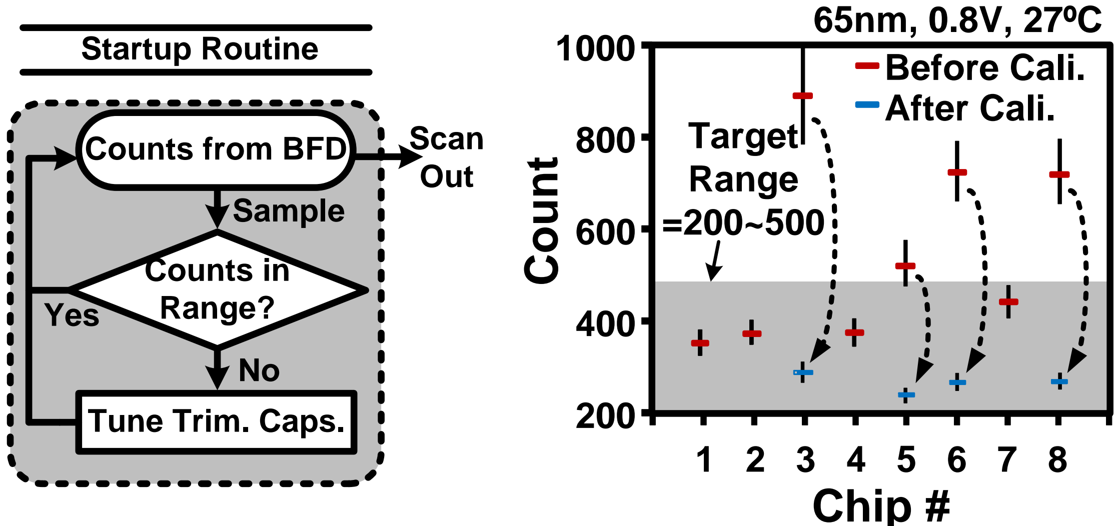


Figure 4.10: One-time calibration of average count during start up.

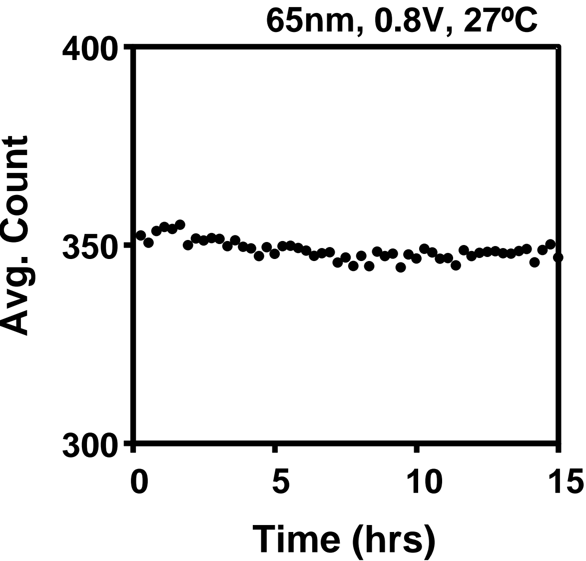


Figure 4.11: Stability under continuous operation.

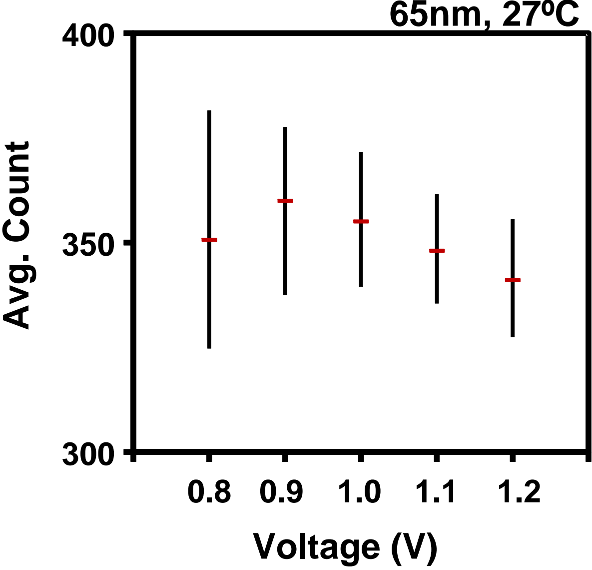


Figure 4.12: Measured count under different voltages.

4.3. Simulation and Modeling

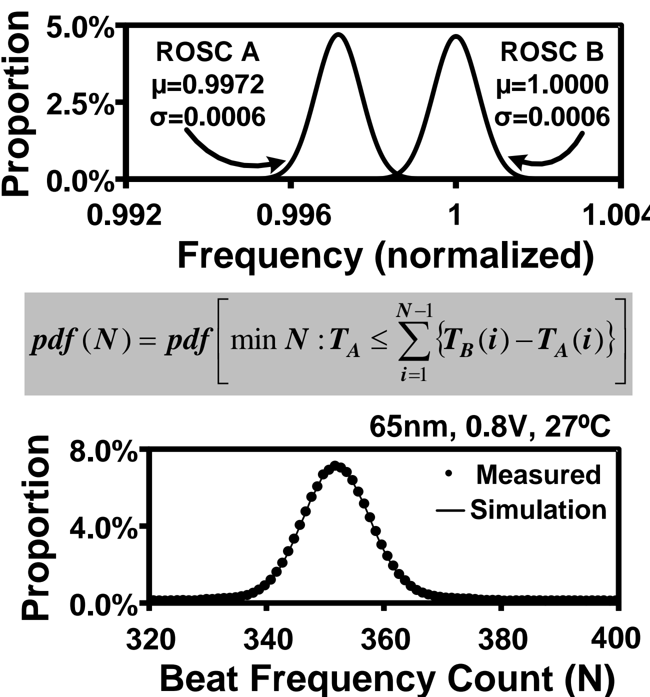


Figure 4.13: (Upper) Individual ROSC frequency distributions estimated using statistical model and measured data. (Lower) Measured count distribution shows good agreement with simulated data.

For better insight into the beat frequency based TRNG circuit, an analysis of the underlying physical noise source model and the entropy harvesting mechanism is required. The randomness source of ROSC based TRNGs is generally the cycle to cycle random noise. In time domain, the random noise is referred as jitter which is defined as the fluctuation of transition time from their ideal position. In BFD TRNG, the ROSC jitters are digitalized into the fluctuated count values. A precise ROSC jitter model should incorporate Gaussian variable, 1/f noise, and a coupling sinusoidal signal [47]. However,

studies also reveal that the dominant component is the independent and identically distributed (i.i.d.) Gaussian variables [48], [49]. Based on this assumption, the oscillation period then can be modeled as Gaussian variable $T \sim N(\mu, \sigma^2)$ [42].

In The BFD TRNG, the two ROSC circuits were formulated as independent Gaussian random variables $T_A \sim N(\mu_A, \sigma_A^2)$ and $T_B \sim N(\mu_B, \sigma_B^2)$. The average values of the two ROSC periods (μ_A and μ_B) on the other hand, can be tuned during initial startup using different trimming capacitor settings for a desired beat frequency count N. Without loss of generality, we always assume ROSC A is faster than ROSC B, i.e., $\mu_A < \mu_B$. Since both ROSC are implemented identical, we can assume that intrinsic noise induced standard deviation σ_A and σ_B under a given operating condition shows the same characteristics ($\sigma_A = \sigma_B$). We can estimate the average count values (N) based on the equation shown in Figure 4.13. Since N does not have a standard probability density function, we performed the Monte Carlo simulation with MATLAB to verify the model. Simulation results from our model using fitting parameters $\mu_B - \mu_A = 0.0028$ and $\sigma_A = \sigma_B = 0.0006$ show excellent agreement with the measured data.

4.4. Multi-phase TRNG for Enhancing the TRNG Generation Efficiency

4.4.1. Circuit Implementation

A multi-phase TRNG capable of sampling the beat frequency from each ROSC stage was implemented in another 65nm test chip. This new design can maximize the number

of random bits generated from a single ROSC pair without increasing the measurement time. The simplified block diagram of the multi-phase TRNG is illustrated in Figure 4.14. As shown in the figure, beat frequency detectors in each ROSC stage sample the frequency difference between the top and bottom ROSCs. Under an ideal condition where no device noise is present, the ROSC signal is simply delayed from one stage to the next by a fixed amount. The multi-phase design in this case does not provide any benefit over the single-phase version as the beat frequency measured from each stage will be identical. In the presence of device noise however, each ROSC stage will introduce a slightly different delay and hence each beat frequency count will be different (Figure 4.15). This noise effect can be captured using simple logic blocks to increase the number of random bits. The counter values are stored in the shift registers and summed up together during the subsequent beat frequency period (i.e. falling edge of next beat signal).

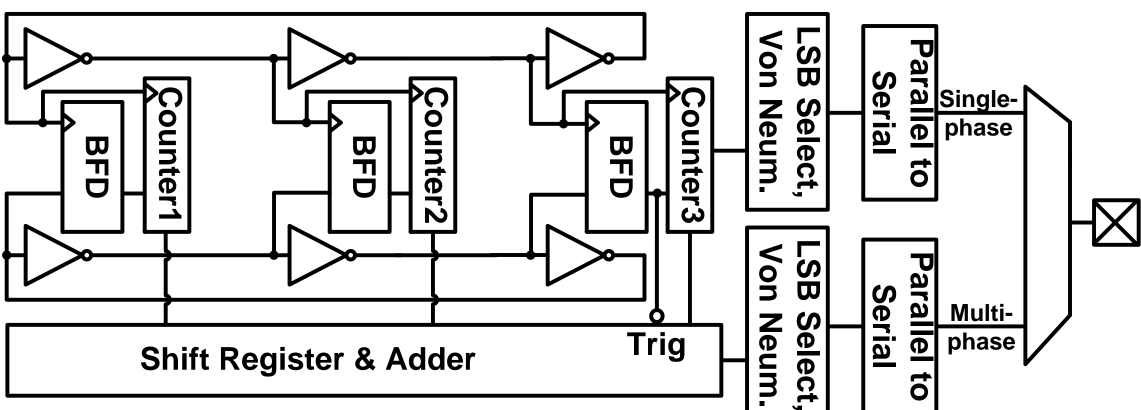


Figure 4.14: Multi-phase TRNG implementation (3 phase example).

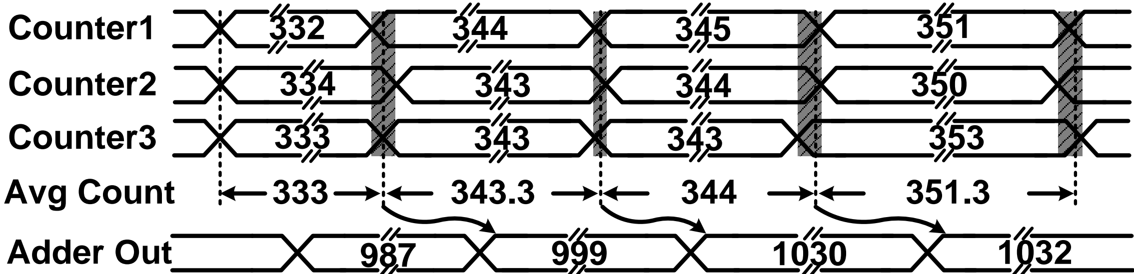


Figure 4.15: The number of LSBs with good randomness increases under the same sampling time as compared to the single-phase version.

4.4.2. Measurement Data

Figure 4.16 compares the measured beat frequency counts from a single-phase and multi-phase TRNG implemented in the same test chip. A larger fluctuation in the count value can be observed for the multi-phase design. Therefore, for the same sampling period, the multi-phase TRNG can provide a higher number of random bits that pass all NIST tests. The number of additional random bits obtained from a multi-phase TRNG has a logarithmic dependency on the number of phases. For example, as shown in Figure 4.17, a multi-phase TRNG using a 3 stage ROSC generates one more random bit while five more bits are generated using a 31 stage ROSC. Even though the power consumption slightly increases due to the extra beat frequency detectors, the overall efficiency still increases due to the improved throughput. Figure 4.18 compares the TRNG performance measured from various single-phase and multi-phase TRNG circuits. A TRNG with fewer ROSC stages achieves a higher bit rate which can be attributed to the higher ROSC frequency. As a result, the TRNG efficiency increases from 2.2 Mbits/mW to 15.1 Mbits/mW as the number of ROSC stages is reduced from 31 to 3.

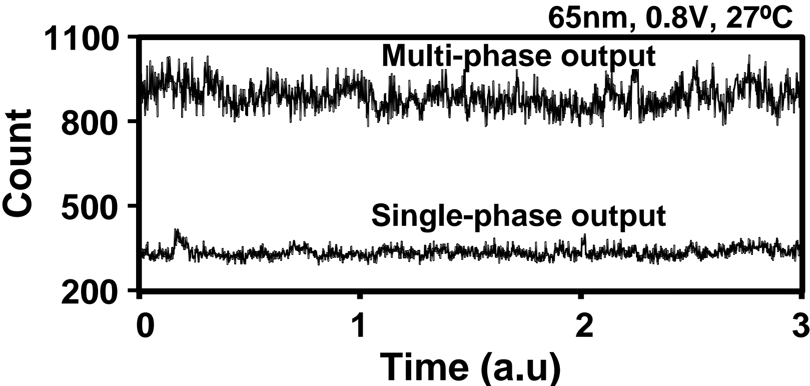


Figure 4.16: Measured count output from single-phase and multi-phase TRNGs.

	3 stage ROSC		31 stage ROSC	
	Single phase	Multi phase	Single phase	Multi phase
# of bits passing NIST	2	3	3	8
Efficiency (Mbits/mW)	13.341	15.114	2.0342	3.176

Figure 4.17: The number of random bits per output that passes all NIST test as well as the TRNG generation efficiency improves using the proposed multi-phase structure.

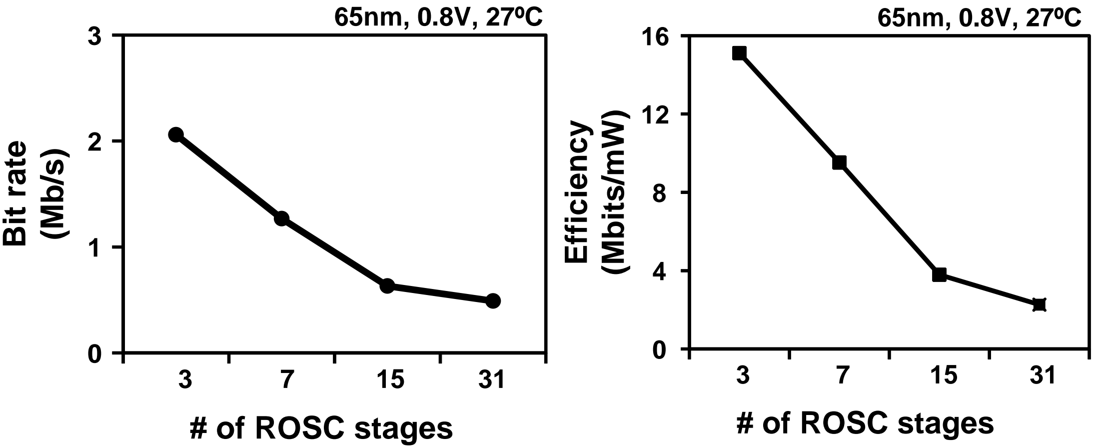
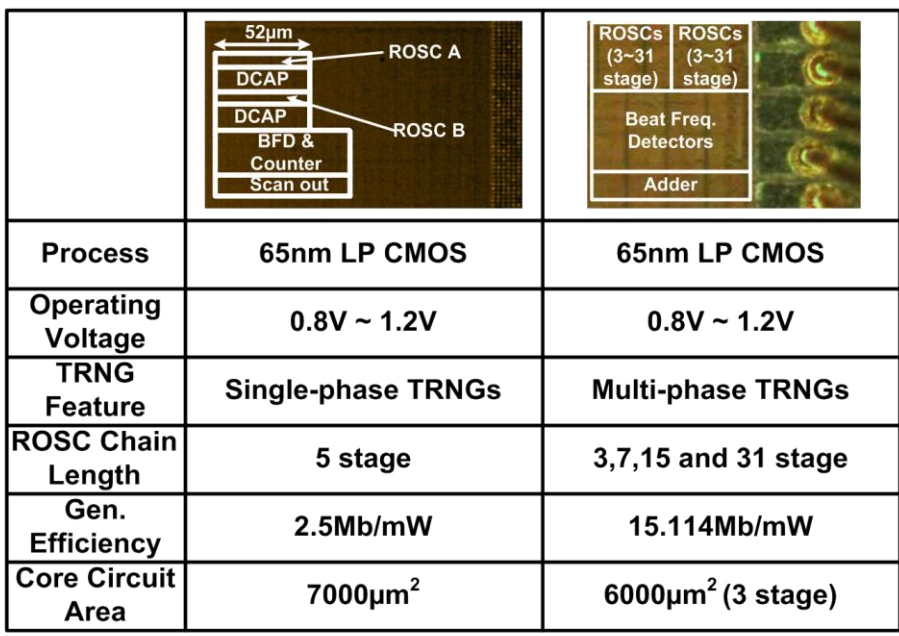


Figure 4.18: Multi-phase TRNG utilizing fewer ROSC stages shows improved bit rate and efficiency.

4.5. Conclusion

ROSC based TRNG circuits utilizing a novel beat frequency detection technique have been demonstrated in two 65nm test chips. Die microphotographs of the two TRNG test chips are shown in Figure 4.19. The random bits generated using the proposed circuit pass all 15 NIST tests under a wide supply voltage range without any feedback scheme. Long-term (>15 hours) tests were performed to confirm good TRNG output under continuous operation. A one-time calibration scheme ensures that ROSC frequency mistmatch across different chips is cancelled out. To further improve the efficiency, a multi-phase TRNG was demonstrated that captures phase noise in each ROSC stage. Experimental data shows a TRNG efficiency of 15.1Mbits/mW for a 3 stage multi-phase design.



Process	65nm LP CMOS	65nm LP CMOS
Operating Voltage	0.8V ~ 1.2V	0.8V ~ 1.2V
TRNG Feature	Single-phase TRNGs	Multi-phase TRNGs
ROSC Chain Length	5 stage	3,7,15 and 31 stage
Gen. Efficiency	2.5Mb/mW	15.114Mb/mW
Core Circuit Area	7000µm^2	6000µm^2 (3 stage)

Figure 4.19: Single-phase and multi-phase TRNG chips in 65nm.

Chapter 5. Physical Unclonable Function

5.1. Introduction

It is reported that with the development of the Internet of Things (IoT), security will be one of the major challenges in this end-to-end communication network that involves billions of digital devices [57]-[62]. Protections from software level alone are proven to be insufficient, especially against physical attacks. Hardware security, on the other hand, provides a low power solution to the Root of Trust (RoT). Most of the current hardware security primitives are based on Non-Volatile Memories (NVMs) such as EEPROM, Flash and Fuse [60]. They have been widely adopted to store the secret keys (e.g. smart cards) and provide authentication source [64]. However, NVMs based security primitives are becoming more vulnerable to invasive attacks as the attacking techniques have been significantly improved [65]. PUF, on the other hand, generates the unique signature for each chip by utilizing the random process variability. PUF outputs only rely on intrinsic physical characteristics which makes it difficult to predict and almost impossible to duplicate. The secret information stored in PUF is only available when the device is powered on. Therefore PUFs are immune to offline attacks.

Based on the applications, PUFs can be categorized into: 1) strong PUFs for low cost authentication; and 2) weak PUFs for chip ID or secret key generation [60]. The key difference between a strong PUF and a weak PUF is whether it can provide large number of CRPs with a moderate hardware cost. Strong PUFs can be directly used for

authentication because the adversaries are unable to figure out all CRPs within polynomial time [60]. A typical example of strong PUF is the arbiter PUF [65]. Weak PUFs, on the other hand, provide limited number of CRPs because the number of CRPs is linearly proportional to the number of PUF unit cells. Most memory or array based PUFs, e.g. SRAM and DRAM PUF [66], [69], are generally referred as the weak PUFs. Compared to arbiter PUF, the memory based PUF requires minimal design overhead because the secret keys are generated with existing blocks which are used as cache in nominal operation. However, this advantage is undermined because it generally requires an excessive on-chip Error Correction (ECC) block for key and ID generations.

5.2. Conventional PUFs

A typical strong PUF is the arbiter PUF as shown in Figure 5.1. The basic unit cell is implemented with two identical multiplexors (MUXs) with swapped input connections which pass the signal in a parallel or crossing pattern. Each arbiter PUF consists of N (typically 128) unit cells and followed by a latch to determine which final output arrives earlier. During the operation, a random N bit challenge (C<0:127>) sets the MUX selection signals. A rising edge is fed to the two inputs in the first stage simultaneously establishing a racing condition between two signal paths. The process variation induced delay difference on each stage is accumulated and comprises the unclonable feature of the arbiter PUF. The randomness of the arbiter PUF relies on the manufacturing induced delay difference between each path. Since each MUX stage is designed identical, an

adversary is unable reproduce the same CRPs even with the knowledge of detail PUF implementation. Theoretically, an arbiter PUF with N stage can provide up to 2^N CRPs.

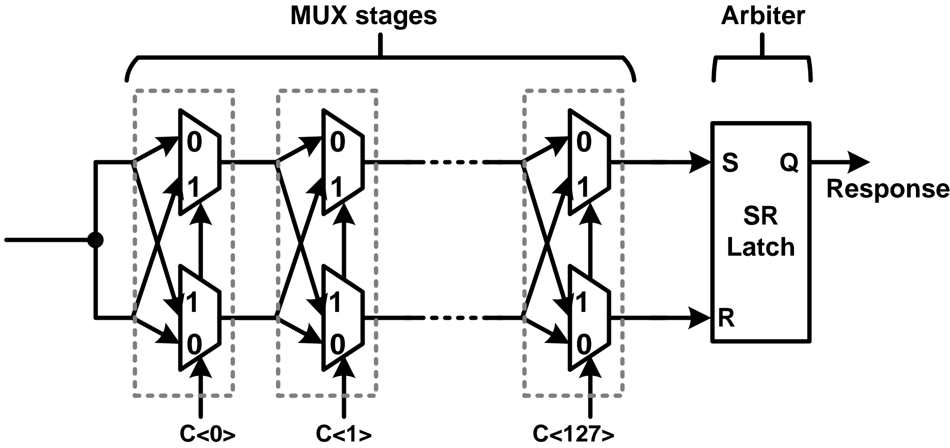


Figure 5.1: Schematic of an arbiter PUF.

One representative of the weak PUFs is the SRAM based PUF. Typical 6T SRAM holds the data through the positive feedback loop in the cross-couple inverter pair. In nominal operation, each cell is written with opposite value on Q and QB. The positive feedback loop forces the cell to hold the data preventing an accidental cell value flip. When used as a PUF unit, the operation involves two phases as shown in Figure 5.2. During the initialization, the cell is completely powered off by setting VDD, BL and BLB to 0V. Once the SRAM cell is powered on, in theory, the initial binary state is unknown because the symmetry of its cross-coupled inverters leads to a metastable state. In actual implementation, Q and QB resolve to a steady state with opposite values due the process variation induced threshold voltage mismatch. The polarity of Q and QB is determined by the inherent asymmetry between the two inverters. This startup value is taken as the PUF unique signature.

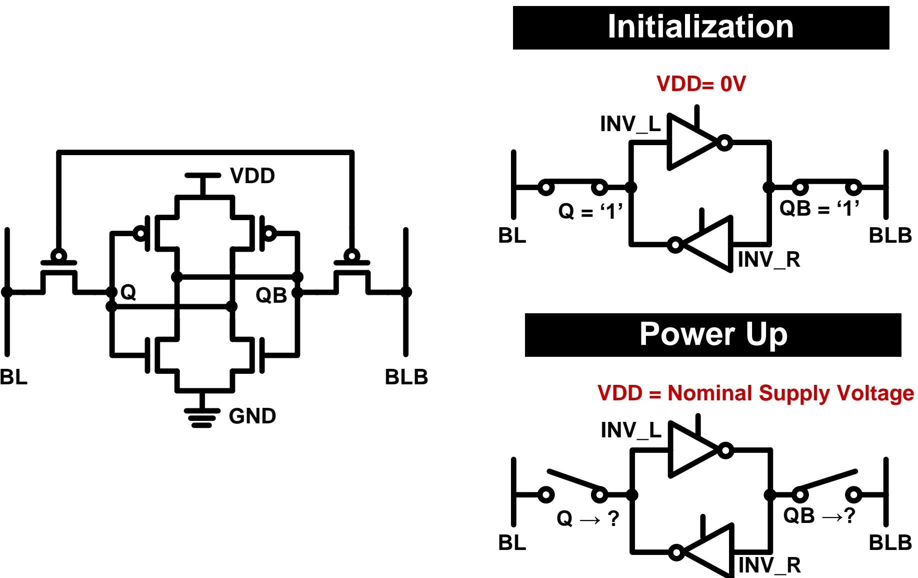


Figure 5.2: SRAM based PUF.

For each PUF consists of N PUF units, the number of available CRPs is 2^N and N for arbiter and SRAM PUF respectively. Therefore, arbiter PUF is categorized into strong PUF while SRAM PUF is a weak PUF. Although the SRAM based PUF are less vulnerable to physical attacks such as cloning and offline attacks as compared to NVMs, SRAM PUF is less reliable as it is sensitive to random noise. For example, if the cross-coupled inverter pair is well balanced, then the response will be determined by the random noise presented at the power up moment. As a result, ECCs are generally required in most weak PUF applications.

5.3. DRAM based PUF for Chip Authentication

Memory based PUFs are attractive [66], [69] as they are readily available in most processors, requiring almost no modifications to the underlying hardware. Moreover, the

array based structure provides a large set of independent entropy sources. Beside the aforementioned SRAM PUF, a 1T1C DRAM based weak PUF prototype was demonstrated by IBM Research in [66]. When used as memory, the data is stored in each DRAM cell by either charging (data '1') or not charging (data '0') the capacitor. These charges, however, will gradually leak away through the access transistor. The leakage current varies from cell to cell resulting in retention time variability among the DRAM array. In PUF application, the bias condition and refresh period is chosen such that there is a certain number of cells flip from '1' to '0'. The flipped cell locations are unique and unclonable for each chip and the bit cell failure map can be taken as the chip ID. Different from SRAM PUF in which '1' and '0' locations are evenly distributed, the percentage of flipped cells (cells that read as '0') in DRAM PUF is generally less than 10% in consideration of the authentication time and intra-chip consistency.

However, one shortcoming of the conventional memory based PUFs compared to arbiter PUFs [60],[70] is that the number of Challenge Response Pairs (CRPs) is linearly proportional to the number of circuit units. Memory PUFs are therefore categorized as "weak" PUFs and are not suitable direct chip authentication applications. To address this shortcoming, we present a novel DRAM based "strong" PUF capable of generating $>10^{32}$ CRPs from a 1Kbit array. The main highlights of this work are: 1) a local encrypting scheme that enhances the authentication security and allows a DRAM to serve as a strong PUF; 2) a repetitive write-back scheme based on existing DRAM refresh circuits for

enhancing PUF stability; and 3) a simple calibration routine to suppress voltage and temperature variation effects.

Figure 5.3 compares the properties of SRAM and DRAM when utilized as PUFs. Unlike SRAM PUFs where the supply voltage is turned off and turned on to generate a response, a DRAM PUF can be accessed anytime during normal operation by writing a '0' or '1' and checking whether the data has flipped or not after a certain retention time. This unique feature allows us to generate an exponentially higher number of CRPs.

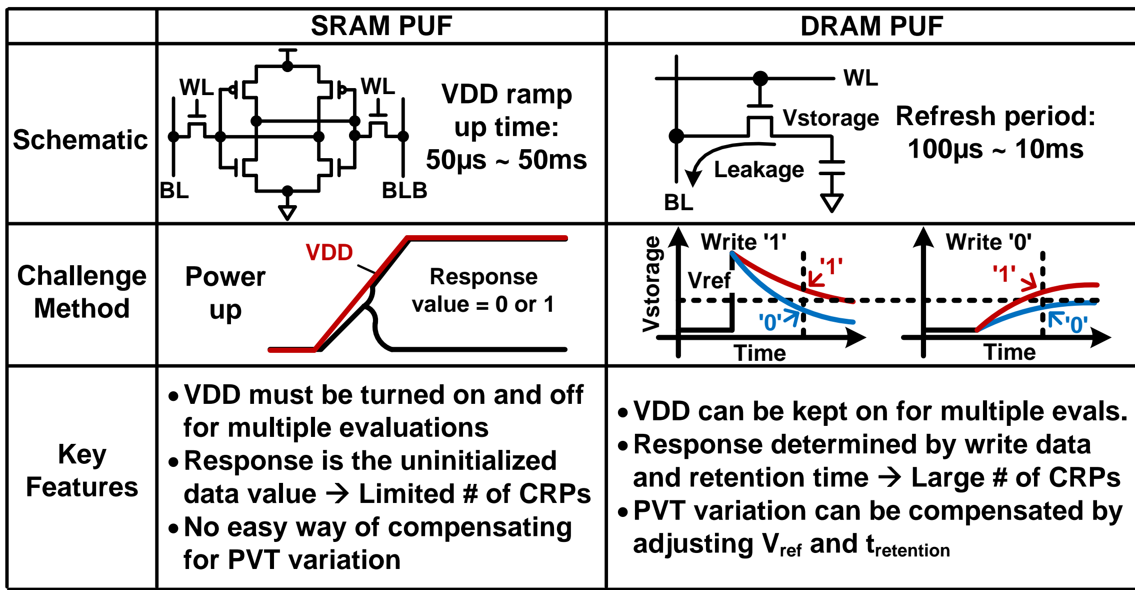


Figure 5.3: Qualitative comparison between SRAM PUF and DRAM PUF.

The schematic of the 2T DRAM [67] used in this work is shown in Figure 5.4 (a). Compared to a 1T1C DRAM cell, 2T DRAM cell does not require a dedicated trench or stacked capacitor process, and has decoupled read and write paths enabling good low voltage margin. The data retention time depends on the storage node capacitance and the leakage current surrounding the storage node. The read reference voltage (V_{ref}) can be

adjusted such that a certain number of cells fail for a given retention time. Figure 5.4(b) shows several retention time failure scenarios. In conventional 1T1C DRAM, each cell in the array is written with data '1' and read out sequentially for certain refresh period. The failure locations (cells that read as data '0') are taken as the unique ID for each chip. The failure in 2T gain cell, on the other hand, depends on not only on the leakage variability but also the initial data been written. A cell with strong pull down leakage will not hold a data '1' value very well, and vice versa. Therefore, the response from 2T DRAM PUF is the function of both location and input data.

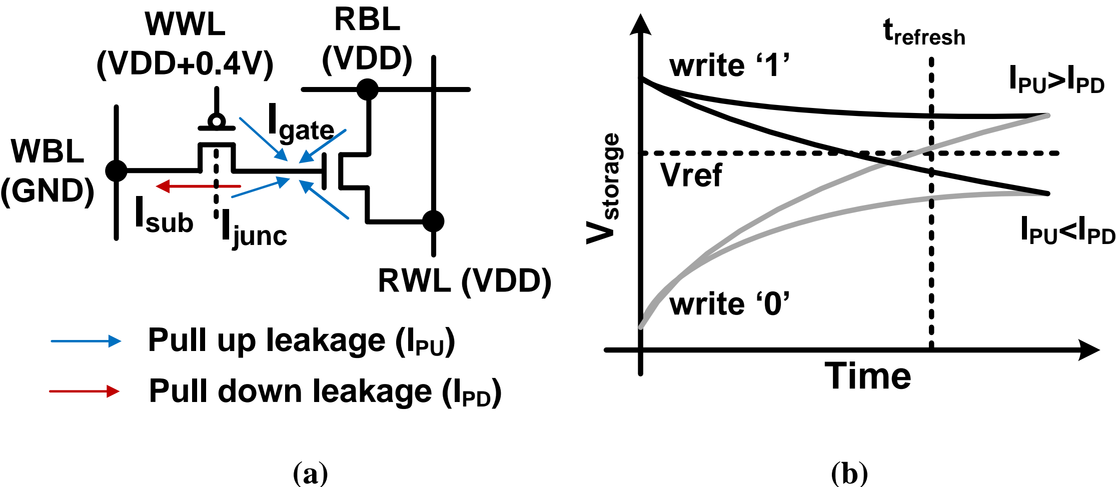


Figure 5.4: (a) 2T DRAM cell schematic and leakage components in hold mode. (b) A DRAM cell generates a different response depending on the write data and retention time.

5.3.1. Proposed DRAM PUF Design

The proposed authentication scheme illustrated in Figure 5.5 comprises of four steps: (1) writing a random pattern (=challenge from server) to a small portion of the DRAM

array; (2) letting 10% of the cells to fail by reading the data after a certain retention time; (3) transferring the data stored in the array to a different location in a random bitwise mapping fashion; and finally (4) repeating step (2). The second step provides local encryption which generates a new random pattern that is hidden from the outside world, providing an added level of security. The local encryption operation cannot be implemented in an SRAM PUF because it requires a random data pattern to be written to the memory array. Access to the local encrypted pattern is only allowed during chip enrollment phase and will be permanently disabled thereafter. Randomly transferring the array data to a different location, combined with the initial random pattern from the server, enables an exponentially higher number of CRPs. The total number CRPs for a 10% retention failure probability can be calculated as follows.

$$N_{CRP} = 2 \cdot n \cdot P(n, i \times 10\%)$$

Here, the pre-factor '2' represents the two values that can be written to a DRAM cell, P(.) is the permutation function, 'n' is the half array size, and 'i' is the number of response bits. According to this equation, the total number of CRPs attainable from a 1Kbit DRAM array for a 128-bit response output is greater than 10^{32}.

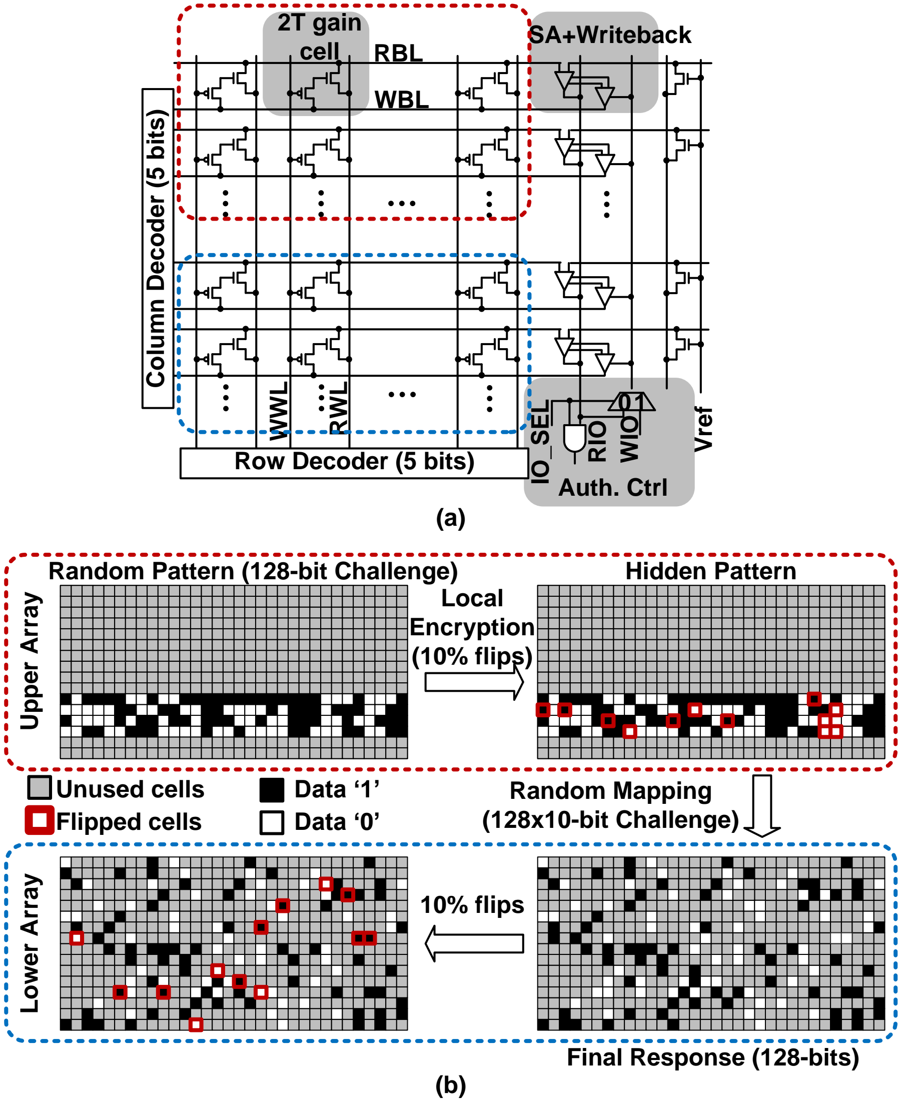


Figure 5.5: The proposed authentication scheme consists of four steps: (1) write random 128 bit challenge to DRAM upper array, (2) allow 10% of bits to flip due to retention failure, (3) transfer data to lower array according to random mapping info from server, and finally (4) repeat step (2). The inherent DRAM retention failure rate is utilized for generating a unique and secure response. For the chip demonstration, we chose a 128-bit random input pattern, a 128 x 10 bit random address mapping info (=128+128x10=1,408 total challenge bits) and a 128-bit response.

The enrollment and authentication procedures of the proposed DRAM PUF are shown in Figure 5.6. Compared to conventional strong PUFs which require an exhaustive test to collect a large number of CRPs, the proposed PUF only needs to store whether a retention failure occurs or not for data '0' and data '1' under a certain retention time. So for a 1Kbit array, the unique PUF information can be stored in just two 1Kbit maps, one for data '1' and one for data '0'. To generate the bit maps, we first write all '1's to the 1Kbit DRAM array, let retention failures occur, and then read the pattern including retention failures. The same procedure is repeated for data '0'. The bit maps are stored on the server as reference key. Figure 5.6 illustrates the authentication flow for generating a 128-bit response from a 1Kbit array. During authentication, the initial 128-bit random pattern along with the 128 x 10 bit random mapping information is generated by the server and sent to the chip as challenge bits. Based on the reference key, the server computes the expected response and compares it with the response from the chip. If the Hamming Distance (HD) between the two responses satisfies the match criterion, access permission is granted to the user.

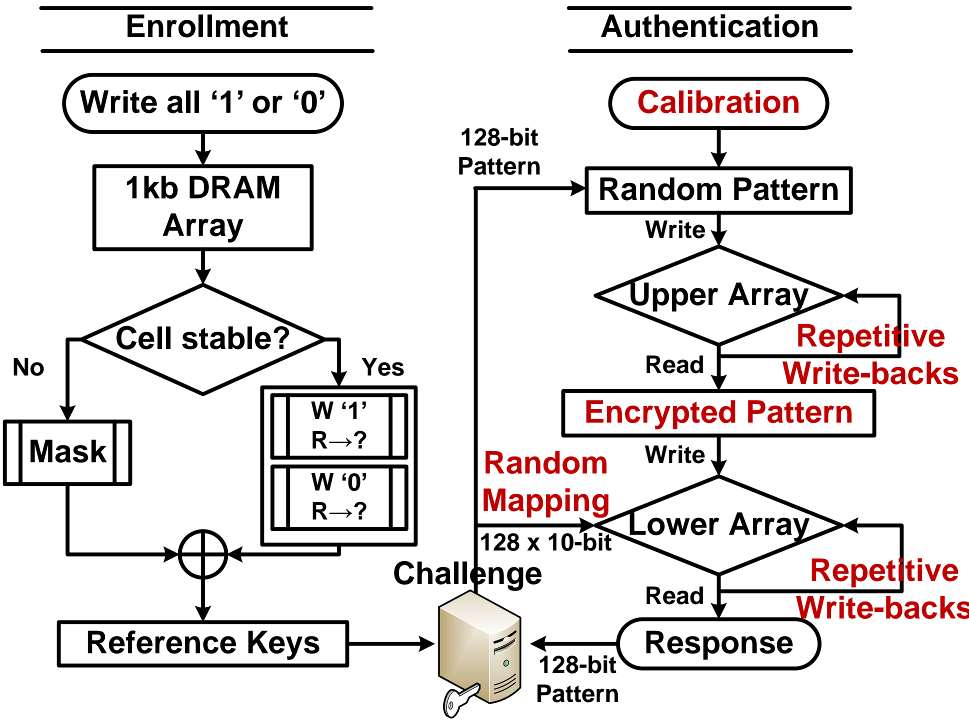


Figure 5.6: Overall enrollment and authentication flow of the proposed DRAM PUF. New techniques proposed in this work are highlighted in red.

5.3.2. Improving DRAM PUF Reliability

Figure 5.7 shows the soft response distribution measured from a 1Kbit DRAM array for 500 trials. Soft response is defined as the average of 500 response values for a particular DRAM cell. For example, if the response is ‘1’ for 90% of the time and ‘0’ for 10% of the time, then the soft response value is 0.9. The left-most and right-most bars represent the stable cells with 0% and 100% retention failures for the entire 500 trial period. The bars in the middle represent unstable cells that generate some retention time failures. Experimental data shows that the percentages of unstable cells (i.e. 0 < soft response < 1) are 8.8% and 7.2% for data ‘1’ and data ‘0’, respectively.

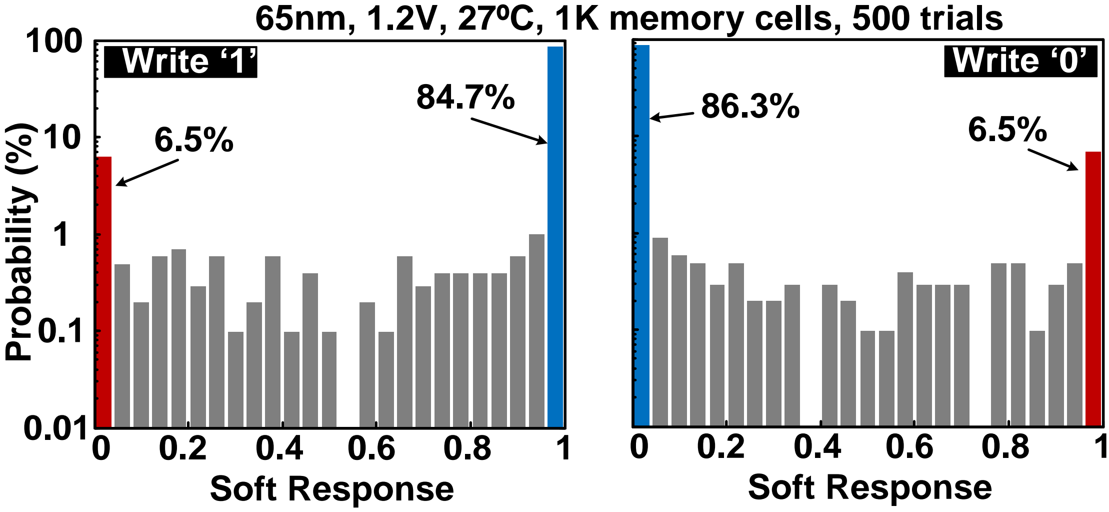


Figure 5.7: DRAM PUF soft response distribution for data '1' and data '0'. Soft response is defined as the average response value over 500 trials. For example, if the output for a particular memory cell is 1 for 90% of the time and 0 for 10% of the time, the soft response for this memory cell is 0.9.

To reduce the percentage of unstable cells, many PUF designs employ Temporal Majority Voting (TMV). TMV is a technique in which a PUF is evaluated multiple times using the same challenge and the majority output value is taken as the final response. The main drawback of TMV is the large area and delay overhead for storing and processing the PUF outputs from each TMV trial. For example, to perform a 15 trial TMV, a 4 bit counter is required for each accessed cell. Sharing a single TMV counter for the entire array will reduce the area overhead, but the authentication time becomes prohibitively long. As an alternative to TMV, we propose a repetitive write-back scheme that can be implemented using existing DRAM refresh circuitry with no hardware overhead. The idea is based on the fact that a cell with a small read margin for data '1', generally has a large read margin for data '0', and vice versa. Based on this unique characteristic, we

propose the repetitive write-back scheme shown in Figure 5.8 where DRAM cells are written with the data read from the previous cycle. Measurement results in Figure 5.9 verify that the percentage of unstable cells decreases, although after 5 cycles it levels out. After 10 write-back cycles, the percentage of unstable cells reduces from 8.8% to 6% for data ‘1’, and from 7.2% to 5.2% for data ‘0’. Although the improvement is not significant, the repetitive write-back scheme is still useful as it incurs no hardware overhead. The server will mask DRAM cells that remain unstable after the repetitive write-back.

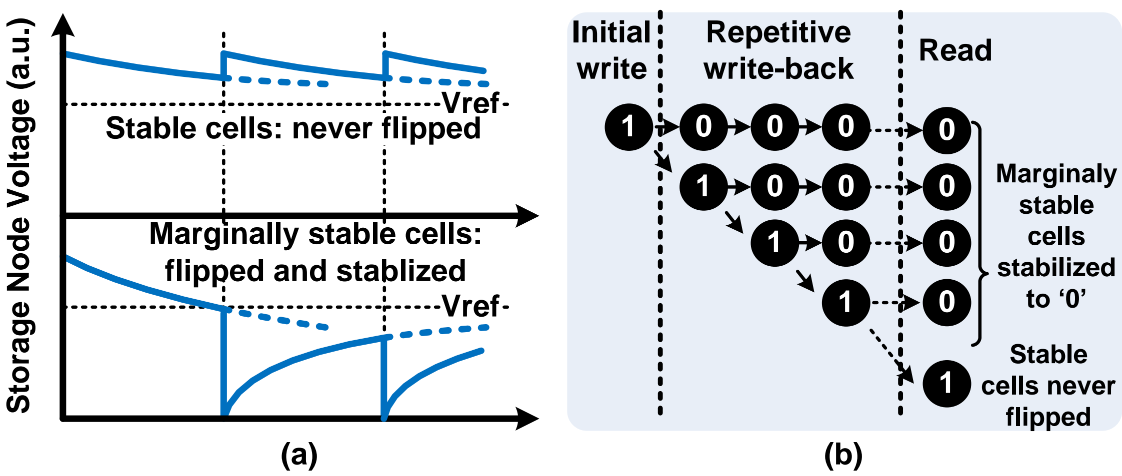
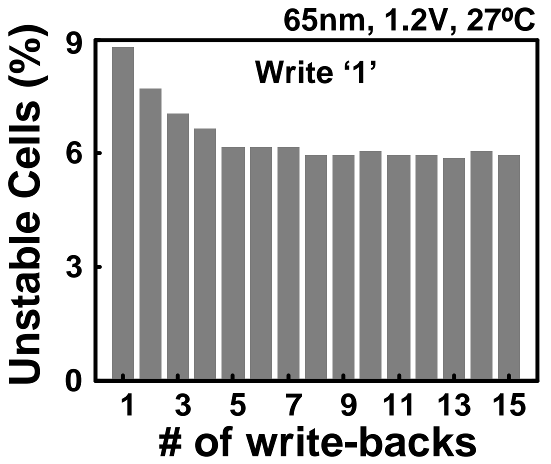


Figure 5.8: Repetitive write back scheme for improving DRAM PUF stability. (a) Waveforms of DRAM cell storage voltage with repetitive write back. (b) Marginally stable bits can be stabilized to the opposite value with repetitive write-back.



	No write-backs		10 write-backs	
	W ‘1’	W ‘0’	W ‘1’	W ‘0’
Always ‘1’	84.7%	86.3%	81.3%	83.2%
Always ‘0’	6.5%	6.5%	12.7%	11.6%
Unstable	8.8%	7.2%	6.0%	5.2%

Final errors
(Bit mask stored on server)

Figure 5.9: Percentage of unstable cells decreases with more write-backs. Cells that remain unstable after 10 write-backs will be flagged and masked by the server during chip authentication.

5.3.3. Test Chip Measurement

A fully functional 64Kb DRAM PUF array was fabricated in a 1.2V, 65nm process for concept verification. Figure 5.10 shows the die photo and key features. We selected a 32x32 (=1Kbit) DRAM subarray to test the proposed PUF authentication scheme. The measured intra-chip HD and inter-chip HD distributions are shown in Figure 5.11. The intra-chip HD was obtained by applying the same challenge 500 times. The intra-chip HD measured under a nominal condition (1.2V and 27˚C) has an average of 2.2% and a standard deviation of 1.03%. After masking the unstable responses, the average and standard deviation of intra-chip HD improves to 0.39% and 0.19%, respectively. Note that bit masking was performed on the server side which obviates the need for an Error Correcting Code (ECC) unit. The inter-chip HD distribution was obtained by applying 10k random challenges to 15 different chips. The distribution has an average of 35.9%

and a standard deviation of 6.35%. The reason why the average inter-chip HD is not centered around 50% is because we deliberately chose a retention failure rate of 10% (and not 50%) to speed up the authentication process. Moving the inter-chip HD distribution to the center is possible but at the expense of a longer authentication time. The margin between the intra-chip and inter-chip HD distributions is 10.2%.

Process	65nm LP CMOS
Cell size	58% of 6T SRAM
Supply	0.8V ~ 1.2V
Temperature	-15ºC ~ 85ºC
Circuit area	350x300µm
Retention time	10.5ms @10% flip (1.2V, 27ºC)
Random cycle	1.2ns
Energy (10 write-backs)	0.89pJ/bit

Figure 5.10: 65nm DRAM PUF chip micrograph and summary table.

Maintaining a narrow intra-chip HD distribution across different voltages and temperatures is imperative for PUFs in real products. As shown in Figure 5.12, the retention failure probability varies significantly under different voltage and temperature conditions. When operated at a relatively lower supply voltage, the 2T1C gain cell pull-up and pull-down leakage currents are reduced contributing to a decrease of failure probabilities. Therefore, a smaller data ‘0’ failure probability is observed. On the other hand, data ‘1’ failure probability increases which is mainly due to reduction of charges being written to the storage node.

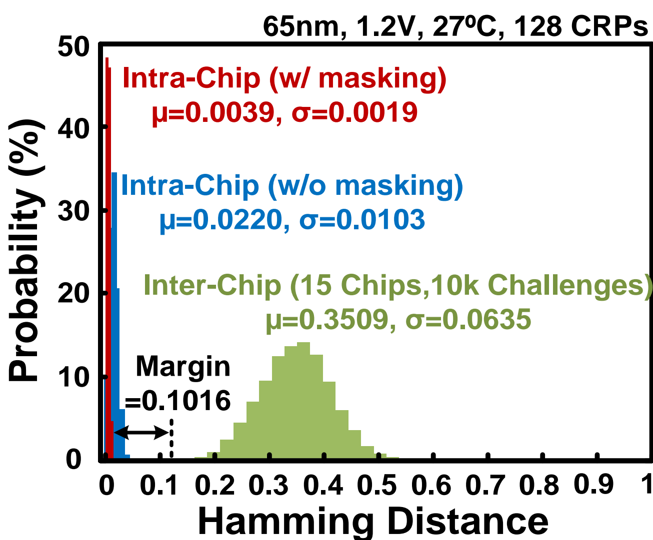


Figure 5.11: Measured inter-chip (15 chips and 10,000 different challenges) and intra-chip Hamming Distance with and without bit masking.

As shown in Figure 5.12, as the supply voltage is reduced from 1.2V to 0.8V, the data '0' flipping probability is reduced from 11% to 0.5% while the data '1' flipping probability is increased from 10.7% to 87.1%. Figure 5.12 also displays failure characteristics of the DRAM array at different temperatures. The leakage current increases at higher temperature resulting in an increase of the failure probability.

To mitigate V and T effects, a calibration scheme is proposed based on the observation that the order of failure locations remains almost the same under different V and T conditions. Although the failure probability shifts under different V, T conditions, the order of failing locations remains the same because the V, T variation affects each cell in the same way. As we discussed in 5.3.1, the PUF response is a function of the failure location and input data. By maintaining the failure probability (e.g. 10%), the failure locations will remain the same therefore generating a consistent response across different

V, T. Basic operation of the calibration scheme is given in Figure 5.13. First, a checkerboard pattern is written into the array. The calibration circuit measures the ratio between '1's and '0's after a certain retention period by counting the number of '1's in the array pattern. If the percentage of '1's does not fall within the desired range (e.g. 49%-51%), V_{ref} is adjusted accordingly. Finally, the calibration circuit adjusts the refresh time $t_{refresh}$ to ensure that the retention failure probability P_{flip} is close to the target (e.g. 10%). Once the calibration is complete for a specific V and T condition, the authentication procedure depicted in Fig. 4 (right) can start.

To establish this, we proposed a dynamic pre-authentication calibration scheme as shown in Figure 5.13(a). A data pattern with alternating '1's and '0's is written to the array. Depending on the read data, the calibration circuit adjust V_{ref} and refresh time ($t_{refresh}$) accordingly in case either '1'/'0' ratio or failure probability fall out of the desired range. Although we conducted this calibration off-chip, it can be implemented with on-chip circuit as shown in Figure 5.13(b). The calibration circuit first detects the '1'/'0' ratio by counting the number of '1's from the read pattern. For example, if the ratio is greater than the specified range (e.g. 0.49~0.51), the reference voltage controller will lower the voltage accordingly, and vice versa. Then the circuit compares the read and write pattern making sure that the overall failure probability is in the desired range.

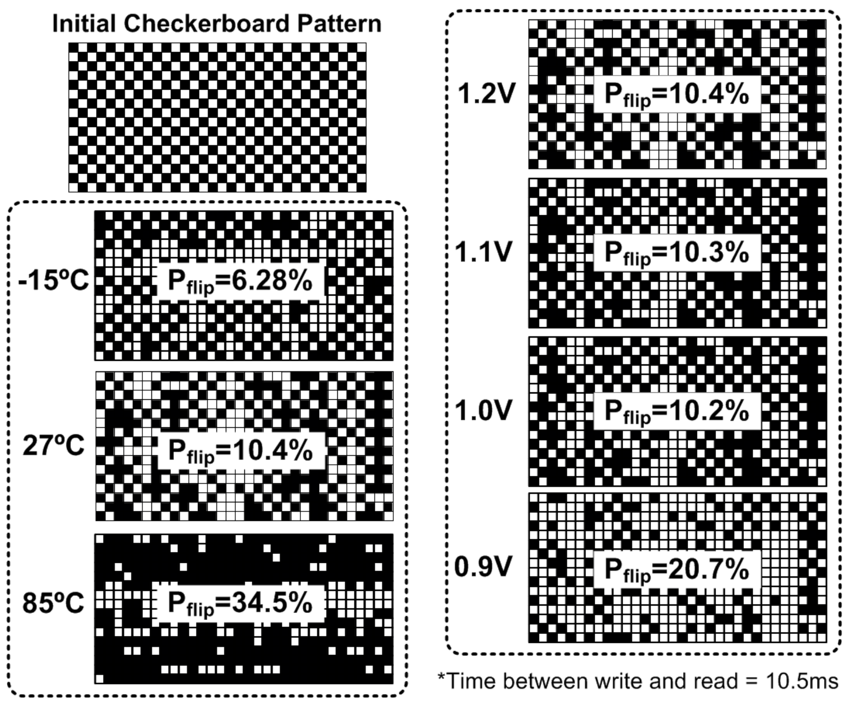


Figure 5.12: DRAM retention failure map measured under different supply voltages and temperatures. The failure probability can be kept within the desired range of 9.5%<P<11% before each authentication test using the calibration scheme described in Figure 5.11.

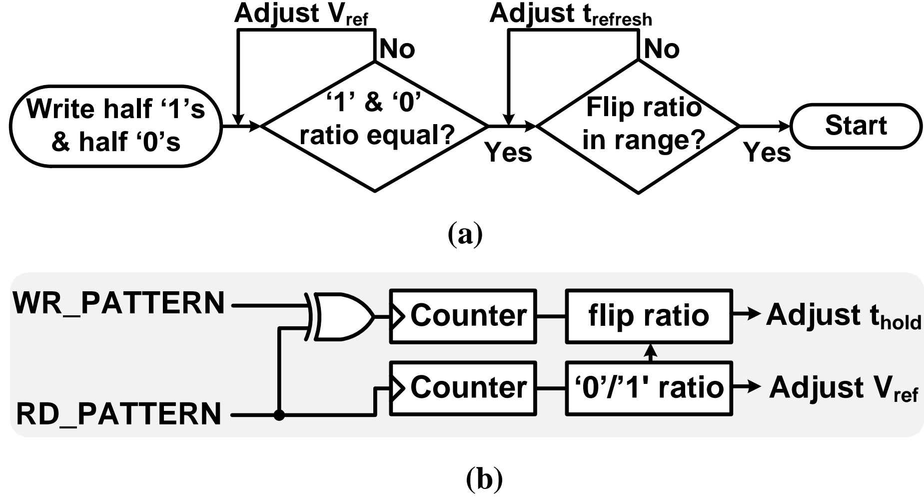


Figure 5.13: A pre-authentication calibration scheme to mitigate the V, T drifts induced intra-chip variation.

Figure 5.14 shows the distributions of intra-chip HD collected from different supply voltage ranges (0.8V to 1.2V). After applying the proposed calibration scheme, the average and standard deviation of the distributions are getting smaller implying a better response stability even with the presence of supply voltage variation. Figure 5.15 displays the maximum intra-chip HDs obtained from different supply voltage ranges. After applying the calibration scheme, the max intra-chip HD is reduced indicating an improvement of the authentication correctness. Similarly, we characterized the distribution and maximum value of the intra-chip HD in a temperature range from -15 ℃ to 85 ℃ as shown in Fig. 14. Measurement results also display that the PUF stability and authentication correctness are improved after applying the calibrations.

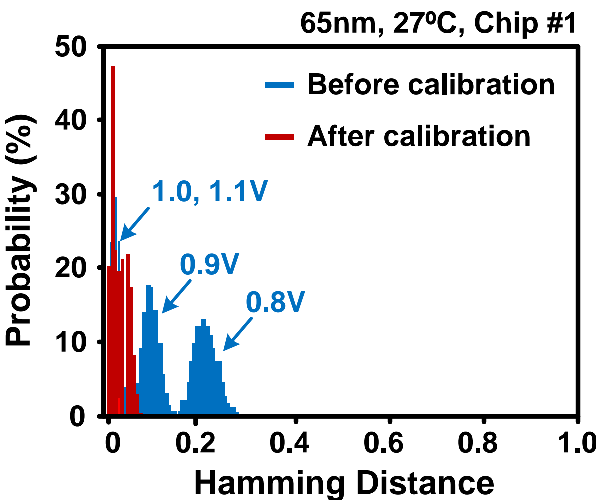


Figure 5.14: Distributions of Hamming distance measured at different supply voltages before and after calibration.

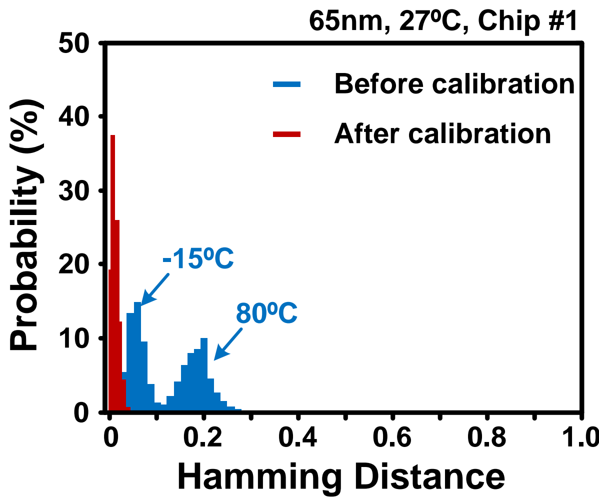


Figure 5.15: Distributions of Hamming distance measured at different temperatures before and after calibration.

5.4. Charge-redistribution based PUF for Chip Authentication using a SAR ADC Circuit

5.4.1. Background

One of the primary principles of the IoT concept is to collect and process the data locally then making the result available throughout the network. The emerging applications in different domains such as healthcare gears, home temperature control and public lighting allow data collected locally through various sensors. At the same time, the increased data flow also brings in concerns on hardware security including unauthorized access and malicious tampering. However, authenticating distributed sensors is more difficult than that in a conventional server network due to limited resource and

unattended operation. Therefore, implementing an authentication scheme with existing building blocks is favorable in such applications.

In a sensor system, ADC is utilized as an interface between the analog front-end and the digital processing units. There are three major types of ADCs that can fit into most of the applications: Success-Approximation (SAR) ADC, sigma-delta ADC and pipeline ADC. Out of the three, sigma-delta ADC can provides the highest resolution however is relatively slow (<1Msps), pipeline ADC can work with high sampling frequency (> 5Msps) but consume large power, SAR ADC consume low power and has a moderate performance as compared to the other two types. Many sensors are powered by batteries or through energy harvesting in which the low power blocks are needed for long time operation. Therefore, SAR ADC is one of the best candidates for such applications due to its low power consumption. In this work, we present a new charge-redistribution based PUF fully utilizing the existing SAR ADC block. Figure 5.16 compares the conventional arbiter, memory and the proposed charge-redistribution PUFs. The proposed PUF advantageous the other two in term of: 1) incurs almost no overhead circuit as compared to the arbiter PUF; 2) provides more CRPs as compared to the memory based PUF therefore can be deployed in direct authentication protocols; 3) more robust towards variations such as aging induced parametric shifts owing to the passive components. It should also be noted that the process induced variation is relatively smaller in a passive device than that in an active device.

	Arbiter PUF	Memory based PUF (e.g. SRAM PUF)	Proposed PUF
Unit Cell Schematic	sel<i> 0 1 1 0		− +
Extracting Variation	Delay difference between two paths	Power up value	Charge redistribution: volt. difference
# of CRPs (N unit cells)	2^N	N	$\binom{N}{k}$
Pros	1. Large # of CRPs 2. Simple	Reuse existed blocks	1. Reuse existed blocks (SAR ADC) 2. Passive device: High tolerance to parametric drfits
Cons	Circuit area overhead	1.Limited # of CRPs 2. VDD must be turned on and off	Small noise margin

Figure 5.16: Comparison between conventional arbiter, memory and the proposed charge sharing PUFs.

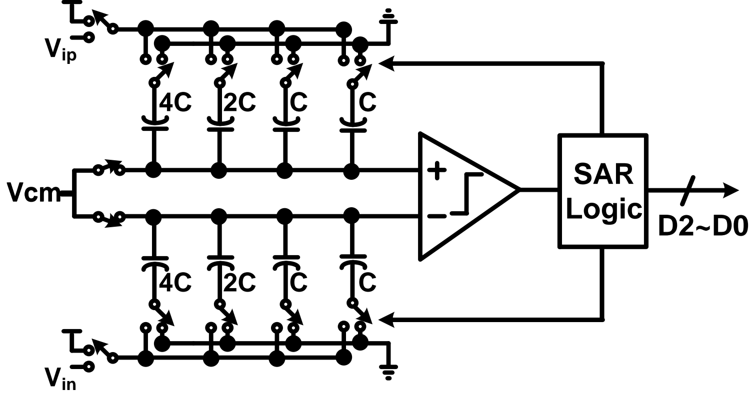
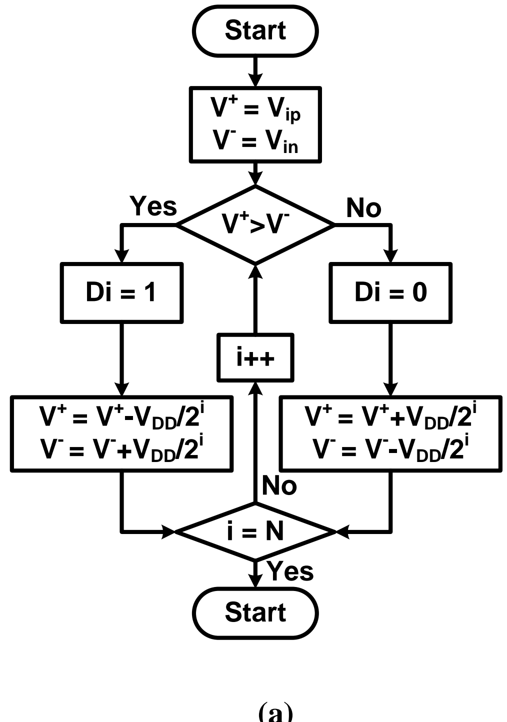


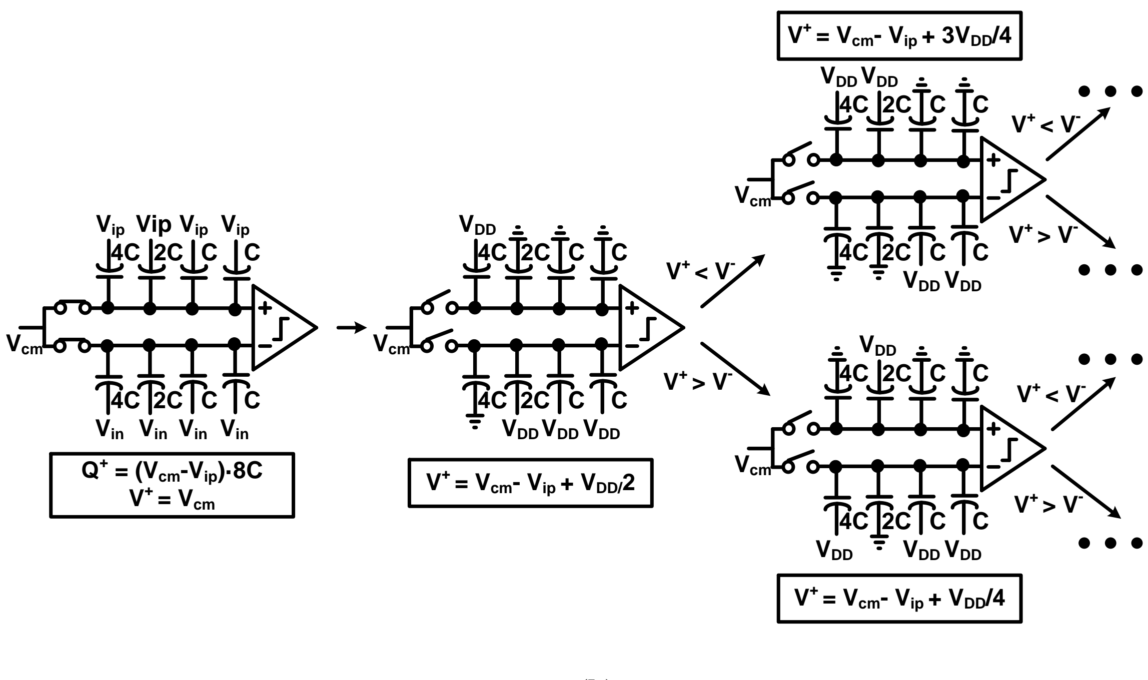
Figure 5.17: 3-bit SAR ADC architecture.

For a simple illustration, Figure 5.17 shows the circuit schematic of a 3-bit single ended SAR ADC. It consists of binary weighted capacitor arrays, a comparator and the successive approximation logic. The capacitor array first samples and stores the input

voltages as charges. These charges will remain constant throughout a sampling cycle however the top plate voltage (V^+ and V^-) will be reconfigured. Depending on the comparator output, which compares V^+ and V^-, the SAR logic determines the corresponding digital code and bottom plate switch connection for the next conversion step. The final goal is to allow V^+ and V approach each other. As the name implies, the SAR logic basically implements a binary search algorithm as shown in Figure 5.18(a). To start, capacitor array samples the differential input voltage V_{in} and V_{ip} on V^+ and V^- respectively. In each bit cycle, if V^+ is greater than V^-, the comparator outputs a logic high and subtracts $V_{DD}/2$ from V^+(add $V_{DD}/2$ to V^-), and vice versa. The SAR control logic then moves to the next bit down, forces that bit high, and does another comparison. The sequence continues until the LSB and the N-bit digital code is available in the SAR logic output. Figure 5.18 (b) shows an example of the switching procedures for a 3-bit SAR ADC. The charge redistribution is achieved by reconfiguring DAC arrays bottom plate connections.



(a)



(b)

Figure 5.18: (a) SAR algorithm; (b) DAC switching procedures [75].

Figure 5.19 shows a transfer function of a 3bit ADC under ideal and real case. The analog input is linearly mapped to the digital output. For ideal DAC array, as indicated with the dotted line, the minimum resolvable steps are uniform and all equal to $V_{DD}/2^N$. However with the presence of capacitor mismatch, the transfer curve is distorted. Random capacitor mismatch is generally induced by the process variation and from the irregular layout. Capacitor mismatch changes the DAC array bit-weight and further shifting the decision level, as shown with the dotted line. Although the nonlinearity of capacitor array is unfavorable in a normal SAR ADC operation, it can be utilized for a potential PUF application.

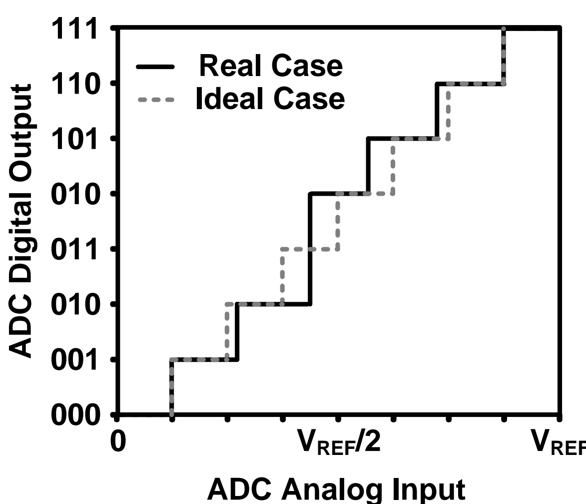


Figure 5.19: Transfer curve for a 3-bit SAR ADC with and without capacitor mismatch.

5.4.2. Charge-redistribution PUF Design

Typically the capacitor array is implemented with unit-element metal-insulator-metal (MIM) or metal-oxide-metal (MOM) capacitors as shown in Figure 5.20 [76], for

example C9 is comprised of 2^9 unit capacitors. In order to maximize capacitor mismatch, the unit-element capacitors are utilized as the PUF elements. This is because the standard deviation of a capacitor σ(C) is inversely proportional to the square root of the occupied area: $\boldsymbol{\sigma}(\boldsymbol{C}) \propto 1/\sqrt{\boldsymbol{W} \cdot \boldsymbol{L}}$. Using a smaller capacitor is able to provide a larger variability. Here in this design, from each DAC array, 63 unit capacitors from the top row are utilized for PUF operation.

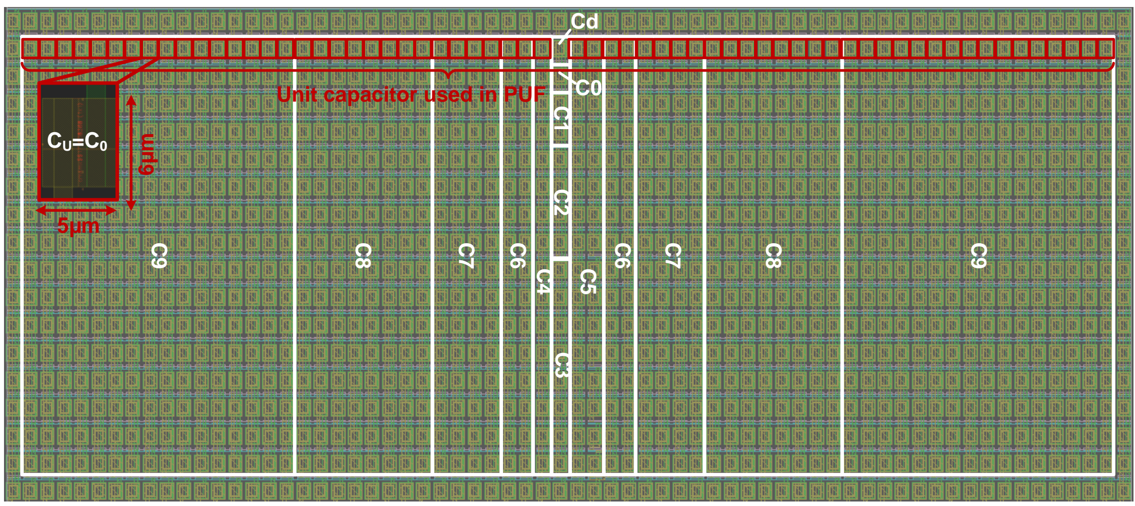


Figure 5.20: Capacitor array layout for a 10-bit SAR ADC [76].

The circuit implementation and timing diagram for the PUF operation is shown in Figure 5.21(a). In additional to the SAR ADC blocks, we implemented a PUF switch control to independently control each unit capacitors and an on-chip counter to collect the PUF soft response. To minimize the comparator offset, we adopted the auto-zero offset cancellation technique [77]. The auto-zero comparator consists of three-stage pre-amplifier and a latch based output structure. During auto-zeroing phase, the input of each

pre-amplifier is shorted and therefore the offset of each stage is sampled and stored on the output capacitors.

The PUF operation is shown Figure 5.21(b). At the rising edge of the PUF_EN signal, V^+ and V^- are initialized to V_{cm}. The first two clock cycles are utilized for voltage sampling and auto-zeroing. During this period, two unit capacitors are enabled and connected in serial between V_{DD} and GND, and the unselected capacitors are left floating. In the first two CLK_CMP cycles, the charge on top plates is:

$$\boldsymbol{Q} = \boldsymbol{V_{cm}} \cdot \boldsymbol{C_{U0}} + (\boldsymbol{V_{cm}} - \boldsymbol{V_{DD}}) \cdot \boldsymbol{C_{U1}}$$

At the third CLK_CMP rising edge, the bottom connections of the two capacitors are swapped, and the top plate voltage, V^+ for example, now is determined by:

$$\boldsymbol{Q^+_{clk1}} = \boldsymbol{Q^+_{clk3}} \Rightarrow \boldsymbol{V^+} = \boldsymbol{V_{CM}} - \frac{\boldsymbol{C^+_{U0}} - \boldsymbol{C^+_{U1}}}{\boldsymbol{C^+_{U0}} + \boldsymbol{C^+_{U1}}} \boldsymbol{V_{DD}}$$

$$\Delta \boldsymbol{V} = \boldsymbol{V^+} - \boldsymbol{V^-} = \left(\frac{\boldsymbol{C^-_{U0}} - \boldsymbol{C^-_{U1}}}{\boldsymbol{C^-_{U0}} + \boldsymbol{C^-_{U1}}} - \frac{\boldsymbol{C^+_{U0}} - \boldsymbol{C^+_{U1}}}{\boldsymbol{C^+_{U0}} + \boldsymbol{C^+_{U1}}} \right) \cdot \boldsymbol{V_{DD}}$$

The capacitance difference between C_{U0} and C_{U1} results in different top plate voltages (ΔV). Depending on the polarity of ΔV, comparator outputs a logic ‘0’ or ‘1’. In the PUF operation, the challenge can be the location of the enabled unit capacitors and the response is the comparator output.

Due to the fact that PUF randomness relies on manufacture variability, which can be very small with some given challenges, there is no guarantee that all challenges will always generate the same response. Therefore, error-correction techniques are employed to improve the PUF reliability. In this charge-redistribution based PUF, a soft-response

based error-correction technique is adopted [55][77] which takes advantage of the probability of a given response-bit instead of its instantaneous output. This is accomplished by repetitively evaluating a given challenge (asserting PUF_EN multiple times) and counting the number of '1's read from the comparator output. The ratio between the count value and the number of evaluation cycles is the probability of the response being '1'.

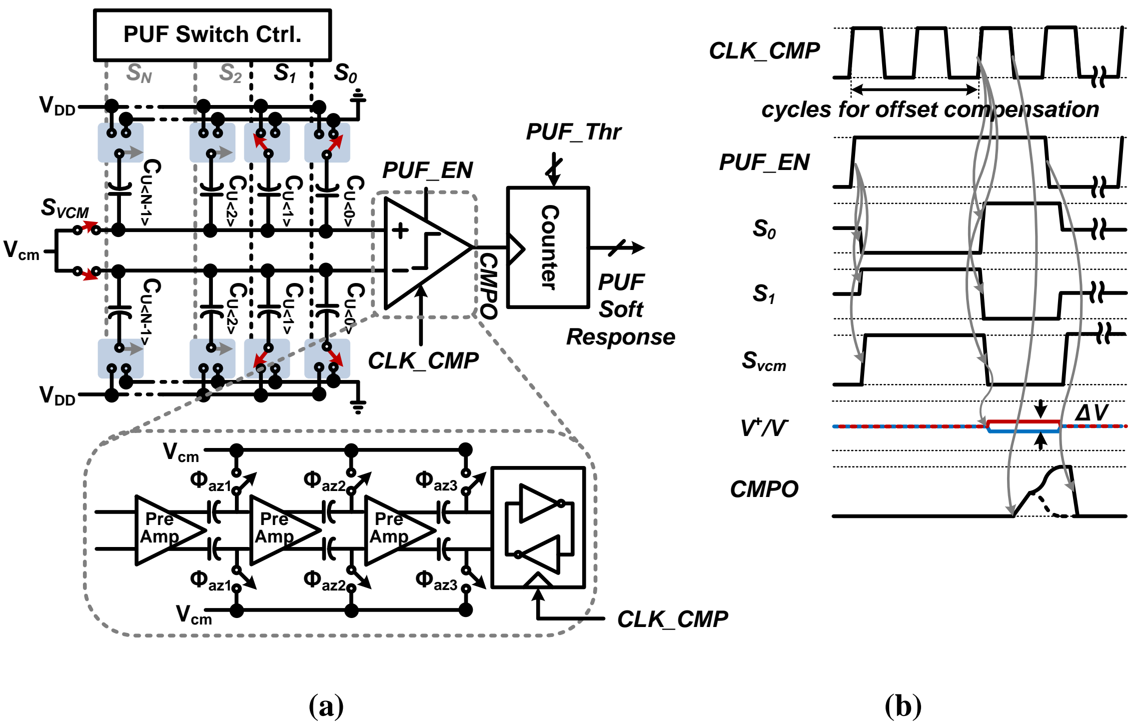


Figure 5.21: (a) Schematic and (b) timing diagram of the proposed charge redistribution PUF.

The number of CRPs that one can obtained from the charge-redistribution based PUF is determined by the total number of available unit capacitors N and the number of enabled unit capacitors k. This is equivalent to randomly choosing k unit capacitors from an N unit capacitor array which is equal to C(N, k), C(.) is the combination function. For

example, in a scenario as shown Figure 5.21, the total number of CRPs is C(63, 2) = 1953. In order to further increase the number of CRPs, we can increase the number of enabled unit capacitors for each comparison. The maximum number of CRPs is obtained when half unit capacitors are enabled.

5.4.3. Test Chip Measurement Result

The soft response obtained from a single charge-redistribution based PUF is shown in Figure 5.22(a) which evaluates 8000 CRPs for 100 times. A soft response equal to 0 or 1 means that the comparator always outputs a '0' or '1' in 100 evaluation cycles. This implies that a large voltage difference is produced from the two capacitor arrays. On the other hand, challenges with soft response close to 0.5 indicate a relative small voltage difference which is largely affected by the random noise. Generally, the soft response is converted to a digital bit by thresholding at 0.5, that is, a soft response greater than 0.5 will be taken as a '1' and vice versa. However, a response recognized as a '1' during the enrollment may flip to '0' during the authentication (e.g. soft response change from 0.51 to 0.49). In order to improve PUF reliability, a dynamic thresholding scheme is applied [55] which utilizes different decision thresholds for the enrollment and authentication. The basic idea is to set a stringent decision threshold during the enrollment, e.g. 0.1~0.9, to identify the stable and unstable CRPs. Unstable CRPs will be discard. A relaxed decision threshold, e.g. 0.5, will be applied during the authentication. This method improves the response stability by reducing the '1'-to-'0' and '0'-to-'1' flipping probability. For example, a soft response with a value of 0.9 is less likely to fall below

the 0.5 threshold as compared to one with a value of 0.51. The PUF reliability with different enrollment threshold is verified by checking the intra-chip HD as shown in Figure 5.23 (a). Both the mean and standard deviation of the intra-chip HD decreases with a smaller enrollment threshold. Although utilizing a more stringent enrollment threshold can improve the PUF reliability, the number of discard CRPs also increases. The percentage of discard CRPs with respect to different enrollment threshold is shown in Figure 5.23(b). The number of stable CRPs increases from 50.6% to 81.7% as the enrollment threshold is relaxed from 0 to 0.2. In our experiment, a 0.1 enrollment threshold is chosen for an optimization of the PUF reliability and number of discard CRPs.

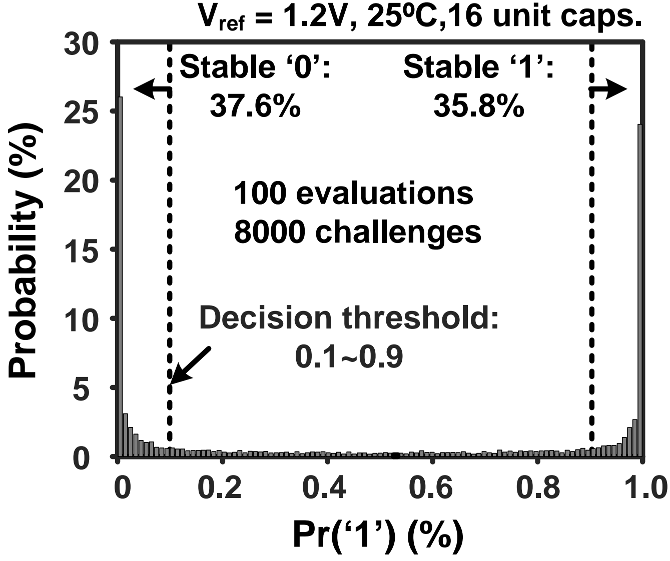
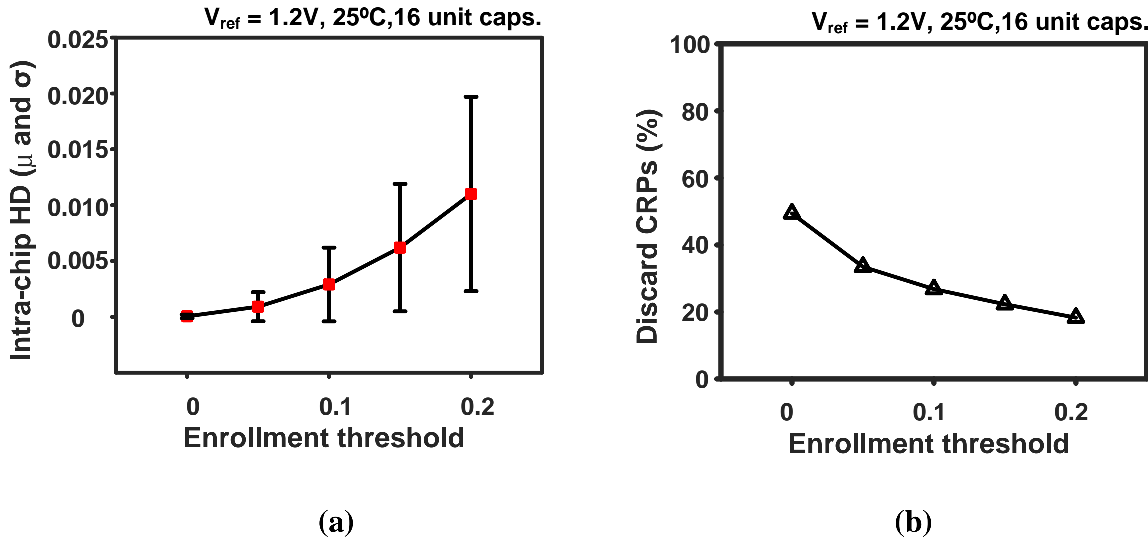


Figure 5.22: Measured soft response distribution for the charge redistribution PUF.

The percentage of discard CRPs also depends on the number of enabled unit capacitors. As shown in Figure 5.24, the percentage of discard CRPs first decreases as the

number of enabled unit capacitors is increased and gradually saturated when more than 16 unit capacitors are enable. For a practical authentication purpose, at least 16 unit capacitors should be enabled in order to provide sufficient number of CRPs. Therefore the rest part of our demonstrations are based on the measurement with 16 unit capacitors enabled.



(a) (b)

Figure 5.23: (a) Intra-chip HD and (b) percentage of discard CRPs for different enrollment threshold.

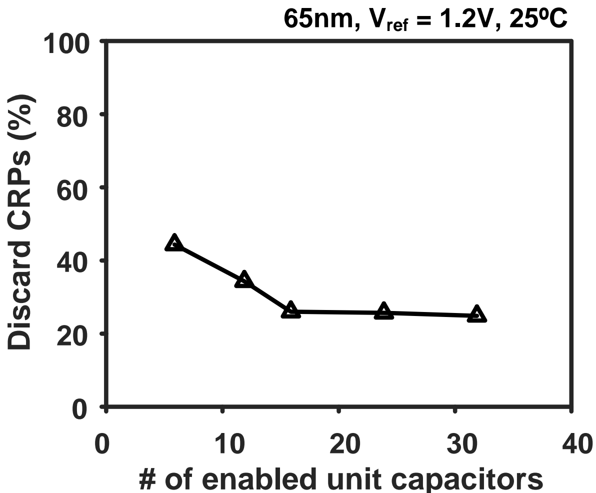


Figure 5.24: Percentage of discard CRPs for different # of enabled unit capacitors.

As lightweight PUFs are distributed and physically exposed to different environments, it should be designed with some resilience to variations such as voltage and temperature drifts. Furthermore, invasive attacks such as side channel attacks can disturb the PUF responses by altering the voltages. Therefore maintaining the PUF stability across different supply voltages is of crucial importance. One benefit of using the passive capacitor as the entropy source is that the change of the "random signature" (e.g. ΔV of the two capacitor array) is proportional to the change of supply voltages [79][80]. This linear scaling feature implies that a stable response at lower supply voltage will become more stable at a higher supply voltage as ΔV is enlarged. By performing an enrollment at a lower supply voltage will guarantee a stable authentication for a wider operation range. To verify this, we evaluated the PUF intra-chip Hamming Distance for a supply voltage from 0.8V to 1.2V with different enrollment voltage as shown in Figure 5.25 (a). The evaluation is performed as follow: 1) obtain soft responses at five supply voltages from

0.8V to 1.2V and determine the valid CRPs for each voltage condition based on the 0.1 enrollment threshold; 2) apply the same challenge set again for multiple times with a varied supply voltage ranging from 0.8V to 1.2V, response bits are determined by the 0.5 authentication threshold; 3) calculate the intra-chip Hamming Distance by comparing responses obtained from 2) and 1), only the valid CRPs will be evaluated. The intra-chip Hamming distance average and sigma values are only 0.56% and 0.6% when the PUF is enrolled at 0.8V whereas they are 9.8% and 5.3% when the PUF is enrolled at 1.2V. It is worth noting though that the intra-chip HD significantly improves with a lower enrollment voltage, the percentage of discard CRPs also increases as shown in Figure 5.25(b). This can attribute to the decreased voltage difference on the two capacitor array top plates with a lower supply voltage.

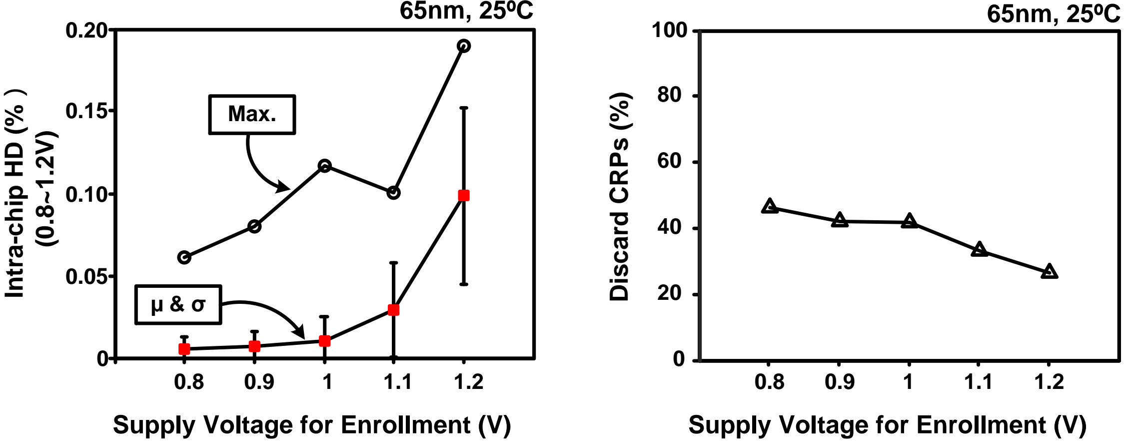
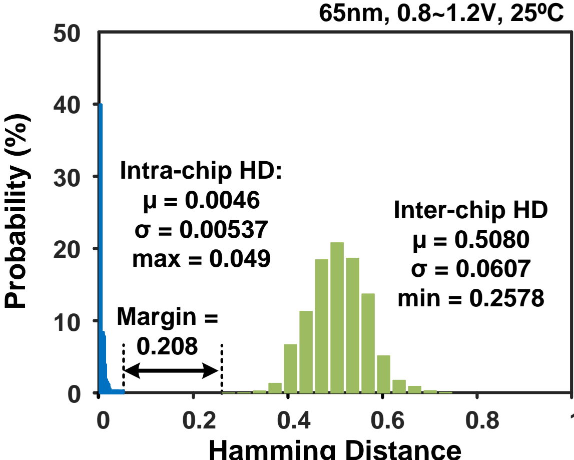


Figure 5.25: (a) Measured mean, standard deviation and maximum values of intra-chip Hamming Distance with a V_{ref} from 0.8~1.2V. A better intra-chip HD is obtained when the enrollment is performed at lower supply voltage. (b) More CRPs are discard at a lower V_{ref} during chip enrollment.

The inter-chip Hamming Distance distributions measured from 10 chips are shown in Figure 5.26. Half of the 80,000 CRPs are discard during enrollment, therefore finally 40,000 CRPs are applied to 10 different chips with a supply voltage ranging from 0.8V to 1.2V. Outputs are grouped into 128-bit responses (312×128bit) in order to guarantee a sufficient bit-stream length to against attacks such as random guessing [56]. The inter-chip Hamming Distance shows a distribution with an average value of 50.6% which is very close to the ideal case (50%) indicating that responses from different PUFs are sufficiently uncorrelated. The minimum inter-chip Hamming Distance and maximum intra-chip Hamming Distance are separated with a margin of 20.8% suggesting a secure authentication can be achieved. Finally, the 65nm test chip die photo is shown in Figure 5.27, the charge-redistribution based PUF is implemented using the existing blocks from a 10 bit SAR ADC.



	Intra-chip	Inter-chip
Threshold	Pr ('1')<0.1 or Pr('1')>0.9	
# of challenges	8k	80k
Discard response	52.2%	
Enrollment range	0.8~1.2V	
# of cycles for eval.	990×10	10
# chips	1	10

Figure 5.26: Measured inter-chip and intra-chip Hamming Distance distributions and evaluation conditions.

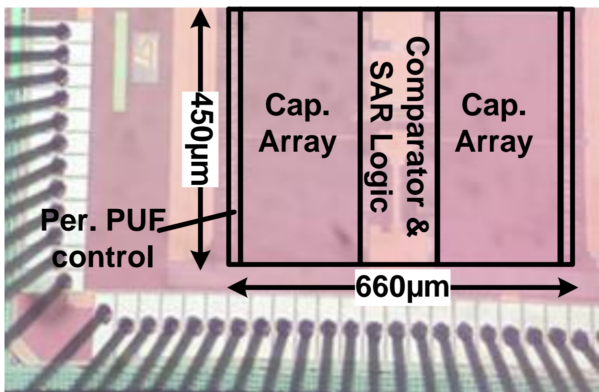


Figure 5.27: 65nm charge-redistribution based PUF chip micrograph.

5.5. Conclusion

This chapter has presented two novel strong PUFs, which are implemented based on the existing circuit blocks DRAM and SAR ADC respectively.

Firstly, we have demonstrated a DRAM PUF utilizing the location of weak retention cells. The proposed PUF can generate more than 10^{32} CRPs from a 1Kbit DRAM array. To improve the consistency of the PUF response, we employed a repetitive write-back scheme along with bit-masking. Intra-chip and inter-chip Hamming distance distributions were measured from a 65nm chip under different supply voltages and temperatures. A calibration routine performed before each authentication operation has shown to effectively suppress voltage and temperature induced instabilities.

Next, we have presented a charge-redistribution PUF that leverages the capacitance mismatch in the SAR ADC capacitor array as the random entropy source. The circuit can work both as a PUF and a SAR ADC circuit by adding a minimal hardware overhead

making it suitable for a resource constraint device. A soft response based dynamic thresholding method is employed to improve the response reliability. Measurement data collected from test chip fabricated in a 65nm process shows an average intra-chip HD of 0.0046 and an average inter-chip HD of 0.508 with a supply voltage ranging from 0.8V to 1.2V. The margin between the maximum inter-chip HD and the minimum intra-chip HD is 0.208 implying a good uniqueness for secure authentication.

Chapter 6. Conclusion

CMOS scaling driven by the need for higher performance has brought in greater process variations, making transistors less reliable. To ensure the circuit performance and functionalities over its intended lifetime, different reliability issues must be characterized and analyzed even before the design stage. This dissertation presents several on-chip monitoring circuits for accurate and efficient collections of the reliability statistics.

In Chapter 2, we have implemented two ring oscillator based monitoring circuit for a realistic characterization of the RTN impact on logic circuit with two generations of process technology. The first RTN sensing scheme was demonstrated in a 65nm LP process which achieves a high frequency measurement resolution (>0.01%) at a short sampling time (>1 μs). Experimental data from a ROSC array displays both single trap and multi-trap RTN behavior. The voltage dependencies of the frequency shift and capture/emission times were measured and analyzed. In order to collect high quality RTN data at a near threshold operation voltage, a dual ring oscillator array test structure has been implemented in a 32nm HKMG process which improves the frequency measurement resolutions of the tested-and-proven beat frequency detection (BFD) technique. RTN induced frequency shifts at different supply voltages, temperatures and stress conditions were measured from the test chips. The impact of RTN on logic and SRAM performance was analyzed based on the measured RTN data. We also present quantitative results of logic timing margin and SRAM noise margin, with and without

RTN. According to this study, RTN appears to have a modest 1% impact on circuit operating frequency in 32nm process, even under pessimistic conditions (i.e. Vdd=0.6V, multiple RTN traps in circuit path).

In Chapter 3, an on-chip compact 2T sensor structure for detecting the radiation induced single event effects has been demonstrated. Measured bit-flip data from a 1Kbit sensor array implemented on a 65nm bulk process has shown a higher measurement sensitivity as compared to 6T SRAM cell under an accelerated alpha particle irradiation. Simulation results also verified that Qcrit of the proposed 2T sensor cell is 17x-60x smaller than that of an inverter chain and SRAM cell.

While the reliability communities are urged to seek for techniques to mitigate the circuit performance instabilities, the security community is trying to extract and utilize these variations for possible hardware-oriented security applications. In this dissertation, we have demonstrated novel hardware building blocks (TRNG and PUFs) that leverage the environmental noise and manufacturing process variation as the entropy sources.

In Chapter 4, we have presented a fully-digital TRNG which measures the frequency difference between two free-running ring oscillators to sample the random frequency jitter. The proposed circuit fabricated in 65nm achieves an energy efficiency of 15.1Mb/mW at 0.8V. Measured data from a batch of TRNG test chips passes all NIST test suites without any feedback scheme for a wide range of operating voltages and temperatures.

In Chapter 5, we have proposed two lightweight PUFs that are based on existing circuit blocks. The first PUF is based on a logic compatible 2T DRAM in a 65nm CMOS. Compared to previous SRAM or DRAM based PUFs, the proposed PUF allows a direct chip authentication by supporting $>10^{32}$ possible challenge response pairs (CRPs) per 1Kbit array using a novel self-encrypting authentication scheme. Hardware data shows a 0.0039 inter-chip HD by utilizing a zero-overhead repetitive write-back technique together with bit-masking. The second PUF employs the capacitor mismatch in the SAR ADC circuit and extract the unique signature with the charge-redistribution operation. Measurement data collected from test chip fabricated in a 65nm process shows an average intra-chip HD of 0.0046 and an average inter-chip HD of 0.508 with a supply voltage ranging from 0.8V to 1.2V. The inter-chip HD and the intra-chip HD is sufficiently separated with a margin of 0.208 showing a good uniqueness for secure authentication.

References

[1] M J Kirtont, M J Urent, S Collinst, et al., "Individual Defects At the Si:SiO_2 Interface," Semiconductor Sci. Technol., 4, pp. 1116-1126, 1989.

[2] A. P. van der Wel, E. A. M. Klumperink, L. K. J. Vandamme, et al. "Modeling random telegraph noise under switched bias conditions using cyclostationary RTS noise." Tran. on Electron Devices, Vol. 50, No. 5, pp.1378-1384, 2003.

[3] H. Miki, M. Yamaoka, N. Tega, et al., "Understanding short-term BTI behavior through comprehensive observation of gate-voltage dependence of RTN in highly scaled high-κ / metal-gate pFETs," Symposium on VLSI Technology, pp. 148-149, 2011.

[4] T. Grasser, K. Rott, H. Reisinger, et al., "A unified perspective of RTN and BTI," International Reliability Physics Symposium, pp. 4A.5.1-4A.5.7, 2014.

[5] H. Miki, M. Yamaoka, D. J. Frank, et al., "Voltage and temperature dependence of random telegraph noise in highly scaled HKMG ETSOI nFETs and its impact on logic delay uncertainty," Symposium on VLSI Technology, pp. 137-138, 2012.

[6] J. Chen, Y. Higashi, I. Hirano, et al., "Experimental study of channel doping concentration impacts on random telegraph signal noise and successful noise suppression by strain induced mobility enhancement," Symposium on VLSI Technology, pp. T184-T185, 2013.

[7] K. Ito, T. Matsumoto, S. Nishizawa, H. Sunagawa, K. Kobayashi and H. Onodera, "The impact of RTN on performance fluctuation in CMOS logic circuits," International Reliability Physics Symposium, pp. CR.5.1-CR.5.4, 2011.

[8] T. Matsumoto, K. Kobayashi and H. Onodera, "Impact of Random Telegraph Noise on CMOS Logic Delay Uncertainty Evaluated with 12,600 Ring Oscillators," International Electron Devices Meeting, 2012, pp. 25.6.1-25.6.4.

[9] S. Realov and K. L. Shepard, "Random telegraph noise in 45-nm CMOS: Analysis using an on-chip test and measurement system," International Electron Devices Meeting, 2010, pp. 28.2.1-28.2.4.

[10] M. Luo, R. Wang, S. Guo, et al., "Impacts of Random Telegraph Noise (RTN) on Digital Circuits," Transactions on Electron Devices, vol. 62, no. 6, pp. 1725-1732, June 2015.

[11] M. L. Fan, V. P. H. Hu, Y. N. Chen, et al., "Analysis of Single-Trap-Induced Random Telegraph Noise on FinFET Devices, 6T SRAM Cell, and Logic Circuits," Transactions on Electron Devices, vol. 59, no. 8, pp. 2227-2234, Aug. 2012.

[12] X. Chen, Y. Wang, Y. Cao, et al., "Statistical analysis of random telegraph noise in digital circuits," Asia and South Pacific Design Automation Conference, pp. 161-166, 2014.

[13] T. Kim, R. Persaud, and C.H. Kim, "Silicon Odometer: An On-Chip Reliability Monitor for Measuring Frequency Degradation of Digital Circuits", Journal of Solid-State Circuits, Vol. 43, No. 4, pp. 874-880, Apr. 2008.

[14] T. Nagumo, K. Takeuchi, S. Yokogawa, K. Imai and Y. Hayashi, "New analysis methods for comprehensive understanding of Random Telegraph Noise," International Electron Devices Meeting, pp. 1-4, 2009.

[15] S. Lee, H. Cho, Y. Son, et al., "Characterization of oxide traps leading to RTN in high-k and metal gate MOSFETs," International Electron Devices Meeting, pp. 1-4, 2009.

[16] T. Grasser, Th. Aichinger, G. Pobegen, et al, "The ‘permanent’ component of NBTI: Composition and annealing," International Reliability Physics Symposium, pp. 6A.2.1-6A.2.9, 2011.

[17] L. Wang and K. Skadron, "Implications of the Power Wall: Dim Cores and Reconfigurable Logic," Micro, vol. 33, no. 5, pp. 40-48, Sept.-Oct. 2013.

[18] I. Wey, P. Lin, B. Wu, et al., "Near-threshold-voltage circuit design: The design challenges and chances," International SoC Design Conference, pp. 138-141, 2014.

[19] N. Tega, H. Miki, F. Pagette "Increasing threshold voltage variation due to random telegraph noise in FETs as gate lengths scale to 20 nm," Symposium on VLSI Technology, pp. 50-51, 2009.

[20] R. C. Baumann, "Radiation-induced soft errors in advanced semiconductor technologies," Transactions on Device and Materials Reliability, vol. 5, no. 3, pp. 305-316, Sept. 2005.

[21] C. Slayman, "Soft error trends and mitigation techniques in memory devices," Annual Reliability and Maintainability Symposium, pp. 1-5, 2011.

[22] A. Dixit and A. Wood, "The impact of new technology on soft error rates," International Reliability Physics Symposium, pp. 5B.4.1-5B.4.7, 2011.

[23] N. Tam, B. L. Bhuva and L. W. Massengill, "Multi-cell soft errors at the 16-nm FinFET technology node," International Reliability Physics Symposium, pp. 4B.3.1-4B.3.5, 2015.

[24] N. N. Mahatme, N. J. Gaspard, T. Assis, et al., "Impact of technology scaling on the combinational logic soft error rate," International Reliability Physics Symposium, pp. 5F.2.1-5F.2.6, 2014.

[25] S. Lee, I. Kim, S. Ha, et al., "Radiation-induced soft error rate analyses for 14 nm FinFET SRAM devices," International Reliability Physics Symposium, pp. 4B.1.1-4B.1.4, 2015.

[26] B. Doucin, C. Poivey, C. Carlotti, et al., "Study of radiation effects on low voltage memories," European Conference on Radiation and its Effects on Components and Systems, pp. 561-569, 1997.

[27] D. F. Heidel, K. P. Rodbell, E. H. Cannon, et al., "Alpha-particle-induced upsets in advanced CMOS circuits and technology," IBM Journal of Research and Development, vol. 52, no. 3, pp. 225-232, May 2008.

[28] "PSTAR and ASTAR Databases for Protons and Helium Ions," Available: http://physics.nist.gov/PhysRefData/Star/Text/programs.html.

[29] L.B. Freeman, “Critical charge calculations for a bipolar SRAM array”, IBM J. Res. Dev., Vol. 40, no. 1, pp. 77-89, Jan. 1996.

[30] Roche, P. et al.; “Determination of key parameters for SEU occurrence using 3-D full cell SRAM simulations,” Transactions on Nuclear Science, Vol 46, Issue 6, pp. 1354 – 1362, Dec. 1999.

[31] T. Heijmen, D. Giot, P. Roche; “Factors That Impact the Critical Charge of Memory Elements”. On-Line Testing Symposium (IOLTS), pp.:57-62, 2006.

[32] Q. Zhou and K. Mohanram, "Gate sizing to radiation harden combinational logic," Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 25, no. 1, pp. 155-166, Jan. 2006.

[33] G. R. Srinivasan, P. C. Murley and H. K. Tang, "Accurate, predictive modeling of soft error rate due to cosmic rays and chip alpha radiation," International Reliability Physics Symposium, USA, pp. 12-16, 1994.

[34] R. Naseer, Y. Boulghassoul, J. Draper, et al., "Critical Charge Characterization for Soft Error Rate Modeling in 90nm SRAM," International Symposium on Circuits and Systems, pp. 1879-1882, 2007.

[35] F. Wrobel, L. Dilillo, A. D. Touboul, et al., "Determining Realistic Parameters for the Double Exponential Law that Models Transient Current Pulses," Transactions on Nuclear Science, vol. 61, no. 4, pp. 1813-1818, Aug. 2014.

[36] A. KleinOsowski, P. Oldiges, R. Q. Williams and P. M. Solomon, "Modeling Single-Event Upsets in 65-nm Silicon-on-Insulator Semiconductor Devices," Transactions on Nuclear Science, vol. 53, no. 6, pp. 3321-3328, Dec. 2006.

[37] R. Brederlow, R. Prakash, C. Paulus, et al., "A low-power true random number generator using random telegraph noise of single oxide-traps," International Solid State Circuits Conference, pp. 1666-1675, 2006.

[38] C. Tokunaga, D. Blaauw and T. Mudge, "True Random Number Generator with a Metastability-Based Quality Control," Journal of Solid-State Circuits, vol. 43, no. 1, pp. 78-85, Jan. 2008.

[39] S. Srinivasan, S. Mathew, R. Ramanarayanan, et al., "2.4GHz 7mW all-digital PVT-variation tolerant True Random Number Generator in 45nm CMOS," Symposium on VLSI Circuits, 2010, pp. 203-204.

[40] K. Yang, D. Fick, M. B. Henry, et al., "16.3 A 23Mb/s 23pJ/b fully synthesized true-random-number generator in 28nm and 65nm CMOS," International Solid-State Circuits Conference, pp. 280-281, 2014.

[41] J. S. Liberty, A. Barrera, D. W. Boerstler, T. B. Chadwick, S. R. Cottier, H. P. Hofstee, J. A. Rosser and M. L. Tsai, “True hardware random number generation

implemented in the 32-nm SOI POWER7+ processor," IBM Journal of Research and Development, Vol. 57, No. 6, pp. 4:1-4:7, Nov. 2013

[42] Y. Lao, Q. Tang, C. H. Kim, et al., "Beat Frequency Detector--Based High-Speed True Random Number Generators: Statistical Modeling and Analysis," Journal on Emerging Technologies in Computing Systems, vol. 13, no. 9, Nov. 2003.

[43] M. Bucci, L. Germani, R. Luzzi, et al., "A high-speed oscillator-based truly random number source for cryptographic applications on a smart card IC," Transactions on Computers, vol. 52, no. 4, pp. 403-409, April 2003.

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

[45] Rukhin, A. et al. "A statistical test suite for random and pseudorandom number generators for cryptographic applications," Special Publication 800–22 Revision 1a, 2010. Available at <http://csrc.nist.gov/publications/PubsSPs.html>.

[46] V. B. Suresh, D. Antonioli and W. P. Burleson, "On-chip lightweight implementation of reduced NIST randomness test suite," International Symposium on Hardware-Oriented Security and Trust, pp. 93-98, 2013.

[47] C. S. Petrie and J. A. Connelly, "A noise-based IC random number generator for applications in cryptography," Transactions on Circuits and Systems I: Fundamental Theory and Applications, 47(5):615–621, 2000.

[48] W. Schindler. "A stochastical model and its analysis for a physical random number generator" Cryptographic Hardware and Embedded Systems, pp. 276–289. 2003.

[49] B. J. Abcunas. Evaluation of random number generators on FPGAs. PhD thesis, Worcester Polytechnic Institute, 2004.

[50] Y. Lao, "Authentication and Obfuscation of Digital Signal Processing Integrated Circuits," PhD Thesis, University of Minnesota, 2015.

[51] R. Pappu, "Physical One-Way Functions," PhD Thesis, MIT, 2001.

[52] Gassend, B., Clarke, D.E., van Dijk, et al., "Silicon physical random functions". Computer and Communications Security (CCS), pp. 148~160, Nov. 2002.

[53] G. Hammouri, E. Ozturk, B. Sunar: A tamper-proof and lightweight authentication scheme. Journal Pervasiveand Mobile Computing, vol. 6, no. 4, 2008.

[54] M. Katagi and S. Moriai, “Lightweight cryptography for the internet of things,” Sony Corporation, Tokyo, Japan, Tech. Rep., 2011. Available: http://www.iab.org/wp-content/IABuploads/2011/03/Kaftan.pdf

[55] C. Zhou, S. Satapathy, Y. Lao, K. Parhi, and C.H. Kim, "Soft Response Generation and Thresholding Strategies for Linear and Feedforward MUX based PUFs," International Symposium on Low Power Electronics and Design, Aug. 2016

[56] J. Delvaux, D. Gu, D. Schellekens, et al., "Secure lightweight entity authentication with strong pufs: Mission impossible?," Cryptographic Hardware and Embedded Systems, pp. 451-475, 2014.

[57] Rolf H. Weber, “Internet of Things – New security and privacy challenges,” Computer Law & Security Review, Vol. 26, No. 1, pp. 23-30, , Jan. 2010.

[58] S. Babar, A. Stango, N. Prasad, J. Sen and R. Prasad, "Proposed embedded security framework for Internet of Things (IoT)," Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic Systems Technology, pp. 1-5, 2011.

[59] Z. K. Zhang, M. C. Y. Cho, C. W. Wang, C. W. Hsu, C. K. Chen and S. Shieh, "IoT Security: Ongoing Challenges and Research Opportunities," International Conference on Service-Oriented Computing and Applications, pp. 230-234, 2014.

[60] C. Herder, M. D. Yu, F. Koushanfar and S. Devadas, "Physical Unclonable Functions and Applications: A Tutorial," Proceedings of the IEEE, vol. 102, no. 8, pp. 1126-1141, Aug. 2014.

[61] A. Ukil, J. Sen and S. Koilakonda, "Embedded security for Internet of Things," Emerging Trends and Applications in Computer Science (NCETACS), pp. 1-6, 2011.

[62] S. L. Keoh, S. S. Kumar and H. Tschofenig, "Securing the Internet of Things: A Standardization Perspective," Internet of Things Journal, vol. 1, no. 3, pp. 265-275, June 2014.

[63] Schellekens, Dries, Pim Tuyls, and Bart Preneel. "Embedded trusted computing with authenticated non-volatile memory." Trusted Computing-Challenges and Applications, pp. 60-74, 2008.

[64] "PUF-Physical Unclonable Functions-Protecting Next-Generation Smart Card ICs with SRAM-based PUFs." NXP Semiconductor N.V. Feb. 2013. Available: http://www.nxp.com/documents/other/75017366.pdf

[65] G. E. Suh and S. Devadas, "Physical Unclonable Functions for Device Authentication and Secret Key Generation," ACM/IEEE Design Automation Conference, pp. 9-14, 2007.

[66] Sami Rosenblatt, Daniel Fainstein, Alberto Cestero, John Safran, Norman Robson, Toshiaki Kirihata, and Subramanian S. Iyer, "Field Tolerant Dynamic Intrinsic Chip ID Using 32 nm High-K/Metal Gate SOI Embedded DRAM," Journal of Solid-State Circuits, vol. 48, no. 4, pp. 940-947, Apr. 2013.

[67] K. C. Chun, W. Zhang, P. Jain and C. H. Kim, "A 2T1C Embedded DRAM Macro With No Boosted Supplies Featuring a 7T SRAM Based Repair and a Cell Storage Monitor," Journal of Solid-State Circuits, vol. 47, no. 10, pp. 2517-2526, Oct. 2012.

[68] J. W. Lee, Daihyun Lim, B. Gassend, G. E. Suh, M. van Dijk and S. Devadas, "A technique to build a secret key in integrated circuits for identification and authentication applications," Symposia on VLSI Circuits, pp. 176-179, 2004.

[69] S. K. Mathew, Sudhir K. Satpathy; Mark A. Anders; Himanshu Kaul; Steven K. Hsu; Amit Agarwal; Gregory K. Chen; Rachael J. Parker; Ram K. Krishnamurthy; Vivek De, "16.2 A 0.19pJ/b PVT-variation-tolerant hybrid physically unclonable function circuit for 100% stable secure key generation in 22nm CMOS," International Solid-State Circuits Conference, pp. 278-279, 2014.

[70] M. Bhargava, et al., "Comparison of bi-stable and delay-based Physical Unclonable Functions from measurements in 65nm bulk CMOS," Custom Integrated Circuits Conference, pp. 1-4, 2012.

[71] Jiangyi Li, Mingoo Seok, “A 3.07μm^2/bitcell Physically Unclonable Function with 3.5% and 1% Bit-Instability across 0 to 80 ℃ and 0.6 to 1.2V in a 65nm CMOS” Symposia on VLSI Circuits, pp. C250-C251, 2015.

[72] A. Maiti, L. McDougall and P. Schaumont, "The Impact of Aging on an FPGA-Based Physical Unclonable Function," International Conference on Field Programmable Logic and Applications, pp. 151-156, 2011.

[73] Sadr and M. Zolfaghari-Nejad, "Weighted Hamming Distance for PUF Performance Evaluation," Electronics Letters, vol. 49, no. 22, pp. 1376-1378, Oct. 24, 2013.

[74] H. Kang, Y. Hori, T. Katashita, et al., "Performance Evaluation for PUF-based Authentication Systems with Shift Post-processing: Additional Experimental Results," Symposium on Cryptography and Information Security, pp. 1-5, 2011.

[75] C. C. Liu, S. J. Chang, G. Y. Huang, et al., "A 10-bit 50-MS/s SAR ADC With a Monotonic Capacitor Switching Procedure," Journal of Solid-State Circuits, vol. 45, no. 4, pp. 731-740, April 2010.

[76] C. Liu, "Design of High-Speed Energy-Efficient Successive-Approximation Analog-to-Digital Converters," PhD Thesis, National Cheng Kung University, June. 2010.

[77] R. Xu, "High-Speed Highly Sensitive CMOS Image Sensors," PhD Thesis, Hong Kong University of Science and Technology (HKUST), Aug. 2013.

[78] R. Maes, P. Tuyls, and I. Verbauwhede, "Low-overhead implementation of a soft decision helper data algorithm for SRAM PUFs," Cryptographic Hardware and Embedded Systems, pp. 332–347, 2009

[79] M. Wan, Z. He, S. Han, K. Dai, et al., "An Invasive-Attack-Resistant PUF Based On Switched-Capacitor Circuit," Transactions on Circuits and Systems I: Regular Papers, vol. 62, no. 8, pp. 2024-2034, Aug. 2015.

[80] J. Ju, R. Chakraborty, C. Lamech and J. Plusquellic, "Stability analysis of a physical unclonable function based on metal resistance variations," Symposium on Hardware-Oriented Security and Trust, pp. 143-150, 2013.

[81] F. Rahman, D. Forte and M. M. Tehranipoor, "Reliability vs. security: Challenges and opportunities for developing reliable and secure integrated circuits," International Reliability Physics Symposium, pp. 4C.6.1-4C.6.10, 2016.