Design of lightweight buffer-free sram and robust ring oscillator based physical unclonable functions

Liu, Chaoqun

2018

Liu, C. (2018). Design of lightweight buffer-free sram and robust ring oscillator based physical unclonable functions. Master's thesis, Nanyang Technological University, Singapore.

<http://hdl.handle.net/10356/73380>

<https://doi.org/10.32657/10356/73380>

**DESIGN OF LIGHTWEIGHT BUFFER-FREE
SRAM AND ROBUST RING OSCILLATOR
BASED PHYSICAL UNCLONABLE
FUNCTIONS**

LIU CHAOQUN

School of Electrical and Electronic Engineering

A Thesis Submitted to The Nanyang Technological University
in Partial Fulfilment of the Requirements for the Degree of
Master of Engineering

2018

Authorization

I hereby declare that I am the sole author of the thesis.

I authorize the Nanyang Technological Univerisity to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I further authorize the Nanyang Technological Univerisity to reproduce the thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Student: Liu Chaoqun

Acknowledgements

The master program and the research work have been a wonderful and fruitful experience. That enjoyment largely resulted from the interaction that I have had with my supervisors, the lab-mates, the university and the people who have helped me so far.

First of all, I would like to express my sincere thanks and great gratitude to my supervisor, Prof Chang Chip Hong for guiding me through this program. He trusted me and encouraged me to maximize my potential. Besides teaching me the essential skills and knowledges of doing research, he also shared with me the joy of discovery and philosophy of life, which is valuable and cultivated my interest in research. It is also very very generous of him to sponsor me to attend the conference, ISCAS2017, which gave me the opportunity to interact and learn from other postgraduates and researchers.

I would also like to thank my seniors Dr. Zhang Le, Dr. Cao Yuan, Dr. Sachin and Sergey for teaching the knowledges and tools about hardware security. They guided me patiently and helped me solve the problems in this projects. I would also like to thank my other team mates, Zheng Yue, Truan, Zhong Huishu, Wang Si and Zhang Yufei, who inspired and accompanied me during the research. Without them, I would not have so many ideas and achievements.

I also want to thank other postgraduates and research staff, including Wang Dong, Jian Sen, Lou Liheng, Hong Yan, Jiang Jize, Qu Yong, Ding Xiangbin, Liu Qianqian, Zhang Ruochong, Liu Siyu and Yang Chuanshi, who provided valuable suggestions, important information and essential equipments for my projects.

I would like to acknowledge the assistance and technical support from the staff in VIRTUS, the IC Design Center of Excellence. I also appreciate MediaTek Singapore Pte Ltd's financial support for my master program. MediaTek also provided a tape-out opportunity for me, which equipped me with essential skills to be an IC designer.

Last but not the least, I also want to express my gratitude to my family, who have been very supportive of my study. They encouraged me a lot and shared my enthusiasm for academic pursuit.

Abstract

In the era of internet of things, electronic devices are becoming ubiquitous. To perform secure tasks such as identification and authentication, security modules need to be embedded in such devices. Physical unclonable function (PUF) is an emerging security primitive to address vulnerability of the traditional data protection scheme that stores the secret keys in non-volatile memories. Despite various kinds of PUFs proposed so far, this dissertation investigates two most popular PUFs, which are Static Random Access Memory (SRAM) and Ring oscillator (RO) based PUFs.

SRAM has recently been developed into a PUF for generating chip-unique signatures for hardware cryptography. The most compelling issue in designing a good SRAM-based PUF (SPUF) is that while maximizing the mismatches between the transistors in the cross-coupled inverters to improve the quality of SPUF, this process itself gives rise to increased memory read/write failures. For this reason, the memory cells of existing SPUFs cannot be reused as storage elements, which increases the overheads of cryptographic system where long signatures and high-density storage are both required. This thesis presents a novel design methodology for dual-mode SRAM cell optimization. The design conflicts are resolved by using word-line voltage modulation, dynamic voltage scaling, negative bit-line and adaptive body bias techniques to compensate for reliability degradation due to transistor downsizing. The augmented circuit-level techniques expand the design space to achieve a good solution to fulfill several otherwise contradicting key design qualities for both modes of operation, as evinced by the statistical analysis and simulation results based on complementary metal–oxide–semiconductor (CMOS) 45 nm bulk Predictive Technology Model.

Using power-up reset to reuse SRAM as PUF suffers from two major constraints. First, it has limited entropy of only one response bit per cell. Second, as power-up reset has a global effect, extra storage is needed to temporarily buffer the original SRAM content before switching to PUF mode. This process has introduced extra area-power overhead as well as potential security leakage. The emerging dual-port (DP) SRAM cell offers attractive multiple access capability for low-power and high-speed memory transfer. By leveraging the forbidden contention state in one of its four multiple access modes, a new DP-SRAM based PUF is proposed to generate two independent response bits per cell and limit data buffering to only those cell content addressed by the challenge. Simulation results based on PTM 22nm LP CMOS model has corroborated its superior reliability, uniqueness and randomness.

RO based PUF is an emerging hardware security primitive but its response reproducibility is susceptible to changes in operating conditions and device aging. Present solutions to increase RO PUF reliability either incur large hardware overhead or require sophisticated RO selection algorithm. This research exploits the optimal biasing of current starved (CS) inverter for the design of RO PUF with very high reliability against both temperature and voltage variations. With two additional transistors, the CS inverter can be adaptively biased at idle time and in active mode to significantly reduce the overall stress, making the proposed CS RO PUF robust against both environmental condition variations and aging. Without using error correction code, the reliability is further improved by a low-cost proximity detector circuit with a small sacrifice on challenge-response-pair (CRP) space. The correlation between successive RO pairs of different input challenges is also broken by an irregular clocking of the linear feedback shift register used to encipher the input challenges. Based on Monte Carlo simulation in TSMC 40nm CMOS process, the reliability of native PUF responses has been raised substantially from 89.78% for the regular RO PUF to 95.88% for the proposed CS RO PUF over a broad temperature range of -40 to 120°C and $\pm 20\%$ supply variation. With the proximity detector, the proposed Aging-resistant Current-starved RO (ACRO) PUF can attain 100% reliability statistically by discarding 34.47% fewer challenge-response pairs than the regular RO PUF. It is also $\sim 3.7\times$ more aging resilient than existing aging-resilient RO PUFs with $\sim 2.38\times$ lower power dissipation.

Table of contents

List of figures	xiii
List of tables	xvii
1 Introduction	1
1.1 Motivation of Research	1
1.2 Main Objective and Direction of Research	2
1.3 Major Contributions	3
1.4 Thesis Organization	3
2 Literature Review	5
2.1 PUF Definition	5
2.2 An Example of PUF	5
2.3 PUF Applications	6
2.3.1 Cryptography Key Generation	6
2.3.2 Low-cost Authentication	7
2.4 PUF Classifications	8
2.4.1 Weak PUF	8
2.4.2 Strong PUF	8
2.5 PUF Quality Metrics	9
2.5.1 Randomness	9
2.5.2 Uniqueness	9
2.5.3 Reliability	10
2.6 Reliability Enhancement Techniques	10
3 Sizing of SRAM Cell with Voltage Biasing Techniques for Memory and PUF Functions	13
3.1 Introduction	13
3.2 Impact of PV on Different Operation Modes	14
3.2.1 Impact of PV on PUF Quality	15
3.2.2 Impact of PV on SRAM Read/write failures	17
3.3 Proposed Design Method for Dual-mode PUF	18

Table of contents

3.3.1	Circuit-level techniques	19
3.3.2	Problem formulation	24
3.4	Simulation Results and Discussions	27
3.4.1	Uniqueness	28
3.4.2	Randomness	28
3.5	Conclusion	28
4	A New Write-Contention Based Dual-Port SRAM PUF with Multiple Response Bits Per Cell	31
4.1	Introduction	31
4.2	Access Modes of Dual-Port SRAM	33
4.3	Proposed DP-SRAM based PUF	34
4.4	Simulation Results and Discussion	36
4.4.1	Correlation between bits in a cell	36
4.4.2	Reliability	37
4.4.3	Uniqueness	37
4.4.4	Randomness	38
4.4.5	Power	38
4.5	Conclusion	39
5	ACRO-PUF: A Low-power, Reliable and Aging-Resilient Current Starved Inverter Based Ring Oscillator Physical Unclonable Function	41
5.1	Introduction	41
5.2	Impact of Environmental Conditions and Aging on RO frequencies	43
5.2.1	Effect of Temperature Variation on RO Frequency	44
5.2.2	Effect of Supply Voltage Variation on RO Frequency	47
5.2.3	Effect of Device Aging on RO frequency	49
5.3	Design of Robust CS RO PUF against Environmental Variations and Aging	50
5.3.1	Reliability Enhancement of RO PUF against Temperature and Voltage Variations	50
5.3.2	Aging-Resistant Biasing Circuit Design	52
5.3.3	Additional Architectural Enhancements of ACRO-PUF	54
5.4	Simulation Results and Discussions	56
5.4.1	Uniqueness	57
5.4.2	Reliability	57
5.4.3	Randomness	59
5.4.4	Aging Resistance	60
5.4.5	Comparison with other RO PUFs	60
5.5	Conclusion	61

6	Conclusion and Future Work	63
6.1	Conclusion	63
6.2	Future Work	64
6.2.1	Dual-Port SRAM PUF for Dual-mode Applications	64
6.2.2	RO PUF fabrication	64
6.2.3	Strong Mono-stable PUF	64
	References	69

List of figures

2.1	PUF device.	5
2.2	An arbiter PUF.	6
2.3	Cryptographic key generation.	7
2.4	PUF-based authentication	7
3.1	(a) Architectural diagram of an SPUF; (b) PUF and memory modes of operation of a 6T SRAM cell in an SPUF.	14
3.2	Design conflict of SRAM cell for dual-mode applications. Increasing the variation of V_{th} improves the quality of PUF at the expense of higher failure rates in regular memory read/write operations.	15
3.3	(a) Changes in $\sigma_{V_{th}}$ with transistor effective channel length. (b) relationship between the percentage of unreliable bits and the scaling factor of SRAM cell template, where the widths of pull-down, pull-up and access transistors of a unit-size SRAM cell are $W_{M1} = 180\text{nm}$, $W_{M2} = 90\text{nm}$ and $W_{M3} = 135\text{nm}$, respectively and all transistors have $L = 45\text{nm}$	16
3.4	(a) Read and (b) write failure effects in an SRAM cell. The original bit preserved in the cell is 1.	17
3.5	The relationship between the failure probability and the transistor width: (a) $M1$, (b) $M2$, (c) $M3$. The default transistor sizes are $W_{M1} = 180\text{nm}$, $W_{M2} = 90\text{nm}$ and $W_{M3} = 135\text{nm}$, respectively.	18
3.6	SRAM array with word-line power supply selector	20
3.7	Dynamic power supply based on the selection of read or write operation. . . .	20
3.8	Implementation of adaptive negative bit-line circuit.	21
3.9	SRAM cell with adaptive body bias	22
3.10	Static noise margins by applying different circuit-level techniques: (a) WLM , (b) DVS , (c) NBL , (d) ABB (NMOS only) (e) ABB (PMOS only), (f) ABB (both NMOS and PMOS).	23
3.11	Design space exploration of SRAM cell for dual-mode SPUF design.	24
3.12	Frequency distribution of the simulated fractional inter-die HD.	28

List of figures

4.1	Single-port and dual-port SRAM cells: (a) 6T-SP-SRAM cell (1RW) (b) 8T-DP-SRAM cell (1R1W) (c) 10T-DP-SRAM cell (1R1W) (d) 8T-DP-SRAM cell (2RW).	32
4.2	Access modes of DP-SRAM: (a) different row and column, (b) different row and common column, (c) common row and different column and (d) common row and column	33
4.3	8T-DP-SRAM operating in PUF mode (a) $BL1 = 1, BL2 = 0$ and (b) $BL1 = 0, BL2 = 1$	34
4.4	(a) Response bit determined mainly by mismatches of access transistors. (b) Response bit determined mainly by mismatches of cross-coupled inverters. . .	35
4.5	Mode switching of conventional SRAM PUF [1].	36
4.6	Response bits generated from the same cell written with complementary data bits	36
4.7	Reliability of (a) single port SRAM PUF and (b) dual port SRAM PUF for simultaneous temperature and voltage variations.	37
4.8	Fractional Hamming distance histograms of (a) SP-SPUF and (b) DP-SPUF. .	37
5.1	(a) Basic architecture of RO PUF. (b) Regular inverter. (c) CS inverter.	43
5.2	(a) Conditions for stable and unstable response bits generated by regular RO pair. (b) Stable response bits are generated by CS RO pair with optimal biasing.	44
5.3	(a) Relative frequency deviations against temperature of CS RO with V_{ctrl} varied from 0.3V to 0.7V. (b) Relative frequency deviations against temperature of regular RO and CS RO with V_{ctrl} varied from 0.8V to 1V.	46
5.4	(a) Conditions for stable and unstable response bits generated against voltage variation. (b) Stable response bits are generated with enlarged frequency variation.	47
5.5	Relationship between V_{bp} and $ V_{GSp} = V_{DD} - V_{bp}$ with V_{DD}	48
5.6	Relative frequency deviation against supply voltage for regular RO and CS RO with V_{ctrl} from 0.3V to 1V.	49
5.7	(a) $V_{bn,ZTC}$ at different supply voltages. (b) Frequency at corresponding $V_{bn,ZTC}$ for different supply voltages.	49
5.8	(a) Conditions for stable and unstable response bits due to aging. (b) Stable response bits are generated by slowing down the aging process.	50
5.9	Coefficient of variation of frequency $\frac{\sigma_f}{\mu_f}$ at different V_{ctrl} for CS RO.	51
5.10	Effect of V_{ctrl} on the trajectories of frequency deviation and frequency variance of CS RO PUF.	52
5.11	Design flow for high reliability CS RO PUF	53
5.12	Worst-case BER against both temperature and voltage variations for different V_{ctrl}	54
5.13	(a) Proposed aging-resilient CS RO design, (b) Equivalent biasing circuit when $EN = '1'$, (c) Equivalent biasing circuit when $EN = '0'$	55

5.14	(a) Block diagram of proposed ACRO-PUF. (b) RO PUF core circuit.	56
5.15	Bidirectional counter with proximity detector circuit.	56
5.16	(a) Regular LFSR clocking with correlated sequence of RO pairs for different challenges. (b) Irregular LFSR clocking with uncorrelated sequence of RO pairs for different challenges.	57
5.17	Interdie fractional HD for (a) regular RO PUF ($\mu = 50.07\%$, $\sigma = 6.91\%$), (b) ACRO-PUF ($\mu = 49.97\%$, $\sigma = 6.98\%$).	58
5.18	Comparison of regular and ACRO-PUF on reliability against (a) temperature variation and (b) voltage variation.	58
5.19	Reliability agianst V_{ctrl} variations.	59
5.20	(a) Frequency degradation of regular RO over time. (b) Frequency degradation of AT RO over time (c) Frequency degradation of ACRO over time.	61
6.1	Architecture of Strong mono-stable PUF.	65

List of tables

3.1	Comparison of different adaptive techniques	24
3.2	Simulation Results @ $V_{DD} = 0.9$ V, temperature= 25°C	27
3.3	NIST test results.	29
4.1	NIST randomness tests for SP-SPUF and DP-SPUF	38
5.1	Power analysis @ $V_{DD} = 1.1$ V, temperature= 27°C	57
5.2	Worst case reliabilities with proximity detector	59
5.3	NIST randomness tests on responses of regular RO PUF and proposed ACRO- PUF.	60
5.4	Comparison of Different RO PUFs.	62

Chapter 1

Introduction

1.1 Motivation of Research

Proliferation of mobile electronic devices that shape the new era of internet of things boost the research and development effort in hardware security to safeguard computing platforms, wherein sensitive data are processed and preserved. Security applications such as identification, device-authentication and counterfeit detection all need a physically unbreakable secrets which are rooted in the embedded device [2]. For the sake of simplicity and efficiency, the primitives for these applications are preferably implemented in an integrated circuit (IC) module, and for security reason, these modules should be resilient against physical intrusions and malicious tamperings [3].

Currently the commonly adopted practice is to place a secret key in the non-volatile memory (NVM)[4] which is encapsulated inside a protection frame such as a coating layer to make NVM tamper-evident. However, such approach has a couple of disadvantages. Firstly, it is difficult and expensive to keep secrets safely in memory. As the secrets always exist in a digital form, the NVM technologies are usually vulnerable to invasive attack [3]. Secondly, for resource constrained platforms such as RFIDs, primitives featured by light-weight properties are often desirable and dedicated IC blocks which are optimized for low power and area consumption are more suitable candidates than the NVMs for key preservation applications [3].

Physical Unclonable Function (PUF) is an emerging primitive based on IC technology for efficient identification and device authentication. It is essentially a physical structure that maps the input challenges to output responses. As the mapping is determined by the uncontrollable process variations that are inherent in each device, the challenge-response pairing characteristics are random, unique and unpredictable, and therefore can be effectively used as a “key”. Similar to the algorithmic one-way hash functions, the mapping of PUFs is easy to compute in the forward direction but is infeasible to be inversely evaluated. This is powerful to address the security vulnerability of traditional data protection schemes based on NVM. There are a number of PUF implementations proposed so far and one of the most

representative approaches is based upon Static Random Access Memory (SRAM)[5–9]. SRAM is commonly used as cache in the computing systems. The mismatch of the inverters in a typical 6-T SRAM cell cannot be avoided due to the uncontrollable process variation. By utilizing the mismatch of the two cross-coupled inverters in an SRAM cell, a unique random bit can be generated. When powering up an SRAM cell, the bit may be skewed to either ‘1’ or ‘0’, depending on the strength of the two inverters. By powering up many cells at the same time, the physical fingerprint of the chip can be extracted for identification or authentication purpose [2]. Due to the inevitable noise and uneven skewness, the extracted bits need to be post-processed by a fuzzy-extractor to produce stable and random secret keys with sufficient entropy.

Another popular PUF candidate is ring oscillator (RO)-based PUF [10–17]. It has the resilience against noise impact and can be implemented for low-power lightweight applications [10]. The response bits are generated by comparing the frequencies of two identically designed ring oscillators. Even though the two oscillators are designed identically, their frequencies are distinct due to the process variations, which is again unpredictable. Due to the slow speed and higher power consumption, RO PUF is better suited to secure processor design and in field-programmable gate array (FPGA) design [3].

It is very hard, expensive and economically not viable to either physically or mathematically create a PUF with a given electronic fingerprint because the randomness of its generated responses is dependent on the uncontrollable process variations [18]. Furthermore, physically invading a PUF to probe or measure random bits will result in nothing but damages to the PUF chip. Thus the security level of PUFs by extracting secrets from sophisticated physical structures is higher than non-volatile memory. Moreover, PUF does not require special manufacturing process or testing procedures [3].

1.2 Main Objective and Direction of Research

The main advantage of SRAM-based PUF is that it can be used as both a normal memory and a PUF. To implement a reliable PUF, large process variations are preferred so that the bits can have a higher probability to be a strong ‘1’ or ‘0’. In contrast, small process variation is favoured by normal memory working mode since failure probabilities of read/write operations due to varying device parameters will be low. Therefore, an SRAM that adequately satisfies these two conflicting constraints is needed, and this requires a novel design methodology. Carefully sizing of transistors in an SRAM based on the analysis of PUF quality metrics and memory operation probabilities may help finding the optimal design point. Besides, dedicated IC techniques such as word-line voltage modulation, dynamic voltage scaling, negative bit-line and adaptive body bias can be used to further optimize the function of the two working modes to meet the specifications. Other constraints of traditional SRAM PUF include generating only one response bit per cell and the requirement of additional buffer storage to store the data

during mode switching. However, this may be solved by exploiting dual-port SRAM instead of single-port SRAM to implement a PUF.

RO PUF has the reliability issue against environmental changes, such as temperature and supply voltage fluctuations, and device aging. Current solutions either incur large hardware overhead or require sophisticated bit selection algorithms, which are not desirable. To address the reliability issue, current-starved (CS) inverters can be used to replace the regular two-transistor inverters. With two additional transistors, the CS inverter can be adaptively biased at idle time and in active mode to significantly reduce the overall stress, making the proposed CS RO PUF robust against both environmental condition variations and aging. Furthermore, the highly unreliable response bits can be discarded using a simple thresholding circuit. In such a way, a highly reliable RO PUF against both environmental changes and aging can be implemented.

1.3 Major Contributions

This thesis presents a novel design methodology to optimize an SRAM cell for dual application modes. The design conflicts are resolved by using word-line voltage modulation, dynamic voltage scaling, negative bit-line and adaptive body bias techniques to compensate for reliability degradation due to transistor downsizing. The augmented circuit-level techniques expand the design space for finding a good solution to fulfil several otherwise contradicting key design qualities for both modes of operation, as evinced by the statistical analysis and simulation results based on CMOS 45 nm bulk Predictive Technology Model.

Another improvement on SRAM PUF is that dual-port SRAM PUF has been proposed to generate multiple response bits per cell. This is achieved by leveraging on the forbidden contention state in one of its four access modes. The proposed PUF also limits the data buffering to only those cell content addressed by the challenge.

In this thesis, CS inverter has been investigated for the design of RO PUF, which has four distinctive contributions over [10]. First, the optimal biasing of CS inverter has been exploited to design the RO PUF with very high reliability against both temperature and voltage variations. Second, the CS inverter is designed to be adaptively biased at idle time and in active mode to significantly reduce the overall stress, making the proposed CS RO PUF robust against aging. Third, a low-cost proximity detector circuit is used to filter out a small fraction of unreliable response bits. Finally, an irregular sequencing of successive RO pairs from different input challenges by modifying the linear feedback shift register (LFSR) clocking scheme is proposed to harden the security of its challenge-response interface.

1.4 Thesis Organization

The thesis is organized into six chapters.

Chapter 1 introduces the research motivation, objectives and research direction, and summarizes the major contributions of this thesis.

Chapter 2 reviews the background of PUFs. The definition of PUF is introduced in an illustrative way. After that, an arbiter PUF is used to explain the CRP generation mechanism. Two PUF applications are also elaborated. Two main categories of PUFs, which are strong PUFs and weak PUFs are then compared. The key figures of merit of PUF are defined and the reliability enhancement techniques are introduced.

Chapter 3 presents the design methodology for performance optimization of SRAM for operation in both PUF and memory modes. The limitations of traditional SRAM PUF are first analyzed. After highlighting the motivation to design SRAM for both modes of operation, the impacts of process variations on the two different operation modes are analyzed separately. Finally, the simulation results for the proposed design are presented and discussed.

Chapter 4 presents a new write-contention based dual-port SRAM PUF with the capability of generating multiple response bits per cell. The disadvantages of the single-port SRAM and the single-port SRAM based PUF are analyzed. Four different types of SRAMs and four access modes of the 8T-DP-SRAM are introduced. Next, the working mechanism and advantages of the proposed DP-SRAM based PUF are explained. Finally, the comparison results for the DP-SRAM based PUF and SP-SRAM based PUF are presented and discussed.

Chapter 5 presents a new CS RO PUF called ACRO-PUF with enhanced reliability and security. The limitations of previous RO PUFs are first presented. Then the impacts of environmental conditions and aging on RO frequencies are analyzed before the CS RO PUF against environmental variations on RO frequency is proposed. Finally, the performance of the proposed RO PUF is evaluated.

Chapter 6 concludes the thesis and proposes prospective works based on current findings. The future work includes improvements on the proposed SRAM PUF and RO PUF, as well as the proposal of a new strong PUF design.

Chapter 2

Literature Review

2.1 PUF Definition

PUF, also known as Physical Random Function, was first introduced by Gassend et al in 2002 [19], which was then called Physical one-way function [20, 21]. It is defined as a function that maps challenges (C) to responses (R). It is embodied by a physical device. It has the properties of easy to evaluate and hard to characterize [19]. As illustrated in Fig. 2.1, the challenges are the input to the PUF and the responses are the output. The mapping of the challenge-response-pairs (CRPs) is determined by the inherent device manufacturing process variations, which are unpredictable and uncontrollable.

Fig. 2.1 PUF device.

PUF can be modelled as a function $R = f(C)$, where the function $f(\cdot)$ is a function of the manufacturing process variation of the device [4].

2.2 An Example of PUF

Fig. 2.2 shows the circuit of an arbiter PUF, which comprises several multiplexers (MUXes) and an arbiter [3]. It has a 128-bit input challenge X ($X[0]$ to $X[127]$) and an 1-bit output Y . Each input bit $X[i]$ controls two MUXes and acts as a switch box (dotted box). If the control bit $X[i]$ is 0, the two delay signals will pass through the switch box from the left side through their respective path (denoted by the arrow pointing to 0). Otherwise, the two signals will switch their paths. In this way, each input vector will produce a pair of identical delay paths.

Due to process variation, the delays of the two paths are different. The output bit is obtained by launching a rising signal to the two paths simultaneously. The two signals will race through the delay paths and the arbiter will decide which signal is faster. The output will be '1' if the signal to the data input of the latch is faster and will be '0' otherwise. As the two delay paths are designed identically and the manufacturing process variation is uncontrollable, the output responses to the same challenge of each chip are different and cannot be predicted. There are

Fig. 2.2 An arbiter PUF.

two methods to create a k -bit response from this simple circuit. One is to evaluate the PUF k times with different inputs. The other one is to duplicate the PUF multiple times and use one input to evaluate them.

2.3 PUF Applications

PUFs normally have two main types of security application: cryptographic key generation and device authentication. Weak PUFs have limited number of responses, thus they can only be used for cryptography key generation. Strong PUFs have large number of CRPs, thus they can be used for authentication application.

2.3.1 Cryptography Key Generation

The cryptography key generation mechanism is shown in Fig. 2.3. It has an initialization phase and a regeneration phase. In the initialization phase, the PUF is challenged and the response will be generated. As the responses are unstable, they will be post-processed by a fuzzy extractor. The fuzzy extractor is used to correct the unreliable bits. Then a key and a helper data will be generated. The helper data will be stored in the non-volatile memory. In the regeneration phase. The response will be generated first. Then the response and the helper data will be sent to the fuzzy extractor for processing. Finally, the same key as the initialization phase will be regenerated with the same helper data. The secret key cannot be generated with the knowledge of helper data alone without the PUF. As the key can always be generated on demand, there is no need to store the key in the memory, which reduces the risks of attacks. This is unlike the traditional way to generate the cryptographic key, in which the keys are generated based on mathematical algorithms and stored in non-volatile memory.

Fig. 2.3 Cryptographic key generation.

2.3.2 Low-cost Authentication

PUF can also be applied to extremely resource constrained platforms such as RFIDs where cryptographic operations may be too costly [3]. The authentication protocol is shown in Fig. 2.4. For PUFs that are used for authentication, they should have exponential number of CRPs. A trusted party A first randomly selects some challenges and apply them to generate unpredictable responses for further authentication. Then the PUF device P is delivered to the user B through the untrusted supply chain or environment. To authenticate the device, A selects a recorded but unused challenge and send to the PUF device. If the response generated by the PUF matches the previously recorded one, then the integrated circuit (IC) is authentic, because only the trusted party and the authentic IC can have the correct CRPs. However, anyone who have access to the PUF can record down the CRPs and put them in the memory. To prevent the PUF from falling into the hands of an adversary who can use it directly to impersonate B , acknowledgment of its receipt by the intended recipient B through a private channel should be received within a reasonably stipulated time before P can be used for subsequent authentications through the unprotected public channel. Meanwhile, the challenges should never be reused to avoid man-in-the middle attack [3].

Fig. 2.4 PUF-based authentication

2.4 PUF Classifications

There are many different kinds of PUF. Their responses can be obtained from memories like SRAM-based PUF or from delay paths like arbiter PUF. However, a more common classification method is to evaluate their CRP spaces, i.e. the number of CRPs, and categorized them as either weak PUFs or strong PUFs.

2.4.1 Weak PUF

The difference of between strong PUF and weak PUF is based on the domain of CRP mapping, which refers to the number of unique challenges possessed by the PUF [4]. Weak PUFs only have small number of CRPs, which can be fully obtained in a short time. They have the following properties:

- Limited number of CRPs, which is normally linearly or polynomially proportional to the number of components [22].
- Impossible to be cloned physically with the same CRPs.
- Responses are stable and reproducible against environmental condition changes.
- Responses are unpredictable, as they are dependent on the uncontrollable manufacturing process variations [4].

Examples of weak PUFs include SRAM PUF, butterfly PUF and coating PUF [22]. Since weak PUFs only have a small number of CRPs, they must be kept secret to avoid the PUF from being emulated by other devices by limiting the access to the PUF outputs [4]. Due to this reason, weak PUFs are more suitable for the generation of secret keys in cryptographic applications such as data encryption and decryption.

2.4.2 Strong PUF

Compared with weak PUFs, strong PUFs have a large number of CRPs, which cannot be fully read out within a reasonable time by an adversary. For this reason, they can be used for authentication without any other cryptographic hardware [4]. Strong PUFs have the following properties:

- Impossible to be cloned physically with the same CRPs.
- A very large number of CRPs, ideally exponentially proportional to the number of components such that the adversary cannot emulate all the CRPs within a realistic time.
- Responses are unpredictable given polynomial number of CRPs, i.e. resilient to modelling attack.
- Responses are stable, with few or even no bit flips during multiple reading.

- No other functionality information except the CRPs are revealed during readout [4].

Examples of strong PUFs include Optical-PUFs and XOR-APUFs [22]. It should be noted that unlike weak PUFs, strong PUFs do not need any additional hardware to protect its CRPs. However, the readout time must be taken into account to avoid the emulation of all CRPs. If the attacker can read out all the CRPs or a large fraction of the CRPs in a realistically short time when he has access to the PUF without being detected and stored them somewhere, he can mount a man-in-the-middle attack to masquerade as the PUF. Unauthorised access to its internal circuit must also be prevented [4].

2.5 PUF Quality Metrics

A good PUF should have responses that are random, unique and reproducible. Thus the quality of PUF can be measured by its randomness, uniqueness and reliability, which are the three most common metrics [22].

2.5.1 Randomness

Randomness means that the PUFs have equal probability of generating response bit of ‘1’ or ‘0’ and each bit is not correlated with other bits. Ideally, the probability to obtain ‘1’ or ‘0’ is 50%. In practice, the NIST (National Institute of Standards and Technology) test can be used to evaluate the cryptographic randomness of PUF responses [23]. This is a Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications. Ideally, the PUFs should pass all the tests. One metric to measure the randomness in a closed system is entropy [23]. The entropy of uncertainty of a discrete random variable X with probabilities $Pr[X = x] = p_x$ is defined as [24]:

$$H(X) = - \sum_{x \in X} p_x \log p_x \quad (2.1)$$

However, as the actual distribution of PUF CRPs is normally unknown, it is difficult to measure the entropy directly. Fortunately, there is an alternative way to estimate the entropy in a PUF by calculating its number of independent output bits [3], which is relatively easy to determine. For example, to determine the possible number of independent bits in a RO PUF, the number of different ordering of ROs based on their frequencies is calculated first, which is $N!$. Since each ordering is equally likely in principle, the entropy is $\log_2(N!)$ bits.

2.5.2 Uniqueness

The uniqueness measures the capability of a PUF instance to distinguish itself from other PUF instances. The uniqueness can be measured by calculating the inter-die Hamming Distance (HD). Ideally, the responses generated from two PUF instances by applying the same challenge should have half of the bits different. Thus ideally the uniqueness should be 50%. Even though

50% is not easy to achieve, some factors such as system bias and correlations between the responses should be minimized, if not eliminated, so that the PUF can have a uniqueness as close as possible to 50%. Let R_u and R_v be the n -bit responses of two different chips, u and v , to the same challenge C , the uniqueness U for m chips is expressed as [10]

$$U = \frac{2}{m(m-1)} \sum_{u=1}^{m-1} \sum_{v=u+1}^m \frac{\text{HD}(R_u, R_v)}{n} \times 100\% \quad (2.2)$$

2.5.3 Reliability

Reliability measures the capability of a PUF to generate the same responses to the same challenges under various environmental conditions. The environmental factors that may affect the PUF reliability are noise, temperature change and supply voltage fluctuations. The response bits may also flip because of the aging effect. Reliability can be measured by the bit error rate (BER). BER is measured by calculating the intra-die HD, which is defined as the mean value of flipped bits among different responses when the same challenge is apply to the same PUF instance. Let R_i be an n -bit response to an input challenge C produced by a PUF chip i under the nominal operating condition. The same set of challenges are then applied k times to obtain the response $R_{i,j}$ for $j = 1, 2, \dots, k$. The reliability S of chip i can be computed by

$$S = 1 - \text{BER} = 1 - \frac{1}{k} \sum_{j=1}^k \frac{\text{HD}(R_i, R_{i,j})}{n} \times 100\% \quad (2.3)$$

Ideally, BER should be 0 and reliability should be 100%. However, it is impossible to achieve this goal [22]. Most research effort in PUF design are devoted to reducing the BER of raw response to as low as possible in order to minimize the cost of error correction.

2.6 Reliability Enhancement Techniques

Among the three quality metrics, high reliability is the most difficult to achieve. There are many techniques to improve the reliability, some of which are:

1. Directed accelerated aging (DAA) or burn-in hardening: the performances and characteristics of ICs change over time due to aging phenomena. The most prevalent aging effect consists of negative bias temperature instability (NBTI), time dependent dielectric breakdown (TDDB), hot carrier injection (HCI), and electromigration (EM) [25]. However, by purposefully aging the ICs directionally, the PUF reliability can be improved. In [26], a method is proposed to increase both the uniformity and reliability of PUF. The design is made with two main steps in the design. First, aging effect is introduced to increase the uniformity to 50%. Then, the device is aged multiple times to strengthen the bits until all the bits are stable. An ideal SRAM PUF can be achieved with this method.

2. Multiple evaluation (ME) or temporal majority voting (TMV): By evaluating the PUF multiple times, a soft version of response can be generated. The response bit to the same challenge may be different at different evaluation. By comparing the number of generated 1s and 0s, a more reproducible bit can be selected for each challenge.
3. Activation control (AC): this technique manipulates the waveform shape of the activation signal, such as the V_{DD} for SRAM and sense enable for SA PUFs [25].
4. Error correction code (ECC): cryptographic primitives requires that every bit of a key is kept constant. ECC is used to deal with the PUF response in both initialization phase and regeneration phase. A BCH (n, k, d) code will generate $b (= n - k)$ bit syndrome (helper data, public information) in the initialization phase. With such syndrome, BCH is able to correct up to $(d - 1)/2$ errors out of n bits to generate a k -bit secret reliably [3].
5. Masking or soft dark bits: some highly unstable bit that cannot be stabilized by TMV or burn-in hardening can be stabilized by using soft dark bitmask [27]. Dark bits can be identified as highly unstable with a specified window and a voltage range. Forcing these bits to a pre-determined value ensures that they will not contribute to PUF instability, hence reducing the overall BER.
6. Dynamic thresholding (DT): Yang et al. [12] proposed a PUF using oscillator collapse and dynamic thresholding technique to enhance the reliability. Responses that collapse slower have much larger average BER. By discarding the response bits that require higher number of cycles than a threshold to collapse, the reliability will be increased. Traditional ring oscillator (RO) PUF responses are generated by comparing the frequency difference of two ROs. The reliability can be enhanced by adding a threshold to the frequency difference.

Chapter 3

Sizing of SRAM Cell with Voltage Biasing Techniques for Memory and PUF Functions

3.1 Introduction

Among the PUF implementations [2, 3, 9, 28–30], SRAM-based PUF (SPUF) has attracted tremendous attention. This is because SRAM, being an integral part of computer memory sub-system, plays a pivotal role in trusted computing platform. The ability to use the storage cells of SRAM inseparably as PUF will replace or augment memory curtaining as a stronger fortification to the roots of trust for memory authentication. Unfortunately, existing SPUF cells cannot be doubled as regular storage elements. The exploitation of process-variation induced device mismatches in the cross-coupled inverter cell for random, unique and reliable response bit generation is detrimental to the regular memory operation, as it will result in increased parametric failures due principally to destructive read and unsuccessful write operations. Archived literatures reported wide varieties of design approaches that either improve the qualities of SPUF or harness the readability and writability of SRAM cell as data storage element, albeit independently [30, 31]. However, no design strategy has been proposed to tackle the conflicting quality requirements to unify the two different modes of operation in an SRAM cell.

In this chapter, the design space exploration of the preliminary proposal in [7] is expanded, by taking into consideration circuit-level techniques that can be used to mitigate conflicting design criteria of both PUF response and memory data stability. The proposed design procedure provides greater opportunity to meet tighter functional and performance specifications of both modes of operation at a small cost of area, delay and power overheads.

The remainder of this chapter is organized as follows. Section 3.2 analyzes the impacts of process variations (PV) and transistor sizing on the two different working modes of SRAM. Section 3.3 presents the proposed design methodology to expand the design space for PUF and

memory stability optimization. Section 3.4 discusses the comparison results for seven different design configurations. The chapter is concluded in Section 3.5.

3.2 Impact of PV on Different Operation Modes

Fig. 3.1 (a) Architectural diagram of an SPUF; (b) PUF and memory modes of operation of a 6T SRAM cell in an SPUF.

The block structure of a complete SPUF is shown in Fig. 3.1(a). The core of the SPUF is a two-dimensional array of storage cells, where the horizontal rows are referred to as word-lines (WL) and the two vertical lines that run through each cell are called the bit-line (BL) and complement bit-line (BLB). An individual cell can be selected by the row and column decoder circuit with an external input address, $addr$. One sense amplifier is attached to each bit-line pair of the column to accelerate read operation while the write driver circuit improves the access time in write operation. When the SPUF is set to PUF mode by the $mode_sel$ input, the SPUF is reset and powered up. Due to mismatch and noise between the cross-coupled inverters formed by the transistor pairs, M1-M2 and M4-M5, a *response* bit is generated in each of the SRAM cells. An input *challenge* is applied to the WL of the addressed cells to turn on the access transistors, M3 and M6. The *response* bits are then read from the bit-lines [2]. In the memory mode, regular read and write operations are performed on the same SRAM array after the SPUF has been fully powered up. For a read operation, the small voltage difference between the true and complement bit-lines is amplified by the sense amplifier and the data stored in the addressed cell is transferred to the output. For a write operation, the bit-lines are initialized with the data that is to be written into the cell with its word-line asserted. These two different modes of operation on a 6T SRAM cell are shown in Fig. 3.1(b).

While both modes of operation seem independent, they cannot be designed independently because the effect of PV influences on them occurs simultaneously. As shown in Fig. 3.2, when the threshold voltage variation $\sigma_{V_{th}}$ increases, the failure rate of memory operations increases whereas the reliability of response bits for PUF operation improves.

Fig. 3.2 Design conflict of SRAM cell for dual-mode applications. Increasing the variation of V_{th} improves the quality of PUF at the expense of higher failure rates in regular memory read/write operations.

3.2.1 Impact of PV on PUF Quality

During the power-up process, all transistors in an SRAM cell operate in the sub-threshold region. The currents that drive the two output nodes, BIT and BITB, in Fig. 3.1(b) are determined by the strength of the two inverters in the cell. The currents flowing through M1-M2 and M4-M5 will simultaneously pull up the voltage at BIT and BITB, respectively. Owing to the transistor mismatches between the two inverters, one of the two nodes may reach the “trip-point” (V_{trip}) of the inverter faster than the other. If the voltage at BIT of INV-1 increases to V_{trip} first, the output node BITB of INV-2 will be pulled down to ground. As the voltage at BITB descends, M2 will be fully turned on eventually and pulls the voltage at BIT to the supply voltage V_{DD} .

Since each response bit is independently generated from an SRAM cell and dominated by independent intra-die variations such as random dopant fluctuation and line edge roughness, the randomness among bits generated is guaranteed. While the response bits produced from individual SRAM cells are generally uncorrelated, their reliability may be jeopardized when the noise effects overwhelm the mismatches between the two inverters in an SRAM cell. Besides, temporal post-manufacturing variations such as unexpected change of environmental conditions in which the devices operate will also impact the reliability.

Mismatch of INV-1 and INV-2

The reliability (or repeatability) of the start-up state of an SRAM cell is highly dependent on the degree of mismatch between INV-1 and INV-2. The mismatch is dominated by the intra-die variation where the variation of V_{th} is the most prevalent. During power-up process, a small deviation of V_{th} , denoted as ΔV_{th} , between the transistors of INV-1 and INV-2 may change the drive currents and shift the intersection of their voltage transfer curves (VTC) [32],

Fig. 3.3 (a) Changes in $\sigma_{V_{th}}$ with transistor effective channel length. (b) relationship between the percentage of unreliable bits and the scaling factor of SRAM cell template, where the widths of pull-down, pull-up and access transistors of a unit-size SRAM cell are $W_{M1} = 180\text{nm}$, $W_{M2} = 90\text{nm}$ and $W_{M3} = 135\text{nm}$, respectively and all transistors have $L = 45\text{nm}$.

causing asymmetric output sensitivity to external stress. The magnitude of ΔV_{th} is primarily determined by $\sigma_{V_{th}}$, which has been proven empirically to be related to the transistor geometry as follows [33]:

$$\sigma_{V_{th}} = \frac{\Lambda \cdot \sigma_{V_{th,max}}}{\sqrt{(W \cdot L)}} \quad (3.1)$$

where Λ is a technology-dependent parameter and $\sigma_{V_{th,max}}$ is the maximum value of $\sigma_{V_{th}}$.

Based on (3.1), $\sigma_{V_{th}}$ can be reduced by sizing up the transistors in an SRAM cell [see Fig. 3.3(a)] [33]. Depending on the transistor sizes, the mismatch between the inverters of a cell may be reduced to an extent that its output state becomes highly susceptible to variations of external condition or in the worst case, completely determined by the noise present in the SPUF. Fig. 3.3(b) depicts the simulation results of the response bits generated by powering up the same SRAM cell 10 times in the presence of transient noise, whose parameters are set as $f_{min} = 10^6\text{Hz}$ and $f_{max} = 10^{11}\text{Hz}$, where f_{min} and f_{max} are respectively the base and maximum frequencies used for modeling frequency dependent noise sources. Transient noise analysis shows the effect of noise on the signal magnitude. It is also useful to see how noise effects the timing of the signal [34]. An SRAM cell that does not generate the same start-up value, i.e., ‘1’ or ‘0’, consistently is regarded as unreliable. The percentage of unreliable bits obtained from 500 cells of an SRAM array is plotted against the size of SRAM cell. The data points can be fitted by a second order polynomial, which indicates that the percentage of unreliable bits increases with the transistor size.

Loop-gain at trip point

The “Loop-gain” at the metastable trip point of a cross-coupled inverter pair is an important factor that affects the reliability of SPUF [35]. The node voltages of SRAM cell stabilize faster

Fig. 3.4 (a) Read and (b) write failure effects in an SRAM cell. The original bit preserved in the cell is 1.

if the loop-gain is higher. According to [36], the loop-gain is formulated as

$$LG(V_{BIT}) = \frac{\partial VTC_2}{\partial V_{BIT}} \cdot \frac{\partial VTC_1}{\partial V_{BITB}} \quad (3.2)$$

where $V_{BIT} = VTC_1(V_{BITB})$ or $V_{BITB} = VTC_2(V_{BIT})$. VTC refers to the voltage transfer characteristic function and can be formulated as follows [32]:

VTC for INV-2:

$$V_{BITB} = V_{DD} + V_T \ln \left\{ \frac{1}{2} \left[1 - G + \sqrt{(G - 1)^2 + 4 \exp \left(-\frac{V_{DD}}{V_T} \right) G} \right] \right\} \quad (3.3)$$

$$G = \exp \left(\frac{2V_{BIT}}{nV_T} - \ln \frac{I_{S5}}{I_{S4}} - \frac{V_{DD}}{nV_T} - \frac{V_{th,4} - V_{th,5}}{nV_T} \right)$$

where I_{Si} and $V_{th,i}$ refer to the saturation current and threshold voltages of the transistor M_i in Fig. 3.1. VTC for INV-1 can be similarly derived.

As pointed out in [35], large loop-gain degrades the reliability of SPUF as the amplified noise in the power-up process overwhelms the transistor mismatches due to process variation. From (3.2) and (3.3), the loop-gain is mainly governed by V_{DD} , V_{th} and V_T .

3.2.2 Impact of PV on SRAM Read/write failures

While the strength mismatches between different transistors of an SRAM cell are good to skew its power-up value to resist SPUF response bit flipping under temporal variations in operational condition, they can increase the read failure (RF) and write failure (WF) rates in the memory-mode. RF occurs when the voltage (V_{read}) sensed at BIT (or BITB) rises above the inverter read trip point (V_{tripRD}) to cause the bit stored in the SRAM cell to flip accidentally [see Fig. 3.4(a)]. As shown in Fig. 3.1(b), when the bit-line BL is precharged, the access

Fig. 3.5 The relationship between the failure probability and the transistor width: (a) $M1$, (b) $M2$, (c) $M3$. The default transistor sizes are $W_{M1} = 180\text{nm}$, $W_{M2} = 90\text{nm}$ and $W_{M3} = 135\text{nm}$, respectively.

transistor $M6$ and the pull-down transistor $M4$ act as a voltage divider. The PV induced voltage deviation at BITB may trigger a flip of the data bit stored in the SRAM cell [37]. To avoid RF, the relative strength of $M4$ and $M6$ [known as beta ratio ($\beta_{npd-nax}$)] should be designed larger, where $BR_{npd-nax}$ is given by

$$BR_{npd-nax} = \frac{\beta_{npd}}{\beta_{nax}} = \frac{\beta_{M4}}{\beta_{M6}} = \frac{\frac{\mu_{eff}C_{ox}W_{M4}}{L_{M4}}}{\frac{\mu_{eff}C_{ox}W_{M6}}{L_{M6}}} \quad (3.4)$$

where μ_{eff} is the effective mobility and C_{ox} is the oxide capacitance. WF refers to the failure of SRAM cell to respond to the data bit written into it within an acceptable time [see Fig. 3.4(b)][37]. Consider the SRAM cell in Fig. 3.1(b), if '0' is to be written into the cell that stores logic '1', the voltage at BIT will be discharged to a low value. If this value is higher than the trip point of INV-2, the write operation fails and the stored data bit remains unchanged. To reduce the probability of WF, the beta ratio of the access transistor and pull-up PMOS ($BR_{nax-pup} = \beta_{nax}/\beta_{pup} = \beta_{M3}/\beta_{M2}$) should be designed such that the write time is less than the word-line turn-on time.

It is possible to reduce the probability $\Pr(\text{RF}, \text{WF})$ of memory access failures by manipulating the transistor sizes. The relationship between $\Pr(\text{RF}, \text{WF})$ and the sizing of each transistor¹ in an SRAM cell is shown in Fig. 3.5. Curve fitting by polynomial functions shows that sizing up the transistors help to reduce failure probability. This is expected since $\sigma_{V_{th}}$ decreases when the size of transistors increases according to (3.1).

3.3 Proposed Design Method for Dual-mode PUF

To design a dual-mode SPUF, the transistors in the SRAM cell need to be sized such that the quality expectations can be met for both modes. For memory mode, the memory failure

¹The readers are referred to [37] for the modelling of these failure probabilities and how they can be affected by the transistor geometry.

probability should be kept as low as possible, whereas for PUF mode, the randomness, uniqueness and reliability of response bits should be made as high as possible. As randomness and uniqueness can be easily enhanced by fuzzy extractor at a smaller cost, this chapter focuses on the reliability, which is more critical and difficult to improve for SPUF. Besides failure probability and PUF reliability, area, leakage current and dynamic power constraints of the target process technology will also be considered.

3.3.1 Circuit-level techniques

In addition to transistor sizing, circuit-level techniques such as Word-Line Voltage Modulation (WLM), Dynamic Voltage Scaling (DVS), Negative Bit-Line (NBL) and Adaptive Body Biasing (ABB) can also be applied to mitigate SRAM failures [38–41]. For ease of exposition, the nominal supply voltage is denoted as ‘SRAMVDD’. Besides, since static noise margin (SNM) is more conveniently used in practice to calibrate the SRAM cell stability against DC noise [32], it will be used as a measure to compare the efficiency of different adaptive techniques.

WLM

The WLM technique proposed in [38] is shown in Fig. 3.6. This technique increases the read stability by increasing $BR_{npd-max}$ to lower the word-line voltage. The WL selector adapts different supply voltages to the WL (V_{DD_WL}). Its control signal is determined by the requirements of speed and stability. By lowering the WL voltage and keeping the cell voltage supply at SRAMVDD, the strength of the access transistors will be lowered to improve $BR_{npd-max}$. However, lowering WL voltage will also degrade the write ability due to the reduced $BR_{nax-pup}$. This problem can be solved by applying different V_{WL} for the read and write operations. This scheme has a drawback as it can only improve read stability but not write ability, as improving V_{WL} during write operation will degrade the SNM of the half-selected cells (which are equivalent to performing a read operation on these cells) even though they are not accessed.

DVS

A column-based DVS technique was proposed in [39] to improve both read and write margins. A differential voltage is created between the SRAM cell and the WL nodes to optimize the read and write margins separately without compromising each other. The mechanism is shown in Fig. 3.7. Each column of memory cells have a common voltage supply, which is either V_{DD_hi} or V_{DD_lo} , depending on the cell operation mode. During the read operation, the column of accessed cells switches its V_{DD} to V_{DD_hi} , which is higher than the word-line voltage SRAMVDD. During the write operation, the column of accessed cell switches its V_{DD} to V_{DD_lo} , which is lower than SRAMVDD. For the columns which are not accessed, their V_{DD} will always be V_{DD_hi} , as they are always experiencing ‘read stress’ even though they are not

Fig. 3.6 SRAM array with word-line power supply selector

selected to be read or written. When a cell is in standby mode, its V_{DD} is switched to V_{DD_lo} to minimize leakage power. In this way, both read and write stability will be increased.

Fig. 3.7 Dynamic power supply based on the selection of read or write operation.

NBL

A write-ability enhancement technique that combines negative bit-line and V_{DD} collapse scheme with high configurability was proposed in [40]. Like DVS, the V_{DD} collapse technique improves $BR_{nax-pup}$ by weakening the pull-up PMOS. The circuit implementation of NBL is shown in Fig. 3.8. During write operation, the bit-line is driven to a negative voltage to write '0'. This will strengthen the driving capability of pass transistors, thus improving $BR_{nax-pup}$ and the write ability. As this technique is only applied during the write operation, it has no influence on the read stability.

Fig. 3.9 SRAM cell with adaptive body bias

Even though ABB can be applied to reduce the read and write failures, it is not without limitations. The ideal case of biasing each transistor independently is not practical due to large amount of voltage supplies and area overhead [43]. As a practical expedient, the same body bias can be applied to all cells to mitigate the degraded cell stability against inter-die variations. In addition, unlike PMOS BB, which can be applied in standard dual-well process, the application of BB to NMOS requires triple-well process [41]. If cost permits, it is preferable to adjust the BBs of both PMOS and NMOS transistors to mitigate the inter-die variations. Under the constraint of manufacturing cost, BB can only be applied to PMOS transistors to reduce the read and write failures despite less efficient [41]. Since the inter-die V_{th} 's of PMOS and NMOS may move in different directions, it is also desirable to control the BB of the two types of transistors individually.

Comparison of different reliability enhancement techniques

The SRAM cells incorporated with the four aforementioned techniques were simulated to compare their effectiveness in improving the SNM. For WLM and DVS, the word-line and supply voltages were swept by $\pm 20\%$ from 0.7V to 1.1V. For NBL, the BL voltage is swept from $-0.2V$ to $0.2V$. For ABB, the SNMs were simulated with BB applied to only NMOS, only PMOS and both NMOS and PMOS transistors, respectively. Due to the non-linear relationship between V_{th} and BB voltage, the BBs were swept for a large range from $-0.5V$ to $0.5V$. By representing the deviation of BB by ΔV , the BB voltages, V_{bn} and V_{bp} , for NMOS and PMOS, respectively are given by:

$$\begin{aligned} V_{bn} &= \Delta V \\ V_{bp} &= \text{SRAMVDD} + \Delta V \end{aligned} \quad (3.7)$$

Fig. 3.10 Static noise margins by applying different circuit-level techniques: (a) *WLM*, (b) *DVS*, (c) *NBL*, (d) *ABB* (NMOS only) (e) *ABB* (PMOS only), (f) *ABB* (both NMOS and PMOS).

The SNM obtained by applying each circuit-level enhancement technique is shown in Fig. 3.10. The efficiency is measured by the SNM sensitivity S as follows:

$$S = \left| \frac{\partial V_{SNM}}{\partial V_X} \right| \quad (3.8)$$

where V_X is the word-line voltage for *WLM*, supply voltage for *DVS*, bit-line voltage for *NBL* or body bias voltage for *ABB*.

It can be seen that all these techniques have improved the cell stability by strengthening or weakening certain transistors. Even though the strengthening or weakening effect can be applied in the row (by modulating WL voltage) or column (by modulating bit-line or cell power supply voltages) direction, column-based approaches are preferred because they can blend well with column multiplexing and write masking [40]. A comparison of the functionalities of different adaptive techniques is shown in Table 3.1.

From the table, it can be seen that the four adaptive techniques improve the SNM in their own ways. *WLM* is not column-based but it is the most efficient in increasing the read SNM. *DVS* is efficient in increasing the read SNM but is inefficient in increasing the write SNM. *NBL* cannot increase read SNM but it is the most efficient in increasing the write SNM. *ABB* is not efficient in increasing either the read SNM or write SNM, but it can be used to mitigate

Sizing of SRAM Cell with Voltage Biasing Techniques for Memory and PUF Functions

Table 3.1 Comparison of different adaptive techniques

Technique	Column-based	$S(\text{Read})$	$S(\text{Write})$	Other drawbacks
WLM	×	0.325	0	-
DVS	✓	0.323	0.006	-
NBL	✓	0	0.908	-
ABB(NMOS only)	×	0.009	0.016	Triple-well process, Large area overhead
ABB(PMOS only)	×	0.026	0.006	
ABB(Both MOSFETs)	×	0.114	0.015	

inter-die variations. Although DVS and ABB may not improve write SNM efficiently, they have the potential to reduce write failures. This is because write SNM is not a strong indicator of dynamic write ability [44]. The latter is better analyzed by the relative strength of access transistors and pull-up PMOS transistors.

3.3.2 Problem formulation

Based on the above analysis, the design space of an SPUF is shown in Fig. 3.11. The objective of the design problem is to find an optimal point $(\mathbf{X}, \mathbf{M})$ that combines the key parameters $\mathbf{X}$ (e.g., W of its transistors) associated with SRAM cell dimension and the aggressiveness $\mathbf{M}$ (e.g. by changing the voltage V_X) of WLM, DVS, NBL adaptation, to minimize

Fig. 3.11 Design space exploration of SRAM cell for dual-mode SPUF design.

$$F(\mathbf{X}, \mathbf{M}) = \min_{\mathbf{X}, \mathbf{M}} \Pr(\text{RF}, \text{WF}) \quad (3.9)$$

subject to the following constraints:

$$\begin{cases} \text{reliability}(\text{ECC}[n, k, t]) \geq \text{reliability}_{\min} \\ \text{area_cell} \leq \text{area_cell}_{\max} \\ \text{leakage_cell} \leq \text{leakage_cell}_{\max} \\ \text{power_cell} \leq \text{power_cell}_{\max} \end{cases} \quad (3.10)$$

where $\text{ECC}[n, k, t]$ is the error correction code specified by the code word length n , number of information bits k and number of correctable errors t . The PUF reliability is measured by the complement of the bit error rate (BER) of its responses. The area_cell , leakage_cell and power_cell refer to the area, standby power and dynamic power dissipated during the read and write operations of the SRAM cell.

Algorithm 1 Design procedure for dual-mode SPUF

- 1: Initialize SRAM array and set $\text{BER}_{\max}$, $\text{area}_{\max}$, $\text{leakage}_{\max}$ and $\text{power}_{\max}$ for an SRAM cell.
 - 2: Size NMOS and PMOS transistors to optimize PUF reliability with the minimum size SRAM cell template $x_{\min} \in \mathbf{X}$.
 - 3: **while** constraints of $\text{BER}_{\max}$, $\text{area}_{\max}$, $\text{leakage}_{\max}$ and $\text{power}_{\max}$ are met **do**
 - 4: Increase x .
 - 5: **end while**
 - 6: Apply circuit-level technique(s) from {WLM, DVS, NBL} within a pre-determined voltage span.
 - 7: **while** power consumed by the SRAM cell in memory-mode with the selected circuit-level technique(s) $< \text{power}_{\max}$ and the adaptive voltage value(s) of applied technique(s) is/are within the span(s) **do**
 - 8: Change the aggressiveness of applied technique to reduce $\Pr(\text{WF}, \text{RF})$.
 - 9: **end while**
 - 10: End of design process.
-

The design problem can be solved by the procedure described in Algorithm 1. It should be noted that the ABB technique is not considered at design time but it is applied in the post-silicon tuning phase to calibrate against inter-die variation in order to reduce the degradation of cell reliability [41]. Its application alongside with other techniques is limited mainly by the constraint of hardware resources and target fabrication process. In Step 6 of Algorithm 1, different combinations of multiple techniques should be considered if one technique alone is inadequate to reduce $\Pr(\text{WF}, \text{RF})$ to meet the target. If $\Pr(\text{RF})$ dominates $\Pr(\text{WF}, \text{RF})$, priority should be given to WLM. Otherwise, if $\Pr(\text{WF})$ dominates, NLB should be included primarily. If $\Pr(\text{RF})$ and $\Pr(\text{WF})$ are comparable, then DVS should be chosen.

The variables and constraints involved can be evaluated as follows:

Failure probability

The overall failure probability of an SRAM cell can be estimated by [37]

$$\Pr(\text{RF}, \text{WF}) = \Pr(\text{RF}) + \Pr(\text{WF}) - \Pr(\text{RF} \cdot \text{WF}) \quad (3.11)$$

Sizing of SRAM Cell with Voltage Biasing Techniques for Memory and PUF Functions

The failure probabilities $\Pr(\text{RF})$ and $\Pr(\text{WF})$ can be efficiently estimated using the method proposed in [37].

Reliability

The reliability (i.e., $1 - \text{BER}$) of the raw PUF response can be estimated by:

$$\text{reliability}_{\text{raw}} = 1 - \frac{1}{m} \sum_{i=1}^m \frac{\text{HD}(R, R')}{n} \quad (3.12)$$

where m is the number of responses generated by the same challenge, n is the bit-length of the response, and $\text{HD}(R, R')$ is the Hamming distance between two responses, R and R' , to the same challenge generated by the same PUF at different time or different environmental conditions. With $\text{ECC}[n, k, t]$, the reliability can be improved to

$$\text{reliability}_{\text{final}} = \sum_{i=0}^t \binom{n}{t} (\text{BER}_{\text{raw}})^i (1 - \text{BER}_{\text{raw}})^{n-i} \quad (3.13)$$

Area

The area of an SRAM array is estimated by

$$\text{area}_{\text{array}} \approx N \times \text{area}_{\text{cell}} \quad (3.14)$$

where N is the number of SRAM cells and $\text{area}_{\text{cell}} = X_{\text{cell}} \times Y_{\text{cell}}$. Based on the layout of [37],

$$\begin{aligned} X_{\text{cell}} &= 5\lambda + 2\max(3\lambda, W_{M3}) + 2L \\ Y_{\text{cell}} &= 9\lambda + \max(3\lambda, W_{M2}) + \max(3\lambda, W_{M1}) + L \end{aligned} \quad (3.15)$$

where λ is half of the feature size. W and L are the channel width and length of the transistor, respectively. The area overhead incurred by any of the abovementioned circuit-level adaptive techniques is negligible compared to $\text{area}_{\text{array}}$.

Power consumption

The power consumption of the SRAM array can be calculated by

$$\text{power}_{\text{array}} = N \times \text{power}_{\text{cell}} \quad (3.16)$$

where $\text{power}_{\text{cell}}$ is the total power consumption of each SRAM cell. Its dynamic and leakage components can be obtained by the built-in function of SPICE simulator. During the read and write operations, dynamic power dominates. Otherwise, leakage current is the major contributor of $\text{power}_{\text{array}}$ [37].

3.4 Simulation Results and Discussions

To demonstrate the efficiency of the proposed optimization strategy, seven different designs were compared. Designs 1 and 2 were optimized for single-mode PUF and single-mode SRAM with the parameters of $\mathbf{X}$ set to the minimum and maximum values, respectively. The transistors of Design 3 were sized for the lowest $\Pr(\text{RF}, \text{WF})$ under the constraints of reliability, area and leakage without applying any circuit-level techniques. Design 4 uses WLM to improve the cell stability, with $V_{WL} = 0.8V$ in read operation and $V_{WL} = 0.9V$ in write operation. Design 5 uses DVS, with $V_{DD} = 1V$ in read operation and $V_{DD} = 0.8V$ in write operation. Design 6 uses NBL, with bit-line voltage $-0.1V$ applied in write operation. Design 7 is a ring oscillator (RO) PUF used for dual-mode application, i.e., as a PUF and a clock. The designs were simulated by HSPICE with CMOS 45 nm bulk Predictive Technology Model. The maximum standard deviation of V_{th} and Λ are assumed to be 15% of its nominal value and 100 nm, respectively. The channel length is set to 45 nm for all transistors. The ECC used is BCH-[127, 15, 27] [45]. The following design constraints were set: $\text{BER}_{\max} = 10^{-6}$, $\text{area_cell}_{\max} = 1\mu\text{m}^2$, $\text{leakage_cell}_{\max} = 50\text{nA}$ and $\text{power_cell}_{\max} = 200\mu\text{W}$.

Table 3.2 Simulation Results @ $V_{DD} = 0.9\text{ V}$, temperature = 25°C

Description	Design	$\mathbf{X}$	$\mathbf{M}$	$\Pr(\text{RF})$	$\Pr(\text{WF})$	$\Pr(\text{RF}, \text{WF})$	BER	Area	Leakage	Power
Single-mode PUF	1	(180, 90, 135)	N.A.	0.131	0.136	0.249	1.37×10^{-7}	0.45	25.2	140
Single-mode memory	2	(500, 250, 375)	N.A.	0.022	0.047	0.068	3.29×10^{-2}	1.28	73.2	405
Dual-mode SPUF	3	(202, 101, 152)	N.A.	0.118	0.125	0.229	9.73×10^{-7}	0.49	28.5	158
	4	(202, 101, 152)	WLM	0.021	0.125	0.143	9.73×10^{-7}	0.49	28.5	154
	5	(202, 101, 152)	DVS	0.027	0.036	0.062	9.73×10^{-7}	0.49	28.5	143
	6	(202, 101, 152)	NBL	0.118	0.004	0.122	9.73×10^{-7}	0.49	28.5	155
Dual-mode RO PUF	7	—	—			0.51	1.46×10^{-10}	—	—	—

¹The transistor width is searched in the range of [45, 500] nm, and the ratio $W_{M1} : W_{M2} : W_{M3}$ is the same for the three designs so that no memory failure occurs with nominal parameter values.

²The power refers to the sum of dynamic power during read and write operations.

³The data of RO PUF was from [46]. The failure probability was calculated with 0.5% frequency tolerance and the BER was obtained with the same ECC as SRAM PUF.

⁴The units for area, leakage and power are μm^2 , nA and μW , respectively.

The results are tabulated in Table 3.2. Design 1 has a low BER, which is good for PUF, but bad for memory mode due to its high $\Pr(\text{RF}, \text{WF})$. Comparing to Design 1, Design 2 has a lower $\Pr(\text{RF}, \text{WF})$ and a higher BER, which makes it a better choice for memory but a poor choice for PUF. Design 3 has a lower $\Pr(\text{RF}, \text{WF})$ than Design 1. Design 4 to Design 6 meet the power constraint but Design 5 has the lowest failure rate, which is due to the fact that both $\Pr(\text{RF})$ and $\Pr(\text{WF})$ are high and they can only be reduced simultaneously by DVS. However, by looking at $\Pr(\text{RF})$ and $\Pr(\text{WF})$ separately, WLM is the most efficient for $\Pr(\text{RF})$ reduction and NBL is the most efficient for $\Pr(\text{WF})$ reduction, which attests the postulation in Section 3.3.2. Compared with the dual-mode SPUF, RO PUF in Design 7 has a lower BER but a much higher functional failure rate. The optimal design (Design 5) reduces the failure rate in memory mode from the baseline design (Design 1) by 75%, while keeping the BER to an

Fig. 3.12 Frequency distribution of the simulated fractional inter-die HD.

acceptable range of $\leq 10^{-6}$ in PUF mode at the cost of a reasonably small area, leakage and power overheads of $\sim 8.9\%$, $\sim 13.1\%$ and $\sim 2.1\%$, respectively.

3.4.1 Uniqueness

The uniqueness of PUF can be measured by (2.2) as defined and described in Section 2.5.2.

To measure the uniqueness of the designed SRAM PUF, two hundred 100-bit-long responses are generated and their fractional HDs are calculated. The distribution is shown in Fig. 3.12. The uniqueness is calculated to be 0.4999, which closely approaches the ideal value of 0.5. The fitted Gaussian curve of the histogram for fractional HDs has a mean value of $\mu = 0.4999$ and a standard deviation of $\sigma = 0.0503$.

3.4.2 Randomness

The Randomness of PUF has been defined and described in Section 2.5.1. To measure the randomness of the augmented SRAM PUF design, one million response bits were generated and split into 100 blocks with 10000 bits each. The NIST randomness test results are shown in Table 3.3, which show good randomness of the PUF responses.

3.5 Conclusion

A novel methodology for designing SPUF for dual application modes has been proposed. The analysis in this chapter shows that the memory and PUF modes of operation are affected in a contradicting way by the device geometry of SRAM cells. Four circuit-level techniques are applied to alleviate the design constraints and expand the design space. With the proposed systematic approach to optimize the SRAM cell design for both memory and PUF requirements, the memory failure probability is reduced by 75% while the BER of PUF response is maintained

Table 3.3 NIST test results.

Test Description	Passed/Total	P-value	Pass?
Frequency	98/100	0.145	✓
Block Frequency ($m = 128$)	100/100	0.262	✓
Cusum-Forward	98/100	0.249	✓
Cusum-Reverse	99/100	0.817	✓
Runs	97/100	0.102	✓
Longest Run of Ones	98/100	0.868	✓
Rank	100/100	0.015	✓
Spectral DFT	100/100	0.024	✓
Non-overlapping Templates ($m = 9, B = 000000001$)	99/100	0.367	✓
Overlapping Templates ($m = 9$)	99/100	0.898	✓
Approximate Entropy ($m = 4$)	98/100	0.475	✓
Linear Complexity ($M = 1000$)	96/100	0.035	✓
Serial ($m = 16, \nabla \Psi_m^2$)	100/100	0.038	✓

below 10^{-6} . Meanwhile, the area, leakage current and dynamic power are kept to an acceptable level.

Chapter 4

A New Write-Contention Based Dual-Port SRAM PUF with Multiple Response Bits Per Cell

4.1 Introduction

Most embedded memories are designed based on six-transistor (6T) single-port SRAM (SP-SRAM) cell, which has only one data access port[47]. Such structure limits access speed and handicaps memory management because it can only support either one read or one write operation at any time. Due to inefficiency in switching a single port from one device to another, less than half of the maximum theoretical bandwidth of SRAM can be utilized by multiplexing. With the reduced price per bit of solid-state memories, there is considerable interest to explore multi-port SRAM to support the escalating data traffic between processor and external memory in high-speed System-on-Chip (SoC). Multi-port SRAM is capable of supporting multiple read and write accesses simultaneously, which greatly minimizes the requirement of on-chip arbitration logic and offers greater flexibility for memory management solutions to tear down the memory wall.

As presented in the previous chapter, SRAM is one of the most popular building blocks used for the construction of PUF for device authentication to ensure reliable management of sensitive data. Compared with other existing PUFs, SRAM-based PUF can reuse the existing circuit without requiring additional components to generate the response[5]. Existing SRAM based PUFs (SPUFs) developed out of 6T SRAM can only generate one response bit per cell by making use of the random power-up reset state. The exception being the current-based SPUF proposed in [5], which can achieve multiple bits per cell. However, additional current sensors are needed to read multiple cells, which introduce extra hardware and power overheads. In [48], a sub-threshold SPUF was designed based on 10T-SRAM cell to achieve low-power operation but it also relies on power-up states for the response bit generation. As power-up reset affects all cells in the memory, when the SPUF is switched from memory to PUF mode

A New Write-Contention Based Dual-Port SRAM PUF with Multiple Response Bits Per Cell

of operation, all data stored in the SRAM block need to be buffered into a temporary storage. This incurs extra area and consumes more power than necessary since not all the SRAM cells are addressed by the challenge. More importantly, any sensitive data originally stored in the SRAM are vulnerable to data remanence side-channel attacks during the data transfer.

In this chapter, a DP-SRAM based PUF (DP-SPUF) is proposed. Compared with other SPUF designs, it has the following advantages: (a) it can generate two independent response bits per cell; (b) there is no need to buffer all the data when switching between memory and PUF modes.

The remainder of this chapter is organized as follows. Section 4.2 introduces the different access modes of dual-port SRAM. Section 4.3 presents the mechanism of proposed DP-SRAM based PUF. Section 4.4 compares the proposed DP-SRAM based PUF against SP-SRAM based PUF. The chapter is concluded in Section 4.5.

Fig. 4.1 Single-port and dual-port SRAM cells: (a) 6T-SP-SRAM cell (1RW) (b) 8T-DP-SRAM cell (1RW) (c) 10T-DP-SRAM cell (1RW) (d) 8T-DP-SRAM cell (2RW).

Fig. 4.2 Access modes of DP-SRAM: (a) different row and column, (b) different row and common column, (c) common row and different column and (d) common row and column

4.2 Access Modes of Dual-Port SRAM

Fig. 4.1 shows four types of SRAM cell. The cell structure of the widely used six-transistor single-port SRAM (6T-SP-SRAM) is shown in Fig. 4.1(a). It has two pull-up PMOS transistors (M_2, M_4), two pull-down NMOS transistors (M_1, M_3) and two access NMOS transistors (M_5, M_6). Since it has only one shared read and write port, at any time, either a read or a write operation but not both can be performed. This data access mode is designated as “1RW” [47]. During power-up reset, when the supply voltage ramps up from 0 to V_{DD} , the voltages of the two nodes (BIT and BITB) will be pulled high but one of them will be pulled down due to the positive feedback of cross-coupled inverters (INV-1 and INV-2). The response bit of the cell is determined by the mismatch of the two inverters. This is the most common way to implement a PUF from SRAM. During PUF operation, the two access transistors (M_5, M_6) are turned off. Thus only four of the six transistors contribute to the response bit generation. Figs. 4.1(b) and (c) show the structures of an 8T SRAM cell and a 10T SRAM cell, respectively. They both have an additional port dedicated for read operation only. Hence, one read and one write operation can be carried out simultaneously at any time. Such access mode is referred to as “1R1W”. Fig. 4.1(d) shows an 8T dual-port SRAM (8T-DP-SRAM) which is capable of performing two read-write (“2RW”) operations concurrently. Compared with 6T SRAM, it has two more access transistors (M_7, M_8), one more word-line ($WL2$) and one more pair of bit-lines ($BL2, BL2B$). As the two ports are separated and both can perform read or write operation independently, it has more access-flexibility than the other three types of cell structure.

Fig. 4.3 8T-DP-SRAM operating in PUF mode (a) $BL1 = 1, BL2 = 0$ and (b) $BL1 = 0, BL2 = 1$.

For the 2RW 8T-DP-SRAM, when both ports are enabled simultaneously, there are four different access modes in total [47]. Figs. 4.2(a) and (b) show two of the access modes, where Ports 1 and 2 select their word-lines from two different rows. These two modes differ in their column access, either the bits in two different columns or the same column are accessed. The two ports will operate as single port access and will not interfere with each other. On the contrary, both word-lines are activated in the same row for the access modes shown in Fig. 4.2(c) (common row and different columns) and (d) (common row and common column). For these two access modes, as the two enabled word-lines will degrade the static noise margin of all cells in the selected row, care must be taken to ensure cell stability. Irrespective of the operation performed by the two ports, the read-stability of the half-selected (only row but not column is selected) cells is critical because this is equivalent to a read operation [47]. The two ports in Fig. 4.2(d) can encounter an access conflict as complementary bits may be written to the same cell, which is normally forbidden for memory operation. This contention can however be leveraged for PUF operation.

4.3 Proposed DP-SRAM based PUF

Unlike the power-up reset of 6T-SP-SRAM based PUF (henceforth abbreviated as SP-SPUF), the use of forbidden write conflict mode of 8T-DP-SRAM to produce the random response bits is proposed. The response bit generation process of an 8T-DP-SRAM based PUF (henceforth abbreviated as DP-SPUF) is illustrated in Fig. 4.3. When the two word-lines are activated at the same time, i.e., $WL1 = WL2 = 1$, a pair of complementary bits will be written onto the bit-lines ($BL1 = 1$ and $BL2 = 0$ or vice verse). In the absence of process variation, the voltages at node BIT and $BITB$ are identical due to symmetry. Because of the ineluctable device mismatches, a small deviation is always present between these two voltages. When both $WL1$ and $WL2$ are disabled, the positive feedback loop formed by $INV-1$ and $INV-2$ will stabilize the data bit to a stable state ($BIT = 1$ or 0). Since two bits of data (either 10 or 01 is written onto $BL1$ and $BL2$) can be written onto one cell, the 8T-DP-SPUF has the potential to generate more than one response bit per cell.

This scheme can harness more entropy from the manufacturing process variations as all the eight transistors of the memory cell are contributing to the arm race as opposed to only two-third of the transistors in the conventional SP-SPUF. The different sources of entropy that reign the final stable state can be analyzed in two phases, as depicted in Fig. 4.4. In the first phase, when the word-lines are enabled, the voltages, V_a and V_b of BIT and BITB, respectively will differ slightly. The entropy in this phase is introduced by the transistors $M_5 \sim M_8$. V_a is determined by the relative strength of M_5 and M_7 while V_b is determined by the relative strength of M_6 and M_8 . In the second phase, the two word-lines are disabled. V_{BIT} and V_{BITB} will be stabilized to either 1 or 0. If V_a is much larger than V_b , then V_{BIT} is more likely to be pulled up to 1 in the second phase. This case is shown in Fig. 4.4(a), where the response bit is mainly determined by the mismatches among $M_5 \sim M_8$. If the difference between V_a and V_b is negligible, the response bit will be determined by the mismatches among $M_1 \sim M_4$. Thus all the 8 transistors have contributed to the response generation process to enrich its entropy.

Fig. 4.4 (a) Response bit determined mainly by mismatches of access transistors. (b) Response bit determined mainly by mismatches of cross-coupled inverters.

Such bit generation scheme has also enhanced the security of the whole system. As shown in Fig. 4.5, to preserve the data originally stored in the 6T-SP-SRAM memory, they have to be temporarily transferred to a buffer before resetting the SRAM to generate the secret key, and the data have to be written back into the SRAM after that [1]. These data buffering operations not only incur additional area and power overheads but also increase the vulnerability to fault-injection attacks. Such attacks make use of remanence decay side channel information [49] by iteratively powering up the SRAM with different power-off times to exhaustively find the key. For the ease of design, most SRAMs are powered by a single supply [50]. The entire SRAM block needs to be powered off and on even though only a small number of cells addressed by the challenge are required to generate the response, which consumes excessive spurious power. The proposed DP-SPUF makes use of write conflict for response generation. Without relying on power-up reset, buffering and hence data remanence decay attacks can be eliminated. As the cells addressed by the challenge can be independently written under write-conflict mode, no spurious power consumption is incurred for response bit generation.

Fig. 4.5 Mode switching of conventional SRAM PUF [1].

Fig. 4.6 Response bits generated from the same cell written with complementary data bits

4.4 Simulation Results and Discussion

The qualities of DP-SPUF and SP-SPUF are simulated using 22nm Predictive Technology Model LP model. For a fair comparison, the same sizing and process variation are assumed for both PUFs. The widths of their pull-up, access and pull-down transistors are set to 44 nm, 66 nm and 88 nm, respectively, to ensure that there is no memory read and write failures. The lengths of all the transistors are set to 22 nm. The noises are set as $f_{min} = 10^6\text{Hz}$ and $f_{max} = 10^{11}\text{Hz}$, where f_{min} and f_{max} are respectively the base and maximum frequencies used for modeling frequency dependent noise sources [34]. The threshold voltage is dependent on transistor geometry and its process variation is determined by (3.1). Λ is assumed to be 5 nm and $\sigma_{V_{th,max}}$ is assumed to be 15% of the nominal value of V_{th} .

4.4.1 Correlation between bits in a cell

To achieve multiple bits per cell, the responses generated from the two different cases shown in Fig. 4.3 should not be correlated. Fig. 4.6 shows the simulation results of the response bits generated from 25 cells, which shows no obvious relationship between the response bits between these two cases. Further simulation results on 1000 cells show that their Pearson correlation coefficient is -0.036 (calculated using MATLAB built-in function "corrcoef"), which is very close to the ideal value of 0. It can be concluded that the cells are uncorrelated and each cell is capable of generating two independent response bits.

4.4.2 Reliability

The reliability of a PUF can be measured by the bit error rate (BER) of its responses when the same challenge is applied to the same PUF instance at different environmental conditions. The main environmental factors that may affect the PUF reliability are temperature and supply voltage. The reliability S of the PUF can be calculated by (2.3) as discussed in Section 2.5.3.

Fig. 4.7 shows the reliability calculated from five hundred challenge-response pairs generated by each type of SPUF under different temperatures and supply voltages. Over a wide range of combinations of temperature variation from -40°C ~ 120°C and voltage variation from 0.75V ~ 1.15V , DP-SPUF exhibits higher reliability than SP-SPUF in most cases. The average reliability of DP-SPUF is 96.2%, which is higher than the 93.7% reliability of SP-SPUF. The worst-case reliability of DP-SRAM is 87.6%, which is comparable though slightly lower than 89.6% of SP-SRAM. This is attributed to the subthreshold operation of access transistors when the supply voltage is too low. When the minimum supply voltage is raised from 0.75V to 0.8V , the worst-case reliability of DP-SPUF is higher than SP-SPUF.

Fig. 4.7 Reliability of (a) single port SRAM PUF and (b) dual port SRAM PUF for simultaneous temperature and voltage variations.

4.4.3 Uniqueness

Fig. 4.8 Fractional Hamming distance histograms of (a) SP-SPUF and (b) DP-SPUF.

Table 4.1 NIST randomness tests for SP-SPUF and DP-SPUF .

Statistical	P VAL		PROP		PASS?	
	SP	DP	SP	DP	SP	DP
Frequency	0.19	0.97	49/50	49/50	Y*	Y
BlockFrequency	0.66	0.62	47/50	49/50	Y	Y
CumulativeSums(forward)	0.11	0.97	49/50	49/50	Y	Y
CumulativeSums(backward)	0.15	0.85	49/50	48/50	Y	Y
Runs	0.53	0.21	48/50	50/50	Y	Y
LongestRun	0.32	0.62	49/50	49/50	Y	Y
Rank	0.24	0.11	50/50	49/50	Y	Y
FFT	0.53	0.38	50/50	50/50	Y	Y
OverlappingTemplate	0.49	0.26	50/50	49/50	Y	Y
ApproximateEntropy	0.07	0.70	47/50	50/50	Y	Y
Serial(forward)	0.70	0.85	49/50	49/50	Y	Y
Serial(backward)	0.96	0.02	49/50	49/50	Y	Y
LinearComplexity	0.62	0.96	48/50	49/50	Y	Y

* Y denotes YES while N denotes NO.

The uniqueness of the PUF is measured by calculating the inter-die HD. For maximum discriminability, half of the response bits generated from any two PUF instances to the same challenge should be different. Thus the ideal uniqueness should be 50%.

Fig. 4.8 shows the distributions of the fractional HD for SP-SPUF and DP-SPUF. Pairwise comparison of two hundred responses were made to evaluate the uniqueness. It can be seen that both SPUFs have uniqueness close to the ideal 50%.

4.4.4 Randomness

The cryptographic randomness of the PUF responses is measured by the National Institute of Standards and Technology (NIST) test. One million response bits were generated and split into 50 blocks with 20,000 bits each for this test. As shown in Table 4.1, thirteen statistical tests were performed. To pass each test, the P-value obtained by applying a chi-square test at 1% level should be larger than 0.01 and the minimum proportion of binary sequence (PROP) that passed the tests is 47/50. The results indicate that both PUFs satisfy the randomness requirements.

4.4.5 Power

To estimate the power consumption for generating a response bit, the rise time and fall time of WL and V_{DD} are both set as 2ns. DP-SRAM consumes $2.42\mu\text{W}$ and SP-SRAM consumes 89.3nW. Even though DP-SRAM consumes more power for generating one response bit, it does not need to transfer all the data to the buffer storage or power up all the cells together. As typically SRAM size is at least a few Kb or Mb, the total power consumption of DP-SRAM should be less than that of SP-SRAM.

4.5 Conclusion

This chapter has presented a new method to generate multiple response bits per cell from 8T-DP-SRAM. Simulation results using PTM 22nm process show that the proposed 8T-DP-SPUF has better reliability than the traditional 6T-SP-SPUF. As the SRAM does not need to be reset to generate the response, data buffering during mode switching is limited to only those cells that are addressed by the challenge, which greatly minimizes its risk of side-channel attacks over the power-up reset based memory PUFs. Such features are important to multiple access memory management for enhanced data integrity.

Chapter 5

ACRO-PUF: A Low-power, Reliable and Aging-Resilient Current Starved Inverter Based Ring Oscillator Physical Unclonable Function

5.1 Introduction

RO PUF is a popular lightweight PUF candidate for device authentication due to its security, low complexity, and ease of implementation and evaluation [3, 51]. The main drawback of RO PUF is its poor response reliability against environmental condition variations. To enhance the reliability, an 1-out-of- k masking scheme is proposed in [3] to select a pair of ROs that has the maximum frequency distance from k pairs. In [14], a RO PUF with multi-level supply voltage is proposed. Instead of having a single power supply, the inverters in a RO are powered by multiple supply voltages so that the voltage configuration with the highest reliability can be selected. Thin transistors and shorter RO stage length have been used to improve the PUF response stability in [15]. To further improve the reliability, the authors proposed feedback based supply voltage control scheme. A temperature-aware cooperative RO PUF was proposed in [17] so that different RO pairs can cooperate to generate reliable bits. The above four methods for RO PUF reliability enhancement incur either high hardware redundancy or large additional memory. Two methods were proposed to increase the stability the RO PUF against temperature in [11]. By applying optimized supply voltage or adding two source feedback resistors with negative temperature coefficient of resistance to the regular inverters, the temperature sensitivity of RO PUF can be reduced. However, to maintain high reliability, noise on power supply lines will have to be substantially minimized. Recently, a lightweight hybrid RO PUF [10] with high thermal stability is proposed, but the reliability is achieved at the expense of degraded reliability against supply voltage variation.

Many techniques have been proposed to improve the RO PUF reliability against variations in temperature and supply voltage. Only a rare few have addressed the reliability degradation due to aging [52, 53, 51, 54, 55]. Software techniques were proposed to eliminate the aging effect on PUFs in [52]. The solutions employed generic protocol-level techniques but there is no implementation to demonstrate their effectiveness. In [53], a configurable RO based on the 1-out-of- k scheme of [3] is proposed to improve the PUF reliability against aging. By using tunable ROs, a pair of ROs with maximum frequency difference from eight pairs is chosen to generate the responses. Redundant ROs are introduced in this approach. In [51], an aging-resistant RO PUF is proposed for reliable key generation by mitigating the NBTI and HCI effects when the PUF is not in use. However, the method did not adequately reduce the aging effect when the PUF is active. Even when the PUF is not in use, the gate-source voltage of PMOS is not zero. As instabilities exist whenever a negative bias is applied to the transistors [56], the PMOS transistors will still experience some NBTI even though less severe than those of the regular RO PUF. In [54], a selection algorithm was proposed to select highly reliable RO pairs by taking into consideration both initial frequency difference and aging rate. To achieve high reliability, a lot of measurements need to be done and redundant RO pairs are required. More recently, the aging tolerant RO-PUF proposed in [55] mitigates the NBTI effect by adding an NMOS to drive the supply voltage into the cascaded RO inverters. The ROs will oscillate at lower frequency due to the lower operating voltage and the inverters will be isolated from V_{DD} when the ROs are in non-oscillating mode. Such design will reduce both NBTI and HCI effects.

This chapter is an extension of the author's work in [57]. Instead of addressing only the reliability problem due to temperature variation, the reliability against supply voltage fluctuation and device aging have been improved. Specifically, the following contributions have been made:

1. A robust RO PUF based on current starved (CS) inverters is proposed by taking into consideration the intricate counteracting effects of temperature and supply voltage fluctuations on deviation of RO frequency from nominal working condition and frequency variance of ROs when the CS inverters are operating in saturation and subthreshold region. A design flow has been proposed to find the optimal biasing point of the CS inverter to achieve a very high overall reliability against both types of environmental variations.
2. A simple proximity detector circuit is proposed to filter out a small fraction of unreliable response bits to bring the reliability to 100% or near 100%.
3. An irregular clocking to the linear feedback shift register (LFSR) is proposed to de-correlate the order of successive RO pairs from different input challenges.
4. A low-cost aging-resistant CS inverter circuit is proposed to enable the proposed RO PUF to operate reliably over a longer period of time.

Fig. 5.1 (a) Basic architecture of RO PUF. (b) Regular inverter. (c) CS inverter.

The rest of the chapter is organized as follows. In Section 5.2, the impacts of temperature, supply voltage and aging on RO frequency and PUF reliability are analyzed. Section 5.3 presents the methodology to alleviate these negative effects, as well as the structural revamp of CS RO PUF to further enhance its reliability and security. Section 5.4 presents the simulation results and comparison of the proposed design. The chapter is concluded in Section 5.5.

5.2 Impact of Environmental Conditions and Aging on RO frequencies

The basic structure of a traditional RO PUF is shown in Fig. 5.1(a). It consists of n identically designed ROs, two multiplexers, two counters and a comparator. Each RO is enabled by a two-input NAND gate (which also acts as an inverter when enabled). Due to the manufacturing process variations, these ROs will oscillate at different frequencies even though they have identical number of inverter stages. Each of the two n -bit wide multiplexers will select one of the n ROs to its counter. The counter counts the number of transitions of the selected RO, and the comparator compares the two counter outputs to produce a response bit [3]. Typically, the inverters used to build the RO are regular inverters as shown in Fig. 5.1(b). By adding two transistors to the regular inverter, a CS inverter is formed, which is shown in Fig. 5.1(c). The extra transistors are used to control the drain current of CS inverter. In this chapter, the regular inverters of the ROs are replaced by CS inverters to improve the PUF reliability. The motivation behind this is best appreciated by analyzing the impacts of operational condition variations on the oscillator frequency due to both types of inverters.

Fig. 5.2 (a) Conditions for stable and unstable response bits generated by regular RO pair. (b) Stable response bits are generated by CS RO pair with optimal biasing.

5.2.1 Effect of Temperature Variation on RO Frequency

Upon manufacturing, the response bit of a RO PUF to a challenge is reliable if one RO in the selected pair always oscillates faster than the other. However, the response bit to the same challenge may not be reproducible when the PUF is operating at a different temperature as illustrated in Fig. 5.2(a). The deviation of RO frequency due to the change in temperature depends on whether its inverters are regular or current-starved. Both cases are analyzed as follows.

For Regular Inverter Based RO

The oscillation frequency of a RO is inversely proportional to the summation of the propagation delay t_d of every inverter stage. The RO frequency can be expressed as [58]:

$$f = \frac{1}{2mt_d} \quad (5.1)$$

where m is the number of inverter stages.

As a first order estimation, t_d can be expressed as [10]:

$$t_d = \frac{C_0 V_{DD}}{\eta I_D} \quad (5.2)$$

where C_0 is the total load capacitance, V_{DD} is the power supply voltage, ηI_D is the mean current, η is a fixed parameter for a given inverter and I_D is the drain current of MOSFET in saturation.

The saturation current I_D can be estimated by [11]

$$I_D = \frac{\mu C_{OX} W}{2L} (V_{GS} - V_{th})^\alpha \quad (5.3)$$

5.2 Impact of Environmental Conditions and Aging on RO frequencies

where $\mu, W, L, C_{OX}, V_{GS}, V_{th}$ and α are the mobility of majority charge carrier, effective channel width, effective channel length, gate capacitance per unit area, gate-to-source voltage, threshold voltage and velocity saturation index, respectively.

The temperature coefficient of current (TCC) [59] is defined as the change in drain current per degree Celsius of temperature change. It can be derived from (5.3) as follows:

$$TCC = \frac{1}{I_D} \frac{dI_D}{dT} = \frac{1}{\mu} \frac{d\mu}{dT} - \frac{\alpha}{V_{GS} - V_{th}} \frac{dV_{th}}{dT} \quad (5.4)$$

The two temperature dependent parameters, V_{th} and μ , can be expressed as [11]:

$$\mu(T) = \mu(T_0) \left(\frac{T}{T_0} \right)^{-\kappa} \quad (5.5)$$

$$V_{th}(T) = V_{th}(T_0) - \sigma(T - T_0) \quad (5.6)$$

where T_0 is the reference temperature, κ is the mobility temperature exponent that assumes a value between 1.2 to 2, and σ is the mobility temperature coefficient of threshold voltage which falls in the range between 0.5 to 3 mV/K.

When temperature increases, $V_{th}(T)$ decreases and the drain current increases. Meantime, the carrier mobility also decreases, which reduces the drain current. For regular inverter, the reduction of carrier mobility dominates. Thus TCC is negative.

For CS Inverter Based RO

By varying its bias, V_{ctrl} or V_{bn} (see Fig. 5.1(c)), the delay of CS inverter can be adjusted to counteract the effect of temperature induced drain current variation. From (5.4), the TCC of a CS inverter changes with V_{GS} . The bias V_{bn} for zero TCC can be calculated by setting the TCC of (5.4) to zero. According to [59], $V_{bn,ZTC}$ for zero TCC can be approximated by

$$V_{bn,ZTC} = V_{th} + \alpha \frac{dV_{th}}{dT} \bigg/ \left(\frac{1}{\mu} \frac{d\mu}{dT} \right) \quad (5.7)$$

The frequency deviations of CS RO against temperature with the V_{ctrl} (V_{bn}) bias varied from 0.3V to 1V are shown in Fig. 5.3. It should be noted that as V_{th} is around 0.5 V, the transistors work in sub-threshold region when V_{ctrl} falls below this voltage. For comparison, the frequency deviation of regular RO against temperature is also shown in this figure. The normal ambient temperature is 27 °C. The frequency increases with temperature when V_{bn} is low, and decreases with temperature when V_{bn} is high. At an optimal bias of $V_{bn,ZTC}$ estimated by (5.7), the RO frequency is almost constant with respect to temperature. It should be noted that when V_{GS} is larger than $V_{bn,ZTC}$, the larger the V_{GS} , the more negative the TCC. Conversely, when V_{GS} is smaller than $V_{bn,ZTC}$, decreasing V_{GS} makes TCC more positive. When TCC becomes too positive or too negative, the RO frequency is susceptible to temperature change and hence the RO PUF reliability may deteriorate. For CS RO, as V_{bn} decreases, TCC first increases from negative to positive, and then becomes increasingly positive, especially when

Fig. 5.3 (a) Relative frequency deviations against temperature of CS RO with V_{ctrl} varied from 0.3V to 0.7V. (b) Relative frequency deviations against temperature of regular RO and CS RO with V_{ctrl} varied from 0.8V to 1V.

the footer and header transistors enter into the sub-threshold region of operation. It can be seen that at $V_{bn} = 0.3$, 400% deviation from the frequency at ambient temperature can be inflicted on the RO. The reason for the continual rise of TCC can be analyzed by deriving the TCC for sub-threshold conduction. The subthreshold drain current can be expressed as [10]:

$$I_{D,sub} = \mu C_{OX} \frac{W}{L} \left(\frac{\kappa_B T}{q} \right)^2 (n-1) e^{\frac{q(V_{GS}-V_{th})}{n\kappa_B T}} \left(1 - e^{-\frac{qV_{DS}}{\kappa_B T}} \right) \quad (5.8)$$

$$n = \frac{1 + (C_S + C_{it})}{C_{OX}}$$

where C_S , C_{it} and C_{OX} are the capacitances of the semiconductor, fast surface states and oxide layer, respectively [59].

Assuming that $V_{DS} \gg \kappa_B T/q$, the temperature coefficient of the switching current TCC_{sub} can be derived as:

$$TCC_{sub} = \frac{1}{\mu} \frac{d\mu}{dT} + \frac{2}{T} - \frac{q}{n\kappa_B T} \left(\frac{dV_{th}}{dT} + \frac{V_{GS} - V_{th}}{T} \right) \quad (5.9)$$

This equation shows that TCC_{sub} increases as long as V_{GS} decreases, which agrees with the simulation results. It can also be seen that TCC of RO will continue to increase as V_{ctrl} decreases regardless of saturation or sub-threshold region of operation. This effect is especially pronounced in the sub-threshold region as the frequency deviation can raise to several times that of the frequency at nominal ambient temperature.

At zero TCC biasing voltage, the oscillation frequency of CS RO is relatively insensitive to temperature variation. Consequently, the reliability of RO PUFs can be improved, as illustrated in Fig. 5.2(b), by using CS inverters with appropriate biasing. When V_{ctrl} is too low, the reliability of RO PUF is likely to suffer from the high frequency deviation from the nominal temperature.

Fig. 5.4 (a) Conditions for stable and unstable response bits generated against voltage variation. (b) Stable response bits are generated with enlarged frequency variation.

5.2.2 Effect of Supply Voltage Variation on RO Frequency

RO frequency also increases with supply voltage. The response bit of a RO PUF can therefore be flipped due to supply voltage change. The bit flipping mechanism is similar to that of temperature variation, which is shown in Fig. 5.4(a). The risk of response bit flip arises when the frequencies of two ROs change at different rates. The deviation of RO frequency due to the change in supply voltages is again different between regular inverter and CS inverter based ROs, as analyzed below.

For Regular Inverter Based RO

From (5.1), (5.2) and (5.3), it can be concluded that

$$f = \frac{\eta \mu C_{ox} W}{4mC_0 L} \frac{(V_{GS} - V_{th})^\alpha}{V_{DD}} \propto \frac{(V_{GS} - V_{th})^\alpha}{V_{DD}} \quad (5.10)$$

For short channel device, due to drain induced barrier lowering (DIBL) effect, V_{th} is also affected by the drain voltage [60] besides temperature. The DIBL effect on V_{th} can be manifested as:

$$V_{th}(V_{DS}) = V_{th}(0) - \lambda_D V_{DS} \quad (5.11)$$

where $V_{th}(0)$ is the threshold voltage at zero drain voltage, and $\lambda_D = -\partial V_{th} / \partial V_{DS}$ is the DIBL coefficient.

For regular RO, V_{GS} increases with V_{DD} but V_{th} decreases with V_{DD} . As α typically lies between 1 and 2, the numerator of (5.10) will increase faster than the denominator, which results in an increase in frequency.

Fig. 5.5 Relationship between V_{bp} and $|V_{GSp}| = V_{DD} - V_{bp}$ with V_{DD} .

For CS Inverter Based RO

For CS RO shown in Fig. 5.1(a) and (c), it is obvious that V_{GS} of NMOS transistors, M4 and M5, does not change. As the PMOS header M6 works as a diode-connected load, its drain current is equal to that of NMOS M5. Consequently, V_{SGp} ($= V_{DD} - V_{bp}$) also remains relatively constant. This can be verified by the simulation result shown in Fig. 5.5.

Since only V_{th} in the numerator of (5.10) decreases, the frequency of CS RO will deviate less than that of regular RO as V_{DD} increases. Due to the limited change in V_{GS} of the footer and header transistors of CS RO, its relative frequency deviation $\Delta f/f$, which is defined as the frequency deviation normalized by the frequency at nominal supply voltage, is lower than that of regular RO, as shown in Fig. 5.6. It should be noted that when V_{ctrl} is less than 0.5 V, the footer and header transistors operate in the sub-threshold region. In this region, the relative frequency deviation increases as V_{ctrl} continues to decrease. Because V_{th} appears in the exponent of (5.8), DIBL dominates other factors in determining the sensitivity of drain current against supply voltage variation.

The above analysis indicates that the frequency deviation against supply voltage variation can be desensitized by lowering the biasing voltage V_{ctrl} in the saturation region. In sub-threshold region, the RO frequency will deviate more from its nominal operating frequency when V_{ctrl} decreases.

In Section 5.2.1, the analysis is made based on the assumption that the supply voltage is constant. The DIBL effect kicks in when the drain-source voltage increases with the supply voltage. This will result in the lowering of V_{th} , which in turn reduces $V_{bn,ZTC}$. The simulated relationship between $V_{bn,ZTC}$ and V_{DD} is shown in Fig. 5.7(a). $V_{bn,ZTC}$ decreases when V_{DD} increases, which can be analyzed from (5.7) and (5.11). Fig. 5.7(b) shows the RO frequency for each $V_{bn,ZTC}$ corresponding to a different V_{DD} . It is evident that the stable frequency for different V_{DD} is different. Thus the method [57] that reduces frequency deviation by $V_{bn,ZTC}$ is inadequate to stabilize the RO PUF response against changes in both temperature and supply voltage.

Fig. 5.6 Relative frequency deviation against supply voltage for regular RO and CS RO with V_{ctrl} from 0.3V to 1V.

Fig. 5.7 (a) $V_{bn,ZTC}$ at different supply voltages. (b) Frequency at corresponding $V_{bn,ZTC}$ for different supply voltages.

5.2.3 Effect of Device Aging on RO frequency

Apart from the supply voltage and temperature variations, aging can also deteriorate the PUF reliability over time. Aging is caused by Negative Bias Temperature Instability (NBTI), Hot Carrier Injection (HCI), Time Dependent Dielectric Breakdown (TDDB) and Electro-migration, etc. [61, 55]. Among which, NBTI and HCI dominates the performance degradation of PMOS and NMOS transistors, respectively. NBTI causes immediate reliability concern in PMOS transistors as they are usually operated with negative gate-source bias. In high-k technology, NMOS can also be degraded by Positive Bias Temperature Instability (PBTI) even though it is less severe than NBTI [61, 26]. Both PBTI and NBTI are induced by stress on the transistor gate (V_{GS} is negative for PMOS and positive for NMOS). The increase in V_{th} for PMOS in the stress phase is given by [62]:

$$\Delta V_{th, stress} = A_{NBTI} \times t_{ox} \times \sqrt{C_{ox}(|V_{GS}| - V_{th})} \times e^{\left(\frac{|V_{GS}| - V_{th}}{t_{ox} E_0} - \frac{E_d}{\kappa T}\right)} \times t_{stress}^{0.25} \quad (5.12)$$

Fig. 5.8 (a) Conditions for stable and unstable response bits due to aging. (b) Stable response bits are generated by slowing down the aging process.

where A_{NBTI} is a constant proportional to the aging rate, t_{ox} is the oxide thickness, E_0 and E_α are the device dependent parameters and t_{stress} is the duration of stress.

When logic ‘1’ is applied to the gate of PMOS, the transistor will be turned off and some of the traps will be eliminated. During the recovery phase when the stress is removed, V_{th} tends to shift back to its original value. Considering both stress and recovery phases, the final increase in V_{th} is given by:

$$\Delta V_{th} = \Delta V_{th, stress} \times \left(1 - \sqrt{\frac{\eta \times t_{rec}}{t_{stress} + t_{rec}}}\right) \quad (5.13)$$

where t_{rec} is time for recovery and η is a constant.

While PMOS is more susceptible to NBTI, NMOS is more vulnerable to HCI aging since electrons has higher mobility than holes [61]. The increase in V_{th} due to HCI can be expressed as:

$$\Delta V_{th} = A_{HCI} \times \alpha \times f \times e^{\frac{V_{DD} - V_{th}}{t_{ox} E_1}} \times t^{0.5} \quad (5.14)$$

where A_{HCI} is a constant proportional to the aging rate, α and f are the activity factor and the frequency, respectively, and E_1 is a device dependent constant.

5.3 Design of Robust CS RO PUF against Environmental Variations and Aging

5.3.1 Reliability Enhancement of RO PUF against Temperature and Voltage Variations

From the analyses of RO frequency deviations against temperature and supply voltage variations in Sections 5.2.1 and 5.2.2, it is intuitive that RO PUF reliability can be enhanced by minimizing the frequency deviation of each RO as shown in Fig. 5.2. Owing to the conflicting criteria,

Fig. 5.9 Coefficient of variation of frequency $\frac{\sigma_f}{\mu_f}$ at different V_{ctrl} for CS RO.

it may not be possible to find a control knob to simultaneously neutralize both effects of temperature and supply voltage variations on RO frequency. A viable alternative is to suitably increase the frequency variance of different ROs within the PUF under the same operating condition, as shown in Fig. 5.4(b). From (5.1) and (5.3), it can be concluded that the RO frequency f is proportional to the drain currents I_D of its inverters. By enlarging the variance of I_D , the variance of f can be increased. Differentiating (5.3) with respect to the threshold voltage,

$$\frac{\partial I_D}{I_D} = \frac{-\alpha \times \partial V_{th}}{V_{GS} - V_{th}} \quad (5.15)$$

By keeping ∂V_{th} constant while decreasing V_{GS} , $|\partial I_D/I_D|$ will increase, which leads to an increase in the coefficient of variation of drain current σ_{I_D}/μ_{I_D} . As f is proportional to I_D , this means that reducing V_{ctrl} will also increase the coefficient of variation of frequency σ_f/μ_f among the identically designed ROs, as evinced by the simulation result of σ_f/μ_f versus V_{ctrl} in Fig. 5.9.

Thus, V_{ctrl} of the CS inverter can be used as a key control knob to improve the CS RO PUF reliability against supply voltage variation as well as temperature variation. However, according to (5.4), TCC will increase as V_{ctrl} decreases. Whether the CS RO PUF reliability against temperature variation will improve with the reduction of V_{ctrl} depends greatly on which counteracting effect is stronger. In sub-threshold region, due to the dominant effect of temperature variation on frequency, the effectiveness of reducing V_{ctrl} to improve the PUF reliability is substantially reduced. In other words, when the footer and header transistors are working in the saturation region, reducing V_{ctrl} increases the frequency variance across different ROs and decreases the frequency deviation of each RO at different operating conditions, leading to improvement in the reliability of RO PUF. As the footer and header transistors enter into the sub-threshold region, reducing V_{ctrl} increases the frequency variance across ROs on one hand, but it also enlarge the frequency deviation of each RO under different operating conditions on the other hand. As a result, this may not necessarily improve the response reliability. The dependence of the trajectories of frequency deviation and frequency variance on V_{ctrl} is shown

Fig. 5.10 Effect of V_{ctrl} on the trajectories of frequency deviation and frequency variance of CS RO PUF.

in Fig. 5.10. An optimal point can be strived to achieve the best overall PUF reliability, where the reliability against temperature variation is higher than against supply voltage variation when V_{ctrl} is higher than this point, and vice versa when V_{ctrl} is lower than this point.

The design flow shown in Fig. 5.11 is proposed to find this optimal V_{ctrl} for a highly reliable CS RO PUF. Given the environmental condition range and V_{ctrl} range, the BERs for each sampled environmental condition and V_{ctrl} are measured. For each V_{ctrl} , the worst-case BERs against temperature and against V_{DD} are found and plotted in a Cartesian plane. Then the points for each coordinate of temperature and V_{DD} are independently fitted using a polynomial of single indeterminate. Based on this observation, third order polynomials are good approximation to minimize under-fitting and over-fitting. The interaction of the two curves is the optimal biasing point and it should fall within the pre-defined V_{ctrl} range unless the pre-defined V_{ctrl} range is too narrow. If there is no intersection due to the narrow V_{ctrl} window, the optimal V_{ctrl} will be selected from a point within the range that has the lowest BER between the two curves. Based on this design flow, it is assumed that the nominal operating condition is 27 °C and 1.1 V. To operate the PUF over a temperature range of -40°C ~ 120 °C, with potential fluctuation of supply voltage V_{DD} between 0.9V ~ 1.3V, the optimal V_{ctrl} is 0.45V over the tuning range of 0.3V ~ 1V, as shown in Fig. 5.12.

5.3.2 Aging-Resistant Biasing Circuit Design

It is observed from (5.12) and (5.13) that ΔV_{th} can be reduced by decreasing $|V_{GS}|$ or improving t_{rec} . Both remedies are not applicable with regular inverter. With the CS inverter structure, $|V_{GS}|$ can be reduced by adding two transistors M7 and M8, as shown in Fig. 5.13, to change the biasing voltages V_{bp} and V_{bn} of CS inverter. The multiplexer MUX in Fig. 5.13 is implemented by a transmission gate. When the ROs are oscillating (i.e., EN = '1'), the transistors will experience AC stress [51]. During this period, M7 is turned on, M8 is turned off and V_{ctrl} is connected to V_{ref} through the multiplexer. The equivalent circuit is shown in Fig. 5.13(b). As V_{bp} and V_{bn} lie above 0 V and below V_{DD} , the stress on M1 and M4 is much less than the

5.3 Design of Robust CS RO PUF against Environmental Variations and Aging

Fig. 5.11 Design flow for high reliability CS RO PUF

corresponding transistors of regular inverter. When CS RO PUF is not working, EN will be connected to ground. The equivalent biasing circuit is shown in Fig. 5.13(c). During this period, V_{bp} is connected to V_{DD} and V_{bn} is connected to ground. Hence no stress is exerted on the transistors, which effectively increases t_{rec} . This is unlike the regular inverter used in typical RO PUF, which will still suffer from DC stress since half of its PMOS transistors will experience NBTI and half of its NMOS transistors will experience PBTI. In summary, the proposed biasing circuit reduces the NBTI and PBTI experienced by the pull-up and pull-down transistors, respectively of the CS inverter, as well as increases their recovery time.

As hot electrons are generated during logic transitions, higher oscillation frequency will cause more severe HCI aging. This can also be inferred from (5.14). For the same number of stages, CS RO PUF has lower oscillation frequency than regular RO PUF, and hence comparatively lower HCI aging is inflicted on it.

Two aging-resilient RO PUFs have been proposed in the literature [55, 51]. Comparing with these designs, the proposed CS RO PUF operates at much lower frequency, which has reduced HCI aging. Comparing with the aging-resistant RO (ARO) PUF in [51], the PMOS transistors in the proposed design are fully turned off, which reduce NBTI aging. It should be noted that the stress on M8 when the PUF is in non-oscillating mode does not affect the PUF reliability because it only acts as a switch. Comparing with the aging-tolerant (AT) RO PUF in [55], not only the PMOS transistors in the proposed design are fully disconnected from the power supply, but the NMOS transistors in the proposed design are also disconnected from ground, which free them from both NBTI and PBTI.

Fig. 5.12 Worst-case BER against both temperature and voltage variations for different V_{ctrl} .

5.3.3 Additional Architectural Enhancements of ACRO-PUF

The architecture of the proposed ACRO-PUF is shown in Fig. 5.14(a). It consists of an 8-bit LFSR, a RO PUF core and a bidirectional counter. The LFSR enciphers the input challenge for the selection of RO pairs. The RO PUF core consists of $256 (= 2^8)$ ROs. Each RO consists of an NAND gate and eight inverters, as shown in Fig. 5.14(b). Every internal challenge generated by the LFSR will select one RO. Unlike the traditional RO PUF architecture shown in Fig. 5.1(a), which has two multiplexers, two counters and one comparator, the proposed design only has one multiplexer and one bidirectional counter. The LFSR will generate two internal challenges to select two ROs sequentially to generate a response bit. The bidirectional counter will count up by the number of transitions of the first selected RO and count down by the number of transitions of the next selected RO to generate an output response bit. The response bit is '0' if the first RO has a higher frequency and '1' otherwise. Such architectural change makes it possible to further enhance the response reliability and obfuscate the input challenges at negligible cost, which will be elucidated below.

Proximity Detector for Reliability Enhancement

Although the thermal stability of RO PUF response can be improved by using CS ROs, some output bits may still be unstable if the oscillation frequencies of their selected RO pairs are too close. With the bidirectional counter, this problem can be easily addressed by incorporating a proximity detector circuit to determine if the frequency difference between the selected pair of ROs is lower than an empirically determined threshold d . To simplify the thresholding circuit, d is set to filter out the response bit generated from a pair of ROs with an absolute difference in count values of less than 2^d . The thresholding logic comprises an $(n - d)$ -bit NAND gate, an $(n - d)$ -bit OR gate and a two-input AND gate. The output 'VALID' of the thresholding circuit is given by

$$\text{VALID} = \begin{cases} 0, & \text{if } |\text{count}_i - \text{count}_j| < 2^d \\ 1, & \text{otherwise} \end{cases} \quad (5.16)$$

Fig. 5.13 (a) Proposed aging-resilient CS RO design, (b) Equivalent biasing circuit when $EN = '1'$, (c) Equivalent biasing circuit when $EN = '0'$.

where $count_i$ and $count_j$ are the numbers of transitions in a given time interval of the selected pair of ROs. When 'VALID' = 0, the response bit obtained from the most significant bit of the bidirectional counter is considered unreliable and discarded. For a 10-bit counter with $d = 2$, the thresholding logic is shown in Fig. 5.15. By discarding the unstable bits, the reliability of the PUF is improved.

Irregular Latency between Successive LFSR Outputs for Security Enhancement

The input challenges enciphered by the LFSR for the RO selection proposed in [10] may leak vital information, which makes the challenge-response pair (CRP) predictable. By directly using the output of LSFR, the challenges are deterministic even though different initial values may be used. Let RO_i denotes the RO selected by the output of the LFSR at the i -th clock cycle. If N_{clk} is fixed, as shown in Fig. 5.16(a), then the ROs are selected successively in a predictable sequence as follows: $RO_1, RO_{N_{clk}+1}, RO_{2N_{clk}+1}, RO_{3N_{clk}+1}, \dots$. For example, if $N_{clk} = 1$, the response bits are generated successively from the following pairs of ROs: $(RO_1, RO_2), (RO_3, RO_4), (RO_5, RO_6), \dots$.

Since the output sequence of LFSR is fixed by its seed value, when the seed of the LFSR is changed by the input challenge, the response bits will be generated from a different starting pairs but in the same order, i.e., $(RO_j, RO_{j+1}), (RO_{j+2}, RO_{j+3}), (RO_{j+4}, RO_{j+5}), \dots$. The two sequences of response are correlated because some of their bits are generated from the same pairs and in the same order. This correlation can be broken by making N_{clk} irregular. For

Fig. 5.14 (a) Block diagram of proposed ACRO-PUF. (b) RO PUF core circuit.

Fig. 5.15 Bidirectional counter with proximity detector circuit.

simplicity, N_{clk} can be increased linearly. As shown in Fig. 5.16(b), if the number of clock cycles between two successive enable (EN) pulses is incremented by one, the successive responses will be generated by the pairs of ROs in the following sequence: (RO₁,RO₂), (RO₄,RO₇), (RO₁₁,RO₁₆), $\dots$. If the challenge changes such that RO_{*j*} is selected first, the responses will be generated from the following successive pairs, (RO_{*j*},RO_{*j*+1}), (RO_{*j*+3},RO_{*j*+6}), (RO_{*j*+10},RO_{*j*+15}), $\dots$. By changing the latency between two successive outputs of LFSR, the correlation between consecutive responses of different challenges is eliminated.

5.4 Simulation Results and Discussions

The proposed ACRO-PUF is designed for TSMC 40nm CMOS process. By using the drain-sharing technique in the layout design, the CS RO only has 25% area overhead compared with regular RO. As the PUF core has 256 ROs and each RO has 8 inverters and an NAND gate, the equivalent additional transistors in the ACRO-PUF core is 1024. Including the biasing circuit and proximity detector, the total number of equivalent additional transistors is less than 1100. With more than 10000 transistors in a regular RO PUF, the area overhead for reliability enhancement is estimated to be less than 11%. Excluding the power consumed in the common LFSR and bidirectional counter, the power consumptions for the same input

Fig. 5.16 (a) Regular LFSR clocking with correlated sequence of RO pairs for different challenges. (b) Irregular LFSR clocking with uncorrelated sequence of RO pairs for different challenges.

Table 5.1 Power analysis @ $V_{DD} = 1.1$ V, temperature = 27°C

Type	Power/CRP	Frequency
Regular	$28.9 \mu\text{W}$	5862 MHz
ARO [51]	$8.71 \mu\text{W}$	999 MHz
AT RO [55]	$3.48 \mu\text{W}$	1015 MHz
ACRO	$1.46 \mu\text{W}$	78 MHz

challenge to different types of RO PUF cores at nominal operating condition are compared in Table 5.1. The results show that the proposed ACRO-PUF consumes significantly ($\sim 2.38\times$) less power than AT RO PUF due mainly to its ability to oscillate at lower frequency with the same number of inverter stages. The uniqueness, reliability and randomness of the proposed ACRO-PUF are also evaluated by Monte Carlo simulations.

5.4.1 Uniqueness

Fig. 5.17 shows the distributions of the fractional HD for the two different types of RO PUFs. Twenty thousand pairwise comparisons of responses were made to evaluate the uniqueness. The simulation results are fitted to a binomial distribution with $n = 64$ and $p = 0.5007$ for regular RO PUF, and $n = 64$ and $p = 0.4997$ for ACRO-PUF. It can be seen that both types of RO PUFs have uniqueness close to the ideal value of 50%.

5.4.2 Reliability

Figs. 5.18(a) and (b) show the reliability calculated from five thousand CRPs generated by each type of RO PUF under different temperatures and supply voltages without the proximity detector. The results show that ACRO-PUF has slightly higher reliability than regular RO PUF

Fig. 5.17 Interdie fractional HD for (a) regular RO PUF ($\mu = 50.07\%$, $\sigma = 6.91\%$), (b) ACRO-PUF ($\mu = 49.97\%$, $\sigma = 6.98\%$).

Fig. 5.18 Comparison of regular and ACRO-PUF on reliability against (a) temperature variation and (b) voltage variation.

under temperature changes, and much more reliable under supply voltage changes. In the worst case environmental condition without proximity detector, the regular RO PUF can only achieve a reliability of 89.78%. In contrast, the proposed ACRO-PUF can still attain a reliability of 95.88%. Using proximity detector can further improve the reliability of ACRO PUF by pruning a fraction of CRPs as shown by the simulation results in Table 5.2. ACRO-PUF achieved a reliability of 100% with a threshold value (2^d) of 64 by discarding only 23.39% CRPs. If the same proposed proximity detector is also applied onto regular RO PUF, 100% reliability can only be achieved by discarding 57.86% CRPs with a threshold value of 16. The fact that a smaller threshold value results in a larger number of unusable CRPs signifies the higher combinations of RO pairs with unreliably close frequency variance and frequency deviation in regular RO PUF than in the proposed CS PUF.

As the biasing voltage V_{ctrl} is provided by a voltage reference, which may also vary under different environmental conditions, its influence on PUF reliability is analyzed. Fig. 5.19 shows the simulation results of reliability with V_{ctrl} varying from 0.33 V to 0.57 V, where the

 Fig. 5.19 Reliability against V_{ctrl} variations.

Table 5.2 Worst case reliabilities with proximity detector

Threshold 2^d	Regular	ACRO
0	89.78%(0)	95.88%(0%)
2	93.37%(8.55%)	96.44%(1.11%)
4	96.61%(16.65%)	97.11%(2.30%)
8	99.31%(29.49%)	98.12%(4.55%)
16	100%(57.86%)	99.19%(8.28%)
32	100%(87.07%)	99.88%(13.45%)
64	100%(99.78%)	100%(23.39%)

(.) indicates the fraction of discarded response bits.

nominal V_{ctrl} is 0.45. Even in the worst case variation of V_{ctrl} to 0.57 V, the reliability is still above 94%. The minimum reliability is 98.46% for V_{ctrl} ranges from 0.42 V to 0.48 V, and can be further improved to 100% by the proximity detector with a threshold value of 32 (i.e., $d = 5$). The voltage proposed in [63] has an average TC of 5.6 ppm/°C and line sensitivity (the parameter that qualifies the V_{ref} sensitivity against supply variation) of 0.1%/V, which is equivalent to less than 0.23 mV deviation from 0.45 V in the temperature range of $-40^\circ\text{C} \sim 120^\circ\text{C}$ and supply voltage range of 0.9 V to 1.3 V. With such voltage reference, the influence of V_{ctrl} on the reliability can basically be ignored.

5.4.3 Randomness

The cryptographic randomness of the proposed PUF responses is measured by the NIST test. Ten thousand response bits were generated and split into 50 blocks with 200 bits each for this test. With a bit string of length 10000, only 7 out of 15 tests are applicable. The results of the 7 statistical tests are shown in Table 5.3. To pass each test, the minimum proportion of binary sequence (PROP) that passed the tests is 47/50. The results indicate that both PUFs passed all the 7 tests.

ACRO-PUF: A Low-power, Reliable and Aging-Resilient Current Starved Inverter Based Ring Oscillator Physical Unclonable Function

Table 5.3 NIST randomness tests on responses of regular RO PUF and proposed ACRO-PUF.

Statistical	PROP		PASS?	
	Regular	ACRO	Regular	ACRO
Frequency	50/50	50/50	Y	Y
BlockFrequency	50/50	49/50	Y	Y
CumulativeSums(forward)	50/50	50/50	Y	Y
CumulativeSums(backward)	50/50	49/50	Y	Y
Runs	49/50	49/50	Y	Y
LongestRun	48/50	50/50	Y	Y
FFT	48/50	48/50	Y	Y
Serial(forward)	49/50	49/50	Y	Y
Serial(backward)	50/50	49/50	Y	Y

5.4.4 Aging Resistance

Based on the analysis in Section 5.3.2, the proposed aging-resilient CS RO design should have higher aging resistance than other ROs. The ability to mitigate the aging effect can be measured by the frequency degradation rate, which is expressed as [55]:

$$\Delta f_t = \frac{f_0 - f_t}{f_0} \quad (5.17)$$

To test the aging effect, the proposed design is simulated with Virtuoso RelXpert Reliability simulator with BSIM4 40nm model provided by Cadence Design Systems, Inc. The reliability model considers both NBTI and HCI effects, but not PBTI effect. As the design in [51] is vulnerable to PBTI effect because of their additional NMOS transistors in the critical path, it is justifiable to only compare the proposed design with regular RO and the AT RO proposed in [55]. This is legitimate since it has been shown in [55] that AT RO PUF is more resilient to aging than the design in [51]. The inverters in each design are sized identically for a fair comparison. Five fractions of activation time (percentage of oscillating time) are chosen, which are 20%, 40%, 60%, 80% and 100%. For each activation time, the ROs were aged from 1 year to 10 years.

The simulation results are shown in Fig. 5.20. It can be seen that longer time and longer activation time will result in greater frequency degradation. In all cases, the proposed design has less frequency degradation than the regular RO and AT RO. In the worst case, which is after 10 years with 100% activation time, the frequency degradation of regular RO, AT RO and CS RO are 21.57%, 4.18% and 1.14%, respectively. Thus the proposed ACRO-PUF is $\sim 3.7\times$ better than AT RO PUF.

5.4.5 Comparison with other RO PUFs

The proposed ACRO-PUF is compared with other types of RO PUFs reported in the literature in Table 5.4. With optimal CS biasing that accounts for frequency variance across ROs besides frequency deviation against temperature variation, the proposed RO PUF exhibits

Fig. 5.20 (a) Frequency degradation of regular RO over time. (b) Frequency degradation of AT RO over time (c) Frequency degradation of ACRO over time.

excellent uniqueness and reliability over a wide range of environmental conditions. With the aging-mitigated CS inverters, the proposed PUF is also resilient to aging effect.

5.5 Conclusion

A low power ACRO-PUF has been proposed with significantly enhanced reliability against temperature and supply voltage variations simultaneously. By leveraging on the extra gate biasing of CS inverter to control its drain current, an optimal bias is determined to account for the intricate counteracting effects of temperature and supply voltage on RO frequency when the transistors of the CS inverter are operating in saturation and subthreshold region. To reduce the frequency degradation over time, two transistors are added into each CS inverter to adaptively suppress NBTI and HCI in both active and idle modes of operation. The reliability is further boosted by a proximity detector circuit to nearly 100% by pruning a fraction of potentially unstable CRPs. The high correlation between the sequences of RO pairs selected by the LFSR for different input challenges is eliminated by the proposed irregular clocking scheme. The robustness and merits of the proposed RO PUF have been corroborated by Monte Carlo simulations and in comparison with reported performances of existing RO PUFs.

Table 5.4 Comparison of Different RO PUFs.

	[3]	[10]	[51]	[55]	[57]*	This work
Tech. (nm)	90	65	90	90	40	40
Uniqueness (%)	46.14	50.42	47	46.22	49.65	49.97
Worst Reliability (%)	99.52	97.28	96.91	95.89	99.92	100%**
Temp range (°C)	20~120	-40~120	0~100	NA	-40~120	-40~120
V_{DD} range (V)	1.08-1.2	1.17-1.23	0.8-1.4	NA	1.05-1.15	0.9-1.3
Aging resilient?	No	No	Yes	Yes	No	Yes

* Only the data of CSp type is compared since it has the best performance in [57].

** Since five thousand CRPs are generated, the worst reliability is at least 99.98%.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

With the increasing popularity of mobile electronic devices, more daily tasks are performed through the integrated circuits, including the storage of secret information and conduct of secure transactions. Comparing with the rapidly improving computing power of processors, hardware security is slower to develop [4]. Currently the best practice to assure data integrity is to place the secret key or sensitive information in the non-volatile memory like electrically erasable programmable read-only memory or battery-backed static random access memory and protect them with some cryptographic function implemented on hardware [4]. Such scheme usually consumes large circuit area and dissipates more power. More importantly, it is vulnerable to both non-invasive and semi-invasive attacks.

PUF serves as a secure and low-cost hardware fingerprint analogous to human biometric to prevent the vulnerability of long-term key presence. The uniqueness and unpredictability of its signature stem from the uncontrollable and random manufacturing process variations. A PUF can only be attacked when it is powered up and running, making them more secure than most other cryptographic primitives [11]. This thesis presented a comprehensive review on the background of PUFs, from the definition to the reliability enhancement techniques. By investigating various kinds of PUFs, some problems on SRAM-based PUF and RO-based PUFs are identified.

In Chapter 3, the design conflict of SRAM used in memory mode and PUF mode has been presented. After analyzing the dependency of process variation on transistor sizing and the augmented voltage biasing techniques, a novel methodology has been proposed. With the proposed design algorithm, the requirements for both modes of operation have been met with other design constraints kept at an acceptable level.

In Chapter 4, 8T-DP-SRAM PUF has been proposed to generate multiple response bits per cell. This scheme greatly expended the CRP space of SRAM PUF and improved its entropy. As only the data in the addressed SRAM cells needs to be buffered during mode switching, the risk of side-channel attacks over the power-up reset based memory PUFs is also minimized.

In Chapter 5, to improve reliability of RO PUFs, regular inverters have been replaced by CS inverters. The additional two transistors are used to control the drain current. The proposed ACRO-PUF can achieve high reliability against both temperature and supply voltage variations when the optimal biasing is achieved. To reduce aging effect, two transistors are added to the biasing circuit to adaptively suppress NBTI and HCI in both active and idle modes of operation. A proximity detector circuit is also proposed to further enhance their reliability to nearly 100%. Furthermore, the security issue of previous RO PUF due to sequence selections is eliminated by the proposed irregular clocking scheme.

6.2 Future Work

6.2.1 Dual-Port SRAM PUF for Dual-mode Applications

Chapter 3 showed a design method to implement a SRAM PUF for dual-mode applications while chapter 4 proposed a DP SRAM PUF that can generate multiple response bits per cell. However, the DP SRAM PUF is not optimized for dual-mode application. As the access mode, which is common row and common column, for PUF response bits generation is normally forbidden in memory mode, it is more flexible to achieve dual-mode applications using DP SRAM. It is therefore worthy to explore the implementation of DP SRAM PUF for dual-mode applications.

6.2.2 RO PUF fabrication

Currently, the concept of the proposed RO PUF is only verified by simulation. The proposed RO PUF will be fabricated and tested in the future to verify its quality. The reliability should be tested under various temperatures and supply voltages. For the aging-resistance test, the fabricated RO PUF should be placed in the environment with high voltage and high temperature.

6.2.3 Strong Mono-stable PUF

Recently a new type of PUF called mono-stable PUF has emerged [64, 65]. Unlike traditional memory based PUF and delay based PUF, mono-stable PUF has the advantage that it only has one stable state. Both delay based PUFs and memory based PUFs need a transient process to generate the response but this is not the case for mono-stable PUFs. The property makes them more resilient to transient noise, thus making it more reliable. However, previous mono-stable PUFs can only generate a limited number of CRPs. In the future, a mono-stable stable PUF with the structure of Fig. 6.1 will be exploited to generate exponential number of CRPs for authentication applications. In this structure, the inverter chains that are working in sub-threshold region will be explored and implemented as a low power strong mono-stable PUF. To further increase the reliability of the PUF, a thresholding circuit that employs transistors with abnormal threshold voltages will be used to filter out the unstable bits.

Fig. 6.1 Architecture of Strong mono-stable PUF.

Author's Publications

Journal Papers:

Chip-Hong Chang, **Chao Qun Liu**, Le Zhang, and Zhi Hui Kong. "Sizing of SRAM Cell with Voltage Biasing Techniques for Reliability Enhancement of Memory and PUF Functions." *Special Issue on Hardware Security – Threats and Countermeasures at the Circuits and Logic Levels, Journal of Low Power Electronics and Applications*, vol. 6, no. 3, Aug. 2016, 16: pp. 1-17 (Invited Featured Article).

Chao Qun Liu, Yuan Cao and Chip-Hong Chang, "ACRO-PUF: A Low-power, Reliable and Aging-Resilient Current Starved Inverter Based Ring Oscillator Physical Unclonable Function," *Special Issue on Selected Papers from 2017 CASS Regional Flagship Conferences, IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 64, no. 12, pp. 3138-3149, Dec. 2017.

Conference Papers:

Chao Qun Liu, Yuan Cao and Chip-Hong Chang, "Low-power, Lightweight and Reliability-enhanced Current Starved Inverter Based RO PUFs," *2016 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS)*, Jeju, South Korea, Oct. 2016, pp. 646–649.

Chao Qun Liu, Yue Zheng and Chip-Hong Chang, "A New Write-Contention Based Dual-Port SRAM PUF with Multiple Response Bits Per Cell", *2017 IEEE International Symposium on Circuits and Systems (ISCAS)*, Baltimore, MD, USA, May 2017, pp. 1712-1715.

References

- [1] L. Zhang *et al.*, “Optimizing emerging nonvolatile memories for dual-mode applications: Data storage and key generator,” *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 34, pp. 1176–1187, July 2015.
- [2] D. Holcomb, W. Burleson, and K. Fu, “Power-up SRAM state as an identifying fingerprint and source of true random numbers,” *IEEE Trans. Comput.*, vol. 58, pp. 1198–1210, Sept. 2009.
- [3] G. Suh and S. Devadas, “Physical unclonable functions for device authentication and secret key generation,” in *Proc. ACM/IEEE Design Automation Conf.*, (San Diego, CA, USA), pp. 9–14, June 2007.
- [4] C. Herder *et al.*, “Physical unclonable functions and applications: A tutorial,” *Proc. IEEE*, vol. 102, pp. 1126–1141, Aug. 2014.
- [5] F. Zhang, S. Yang, J. Plusquellic, and S. Bhunia, “Current based puf exploiting random variations in sram cells,” in *Proc. Design, Automation and Test Europe Conf. (DATE)*, pp. 277–280, Mar. 2016.
- [6] E. I. Vatajelu, G. D. Natale, and P. Prinetto, “Towards a highly reliable SRAM-based PUFs,” in *Proc. Design, Automation and Test Europe Conf. (DATE)*, pp. 273–276, Mar. 2016.
- [7] L. Zhang, C. H. Chang, Z. H. Kong, and C. Q. Liu, “Statistical analysis and design of 6T SRAM cell for physical unclonable function with dual application modes,” in *Proc. IEEE Int. Symp. Circuits and Syst.*, (Lisbon, Portugal), pp. 1410–1413, May 2015.
- [8] X. Xu, A. Rahmati, D. E. Holcomb, K. Fu, and W. Burleson, “Reliable physical unclonable functions using data retention voltage of sram cells,” *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 34, pp. 903–914, June 2015.
- [9] A. R. Krishna, S. Narasimhan, X. Wang, and S. Bhunia, “MECCA: a robust low-overhead puf using embedded memory array,” in *Cryptographic Hardware and Embedded Systems—CHES 2011*, pp. 407–420, Springer, 2011.
- [10] Y. Cao, L. Zhang, C. H. Chang, and S. Chen, “A low-power hybrid RO PUF with improved thermal stability for lightweight applications,” vol. 34, pp. 1143–1147, July 2015.
- [11] R. Kumar *et al.*, “On design of temperature invariant physically unclonable functions based on ring oscillators,” in *Proc. IEEE Comput. Soc. Annu. Symp. VLSI*, (Amherst, MA, USA), pp. 165–170, Aug. 2012.
- [12] K. Yang, Q. Dong, D. Blaauw, and D. Sylvester, “14.2 A physically unclonable function with BER<10^{−8} for robust chip authentication using oscillator collapse in 40nm CMOS,” in *Proc. IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, pp. 1–3, Feb. 2015.

References

- [13] A. Cherkaoui, L. Bossuet, and C. Marchand, "Design, evaluation, and optimization of physical unclonable functions based on transient effect ring oscillators," *IEEE Trans. Inform. Forensics Security*, vol. 11, pp. 1291–1305, June 2016.
- [14] S. S. Mansouri and E. Dubrova, "Ring oscillator physical unclonable function with multi level supply voltages," in *Proc. IEEE Int. Conf. Comput. Design*, (Montreal, QC, Canada), pp. 520–521, Sept. 2012.
- [15] V. Vivekraj and L. Nazhandali, "Feedback based supply voltage control for temperature variation tolerant PUFs," in *Proc. IEEE Int. Conf. VLSI Design*, pp. 214–219, Jan. 2011.
- [16] C. E. D. Yin and G. Qu, "LISA: Maximizing RO PUF's secret extraction," in *Proc. IEEE Int. Symp. Hardware-Oriented Security and Trust (HOST)*, pp. 100–105, June 2010.
- [17] C.-E. Yin and G. Qu, "Temperature-aware cooperative ring oscillator PUF," in *Proc. IEEE Int. Symp. Hardware-Oriented Security and Trust (HOST)*, pp. 36–42, July 2009.
- [18] M. Cortez, A. Dargar, S. Hamdioui, and G. J. Schrijen, "Modeling SRAM start-up behavior for physical unclonable functions," in *Proc. IEEE Int. Symp. Defect Fault Tolerance VLSI Nanotechnology Syst. (DFT)*, pp. 1–6, Oct. 2012.
- [19] B. Gassend *et al.*, "Silicon physical random functions," in *ACM Conf. on Computer and Communications Security*, (Washington, DC, US), Nov. 2002.
- [20] P. S. Ravikanth, *Physical One-Way Functions*. PhD thesis, Massachusetts Institute of Technology, Mar. 2001.
- [21] R. Pappu, B. Recht, J. Taylor, and N. Gershenfeld, "Physical one-way functions," *Science*, vol. 297, no. 5589, pp. 2026–2030, 2002.
- [22] Y. Gao, D. C. Ranasinghe, S. F. Al-Sarawi, O. Kavehei, and D. Abbott, "Emerging physical unclonable functions with nanotechnology," *IEEE Access*, vol. 4, pp. 61–80, 2016.
- [23] A. Rukhin, J. Soto, J. Nechvatal, E. Barker, S. Leigh, M. Levenson, D. Banks, A. Heckert, J. Dray, S. Vo, *et al.*, "Statistical test suite for random and pseudorandom number generators for cryptographic applications, NIST special publication," 2010.
- [24] Y. Cao, L. Zhang, S. S. Zalivaka, C. H. Chang, and S. Chen, "CMOS image sensor based physical unclonable function for coherent sensor-level authentication," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 62, pp. 2629–2640, Nov. 2015.
- [25] M. Bhargava, C. Cakir, and K. Mai, "Reliability enhancement of bi-stable PUFs in 65nm bulk CMOS," in *Proc. IEEE Int. Symp. Hardware-Oriented Security and Trust (HOST)*, pp. 25–30, June 2012.
- [26] A. Garg and T. T. Kim, "Design of SRAM PUF with improved uniformity and reliability utilizing device aging effect," in *Proc. IEEE Int. Symp. Circuits and Syst.*, pp. 1941–1944, June 2014.
- [27] S. K. Mathew, S. K. Satpathy, M. A. Anders, H. Kaul, S. K. Hsu, A. Agarwal, G. K. Chen, R. J. Parker, R. K. Krishnamurthy, and V. De, "16.2 A 0.19pJ/b PVT-variation-tolerant hybrid physically unclonable function circuit for 100% stable secure key generation in 22nm CMOS," in *Proc. IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, pp. 278–279, Feb. 2014.
- [28] L. Zhang, Z. H. Kong, and C. H. Chang, "PCKGen: A phase change memory based cryptographic key generator," in *Proc. IEEE Int. Symp. Circuits and Syst.*, (Beijing, China), pp. 1444–1447, May 2013.

-
- [29] P. Van Aubel, D. J. Bernstein, and R. Niederhagen, "Investigating SRAM PUFs in large CPUs and GPUs," in *Security, Privacy, and Applied Cryptography Engineering*, pp. 228–247, Springer, 2015.
- [30] S. Chellappa and L. Clark, "SRAM-based unique chip identifier techniques," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. PP, no. 99, pp. 1–1, 2015.
- [31] B. Zimmer *et al.*, "SRAM assist techniques for operation in a wide voltage range in 28-nm CMOS," *IEEE Trans. Circuits and Sys. II: Express Briefs*, vol. 59, no. 12, pp. 853–857, 2012.
- [32] B. H. Calhoun and A. P. Chandrakasan, "Static noise margin variation for sub-threshold SRAM in 65-nm CMOS," *IEEE J. Solid-State Circuits*, vol. 41, pp. 1673–1679, July 2006.
- [33] M. Orshansky *et al.*, *Design for Manufacturability and Statistical Design: A Constructive Approach*. Springer, 2008.
- [34] Synopsys, Inc, "Hspice® user guide: RF analysis," 2010.
- [35] S. Chellappa, A. Dey, and L. Clark, "Improved circuits for microchip identification using SRAM mismatch," in *Proc. IEEE Custom Integr. Circuits Conf.*, (San Jose, CA, USA), pp. 1–4, Sept. 2011.
- [36] K. Agarwal and S. Nassif, "Statistical analysis of SRAM cell stability," in *Proc. ACM/IEEE Design Automation Conf.*, (San Francisco, CA, USA), pp. 57–62, 2006.
- [37] S. Mukhopadhyay, H. Mahmoodi, and K. Roy, "Modeling of failure probability and statistical design of SRAM array for yield enhancement in nanoscaled CMOS," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 24, pp. 1859–1880, Dec. 2005.
- [38] B. Alorda, C. Carmona, and S. Bota, "Word-line power supply selector for stability improvement of embedded SRAMs in high reliability applications," in *Proc. Design, Automation and Test Europe Conf.*, (Dresden, Germany), pp. 1–6, Mar. 2014.
- [39] K. Zhang *et al.*, "A 3-GHz 70-Mb SRAM in 65-nm CMOS technology with integrated column-based dynamic power supply," *IEEE J. Solid-State Circuits*, vol. 41, pp. 146–151, Jan. 2006.
- [40] M. E. Sinangil *et al.*, "A 28 nm 2 Mbit 6T SRAM with highly configurable low-voltage write-ability assist implementation and capacitor-based sense-amplifier input offset compensation," *IEEE J. Solid-State Circuits*, vol. 51, pp. 557–567, Feb. 2016.
- [41] S. Mukhopadhyay, H. Mahmoodi, and K. Roy, "Reduction of parametric failures in sub-100-nm SRAM array using body bias," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 27, pp. 174–183, Jan. 2008.
- [42] A. Hokazono *et al.*, "Mosfet design for forward body biasing scheme," *IEEE Electron Device Letters*, vol. 27, pp. 387–389, May 2006.
- [43] H. Mostafa, M. Anis, and M. Elmasry, "Adaptive body bias for reducing the impacts of NBTI and process variations on 6T SRAM cells," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 58, pp. 2859–2871, Dec. 2011.
- [44] J. Wang, S. Nalam, and B. H. Calhoun, "Analyzing static and dynamic write margin for nanometer SRAMs," in *Proc. ACM/IEEE Int. Symp. Low Power Electronics and Design (ISLPED)*, pp. 129–134, Aug. 2008.

References

- [45] S. Lin and D. Costello, *Error Control Coding*. Englewood Cliffs, NJ: Prentice-Hall, 2004.
- [46] A. Maiti, J. Casarona, L. McHale, and P. Schaumont, "A large scale characterization of RO-PUF," in *Proc. IEEE Int. Symp. Hardware-Oriented Security and Trust (HOST)*, (Anaheim, CA, USA), pp. 94–99, June 2010.
- [47] K. Nii, Y. Tsukamoto, M. Yabuuchi, Y. Masuda, S. Imaoka, K. Usui, S. Ohbayashi, H. Makino, and H. Shinohara, "Synchronous ultra-high-density 2RW dual-port 8T-SRAM with circumvention of simultaneous common-row-access," *IEEE J. Solid-State Circuits*, vol. 44, pp. 977–986, Mar. 2009.
- [48] M. Kassem *et al.*, "A sub-threshold SRAM based PUF," in *Proc. 2010 Int. Conf. Energy Aware Computing*, (Cairo, Egypt), pp. 1–4, Dec. 2010.
- [49] S. Zeitouni *et al.*, "Remanence decay side-channel: The PUF case," *IEEE Trans. Inform. Forensics and Security*, vol. 11, pp. 1106–1116, June 2016.
- [50] S. Ohbayashi, M. Yabuuchi, K. Nii, Y. Tsukamoto, S. Imaoka, Y. Oda, T. Yoshihara, M. Igarashi, M. Takeuchi, H. Kawashima, Y. Yamaguchi, K. Tsukamoto, M. Inuishi, H. Makino, K. Ishibashi, and H. Shinohara, "A 65-nm SoC embedded 6T-SRAM designed for manufacturability with read and write operation stabilizing circuits," *IEEE J. Solid-State Circuits*, vol. 42, pp. 820–829, April 2007.
- [51] M. T. Rahman, F. Rahman, D. Forte, and M. Tehranipoor, "An aging-resistant RO-PUF for reliable key generation," *IEEE Trans. Emerging Topics Computing*, vol. 4, pp. 335–348, July 2016.
- [52] M. S. Kirkpatrick and E. Bertino, "Software techniques to combat drift in PUF-based authentication systems," in *Proc. Workshop Secure Component and System Identification (SECSI 2010)*, (Cologne, Germany), pp. 1–9, 2010.
- [53] A. Maiti and P. Schaumont, "The impact of aging on a physical unclonable function," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 22, pp. 1854–1864, Sept 2014.
- [54] M. T. Rahman, D. Forte, F. Rahman, and M. Tehranipoor, "A pair selection algorithm for robust RO-PUF against environmental variations and aging," in *Proc. 2015 33rd IEEE Int. Conf. Comput. Design (ICCD)*, (New York, NY, USA), pp. 415–418, Oct 2015.
- [55] S. R. Sahoo, S. Kumar, K. Mahapatra, and A. Swain, "A novel aging tolerant RO-PUF for low power application," in *Proc. 2016 IEEE Int. Symp. Nanoelectronic and Inform. Systems (iNIS)*, (Gwalior, India), pp. 187–192, Dec 2016.
- [56] V. Huard, M. Denais, and C. Parthasarathy, "NBTI degradation: From physical mechanisms to modelling," *Microelectronics Reliability*, vol. 46, no. 1, pp. 1–23, 2006.
- [57] C. Q. Liu, Y. Cao, and C. H. Chang, "Low-power, lightweight and reliability-enhanced current starved inverter based RO PUFs," in *Proc. 2016 IEEE Asia Pacific Conf. Circuits and Syst. (APCCAS)*, (Jeju, South Korea), pp. 646–649, Oct 2016.
- [58] M. Mandal and B. Sarkar, "Ring oscillators: Characteristics and applications," *Indian J. Pure Appl. Physics*, vol. 48, pp. 136–145, 2010.
- [59] E. Socher *et al.*, "Temperature sensitivity of SOI-CMOS transistors for use in uncooled thermal sensing," *IEEE Tran. Electron Devices*, vol. 52, pp. 2784–2790, Dec. 2005.

-
- [60] Y. Zheng, Y. Cao, and C. H. Chang, "A new event-driven dynamic vision sensor based physical unclonable function for camera authentication in reactive monitoring system," in *Proc. 2016 IEEE Asian Hardware-Oriented Security and Trust (AsianHOST)*, (Yilan, Taiwan), pp. 1–6, Dec 2016.
 - [61] D. Ganta and L. Nazhandali, "Study of IC aging on ring oscillator physical unclonable functions," in *Proc. 15th Int. Symp. Quality Electron. Design*, (Santa Clara, CA, USA), pp. 461–466, March 2014.
 - [62] A. Tiwari and J. Torrellas, "Facelift: Hiding and slowing down aging in multicores," in *Proc. 2008 41st IEEE/ACM Int. Symp. Microarchitecture*, (Lake Como, Italy), pp. 129–140, Nov 2008.
 - [63] J. Jiang, W. Shu, and J. S. Chang, "A 5.6 ppm/°C temperature coefficient, 87-dB PSRR, sub-1-V voltage reference in 65-nm CMOS exploiting the zero-temperature-coefficient point," *IEEE J. Solid-State Circuits*, vol. 52, pp. 623–633, March 2017.
 - [64] A. Alvarez, W. Zhao, and M. Alioto, "14.3 15fJ/b static physically unclonable functions for secure chip identification with <2% native bit instability and 140x inter/intra PUF hamming distance separation in 65nm," in *Proc. 2015 IEEE International Solid-State Circuits Conference - (ISSCC) Digest of Technical Papers*, (San Francisco, CA, USA), pp. 1–3, Feb 2015.
 - [65] B. Karpinsky, Y. Lee, Y. Choi, Y. Kim, M. Noh, and S. Lee, "8.7 physically unclonable function for secure key generation with a key error rate of $2e-38$ in 45nm smart-card chips," in *Proc. 2016 IEEE International Solid-State Circuits Conference (ISSCC)*, (San Francisco, CA, USA), pp. 158–160, Jan 2016.