Ultra-low power SRAM and SRAM based PUF design

Lu, Lu

2019

Lu, L. (2019). Ultra-low power SRAM and SRAM based PUF design. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/86213>

<https://doi.org/10.32657/10356/86213>

**NANYANG
TECHNOLOGICAL
UNIVERSITY**

SINGAPORE

Ultra-Low Power SRAM and SRAM based PUF Design

LU LU

SCHOOL OF ELECTRICAL & ELECTRONIC ENGINEERING

2019

Ultra-Low Power SRAM and SRAM based PUF Design

LU LU

School of Electrical & Electronic Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirement for the degree of
Doctor of Philosophy

2019

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

4th Sept. 2019

.....
Date

.....
Lu Lu

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

5th Sept. 2019

.....
Date

.....
Assoc. Prof. Kim Tae Hyung

Authorship Attribution Statement

This thesis contains material from 1 paper published in the following peer-reviewed journal where I was the first and/or corresponding author.

Chapter 3 is published as Lu Lu, Taegeun Yoo, Van Loi Le, and Tony Tae-Hyoung Kim. in Proc. IEEE Asian Solid-State Circuits Conference (A-SSCC), pp. 143-144, Taiwan, Nov. 2018. Invited and submitted in IEEE Journal of Solid-State Circuits (JSSC).

The contributions of the co-authors are as follows:

- Assoc. Prof. Kim provided the initial project direction and edited the manuscript drafts.
- I co-designed the study with Assoc. Prof. Kim and performed all the laboratory work at the School of Electrical and Electronic Engineering. I also analyzed the data.
- Dr. Yoo assisted in the design of IO pad layout in this chip.
- Van Loi Le co-designed the PCB of the testing chip.

Chapter 4 is published as Lu Lu, Tony Tae-Hyoung Kim, in IEEE International Symposium on Circuits and Systems (ISCAS) 2019.

The contributions of the co-authors are as follows:

- Assoc. Prof. Kim provided the initial project direction.
- I wrote the drafts of the manuscript. The manuscript was revised by Assoc. Prof. Kim.

4th Sept. 2019

.....
Date

.....
[Student's Name Here]

Acknowledgement

First and foremost, I would like to express my special thanks and gratitude to my supervisor, Professor Tony Tae-Hyoung Kim. He gave me an opportunity to join our group, guided me to explore on my research path in my whole Ph.D. period. He decided to take me as his apprentice and to mentor me, taught me a lot of valuable knowledge. His patience makes students feel comfortable and his leadership gave me the opportunity to attend an international conference to present my own work and meet many people from this industry. He encouraged me when I was going through tough times and took care of me as a family. Without his wisdom and guidance, I wouldn't be able to achieve my goal and make a big step in my life. Thank you for everything.

Secondly, I would like to express my gratitude to Dr. Mohammed Sultan Mohiuddin Siddiqui. He passed on to me a lot of specialized knowledge about SRAM and taught me how to solve issues. His kindness, patience, and friendliness are still warming my heart.

I would like to thank my co-supervisor Dr. Wang Lei and the company NXP. NXP provided me a scholarship for four years, allowed me to concentrate on my research. Dr. Wang Lei gave me many pieces of advice about SRAM design.

I would also like to thank all my colleagues in Virtus at NTU. Dr. Yoo Taegeun, Dr. Do Anh Tuan, Dr. Karim Hany Mohamed Mohamed Raw,

Dr. Seyed Mohammad Ali Zeinolabedin, Dr. Le Ba Ngoc, Dr. Neelakantan Narasimman, Loi Van Le, Dr. Bo Wang, Dr. Thu Linn Aung Myat, Dr. Abhik Das, Dr. Zhao Chuan Lee, Ruchi Sharma, Reenu James, and Kyungchan An. I had a wonderful time with you, I feel very lucky to meet you here.

Last but not least, I want to give my sincere thank you to my parents, they supported me with love to go through thin and thick. They are the greatest mom and dad in this world.

Table of Content

STATEMENT OF ORIGINALITY	5
SUPERVISOR DECLARATION STATEMENT	6
AUTHORSHIP ATTRIBUTION STATEMENT.....	I
ACKNOWLEDGEMENT.....	I
TABLE OF CONTENT.....	III
SUMMARY VI	
LIST OF FIGURES	IX
LIST OF TABLES.....	XIV
CHAPTER 1 INTRODUCTION	1
1.1 Motivation.....	1
1.2 Objectives and Major Contributions	6
1.3 Thesis Organization	10
CHAPTER 2 BACKGROUND AND LITERATURE REVIEW	12
2.1 Conventional 6T Single-port SRAM	12
2.2 The Conventional 8T SRAM.....	17
2.3 Techniques for the Write and Read Operations.....	19
2.3.1 Write Assist Techniques	19
2.3.2 Read Assist Techniques.....	23
2.3.3 Implementing of Techniques	26
2.4 The SRAM based PUF	38
2.4.1 Error-Correcting Codes (ECC)	39
2.4.2 Temporal Majority Voting (TMV).....	41
2.4.3 Aging	43
2.4.4 Analog PUF.....	45
2.4.5 Connecting multiple bit cells	46

CHAPTER 3 LOW POWER LOW LEAKAGE 8T WITH VERTICAL READ WORD-LINES AND SELECTIVE DUAL SPLIT POWER LINES.....	49
3.1 Background.....	49
3.2 Proposed SRAM Energy Consumption Improvement Techniques	52
3.2.1 Read Assist Techniques.....	52
3.2.2 Energy reduction with vertical RWL.....	54
3.2.3 Read bitline sensing with vertical RWL	57
3.2.4 Sensing amplifier with reference signal	63
3.3 Proposed SRAM Ultra-Low Voltage Techniques	64
3.3.1 Write operation with selective dual split power lines.....	64
3.3.2 VSSUP vs. VCC_BST for stability	70
3.3.3 Write-completion flag generation.....	73
3.3.4 Architecture of 16k SRAM	75
3.3.5 Write-assists comparison	75
3.4 Measurement Results.....	78
3.5 Conclusion	82
CHAPTER 4 A SEQUENCE DEPENDENT CONFIGURABLE PUF BASED ON 6T SRAM FOR ENHANCED CHALLENGE RESPONSE SPACE	83
4.1 Background.....	83
4.2 Proposed Sequence-Dependent PUF	90
4.3 The Enhancement of Challenge Response Space.....	97
4.3.1 CRPs enhancement in two dimension movement.....	97
4.3.2 Security enhancement analysis.....	100
4.4 SRAM Function Analysis Result	102
4.5 Measurement Results and Discussion	106
4.5.1 Reliability	108
4.5.2 Uniqueness	109
4.5.3 Randomness	113
4.6 Conclusion	117
CHAPTER 5 A SEQUENCE DEPENDENT CONFIGURABLE PUF BASED ON 6T SRAM FOR HIGH RELIABILITY	118
5.1 Background.....	118
5.2 Proposed Reliable PUF	120
5.2.1 Proposed SPUF cell	120
5.2.2 Reliability enhancement	124
5.2.3 Sequence Dependent SPUF Analysis.....	126

5.2.4 CRPs enhancement	129
5.2.5 SRAM Architecture and Operation Flow.....	131
5.3 Simulation and Measurement Result	134
5.4 Conclusion	139
CHAPTER 6 CONCLUSION AND FUTURE WORK	140
6.1 Conclusion	140
6.2 Future Work	142
LIST OF AUTHOR’S PUBLICATIONS	143
BIBLIOGRAPHY	144

Summary

With the recent development of portable devices and wearable technology, the requirements of long battery life-time, ultra-low power consumption and low cost in system-on-a-chip (SoC) are becoming increasingly significant. However, the aggressive CMOS technology which integrates with more devices and larger capacitor cache cause high power density. As a crucial approach to meet the ultra-low-power and energy usage, a supply voltage scaling technique has been commonly employed. The power consumption is being effectively suppressed with optimized energy efficiency under near- / sub-threshold voltage region. Meanwhile, static-random-access-memory (SRAM) designs are also facing some challenges when it comes for voltage scaling. In conventional 6T SRAMs, the minimum operating voltage is limited by the conflicting requirements from the ability to write and read stably. In addition, operating SRAMs in the near- or sub-threshold voltage regions have significantly degraded I_{on} -to- I_{off} ratios and exponentially increased device variations, which deteriorates various SRAM design parameters, like read margin, static noise margin and write ability, etc.

In this thesis, we investigate the characteristics of transistors under near- and sub-threshold voltage regions, to analyze some efficient methods for reducing the power consumption. Based on these analyses, we have developed an ultra-low power dissipation circuit with enhanced performance.

The ultimate goal of this research work is to develop and demonstrate an ultra-

low-power SRAM design, and an SRAM based physical-unclonable-function (PUF) with improved performance compared to commonly used PUF.

1. An 8T SRAM macro with vertical read word-line (RWL) and selective dual split power lines techniques is proposed. The proposed vertical RWL reduces dynamic energy consumption. The data-aware dual split power lines technique combined with vertical write bit-lines enhances both the write margin (WM) and the static noise margin (SNM). A 16kb SRAM test chip that was fabricated using 65nm CMOS technology demonstrates the minimum energy consumption of 0.506 pJ at 0.4 V. This achieves the smallest normalized energy per bit when compared to other state-of-the-art works. The SRAM could operate functionally down to 0.26 V.
2. A 2D sequence-dependent PUF based on SRAM is proposed. It expands the challenge to response pairs (CRPs) by the order of $\text{rows}^{(\text{sequence length} - 1)} \times \text{columns}^{(\text{sequence length} - 1)}$ for reliable authentication. This is achieved by configuring the sequence of SRAM cell selection. The proposed technique allows us to generate multiple data maps from a single chip. The bit error rate is 3% at the nominal point (0.8 V/20 °C) and the inter-hamming distance between chips is 0.4869. The CRPs is improved from 1.17×10^{11} in the previous work to 8.37×10^{17} .
3. A sequence-dependent SRAM based PUF with high reliability is proposed. It expands the CPRs by the order of $\text{rows}^{(\text{sequence length} - 1)} \times \text{columns}$ for reliable authentication, and for each step, it improves the reliability from $\frac{1}{2}$ to $\frac{3}{4}$. The bit error rate is improved 3% at the nominal point (1 V/20 °C) by a one-step operation. For the same sequence with a

different order in one chip, the hamming distance achieved 0.441 when sequence length is 5. The measurement results demonstrate that the proposed PUF is configurable and reliable for the physical security keys generation.

List of Figures

Fig. 1.1 Process technology size trend [1]	1
Fig. 1.2 Area weight of SRAM in SoC System [5]	3
Fig. 1.3 Full chip layout and the memory array [8].	4
Fig. 1.4 Power dissipation as a function of VDD for the 16 kb 9T SRAM [6]	4
Fig. 2.1 Conventional 6T SRAM bitcell.	13
Fig. 2.2 Conventional 6T SRAM $m \times n$ array.	13
Fig. 2.3 The β ratio measures the relative strengths of the pull-down and pass- gate transistor [12].....	14
Fig. 2.4 Active pattern for a) Thincell [12] and b) straight-line layout. c) The straight-line layout yields higher SNM.	14
Fig. 2.5 Optimal SNM lies at $\beta \sim 2-3$. Cell sizes range from $\beta = 1$ to $\beta = 4.4$	16
Fig. 2.6 Conventional 8T SRAM bitcell.	18
Fig. 2.7 Diagram of cell VDD lowering. (a) schematic (b) timing diagram.....	20
Fig. 2.8 Diagram of WL boosting. (a) schematic (b) timing diagram.....	20
Fig. 2.9 Diagram of VSS raising. (a) schematic (b) timing diagram.	21
Fig. 2.10 Diagram of VSS raising. (a) schematic (b) timing diagram.	22
Fig. 2.11 Diagram of WLUD. (a) schematic (b) timing diagram.....	23
Fig. 2.12 Diagram of suppressed BL. (a) schematic (b) timing diagram.	24
Fig. 2.13 Diagram of negative unselected WL. (a) schematic (b) timing diagram.....	24
Fig. 2.14 The SRAM bitcell in [19].	26
Fig. 2.15 The reported assist circuit in [19].	27
Fig. 2.16 BL pre-charge & discharge circuit [19].	28
Fig. 2.17 The timing diagram of read assist, WLUD [19].	28
Fig. 2.18 Reported WRE-LCV write assist scheme in [13]	30
Fig. 2.19 (a) simulation waveforms of the WRE-LCV scheme. (b) simulated write	31

Fig. 2.20 Reported SCS-NBL write assist scheme [13].....	32
Fig. 2.21 Simulation waveforms of the SCS-NBL write assist scheme [13].....	32
Fig. 2.22 6T SRAM bitcell with PMOS access transistors structure [20].....	33
Fig. 2.23 Simulated timing of WL and BLs for boosted and non-boosted	34
Fig. 2.24 6T SRAM with dual split control assist bitcell structure [18].....	35
Fig. 2.25 Write operation of DSC6T [18].....	35
Fig. 2.26 Read operation of DSC6T [18].....	36
Fig. 2.27 Operations of a PUF-based key storage system. (a) Enrollment. (b) Reconstruction [26, 27].....	40
Fig. 2.28 TMV circuit [27].....	42
Fig. 2.29 TMV measurement result [28].....	42
Fig. 2.30 Block Diagram of reported reliability enhancement scheme [32]	44
Fig. 2.31 Architecture of the analog PUF schematic [33]	45
Fig. 2.32 8T-DP-SRAM operating in PUF mode: (a) BL1 = 1, BL2 = 0 and (b) BL1 = 0, BL2 = 1. [34].....	46
Fig. 2.33 Sequence dependent PUF – Small array example. [35].....	47
Fig. 3.1 (a) Proposed 8T SRAM cell with HVT devices in the latch and RVT devices in the access transistors and the read port and (b) layout of the 8T SRAM cell based on the 65nm logic design rules.	53
Fig. 3.2 (a) Conventional read operation in 8T SRAM, (b) proposed read operation in 8T SRAM, and (c) energy consumption comparison of conventional one and proposed one. Note that the array size of 256 rows × 128 columns with the multiplexer ratio of 16-to-1 is assumed.	55
Fig. 3.3 (a) Principle of the conventional read operation. (b) Sensing margin at 0.6 V and (c) Sensing margin at VDD = 0.25 V. The number of cells per bitline is 256.	58
Fig. 3.4 Principle of the proposed read operation.	59
Fig. 3.5 Impact of the proposed vertical RWL on the RBL swing when compared with the conventional RBL structure.....	60

Fig. 3.6 (a) Read times of the proposed and conventional, RPB=256, CMUX = 16, in TT corner with varying temperature and supply voltage. (b) Normalized read time in varying voltage and temperature. (c) Read energy consumption of the proposed and conventional one, RPB=256, CMUX = 16, in TT corner with varying temperature and supply voltage.	61
Fig. 3.7 Diagram of sensing amplifier and reference signal generator	63
Fig. 3.8 Principle of the proposed write operation: (a) the topology of writing '0' operation, (b) the topology of writing '1' operation, (c) timing diagram of writing '0' operation and (d) timing diagram of writing '1' operation.....	66
Fig. 3.9 The impact of data-aware split power lines to writing '0' operation in selected and half-selected cells.	67
Fig. 3.10 Comparison of WM (a) and SNM (b) between the conventional and proposed SRAM.	69
Fig. 3.11 (a) impact of VSSUP to SNM under 0.4 V, (b) impact of VCC_BST to SNM under 0.4 V, and (c) impact comparison of VCCUP and VSSUP on SNM.	71
Fig. 3.12 The circuit in column to generate VCC_BST.....	72
Fig. 3.13 The feedback signal generator circuit.	73
Fig. 3.14 Schematic with the proposed assist circuit.....	74
Fig. 3.15 the comparison of WM and SNM between NBL and SDSP.....	76
Fig. 3.16 (a) The die photo of proposed 16 kb SRAM. (b) The waveform of test chip at 0.26V voltage supply.	78
Fig. 3.17 Measured results: (a) minimum clock period, (b) leakage current and total current, and (c) energy consumption.	79
Fig. 4.1 Random data in SPUF generation mechanism [53].	84
Fig. 4.2 Conventional PUF cells: (a) unit PUF cell and (b) voltage transfer curve of noise margin. [68].	86
Fig. 4.3 Proposed PUF cells: (a) unit SPUF cell, (b) the layout of the SPUF cell and (c) 2×2 array for on bit SPUF data generation.	91

Fig. 4.4 Principle of the PUF mode operation.	92
Fig. 4.5 Operation principle of the proposed SPUF data generation.	94
Fig. 4.6 Dependency of SPUF data on selection sequence order.	95
Fig. 4.7 Previous technique for changing sequence lengths [34].	97
Fig. 4.8 Different cell selection combinations in the proposed PUF. (a) with an adjacent four cells example (b) with a non-adjacent four cells example.	98
Fig. 4.9 Proposed SPUF cells in SRAM mode: (a) Write operation, (b) Read operation. (c) SNM for both read and write operations comparison between proposed SPUF cell and conventional SRAM cell.	103
Fig. 4.10 Proposed SRAM PUF architecture.	104
Fig. 4.11 Die photo of the proposed SPUF.	106
Fig. 4.12 Measurement environment.	106
Fig. 4.13 Measured BER sweeping: (a) voltage and (b) temperature.	107
Fig. 4.14 Measured HD for sequence #1 with order #1 ~ order #38.	110
Fig. 4.15 Measured HD for sequence #1 with different chips.	110
Fig. 4.16 Measured PUF data maps from one chip.	111
Fig. 4.17 Measured hamming distances.	112
Fig. 4.18 NIST test result.	113
Fig. 4.19 Measured hamming distances.	114
Fig. 5.1 Proposed SPUF cells: (a) unit SPUF cell and (b) layout of the SPUF cell. .	121
Fig. 5.2 3 × 3 array of the proposed SPUF in PUF mode.	122
Fig. 5.3 Reliability enhancement for PUF operation.	123
Fig. 5.4 Waveforms of PUF operation under different situations.	126
Fig. 5.5 Dependency of SPUF data on selection sequence order.	127
Fig. 5.7 Changing sequence length in proposed SPUF.	130
Fig. 5.6 Previous technique for changing sequence lengths [35].	129
Fig. 5.8 Proposed SRAM PUF architecture.	132
Fig. 5.9 The operation flow to generate reliable map.	133

Fig. 5.10 Die photo of the proposed SPUF.	134
Fig. 5.11 Simulated BER of the proposed SPUF with the golden value of 1V / 20 °C. (a) the BER with varying voltage, (b) the BER with varying temperature.....	135
Fig. 5.12 Measured BER of SPUF with the golden value (1V / 20 °C). (a) the raw data, (b) the PUF data (sequence length = 1).....	136
Fig. 5.13 Inter sequence hamming distance	138

List of Tables

Table 2.1 Comparison of different techniques	25
Table 2.2 Comparison of different types of SRAM.....	37
Table 3.1 Vertical RWL read operation.....	52
Table 3.2 Data-aware write operation in the selected cell	52
Table 3.3 Comparison of write assist with previous work	77
Table 3.4 Comparison of SRAM performance.....	80
Table 4.1 Comparison with Previous Works	115

Chapter 1 Introduction

1.1 Motivation

For many forefronts, technological equipment, from the medical devices, smartphones, monitor & television sets to military systems, the semiconductors are the crucial core to realize the portability and functionality. Nowadays, the challenges that the semiconductor industry is facing in recent years, the growth of ever-demanding markets of portable and implantable technologies, their requirements for a smaller size, lower power consumption and better performance. The high-tech innovations, such as the R&D and Internet of Things, play key roles in the burgeoning competitions.

Semiconductor process of miniaturization is down-scaling 0.7 times every two years in recent decades as shown in Fig. 1.1 in [1], the aggressive technology

Fig. 1.1 Process technology size trend [1]

development is driven by the demand in expense reduction. The high-density technology allows integration of an incredible number of transistors into an area of only dozens of square millimeters with optimized performance and minimized power dissipation [1 - 4]. Depending on the application, System-on-Chips (SoCs) could integrate all kinds of components of the electronic system, including digital, analog, RF processing circuit or mixed-signal. Under these circumstances, the power consumption becomes a critical issue due to the limited chip size and battery lifetime.

Static-random-access-memory (SRAM) occupies a significant area and has high power consumption in the modern SoC systems. As shown in Fig. 1.2, as the technology improves, the size of the chip keeps scaling down, and the area section and power dissipation of SRAM are getting more and more significant. SRAM retains data bits for as long as power is being supplied to the memory. Compared to the dynamic RAM (DRAM) which stores the bits in cells consisting of a capacitor and a transistor, SRAM does not have to be refreshed periodically. SRAM provides faster access to data but is also much more expensive than DRAM. With the proliferation of Nanodevices and portable devices that demand high performance on little real estate, there has been a major desire for improvement of power consumption for these devices, and in particular in the reduction of voltage-per-bit-of-memory in SRAM chips [7]. Conventional 6T SRAM cells remain popular due to their optimum structure and high density, but they suffer from low write ability and half-select (HS) disturbances at a low supply voltage (V_{DD}). This impairs the efficiency of having a lowered minimum supply voltage ($V_{DD_{min}}$) in order to reduce power consumption. While alternative setups have seen an increased application, there are some tradeoffs observed.

Fig. 1.2 Area weight of SRAM in SoC System [5]

In conventional 8T SRAM cells, the data nodes are fully decoupled from the read bit-line (BL) to significantly improve read stability, as compared to a 6T cells' susceptibility to noise during the reading process. However, 8T cells are prone to drops in voltage at the read bit-line (BL), which causes high power dissipation.

A full SRAM chip for ultralow-power application in 22nm CMOS layout is shown in Fig. 1.3 [8]. Except for the bit-cells, SRAM chip also needs a pre-decode circuit, a row decoder, and a column decoder. Nowadays, since the operational voltage is scaled to an ultralow value, from nominal level down to near or beneath transistor near-/sub- threshold voltage, the power is substantially reduced by several orders of magnitude. As shown in Fig. 1.4, the active power of the 16 kb 9T SRAM decreases from 1.3 mW to 1.2 μ W when supply voltage changes from 1.2 V to 0.24 V with a power saving of more than 1000 times [9]. However, various challenging issues like frail sub-threshold operations, lack of systematic design methodology and high process-voltage-temperature (PVT)

Fig. 1.3 Full chip layout and the memory array [8].

Fig. 1.4 Power dissipation as a function of VDD for the 16 kb 9T SRAM [6]

variations sensitivity, defeat the utility of ultra-low-voltage circuits. New design methodologies with novel circuit techniques to enhance performance, combat variability and suppress leakage are worthy of extensive and in-depth exploration.

In order to reduce the area overhead, it is a trend to use existing physical setting to implement multiple functions. In this context, the physical unclonable function (PUF) is used to generate key information for the security application. Particularly, SRAM based PUF (SPUF) utilizes the current hardware to realize both SRAM and PUF functionalities. For the SPUF, normally the challenge-response pair (CRP) is equal to 1, indicating that only one response is generated based on the same address. The data map generated from one device is unique, which is caused by the mismatch from the process variations during the fabrication, it is called a chip ID. However, since the CRP is limited, the physical encryption easily becomes a target of an attack. Under the circumstance, increasing the challenge-response space is very important to improve the security of a SPUF design.

1.2 Objectives and Major Contributions

The demand for ultra-low voltage SRAM has been increasing due to their wide applications in power or energy-constrained SOC. Ultra-low-voltage SRAMs suffer from degraded static noise margin, small I_{ON} -to- I_{OFF} current ratio translates to a low bit line sensing margin and large device variations. This limits the number of bit cells per bit line, minimum operating voltage and temperature. Conventional 8T SRAM bit cell [10, 11] with decoupled read port have been employed to eliminate read disturbing current into data storage nodes during read operation to reduce static noise margin. However conventional 8T SRAM cell suffers from poor read bit line sensing margin due to more read port leakage and cell stability issue in the half-selected condition which is essential for ultra-low voltage environment for energy efficiency.

The greatest challenge of low power SRAM is how to reduce the minimal supply voltage V_{min} without performance deterioration. For the write port, the V_{min} is constrained by the write margin and noise margin. Under the ultra-low voltage supply, we need the assist circuit to help writing the data. If we want to strengthen the access transistor, the data in half selected condition can easily be flipped, so we need to avoid write failure while guaranteeing that the static noise margin is large enough. For the read port, the challenge is that the ratio of I_{on} and I_{off} is too small, which causes the limited sensing margin, the small sweeping might lead to reading failure.

For the SRAM based PUF, the data is generated based on the mismatch between the PVT variations. If the mismatch is significant, it tends to reproduce

the same data over time, meaning the device has high reliability. However, for the SRAM mode, the large mismatch between two cross-coupled inverters causes weak stability, which degrades the performance. The common approaches for PUF to strengthen the mismatch, such as aging and burn-in techniques, are facing these issues. The previous works address the issues to generate reliable SPUF with large CRPs, to improve the reliability and security in SPUF design.

The research outcomes are demonstrated through ultra-low-voltage SRAMs, sequence-dependent PUF and reliable SRAM based PUF circuits. Specifically, the contributions of our research works are as followings:

1. The first work presents an 8T SRAM macro with vertical read word-line (RWL) and selective dual split power lines techniques. The proposed vertical RWL reduces dynamic energy consumption during a read operation by charging and discharging only selected read bit-lines (RBLs). The data-aware dual split power lines technique combined with vertical write bit-lines enhances both the write margin (WM) and the static noise margin (SNM). A 16kb SRAM test chip fabricated in 65nm CMOS technology demonstrates the minimum energy consumption of 0.506 pJ at 0.4 V. This achieves the smallest normalized energy per bit when compared to other state-of-the-art works. The SRAM could operate functionally down to 0.26 V.
2. The second work proposes a 2D sequence-dependent PUF based on SRAM. It expands the challenge to response pairs (CRPs) by the order of rows^(sequence length - 1) × columns^(sequence length - 1) for reliable authentication.

This is achieved by configuring the sequence of SRAM cell selection. Each bit cell has a vertical word-line and a horizontal word-line to utilize the orthogonal word-lines to connect four cells simultaneously to generate one-bit data. The proposed technique allows us to generate multiple data maps from one chip. This non-linear behavior also makes the chip more resistant to attacks using machine learning. A test chip was fabricated in 65 nm CMOS technology with the area of the chip is 12580 μm^2 . The bit error rate is 3% at the nominal point (0.8 V/20 °C) and the inter-hamming distance between chips is 0.4869. The hamming distance of 0.427 was measured when using the same sequence length with different orders. The CRPs is improved from 1.17×10^{11} in the previous work to 8.37×10^{17} .

3. The third work proposes a sequence-dependent SRAM based PUF with high reliability. It expands the CPRs by the order of rows^(sequence length - 1) $\times$ columns for reliable authentication, for each step it improves the reliability from $\frac{1}{2}$ to $\frac{3}{4}$. Each bit cell implements separated word-lines to control the two access transistors, during the PUF mode, turns on two left word-lines and two right word-lines simultaneously. 18 transistors are involved to generate one-bit data. The bit error rate is improved 3% at the nominal point (1 V/20 °C) by a one-step operation. For the same sequence with a different order in one chip, the hamming distance achieved 0.441 when sequence length is 5. The measurement results demonstrate that the proposed PUF is configurable and reliable for the physical security keys generation. A test chip was fabricated in 40 nm CMOS technology with the area of the chip is $154 \times 43 \mu\text{m}^2$.

In summary, the main contribution of this thesis is exploration and demonstration of optimal solutions for SRAM and PUF circuits to overcome the limitations of ultra-low-voltage operation while satisfying the requirements of energy-constrained applications.

1.3 Thesis Organization

The organization of this report is as follows.

Chapter 2 presents the view of different types of SRAM bit cell, analyzing the write and read assist circuit to improve the performance, and compares the advantages and drawback between different approaches. In addition, this chapter presents some previous works which integrate multiple techniques together to optimize the design. Meanwhile, this part discusses some state-of-the-art of PUF, with their own pros and cons.

Chapter 3 proposes an 8T SRAM cell for ultra-low-power application. This chapter proposes a vertical read word line (RWL) scheme for read operation and a dual split power lines technique for write operation. To reduce leakage, we adopt multiple threshold devices and virtual read ground to optimize the performance. This chapter also analyzes the simulation data and presents the measurement result.

In Chapter 4, a sequence-dependent SRAM based PUF is proposed. This chapter utilizes four cells to generate one data at one time, which could implement multiple operation steps (sequence length). This part elaborates the principle of the proposed PUF operation and the analysis of the enhanced CRPs. Additionally, it presents the results of the measurement to demonstrate the performance.

Chapter 5 presents a proposed 6T SRAM based PUF with high reliability and enhanced CPRs. This chapter uses three cells to generate one data at one time,

with multiple operation steps, it analyzes the improvement of the reliability of each step with CPR enhancement and the area/power overhead. For one device, we use varying temperature and voltage to generate different maps to present reliability results.

Chapter 6 summarizes all the works presented in this thesis and proposes the recommendations for further research.

Chapter 2 Background and Literature Review

The growth in demand for portable devices and wearable technologies requires an extension in a battery lifetime, so the importance of low power consumption has increased significantly. Supply voltage reduction is the predominant way to achieve power reduction in system-on-a-chip (SoC) designs. However, the high density of SRAM bit-cells with minimum size transistors restrict the minimum operating voltage. Since the robustness of SRAM is substantially degraded in near-threshold voltage, the fundamental concern of the stability is limiting the minimum voltage.

2.1 Conventional 6T Single-port SRAM

The conventional 6T SRAM (Fig.2.1) uses latching circuitry (flip-flop) to store one-bit of information, the core of the SRAM is formed by two coupled CMOS inverters (M1 and M2 as pull up transistors, M3 and M4 are pull-down transistors), each output of inverter as the input to feed the other inverter, so the data is be stored stably in this feedback loop. As long as the power supply is maintained, the cross-coupled structure won't lose the data. The data stored in Q can be read and written by the access transistor M5 and M6.

The structure of $m \times n$ SRAM array is shown in Fig.2.2, the array is sharing the word line (WL) in a horizontal way and sharing the bit line in the vertical direction. During the hold period, the word lines will be '0', and all of the bit lines

Fig. 2.1 Conventional 6T SRAM bitcell.

Fig. 2.2 Conventional 6T SRAM $m \times n$ array.

Fig. 2.3 The β ratio measures the relative strengths of the pull-down and pass-gate transistor [12]

Fig. 2.4 Active pattern for a) Thincell [12] and b) straight-line layout. c) The straight-line layout yields higher SNM.

will be charged to VDD. During the read operation, in the selected column, the BL and BLB will be charged to VDD then release floating, and the WL in the selected row will be turned to high voltage. The access transistors M5 and M6

will be turned on, and then the BL or BLB will be discharged which regarding the data stored in the selected cell is '0' or '1'. In the write operation, if the input data is '0', the BL will be GND, and BLB will be VDD, also turned on the WL. Then the Q will be discharged by BL and QB will be charged by BLB and the PU transistor. For the write operation, the unselected cells in the same row are suffering from the asserted access transistors, we call it half-select condition. If the access transistors are too strong, the half-selected cells will easily be flipped. On the contrary, when access transistors are too weak, it may cause write or read failures. Stability is a critical issue during the read and writes operations, but it could be optimized by using the device with a different threshold voltage, choosing the cell layout and adjusting the β ratio (Fig.2.3) [12]. One way would be to choose low threshold PMOS for pull up transistors and high threshold voltage CMOS for access transistors, which could improve the static noise margin (SNM), meanwhile degrade the write margin (WM). Or we could adjust the V_{th} by changing the length of the transistor, which is effective at a given VDD, but not suitable for a wide voltage range, and it's also technology dependent. The second way, the cell layout can be optimized to minimize variability by converting the "Thin-cell" active pattern to a straight-line layout (Fig.2.4), which could eliminate jogs and corners, thereby reducing reliance on interconnections of metal. This facilitates lithography and reduces sensitivity to overlay errors, which improves mismatch and critical dimension control, thus increasing SNM. Third, it could also change the β ratio, which is a trade-off for area balance and performance. Increasing the β could reduce the disturbing voltage at 0 nodes of the cell while augmenting the β could reduce the switch point of the inverters. These two factors affect the SNM in the opposite way.

Fig. 2.5 Optimal SNM lies at $\beta \sim 2-3$. Cell sizes range from $\beta = 1$ to $\beta = 4.4$.

Finally, the optimal stability of β is from 2 to 3 (Fig.2.5), but the impact of β for the SNM is limited. If considering the area overhead, normally will choose β equal to 2. The minimum operating voltage of convention 6T SRAM is around 700 to 800 mV and needs assisting circuits to improve the stability which also consumes power and occupies a portion of the area.

2.2 The Conventional 8T SRAM

In conventional 6T SRAM, the stability is fundamentally more vulnerable during a read operation as compared to write operation, so the low read SNM may lead to the failure of the operation, which restricts the minimum operation voltage significantly. Meanwhile, the area overhead compared to the 6T SRAM is only 30% larger with the same β value. The conventional 8T SRAM (Fig.2.6) could effectively improve stability without degrading the performance. Because the 8T SRAM eliminates the disturbing current, which flows from access to the 0 node side and pulls it up to nonzero value. But the 8T SRAM needs to split up the write and read word lines and the bit line for dual-port operation. The write SNM is determined by the cross-coupled inverters, which are significantly larger than the 6T ones, and for the read operation, we don't need to be concerned about the stability, as there is no current drop into the inverters, the same as for the half-selected cells. For 8T SRAM the bit line is discharged when reading 0, and when reading data 1 the read bit line is only discharged by the leakage current. During the large voltage supply, the leakage current is too small comparing to the discharging current, it is negligible. Under the near-threshold voltage, the ratio of I_{on} and I_{off} will be reduced significantly, then the value of RBL 0 and RBL 1 will be close to each other. Since the uncoupled write and read ports don't have differential bit line for the read operation, it is necessary to generate an accurate reference signal to distinguish data 0 and 1, otherwise the read operation will fail.

Fig. 2.6 Conventional 8T SRAM bitcell.

For both conventional 6T and 8T SRAM we could use assisting circuit to improve robustness and performance, but for the near-threshold voltage the area overhead of 6T increases faster than 8T. So, the 8T structure provides extra benefits with ultralow voltage supply and could allow continually further scaling beyond the possibility of traditional 6T SRAM.

2.3 Techniques for the Write and Read Operations

Implementing assisting circuits is a common practice to improve the performance and stability of traditional 6T and 8T SRAM. Each technique aims at either write or read operation. Usually, for one design, it is necessary to combine more than one technique together to achieve better performance. Each technique brings area and power disadvantage.

2.3.1 Write Assist Techniques

(i). Cell VDD Lowering

Decreasing the VDD_{cell} lower than supply voltage as shown in Fig. 2.7, will cause the pull-up and pull-down transistors to be weaker than the access transistors [13-15].

Since the cell voltage drops to a lower level, the pull up transistors become weak. For the write operation, the write ability depends on the side which writes '0' in. If the pull up current is reduced when BLB discharges QB node, the QB is easier to be discharged to logic low. So that the lowering VDD_{cell} can improve the write ability. However, for the half-selected cells, the disturbing current is also easier to overwrite the stored data. Moreover, static current flow is generated by the voltage divider when clamping VDD_{cell} , which consumes more power.

Fig. 2.7 Diagram of cell VDD lowering. (a) schematic (b) timing diagram.

(ii). Word Line Boosting

Overdrive the word-line (Fig. 2.8) [16, 17] increases stress on the gate of access transistors, which raises the current. For the selected cell, since the effect is the same as reducing the value of β , more current injects into Q and discharges from QB, the data is easy to write into the cell. The write ability is improved from word-line boosting effectively. However, for the half-selected cells, the

Fig. 2.8 Diagram of WL boosting. (a) schematic (b) timing diagram.

disturbing current is also increased. Although the BL and BLB are floating on the unselected column, the injected current into logic low side node is still larger, if the voltage of the lower node hits the switch threshold point of the inverter, the data stored inside will flip over. So that this technique improves the write ability while degrades the stability.

(iii). VSS Raising

Another way to improve the write ability is by raising the VSS_{cell} as shown in Fig. 2.9, which weakens the pull-down transistors [18]. Raising the voltage to 150mV in VSS_{cell} will result in 300mV VDD_{min} improvement.

Fig. 2.9 Diagram of VSS raising. (a) schematic (b) timing diagram.

Since VSS_{cell} is pull up, the pull-down NMOS is impaired by reducing V_{GS} . The impact is the same as β reduction. In the normal case, during the write operation, the pull-down transistors compete with access transistors, if the pull-down NMOS becomes weak, the voltage is easier to accumulate on the node.

Boosting Bit Line/Negative BL raises/lowers node above/below the inverter trip

2.3.2 Read Assist Techniques

(i). Word Line Under-Driving

Under-drive the word-line (Fig. 2.11) making the access transistor weaker by decreasing the voltage of the gate to reduce the disturbing current [19]. For the half-selected cells, since the disturbing current is reduced, the stability could be improved. Nonetheless, the lower voltage on access transistors also causes the smaller discharging current. The read time is impaired by implementing this technique. On the other hand, the voltage divider which uses to generate the lower word-line voltage also consumes more power.

Fig. 2.11 Diagram of WLUD. (a) schematic (b) timing diagram.

(ii). Suppressed Bit Line

Utilizing suppressed bit-line (Fig. 2.12) could lead the access transistor weaker by increasing the voltage of source to reduce the disturbing current [19]. Since the disturbing current is reduced, the stability of a bitcell is improved. Not only

can get lower VDD_{min} , but also the suppressed bit-line could reduce the leakage current on the unselected cells. However, because reducing the voltage on bit-line, the read speed will be reduced too, and the swing on the bit-line becomes smaller.

Fig. 2.12 Diagram of suppressed BL. (a) schematic (b) timing diagram.

(iii). Negative Unselected Word Line

Fig. 2.13 Diagram of negative unselected WL. (a) schematic (b) timing diagram.

Using a negative voltage (Fig. 2.13) on the gate of the access transistor to for the unselected rows (US-row) reduce the leakage current. This technique could reduce the leakage current on the unselected cell, and get lower VDD_{min} too. However, it consumes a large power to generate a negative voltage.

Compared to the reported write and read assist techniques, Table 2.1 is shown to analyze the pros and cons of each approach. The write ability - stability is a tradeoff, improve one will degrade the other one. Since no one is helpful for all of the aspects, we need to combine multiple techniques to optimize performance.

Table 2.1 Comparison of different techniques

Cell Metric	Cell VDD lowering	WL boosting	VSS raising	Negative BL	WL under driving	Suppressed BL
Write ability	Improves	Improves	Improves	Improves	Degrades	NI
Stability	Degrades	Degrades	Degrades	Degrades	Improves	Improves
Write time	Degrades	Limited improves	NI	Improves	Degrades	NI
Read speed	Degrades	Improves	Degrades	NI	Degrades	Degrade
Leakage	NI	NI	NI	NI	NI	Improve
Reliability	NI	Degrades when use at high voltage	NI	Degrades when use at high voltage	NI	NI

NI: negligible impact.

2.3.3 Implementing of Techniques

There are several ways to improve the write and read operations of SRAM, but usually, we will not use just one of the techniques to improve the performance, but based on the characteristic of individual assisted circuits, most SRAM will combine two to three techniques to achieve a better result. We will provide some typical examples to explain how to implement different techniques in SRAM applications.

(i). 6T SRAM with the disturbance-noise reduction (DNR) scheme

As shown in Fig.2.14, the 6T SRAM keeps the conventional structure, just by using the assisted circuit to improve the performance. For the write operation, the reported solution uses negative bit line (NBL) to reduce the source voltage of PG1 (PG2) to increase the current flowing through the access transistors, then lower the data below inverter trip point to write data into the selected bitcell.

Fig. 2.14 The SRAM bitcell in [19].

Fig. 2.15 The reported assist circuit in [19].

The whole assisted circuit including the read and write port schematic is shown in Fig.2.15. The NBL write driver will generate the negative voltage which will emit the signal to the bit lines through the transmission gate.

The main contribution of this publication is the read assisted circuit, which is called the disturbance-noise reduction (DNR) scheme, for which the schematic is shown in Fig.2.16. In the DNR the pre-charge and discharge circuits are used

Fig. 2.16 BL pre-charge & discharge circuit [19].

Fig. 2.17 The timing diagram of read assist, WLUD [19].

to reduce bit line noise dynamically by minimizing a timing-penalty. The purpose of discharging BL to a lower level is to suppress the bit line to make access transistor weaker by increasing the voltage of the source, which in turn could reduce the disturbing current. The clamping PMOS circuit is used to optimize the voltage of bit-line with minimum safe-voltage level by trading-off with the

discharge timing. The functionality of the cross-coupled latch is to keep the bit-line high with complementary bit-line.

For the read operation, excluding the DNR scheme, they also used WLUD to reduce the disturbing current from bit-line, which need to generate a lower voltage. The DC current caused by the voltage divider will consume extra power which is inevitable. The timing diagram is shown in the Fig.2.17.

In this reported SRAM, they combined three separate techniques to achieve the desired performance. The NBL is used for write assist and the SBL and WLUD are used for the read assist. The reason why WLOD and VDCL were not adopted in this circuit, is that the higher WL also increases the risk of impairing the stability contrarily for half-selected bit-cells. And although lowering the cell's VDD could improve the write margin, it also affects the SNM of unselected bit-cells in the same column.

(ii). 6T SRAM with SCS-NBL and WRE-LCV technique

This paper also used conventional 6T SRAM bit-cell to reduce the area overhead. For the write operation, this paper implemented write recovery enhanced lower cell-VDD (WRE-LCV) technique to improve the write margin, which in turn could reduce the minimum supply voltage. The reported WRE-LCV schematic is shown in Fig.2.18. The lower voltage is provided by the WRE-LCV voltage generator. The cell voltage is controlled by the column selection signal, where only the selected column will be affected by the lower voltage. The lower cell voltage could weaken both the pull up and pull down transistors [13].

Fig. 2.18 Reported WRE-LCV write assist scheme in [13]

reduces β value. So, if we drop the cell voltage even further, we could get the smaller minimal supply voltage. However, the write minimum voltage improvement will become saturated at 50% VDD LCV power level, because when the VDD drops to a value which is approximately 50% of VDD, the pull-up transistor will be turned off. Afterward, the $V_{\min}$ can't be reduced by lowering the cell VDD. The simulation waveform and simulated write $V_{\min}$ is shown in the Fig.2.19.

Beyond that, this paper also adopted the suppressed coupling signal negative bit-line (SCS-NBL) technique to provide the write assist. The schematic of the reported SCS-NBL circuit is shown in Fig.2.20. The charge pump is used for the negative voltage generator, the optimum size of the C1 capacitor is determined by the BL parasitic capacitor. For the normal charge pump without the NBL_FIRE, the negative voltage will have peak value during the signal switching.

Fig. 2.19 (a) simulation waveforms of the WRE-LCV scheme. (b) simulated write

It is dangerous for the negative value to get too large, as it may break down the diode in BE junction. In order to avoid this situation, they have used coupling voltage suppression block to reduce the coupled negative bias at high supply voltage operation regions, and also to mitigate the pull gate transistors overstress condition. The comparison of the negative generator with and without NBL_FIRE block with the simulated waveforms is shown in Fig.2.21. The suppression block could reduce the dropping speed of the coupling signal, which can relieve the coupled negative voltage.

Fig. 2.20 Reported SCS-NBL write assist scheme [13].

Fig. 2.21 Simulation waveforms of the SCS-NBL write assist scheme [13].

In this reported SRAM, the two combined techniques for the write performance improvement, they combined the lowering of the cell's VDD and negative bit-line together, to realize the low power operation. The area overheads of the two

- (iii). 6T SRAM with PMOS access transistor and negative boosted WL [20]

Since PMOS is used, the access transistors are weaker than the conventional one. So that the reported 6T SRAM has higher static noise margin for the unselected cells in the write operation and better reading stability in the read operation. However, the write ability and read speed are degraded at the same time. In order to improve the write ability and read speed, authors are using negative boosted WL to increase the current on access transistor.

Fig. 2.23 Simulated timing of WL and BLs for boosted and non-boosted options at 300 mV, TT corner, 25 °C.

Fig. 2.23 shows the transient simulation of WL with and without boosting. When WL is negatively boosted, the time required to develop 100 mV of differential voltage is reduced by 10 ns [20].

(iv). 6T SRAM with dual split control assist scheme

As shown in Fig.2.24, the schematic of the reported SRAM is using a dual split control (DSC) scheme to improve the minimal supply voltage. One is the VDD supplied split WL (SWL) to reduce the half select disturbance without degrading the current to charge or discharge the cell, the other one is row-based split cell VSS (S-CVSS) controls to increase the half select noise margin and write margin under the S-WL scheme.

Fig. 2.24 6T SRAM with dual split control assist bitcell structure [18].

Fig. 2.25 Write operation of DSC6T [18].

For the write operation, we can see in the Fig.2.25, the word line and cell VSS are data-dependent. If the input data is '0', only the WL1 will be raised to VDD, and then PG1 will be turned on, and BL will have negative voltage (NBL). Meanwhile, the CVSS2 will raise a certain level to help writing the data. Since the word line is shared by the whole row, the cells in the same row will suffer

Fig. 2.26 Read operation of DSC6T [18].

the half select condition. When PG1 of HS cell turns on, the CVSS also runs horizontally, and the CVSS will mitigate the noise margin, meaning the split cell VSS can improve both the write margin and the noise margin at the same time.

For the read operation, only turn on the PG1 to read data. Compared to WLUD, they chose to raise the CVSS2 to improve the noise margin in the half-selected cells. It was done to achieve a low supply voltage operation without degrading the read speed (like WLUD will reduce the discharging current).

In this reported SRAM, the combined split WL with two separate techniques were used to improve the write and read performance. The split WL, NBL, and S-CVSS help to improve the WM and SNM, and the raised VSS also improved the read port which is also based on the split WL. DSC6T-CSVG consumes 76% less write energy than WLUD did, and it has 2.5% area overhead for a 256-kb SRAM macro in low voltage operation. Thus, the reported SRAM could achieve the characteristic for low power applications.

(v). Comparison of Different Types of SRAM

There are multiple types of SRAM besides 6T and 8T bit-cell structures, like 5T, 7T or more transistors circuits. Most of the 5T and 7T solutions have some obvious drawbacks which are unstable. The more transistors circuits like 9T to 12T sacrifice too much area which is not economical.

Performance comparisons of several SRAMs are summarized in the following Table 2.2. The SRAM in [20] combines PMOS access transistors and negative boosted WL to achieve the lowest minimum operating voltage while the SRAM in [13] lowers cell VDD and uses negative BL to get the minimum operating voltage of 300 mV. The minimum normalized energy consumption is 61 aJ/b in [11] which uses a modified structure with negative BL and virtual GND.

Table 2.2 Comparison of different types of SRAM

	[19]	[22]	[13]	[11]	[21]	[20]
Technology	14nm Bulk FinFET CMOS	14nm FinFET CMOS	16nm FinFET CMOS	65nm CMOS	22nm FinFET CMOS	65nm CMOS
Transistor Count	6T	10T	6T	9T	6T	6T
Capacity	128Mb	84Mb	128Mb	72kb	32kb	2kb
Size	0.08 μm^2	0.0588 μm^2	0.07 μm^2	420F ²	—	508F ²
V _{min}	500mV	600mV	300mV	350mV	300mV	290mV
Frequency	—	1.5GHz @600mV	200MHz @300mV	229KHz @350mV	—	3.34MHz
Min. Power	—	—	—	4.05uW @350mV	—	—
Min. Energy	—	—	—	4.5pJ @350mV	5.4pJ @300mV	2.25pJ @400mV

2.4 The SRAM based PUF

The PUF is a critical part in a modern hardware security system, it is required to generate reliable and unique secret keys which have low cost and are easily implemented. In order to reduce the area overhead from integrating a new functionality, SRAM based PUF (SPUF) is implemented on the existing SRAM structure hardware without adding an assist circuit. Since all memory-based PUF are weak PUF, the challenge-response pairs (CRPs) have limited complexity growth with an increasing number of basic bit cells. Typically, the SPUF uses the address as a challenge and the stored data as a response, the CRP is equal to 1. The generated data is called a golden key, which is treated as the chip ID.

The memory-based PUF are usually used for chip identification and cryptographic key generation [23-25]. To produce a key with practically enough implementations, an electronic PUF should meet at least some of the following requirements:

- 1) Randomness – the security key is generated from an origin of arbitrary physical parameters, like the CMOS technology process variations, voltage, and temperature changes. The generated data should be reproducible mathematically or physically, which is produced by the stochasticity of manufacturing.
- 2) Reliability – the secure data generated randomly must be reproducible over time and in various environmental conditions, such as different voltage,

temperature variations, and aging. The responses have to be consistent using the same challenge.

3) Uniqueness – for each challenge the response must be unique, could not be produced by another challenge in the same device nor other dies. The security keymap should be unique and unpredictable.

Nowadays, these criteria are used to evaluate the PUF design, several techniques are adopted to cater to these standards.

2.4.1 Error-Correcting Codes (ECC)

In order to generate a reliable security key, the bit error rate (BER) should be as low as possible. For the conventional SPUF, if the mismatches between two cross-coupled inverters are too small, the secure data is easily affected by the noise. To address this issue, a key generator based on fuzzy extractor (FE) is utilized to produce a cryptography key and helper data. It uses the helper data to recover the key from PUF responses which is stained by the noise, to improve the reliability.

Although ECC is very useful for enhancing the reliability in the security application, the helper data is inevitably generated to immune the noise. Therefore, the helper data could expose the information from security key more or less, which diminishes the unpredictability of the PUF device. Furthermore, the randomness of the PUF will lose due to the error-correcting code's capability for helper data scaling. Since the complexity of PUF decoder is increased, the

assist circuit needs to occupy a large area, which raises the cost of the PUF device.

Fig. 2.27 shows a basic fuzzy extractor based PUF based key storage and generator system [26]. It consists of two parts, enrollment and reconstruction.

Fig. 2.27 Operations of a PUF-based key storage system. (a) Enrollment. (b) Reconstruction [26, 27].

Enrollment is an operation to generate a cryptographic key. The measured secure data is recorded as a PUF reference response (PRR), and an extra arbitrary seed is adopted. Using a fuzzy extractor to process the PRR and the extra random data is used to generate a high-security cryptographic key and helper data. However, this helper data needs to be stored in non-volatile memory as public information.

Reconstruction is a reverse procedure of the enrollment operation to recover the enrolled cryptographic key. It first needs to take the measured PUF data,

then records it as a PUF response (PR), which is a noisy version of PRR. In the next step, takes the helper data from the extra non-volatile memory, using the fuzzy extractor to process the PR and the helper data simultaneously. If the processed PR is quite similar to PRR, it means the fuzzy extractor successfully decoded the cryptographic key.

2.4.2 Temporal Majority Voting (TMV)

Temporal majority voting is a very useful method to improve the reliability of a PUF device. If the response of a PUF is mostly affected by a transient noise, the majority voting technique could effectively improve the stability. It collects significant quantities of responses from the same PUF device by repeating the operation multiple times.

If a PUF device has an initial bit error rate as δ , with an n -way temporal majority voting, the error rate could be effectively reduced to δ' , where $\delta' = 1 - \sum_{i=1}^{(n-1)/2} \delta^i \times (1 - \delta)^i \times C_i$ [28]. However, the TMV could only help to correct the error rate up to 8% based on the 15-way voter [29].

Fig. 2.28 displays the structure of TMV circuit in [28]. The circuit has a 4-bit counter, which is used for 15-way TMV. Fig. 2.29 shows the results of measurement in this paper. It clearly shows that the 15-way TMV technique could effectively reduce the unstable bits 53% of the time from the raw data.

Except for the temporal majority voting, if we repeat the PUF operation on multiple devices at the same time, then it is called spatial majority voting.

Fig. 2.28 TMV circuit [27]

Fig. 2.29 TMV measurement result [28]

2.4.3 Aging

Negative-bias temperature instability (NBTI) is a crucial reliability issue in CMOS technology. It stresses negative bias on the gate of the PMOS transistor, to modify the charge trap inside. As a result, the threshold voltage of the transistor is increased, consequently degrades the drain of the source current and decreased the transconductance [30]. If the negative bias on the gate of PMOS is removed, the energy of the charge trap will be also released. The degradation of some components is recoverable, but also could be permanent for others. When we apply reversed bias on the gate of PMOS, it could help to mitigate the degradation of the permanent ones [31]. The similar process on the NMOS is called positive-bias temperature instability (PBTI), which is much weaker than NBTI, so it is always ignored.

On the SPUF, the start-up data is depended on the mismatch between cross-coupled inverters. Sometimes, it is possible that the majority of the bit cells are biased towards '1' or '0'. In this case, the NBTI technique could be used for aging the transistors. Negative stress could be applied on one side of the PMOS pull-up transistor, to reduce the pull-up current. Moreover, if the mismatches between the individual cells are too small, the security data can easily be affected by low noise, making the PUF unstable. Under this situation, we also could apply aging injection on the cells, to increase the mismatch between two sides [30].

The aging technique could improve the reliability and uniqueness significantly, which is becoming the most common way to improve the performance of a PUF

device. However, aging could damage the transistors irreversibly. This is a serious drawback especially for the SPUF because the device still needs to work in SRAM mode. The increased mismatch degrades the performance of SRAM functionality. And also, accelerated aging shortens the lifetime of a device.

In order to alleviate the damage, [32] reported a method to age gracefully. As shown in Fig. 2.30, they combined ECC, TMV and aging techniques together, to improve the performance with minimum damage to the transistors. They used a control circuit to supervise the results from TMV, and if the BER hits the optimal standard of the PUF, the control circuit will stop burn-in operation.

Fig. 2.30 Block Diagram of [reported](#) reliability enhancement scheme [32]

2.4.4 Analog PUF

Instead of relying on the mismatch between two cross-coupled inverters, [33] seeks for a new technique to transfer the existing SRAM circuit to analog PUF. They use the manufactory mismatch between two access transistors, utilize the native device to measure the threshold voltage of two access transistors. Comparing to the conventional way, the reported technique is highly reliable. Because it also compares the value of threshold voltage between two transistors and utilizes the native device to build a voltage divider, converts the threshold voltage into an analog voltage.

Fig. 2.31 Architecture of the analog PUF schematic [33]

This SPUF design uses a differential sensing amplifier to detect the difference between converted voltages. However, based on the architecture which is shown in Fig. 2.31, the converted analog voltage from the threshold voltage of two access transistor couldn't be generated at the same time. The point of this work is, it is difficult to maintain an analog voltage, and waiting for the comparison with another one.

2.4.5 Connecting multiple bit cells

Another trend for SPUF is improving the randomness by increasing the number of transistors which are involved to generate a one-bit data. Since the randomness comes from PVT variations created during manufacturing, and

Fig. 2.32 8T-DP-SRAM operating in PUF mode: (a) BL1 = 1, BL2 = 0 and (b) BL1 = 0, BL2 = 1. [34]

Fig. 2.33 Sequence dependent PUF – Small array example. [35]

more devices are involved to generate data, means it contains more uncontrollable parameters. Creating conflicts also enhances the arbitrary data generation.

[34] provides a solution to utilize the dual ports SRAM bit-cell. Compared to the conventional dual-port bit-cell, which tries to improve the overall capability of write and read operations, the reported scheme turns on both ports on one-bit-cell simultaneously. As shown in Fig. 2.32, it creates conflicts between two access transistors on each side. In total it uses 8 transistors to produce one-bit data.

Also [35] chooses to connect two cells together, to involve 12 transistors to generate one-bit data, the schematic is shown in Fig. 2.33. Furthermore, this reported work not only improves the randomness but also enhances the CRPs space. For the SPUF device, the number of CRPs is limited, in the worst case, it is only 1. This design utilizes sequence-dependent PUF operation, to improve the CRPs significantly, which makes the PUF immune to the machine learning attacks. We will elaborate more on this work in chapter 4 and chapter 5.

Chapter 3 Low Power Low leakage 8T with Vertical Read Word-Lines and Selective Dual Split Power Lines

3.1 Background

With the recent development of portable devices and wearable apparatus requiring long battery lifetime, ultra-low power consumption has been increasingly significant. In these applications, Static Random Access Memories (SRAMs) play a vital role in the requirements of performance, power, and energy [36]. As a crucial approach to meet ultra-low power and energy, supply voltage scaling has been commonly employed. However, SRAMs with minimum size transistors restrict the minimum operating voltage due to various key metrics such as stability, sensing margin, write margin, etc. In conventional 6T SRAMs, the minimum operating voltage is limited by the conflicting requirements from write ability and read stability [37-38]. Therefore, SRAM cells with decoupled read ports are widely employed for supply voltage scaling since they can eliminate read disturbing current and allow more flexibility in the cell design [39]. In addition, SRAMs operating in the near- or sub-threshold have significantly degraded Ion-to-Ioff ratios and exponentially increased variations, which deteriorates various SRAM design parameters. Many design techniques have been proposed to partially address these SRAM design issues [40-49].

For the read operation, several techniques could be used to enhance the Ion-

to-loff ratio between active discharge current and leakage current, including i) Increasing the discharging current such as boosted RWL on the selected cells or using negative cell virtual ground voltage [40]. Both of them incur the power and area overhead. ii) Providing a pull-up current against the bit-line leakage to improve the sensing margin and sensing window (such as using an extra transistor to equalize the leakage [41]). However, the bitline leakage is equalized to maximum leakage, which increases power consumption. iii) Adopting hierarchical bit-line to mitigate cell disturbance by reducing the amount of loading [42]. iv) Generating the self-adjusted reference voltage for sensing amplifiers to enhance the sensitivity to small bitline swing [43]. At higher supply voltage, the dynamic power contributes the most consumption in the read operation. In the conventional SRAMs, majority of the switching power is consumed to pre-charge the unselected bit-lines due to the horizontal read word-lines [41-44]. The proposed vertical read word-lines reduce the dynamic power significantly by eliminating the current flowing in unselected columns. At lower supply voltage, the leakage power becomes equally significant in the total power consumption. To address this issue, higher threshold voltage (HVT) devices are used in the non-critical paths to reduce the leakage current without performance degradation. Virtual ground is widely used to suppress bitline leakage current. In [44], raised the virtual ground and floating bitlines with programmable NMOS sleep transistors are utilized to reduce the leakage current.

For a reliable write operation, write access transistors need to be stronger than pull-up PMOS transistors in SRAM cells. However, the conflicting requirements between SNM and WM restrict supply voltage scaling [45]. Various techniques

have been proposed to ameliorate these issues. For example, word-line underdrive (WLUD) improves SNM while degrading WM [46, 14]. On the contrary, lowering cell supply voltage [47, 48], word-line overdrive (WLOD), and raising VSS [38] will enhance WM while leading to deteriorated SNM. Another technique is using negative bit-line (NBL) [13, 49] for WM enhancement. However, this suffers from area penalty caused by negative voltage generation, and power overhead. Therefore, most works combine two or more techniques [47-50] to achieve the target SNM and WM. In [51], split control of write access transistors is proposed while raising VSS and NBL to enhance the write ability. Meanwhile, the SNM for inactive cells is weakened. The work reported in [51] proposes vertical power lines to improve the write ability also at the cost of stability degradation. [52] is also combining the split power line with forwarding body bias to improve the WM and SNM, it scarifies the cell area. Overall, no simple technique can improve WM without SNM degradation and vice versa. Hence, it is highly necessary to develop techniques for improving WM without SNM degradation or enhancing SNM without WM degradation.

This work presents two techniques to reduce energy consumption and improve the SNM and the WM. The vertical RWL reduces the read dynamic energy. The dual split power lines with selective boosted cell supply improve both WM and SNM compared to the conventional 6T or 8T SRAMs [53]. The remainder of the paper is organized as follows. Section II explains the key principles of the proposed design. Measurement results are presented in Section III. Finally, the conclusion is drawn in Section IV.

3.2 Proposed SRAM Energy Consumption Improvement Techniques

3.2.1 Read Assist Techniques

Fig.3.1 (a) shows the schematic of the proposed 8T SRAM. The 8T SRAM cell consists of a cross-coupled latch (M1-M4), write access devices (M5-M6), and a read port (M7-M8). The read word-line (RWL), the read bitline (RBL), and the write bitlines (WBL and WBLB) are controlled vertically while the read virtual

Table 3.1 Vertical RWL read operation

	Selected	Unselected	Holding
RD_VGND	Low	High	High
RWL	High	Low	Low

Table 3.2 Data-aware write operation in the selected cell

	Writing '0'	Writing '1'	Holding
WBL	High	Low	Low
WBLB	Low	High	Low
WR_VGND	Low	Low	High
VCC1	High	Boosted	High
VCC2	Boosted	High	High

Fig. 3.1 (a) Proposed 8T SRAM cell with HVT devices in the latch and RVT devices in the access transistors and the read port and (b) layout of the 8T SRAM cell based on the 65nm logic design rules.

ground (RD_VGND) (Table 3.1), the write virtual ground (WR_VGND) are controlled horizontally. M7 is stressed by the QB instead of RWL to avoid the coupling effect between RWL and RBL. In write operation, WBL, WBLB, VCC1, and VCC2 are biased separately depending upon the write data (Table 3.2). WR_VGND of the selected cell is grounded while those of the unselected cells are kept at VDD. The cross-coupled latch is implemented with HVT devices for leakage reduction while the write access paths and the read port employ regular threshold (RVT) devices for better performance. The novelty of the proposed 8T

SRAM cell compared to the conventional 8T SRAM cell includes the vertical RWL and the data-aware control of VCC1, VCC2, WBL, and WBLB. The detailed operation of the cell will be explained in the following sections.

Fig.3.1 (b) shows the cell layout of the proposed 8T SRAM. The placement of transistors is the same as the conventional 8T cell [54] and follows a similar layout style except the metal routing. The power lines in the conventional cell are split to VCC1 and VCC2 running vertically. The RD_VGND and WR_VGND are implemented with horizontal metal-3 lines, and the RWL is realized in a vertical metal-2 line. The cell layout occupies $2.23 \times 1.02 \mu\text{m}^2$ when designed in the logic rule of the employed 65nm CMOS technology.

3.2.2 Energy reduction with vertical RWL

SRAM energy consists of dynamic energy (switching) and static one (leakage). Several techniques are available for leakage reduction. Floating bitlines [44] could reduce the leakage current through access transistors by automatically generating bitline voltage providing minimum bitline leakage. Body biasing [55] is another method for leakage reduction. Since leakage current is an exponential function of the device threshold voltage, applying reverse body bias can reduce leakage current effectively. Another common idea is to adopt multiple V_{th} devices (MTCMOS) [56] so that the leakage in the non-critical paths can be reduced. Virtual ground (RD_VGND) [44] is also widely used to reduce the leakage flowing through read ports. This work also employs MTCMOS and RD_VGND to reduce the leakage in the hold mode.

One of the most significant dynamic read energy in the conventional SRAMs comes from bitline discharging and pre-charging. Even though only a few columns are accessed for read operation, activating a horizontal RWL can allow all RBLs to be discharged in the worst case. This work tackles this issue by utilizing a vertical RWL scheme. Fig. 3.2 compares the energy consumption of the proposed RWL scheme with that of the conventional counterpart. Note that the array size of 256 rows $\times$ 128 columns with the multiplexer ratio of 16-to-1 is assumed. In the conventional scheme (Fig. 3.2(a)), all RBLs are discharged in the worst scenario where all QBs in the selected row are '1'. This is not the case in the proposed SRAM where only a selected RBL can be discharged to a lower level. This is possible since the vertical RWLs of unselected columns are grounded (Fig. 3.2(b)). This reduces the read energy significantly. In the proposed SRAM, the discharging current in the selected column is the pull-down cell current minus the pull-up leakage current. In the conventional SRAM, the discharging current consists of the pull-down cell read current and the pull-down leakage current. The dynamic discharging energy in the selected column of the proposed scheme is 10% ~ 35% larger than that of the conventional one. Since RD_VGND is used, it increases leakage by 17 %, power by 3%, and area by 4.8%. However, the proposed SRAM eliminates the dynamic energy from the unselected columns. As shown in Fig. 2(c), in the SRAM array, assuming the data pattern is 50% "1" and 50% "0", under room temperature and 0.5 V supply voltage, the energy consumption in the array without peripheral circuit for the conventional read port is 5 times of the proposed one. Therefore, the overall energy of the proposed SRAM is smaller than that of the conventional 8T SRAM.

3.2.3 Read bitline sensing with vertical RWL

Reliable read operation depends on the sensing margin, which is defined as the difference between the RBL levels of data '0' and '1'. The RBL level is determined by several parameters such as bitline leakage, bitline length, supply voltage, etc. In the conventional SRAM (Fig. 3.3(a)), the worst case scenario for reading '0' is with minimum pull-down leakage current, which occurs when all QBs in the unselected cells are '0'. This will produce the lowest RBL discharging speed. Likewise, the worst case scenario of reading '1' is with maximum leakage current, which occurs when all QBs in the unselected cells hold data '1'. Maximum pull-down leakage discharging RBL for data '1'. The necessary

condition for successful reading is that the cell discharging current with the minimum leakage current ($I_{\text{cell}} + I_{\text{leak_min}}$) needs to be larger than the maximum leakage current ($I_{\text{leak_max}}$). Supply voltage scaling will make I_{cell} close to I_{leak} and degrades the sensing margin. Fig. 3.3 (b) and (c) shows the impact of supply scaling and bitline leakage on RBL sensing. At higher level supply voltage (Fig. 3.3(b)), the bitline leakage. Therefore, the sensing margin is large enough for single-ended sense amplifiers to be adopted. However, when the supply voltage scales to the subthreshold region, bitline leakage becomes close to the pull-down cell current. This increases the impact of the bitline leakage on the sensing margin. As presented in Fig. 3.3(c), the RBL level for data '1' could be lower than that of data '0' because of the significant data-dependent bitline leakage.

Fig. 3.4 illustrates the worst case read conditions of the proposed vertical RWL technique. When the read data is '0' (QB = '1'), RD_VGND of the selected row

For successful sensing, $I_{cell} + I_{leak_min} \gg I_{leak_max}$

(a)

Fig. 3.3 (a) Principle of the conventional read operation. (b) Sensing margin at 0.6 V and (c) Sensing margin at VDD = 0.25 V. The number of cells per bitline is 256.

Fig. 3.4 Principle of the proposed read operation.

is grounded and RBL is discharged through M7a and M8a. The half-selected cells in the same column generate pull-up leakage and prevent RBL from being fully discharged to ground. When all the half-selected cells store '0' ($QB = '1'$), the pull-up leakage will be maximized forming the highest RBL level for data '0'. Similarly, if the read data is '1', the worst case RBL level is found when all the half-selected cells store '1' ($QB = '0'$). This will lead to the lowest RBL level for '1'. Unlike the conventional SRAMs only with pull-down current paths, the proposed scheme has both pull-down and pull-up current paths in RBL. This forms static RBL levels for data '0' and '1', which widens the sensing timing window. Fig.3.5 compares the RBL waveforms of the proposed vertical RWL with that of the conventional scheme. It is obvious that the proposed scheme produces better sensing timing window and sensing margin.

Fig. 3.5 Impact of the proposed vertical RWL on the RBL swing when compared with the conventional RBL structure.

For the conventional horizontal RWL scheme with row based VGND [57][58], the leakage current keeps charging the RBL to maintain the high potential. At the same time, if the RWL is 0 in the unselected rows, RBL could be discharged to very low voltage. Lastly, the sensing margin of the horizontal RWL with row based VGND is better than both conventional SRAM and the proposed scheme.

(a)

(b)

(c)

Fig. 3.6 (a) Read times of the proposed and conventional, RPB=256, CMUX = 16, in TT corner with varying temperature and supply voltage. (b) Normalized read time in varying voltage and temperature. (c) Read energy consumption of the proposed and conventional one, RPB=256, CMUX = 16, in TT corner with varying temperature and supply voltage.

However, since the leakage is charging the bit-line during the read operation, for each column it consumes more energy compared to the conventional scheme. All RBLs are consuming energy, and eventually, the read energy consumption is larger than the convention scheme.

Fig. 3.6(a) the read time comparison between the proposed vertical RWL scheme and conventional horizontal RWL scheme. 256 cells are loaded on one RBL, in TT corner and CMUX is 16. The data pattern is 50% “0” and 50% “1”.

Fig. 3.6(b) shows the normalized read time, it shows the read time ratio between conventional and proposed SRAM. Overall, the proposed read time is only slightly longer than the conventional one. Only at 0.4V and when the temperature is 125°C, the ratio reaches 1.56.

Fig. 3.6(c) shows the energy consumption comparison between the proposed vertical RWL and conventional horizontal RWL. 256 cells are loaded on one RBL, in TT corner and CMUX is 16. When the data pattern is 50% “0” and 50% “1”, the conventional energy is almost half compared with the data pattern is 100% “0”. Because for the conventional one, RBL needs to consume a large amount of energy when data stored is “0”. On the contrary, if the data stored is “1”, only leakage current consumes the energy. For the proposed vertical RWL scheme, no matter the data stored in the unselected column is “0” or “1”, since the RWL is turned off, the energy consumption is almost equal.

For the data pattern with 50% “0” and 50% “1”, under 1V supply voltage, the energy consumption of conventional horizontal RWL SRAM is 5 times higher than the proposed one. And when the supply voltage is 0.4V, the proposed one has 2/3 reduced energy consumption compared to the conventional SRAM. It

demonstrates that the proposed vertical RWL design could save significant dynamic power. Note this energy consumption is without any peripheral circuits.

3.2.4 Sensing amplifier with reference signal

The proposed RBL cannot be fully discharged to GND due to the leakage current. To solve it a single ended sensing amplifier is introduced. In the whole voltage range, the proposed RBL '0' is always slightly above half VDD. To maintain the reference voltage at a high level, we have adopted three PMOSs to generate the reference voltage as shown in Fig. 3.7. To reduce the leakage current, the three PMOSs are minimum width with increased length (80nm).

Fig. 3.7 Diagram of sensing amplifier and reference signal generator

3.3 Proposed SRAM Ultra-Low Voltage Techniques

3.3.1 Write operation with selective dual split power lines

HVT devices are adopted at the cross-coupled latch of the proposed SRAM cell to reduce the leakage current. However, this degrades the write speed at ultra-low voltage operation, which makes the write speed dominant and increases the overall energy. To ameliorate this issue, we adopt RVT devices on the access transistors. However, this will deteriorate the stability of the half-selected cells in a selected row. [52] presents the split power line to improve the write ability, and forward body bias is used on PU transistor to compensate the stability degradation. To control the body voltage, the area overhead is significant. In this work, we utilize data-aware VCC1/VCC2 control to improve both write speed and write margin. Furthermore, split boosted power lines are proposed to improve the SNM of the half-selected cells. Thereby, the proposed SRAM can operate reliably at ultra-low supply voltage, consequently reducing the total write energy.

Fig. 3.8 illustrates the principle of the proposed SRAM write operation with selective dual split power lines, and Fig.3.9 indicates the approach of VCC1/2 to improve write ability and stability using written '0' as an example. In the selected row, WR_VGND[0] is connected to GND. If the write data is '0', WBL[0] is raised to VDD and Q00 is discharged to WR_VGND[0] (Fig. 3.8(a)). Meanwhile, a charge pump circuit generates boosted voltage on VCC2[0]. As shown in Fig. 3.9 (Q[0] = '1'), the single ended writing has less write margin,

(a)

(b)

Fig. 3.8 Principle of the proposed write operation: (a) the topology of writing '0' operation, (b) the topology of writing '1' operation, (c) timing diagram of writing '0' operation and (d) timing diagram of writing '1' operation.

will be set to VDD (as Q10 in Fig. 3.8(b)). Since the unselected WR_VGNDs are pre-charged to high, the current flows from WR_VGNDs to Q_{x0} ($x \in (1, (m-1))$), and the voltage of Q_{x0} goes up. The data in Q_{x0} and Q_{Bx0} could be flipped because of the disturbing current. However, the proposed write technique utilizes boosted VCC2 to overdrive the pull-down transistor on the Q_{x0} side. This discharges the disturbing current on Q_{x0} quickly. As shown in Fig. 3.9(middle), if the $Q[m]$ reaches the switching threshold voltage of INV II, data will flip. In this case, Vst2 could be improved by boosting VCC2, avoiding the data flips in the half-selected condition. If the data stored in the half-selected cell is '0' while writing '1' then the WBLB[0] will be asserted (as Q_{B00} in Fig. 3.8(a)), the unselected WR_VGNDs are pre-charged to logic '1' then floating. No disturbing current flows from Q_{B00} to WR_VGND during the write operation. Even with boosted VCC1[0], the pull-up transistor on the Q_{x0} side is open since Q_{Bx0} is '1'. Therefore, the voltage at Q_{x0} is not affected by boosted VCC1[0]. As shown in Fig. 3.9(bottom), it's a mirrored case for Q_{B00} . If we boost VCC2, because $Q[n]$ is 1, the M2 is turned off, the value of $Q_{B[n]}$ maintains at GND. In order to minimize the write time and avoid the fluctuation on the boosted signal at the beginning, the VCC1/2 is turned on before WBL/WBLB enable. Consequently, the dual split data-aware power lines scheme improves the write ability and strengthen the stability, simultaneously. For the hold period, both VCC1 and VCC2 are at VDD, the WBL/WBLB is '0' to disable the access transistors, and all of the WR_VGND are pre-charged to VDD. Since the surrounding nodes are held at the same potential, the overall leakage current is similar to the conventional one implemented with the same size and the same threshold devices.

Fig. 3.10 Comparison of WM (a) and SNM (b) between the conventional and proposed SRAM.

Fig. 3.10 presents the statistical analysis for the write ability and stability between the proposed and the conventional SRAM simulated on the schematic level. The number of sample points is 10,000, the temperature is 27°C, the VCC1/2 is boosted to 125% VDD. For the write margin in Fig. 10(a), in the conventional write operation, write failures start to occur when the supply

voltage drops to 0.6 V. The minimum write margin drops to -69.4 mV and -75.4 mV when the supply voltage is 0.3 V and 0.25 V respectively. In the SDSP scheme, write failures start to appear from 0.3 V. The rate is 0.01% in 10,000 samples. When the voltage is decreased to 0.25 V, the write fails rate increases to 0.24%, and the minimum write margin is -7.4 mV. For the static noise margin in Fig. 10(b), in the conventional SRAM, data begins to flip when the voltage reduces to 0.8 V. The worst cases of SNM at 0.3 V and 0.25 V are -101 mV and -110 mV respectively. In the SDSP technique, the data in half-selected cells is fully maintained until the supply voltage drops below 0.3 V. At 0.25 V, 0.71% of samples are flipped. This demonstration shows that SDSP improves the static noise margin significantly.

3.3.2 VSSUP vs. VCC_BST for stability

In the conventional SRAM cell, stability and write ability have contradictory requirements in the write operation. Stronger write access transistors lead to larger write margin while they worsen SNM. However, in the proposed SRAM, both data-aware split ground lines VSS1, 2 (raising VSS value signed as VSSUP) [14, 47] and power lines VCC1, 2 (boosting the VCC value signed as VCC_BST) could improve WM and HS-SNM, simultaneously. Fig. 3.11 presents simulated SNM using VSSUP and VCC_BST. As shown in Fig. 3.11(a), the half-selected cell static noise margin for data storing '0' (HS-SNM-0) is improved by raising VSS while the half-selected cell static noise margin for data storing '1' (HS-SNM-1) degrades. However, since HS-SNM-1 is much larger than HS-

(a)

(b)

Fig. 3.11 (a) impact of VSSUP to SNM under 0.4 V, (b) impact of VCC_BST to SNM under 0.4 V, and (c) impact comparison of VCCUP and VSSUP on SNM.

Fig. 3.12 The circuit in column to generate VCC_BST .

SNM-0, the worst case SNM is still improved. Note that if $VSSUP$ is raised beyond 20% of VDD in this work, HS-SNM-1 becomes dominant and the worst case SNM starts to drop. For the VCC_BST scheme, as shown in Fig. 3.11(b), HS-SNM-0 also improves with VCC boosting while HS-SNM-1 decreases slowly compared to the $VSSUP$ scheme. In this case, HS-SNM-0 is still limiting when VCC is boosted by 20%. Fig. 3.11(c) compares the impacts of the $VSSUP$ and VCC_BST schemes on stability. Since VCC_BST shows better stability improvement, we utilized VCC_BST in this work.

Each power lines ($VCC1$ or $VCC2$) drives 256 cells in one column and 16

columns share one VCC_BST generator. Fig. 3.12 shows the schematic in the column of the implemented boosted voltage generator. VCC1 and VCC2 are controlled by the data input Din and write column signal WR_Col. S0 and S1 are controlling the boosted voltage generators. S0 is used during all write operations, S1 is activated for ultra-low voltage only (below 0.4 V). Note that either VCC1 or VCC2 is connected to the VCC_BST node during write operation. If not, WBL and WBLB are set to VDD, driving VCC1 and VCC2 with VDD.

3.3.3 Write-completion flag generation

Fig. 3.13 The feedback signal generator circuit.

The boosted VCC induces power overhead, and also reduces the write time. To minimize the energy overhead, this work employs a write-completion flag generator for resetting WBL/WBLB once a write operation is completed. A modified SRAM cell is duplicated in each column as shown in Fig. 3.13. The write access transistors are controlled by the write enable signals (WR and

Fig. 3.14 Schematic with the proposed assist circuit.

WRB). During write operation, WR is enabled, the node Q_{feed} is discharged. Once the state of QB_{feed} is flipped, the write-completion flag signal (FB) will disable the asserted WBL/WBLB to reduce the redundant energy consumption. Since QB_{feed} is connected to an inverter, QB_{feed} is flipped later than the cells in the array due to the parasitic capacitance. In addition, the delay of additional NAND2 gate provides enough timing margin for write completion in the main array. Overall the energy consumption is similar to conventional write operation at the same supply voltage.

3.3.4 The architecture of 16k SRAM

Fig. 3.14 shows the architecture of the proposed 16kb SRAM. The test chip is configured of 4×4 kb SRAM macros, each 4 kb SRAM macro is implemented with a 256×16 array, WWL_VGND, RWL_VGND and column decoder array, data-aware VCC_BST switch circuit, local amplifier, feedback loop and boosted voltage generator. WR_VGND (RD_VGND) is driven by a WWL_VGND (RD_VGND) decoder for 4 switch circuits, each one is for 16 columns. VCC1s and VCC2s are generated from a column decoder using write data. A boosted voltage generator is placed at the middle and bottom of each macro with the data-in and data-out buffers. The proposed write assist technique creates 1.37% area overhead in the implemented 16 kb SRAM chip (the area for data-aware power lines boosting in the whole chip layout).

3.3.5 Write-assists comparison

The NBL is one of the widely used write assist techniques [45][47][48], it improves the WM without degrading the SNM. The comparisons of WM and SNM between SDSP and NBL are shown in Fig. 3.15. NBL has better WM in the super-threshold voltage range, becomes very similar at the sub-threshold and worse close to 0.25V. The SNM of NBL is lower than SDSP in the whole voltage range. SDSP consumes more power than NBL in the selected column due to the capacitor of data-dependent internal nodes. However, the unselected

(a)

(b)

Fig. 3.15 the comparison of WM and SNM between NBL and SDSP.

columns don't consume power in the SDSP scheme. Overall, when using 16:1 multiplexer the power consumptions of the SDSP and NBL are similar.

Table. 3.3 presents a comparison of write assists of SDSP and previous write assist schemes. In this table, all the write assist schemes are simulated in the same technology node and under the same conditions: every single cell under 0.5 V, 27°C, testing the WM in SF corner and HS-SNM in FS corner. Furthermore, the same deviation values are implemented: $\Delta\text{NBL} = 0.2 \cdot \text{VDD}$,

Table 3.3 Comparison of write assist with previous work

Assist types	Conv. 6T	DTWL ^{*1} + NBL [52]	DDDS ^{*1} +BM ^{*1} [48]	DSC ^{*1} +NBL [45]	DSC ^{*1} +DS PL This Work
schematic
Cell Size	1x	1x	1.75x	1x	1x
HS assist	No	DTWL	FBB	S-WL + S-VSS	S-WL + S-VCC
Write assist	No	DTWL + NBL	DDDS	S-VSS + NBL	S-WL + S-VCC
Leak age	NI ^{*2}	NI	Degrades	NI	NI
WM ^{*3} ₄	Fail	x0.63	x2.48	x2.5	x1
HS-SNM ^{*3} ₄	Fail	x0.23	x0.34	x0.34	x1

*1 DTWL: Dual transient wordline; DDS: Data dependent differential supply; BM: Body modulation; FBB: Forward body bias; DSC: Dual split control.

*2 NI: No/negligible impact.

*3 Data is under 0.5V supply voltage, 27°C, WM is in SF corner, SNM is in FS corner.

*4 $\Delta NBL=0.2*VDD$, $\Delta WLUD=0.2*VDD$, $\Delta WL0D=0.2*VDD$, $\Delta DDS=0.2*VDD$, $\Delta VSS=0.2*VDD$, $\Delta VCC=0.2*VDD$.

$\Delta WLUD = 0.2*VDD$, $\Delta WL0D = 0.2*VDD$, $\Delta DDS = 0.2*VDD$, $\Delta VSS = 0.2*VDD$, $\Delta VCC = 0.2*VDD$. The WM and SNM can change based on the technology and assist voltage.

3.4 Measurement Results

We fabricated a 16kb SRAM in 65nm CMOS technology. It comprises 4 sub-blocks (each block consists of 256 rows $\times$ 16 columns) and generates four-bit data for test purposes. S0 is asserted in the write operation, to generate the

(a)

(b)

Fig. 3.16 (a) The die photo of proposed 16 kb SRAM. (b) The waveform of test chip at 0.26V voltage supply.

(a)

Total Current

Leakage Current

(b)

(c)

Fig. 3.17 Measured results: (a) minimum clock period, (b) leakage current and total current, and (c) energy consumption.

Table 3.4 Comparison of SRAM performance

	JSSC 2013 [51]	TCAS-I 2016 [41]	JSSC 2017 [45]	JSSC 2017 [49]	JSSC 2018 [59]	This Work
Technology	65nm	65nm	28nm	28nm	65nm	65nm
Density	2kb	16kb	256kb	32kb	2kb	16kb
Transistor count	9T	9T	6T	8T	6T	8T
Cell area	675.03 F ²	448.19 F ²	161.99 F ²	708.56 F ²	508.88 F ²	538.37 F ²
VDDmin	0.28V	0.26V	0.5V	0.358V	0.29V	0.26V
Frequency	330kHz (0.32V)	N.A.	20MHz (0.5V)	50MHz	3.34MHz	1.4MHz (0.4V)
Leakage current	0.05μA (0.4V)	1.4μA (0.1V)	120μA (0.5V)	N.A.	8.7uA (0.29V)	0.22μA (0.26V)
Normalized Leakage	24.4pA/b (0.4V)	85.44pA/ b (0.1V)	457.8pA/ b (0.5V)	N.A.	4.25nA/b (0.29V)	14pA/b (0.26V)
Min. energy	0.57pJ	2.07pJ	30pJ	N.A.	2.25pJ (0.4V)	0.506pJ (0.4V)
Normalized energy	278aJ/b	126aJ/b	120aJ/b	N.A.	1.1fJ/b	31aJ/b

boosted VCC1/2. The S1 only enable to launch the second charge pump when the supply voltage scales under 400 mV. Fig. 3.16(a) shows the test chip micrograph. It occupies $327 \times 181 \mu\text{m}^2$. The sample measured waveforms in Fig. 3.16(b) demonstrates the successful operation of the proposed SRAM at 0.26 V. At 0.26 V, our fabricated chip achieved maximum frequency and energy consumption of 60 kHz and 78 nW, respectively (shown in Fig. 3.17(a)). Fig.

3.17(b) shows the leakage current and the total current at room temperature. The leakage was measured in the standby mode after disabling the clock signal. It is reduced significantly because of the virtual ground, eliminating the sub-threshold leakage in the read port. At ultra-low voltage (under 0.35 V), the leakage current becomes dominant. The measured minimum energy dissipation is 0.506 pJ at 0.4 V (Fig. 3.17(c)), while the operation frequency achieves 1.4MHz. Note that the total current and the energy were measured using balanced read and write operations with balanced data patterns. The proposed 8T SRAM is compared with other state-of-the-art works in Table 3.4. The proposed vertical RWL reduces the dynamic energy significantly without substantial performance degradation. Therefore, we could achieve an energy efficiency SRAM design (31 aJ/b at 0.4 V).

3.5 Conclusion

A 16kb 8T SRAM array with vertical RWL and SDSP for ultra-low-power application has been presented. The column-based RWL reduces the dynamic energy consumption by only discharging the selected RBL in one CMUX. A single-ended write operation is developed with vertical control. Finally, the proposed data-aware SDSP enhance the write margin (WM) and the static noise margin (SNM), simultaneously. The test chip of the proposed SRAM is fabricated in 65nm CMOS technology. Based on the experimental result, the minimum energy per bit of 31 aJ is achieved at 0.4 V. This achieves the smallest normalized energy per bit when compared to other state-of-the-art works. The SRAM is successfully functional down to 0.26 V.

Chapter 4 A Sequence Dependent

Configurable PUF Based on 6T SRAM for

Enhanced Challenge Response Space

4.1 Background

The trend of implementing security in IoT system has currently no consensus, the cyber defensive solutions are limited by the physical constraints and different traits. Software security implementations are flexible, which are easier to maintain and update. However, software security mainly relies on the protection of a physical layer, which is generally vulnerable [60]. Although software implementation is cheaper, it's restricted by the operating system and it is easy to be tampered with or attacked. As a consequence, cybersecurity is shifting from software methods to hardware methods. Whereas the hardware security modules are using some complicated mathematical models to encipher, they become an interesting target for the rival companies and attackers [61]. Physically unclonable function (PUF) was built in the early 2000s [62, 63], which relies on the intrinsic difference and irreproducibility of the security system, to replace complex mathematics puzzles with hardware modules. The critical property of PUF is a unique reproducible response, served to generate and store a secret golden key for secure cryptography. Since it is difficult to predict or characterize uncontrollable variations caused by manufacturing process, PUFs are attractive for security applications such as banking infrastructure, critical communication links, etc. [64].

Fig. 4.1 Random data in SPUF generation mechanism [53].

An ideal on-chip PUF needs to meet three requirements: uniqueness, reliability, and randomness [65]. Uniqueness means that each PUF needs to be unique, even going through the same process, it must be different from die to die. Hamming distance is used to measure the differentials between dies. Reliability implies that the chip is able to reproduce the key pattern, has a tolerance to the environmental noise and non-technology related parameters, such as supply voltage or temperature, and even aging. Lastly, randomness requires the PUF cell to be non-reproducible, the generated data is based on the deviation in hardware manufacturing process mismatch. PUFs could be categorized in multiple ways, such as different circuit structures, the declination analog parameters, etc. The various implementations of the new PUFs influence the entropy of fabricating process mismatches. Because the number of CRPs influences on different applications and models directly, PUFs are divided into strong and weak types. The strong PUF is defined as the one where CRPs is exponentially increasing in direct correlation with the number of basic blocks. It is usually implemented with the authentication protocols since the size of CRP

is large, the challenges are literally impossible to be evaluated in an operable period. However, the response could be affected by a random noise because of the complex combinations of the inherent mismatches. The weak PUF means it is constrained by the limited space of CRPs. The size of CRPs grows polynomially with the number of basic blocks, or slower. In the extreme case, the CRP could be only 1, which works as an identical chip ID. Since the weak PUF merely rely on the inaccuracies in the physical manufacturing process, every individual device is uncorrelated with each other. So that, normally it uses for golden key generation, which is difficult to be interfered and go through all challenges to test responses in real-time. SRAM based PUF (SPUF), as one kind of weak PUF, uses the existing memory array on-chip to generate a PUF function, is devoid of additional components to minimize the area overhead [66].

SRAM based PUF generates random data using a start-up procedure. As shown in Fig. 4.1, when the power ramps up, the pull-up strength of two inverters race to decide one side becoming logic '1' and the other side is '0', then the raw data could be read out through the IO interface. Basically, SPUF should work alternatively in PUF mode and memory mode. The specifications of SPUF are 1) the hardware shouldn't need to be changed in alternative modes; 2) in dual-mode, applications should meet the requirements of the functional system; 3) either mode doesn't affect the implementation of the other one [67]. Ideally, it is expected that the performance of one mode isn't degraded by the other one. However, PUF needs larger variations in fabrication to increase the randomness for security, it is antipole for the SRAM mode, which has a lower tolerance for the process mismatch to strengthen the stability. Therefore, a novel strategy is

(a)

(b)

Fig. 4.2 Conventional PUF cells: (a) unit PUF cell and (b) voltage transfer curve of noise margin. [68].

required to avoid the deterioration of the SRAM characteristic. Depending on the mismatch between the cross-coupled inverters, the SPUF cells could be classified into three types: non-skewed cells, partially skewed cells, and fully skewed cells. It indicates the impacts of mismatch of two inverters are from neutralized to be biased. For the start-up values (SUVs), a non-skewed cell is easily influenced by the noise, and a fully skewed cell is reliable to reproduce

the code. We assume the SUV is only affected by the noise and mismatch between two inverters, the PUF static noise margin (PSNM) [9] is involved to evaluate the impacts of mismatches to SUVs generation.

Fig. 4.2 (a) shows the structure of conventional 6T SPUF bit-cell, the horizontal word-line (WL) controls two access transistors at the same time, and bit-lines BL and BLB are connected to each access transistor. Fig. 4.2 (b) shows the voltage transfer curve of noise margin, the solid line, and dash line represent the V_{in} in INV1 and V_{out} in INV2 respectively. The largest squares inside the curves show the SNM0 and SNM1 of this SRAM cell. However the value of SNM is just for read and write operation in SRAM mode, it doesn't decide the SUVs which are affected by the devices and process variations. So that, PSNM is needed to calculate the noise tolerance of the SPUF cell. In this case, we could get minimum values of noise margin SNM0 and SNM1 as following [68]:

$$SNM0 = \min(NM_H = V_{OH} - V'_{IH}, NM_L = V_{IL} - V'_{OL})$$

$$SNM1 = \min(NM'_H = V'_{OH} - V_{IH}, NM'_L = V'_{IL} - V_{OL}) \quad (1)$$

V_{xy} indicates the voltage of V_{in} on the critical points and V_{xy} points to V_{out} . And the $PSNM_{ratio}$ is calculated as

$$PSNM_{ratio} = SNM0 / SNM1 \quad (2)$$

The $PSNM_{ratio}$ represents the reliability of SUVs, if the value is larger than 1, means the raw data has a higher chance to be 1. If the value is less than 1, the raw data tends to be 0. The bit-cell is more reliable while the value of $PSNM_{ratio}$ is further away from 1.

For the conventional 6T SPUF, only uses raw data as the response, if the $PSNM_{ratio}$ is far away from 1, it means the SNM in SRAM mode is very small, but the cells are very reliable when working under the PUF mode. If the value of $PSNM_{ratio}$ is close to 1, the security code will be highly affected by the environment noise. That means, it is hard to meet the conflicting requirements from both modes. On the other hand, the conventional SPUF only has one response according to one address challenge. It becomes a target exploited by machine learning or other attack techniques since all of the possibilities could be easily evaluated in a short time. To address these issues, we propose a new SPUF structure. It keeps six transistors topology, to guarantee no performance degrading in reading and write operations. Besides, instead of raw data, the code is generated by configurable multiple steps operations. If $PSNM_{ratio}$ is close to 1, the fully skewed cells nearby also could help to improve the reliability of the non-skewed cells. Furthermore, compared to conventional chip ID, this process is a two-dimensional sequence-dependent SPUF cell, which expands the challenge-response space significantly. It increases the difficulty to test all of the combinations in limited time, making it immune to attacks.

This work not only changes the order of selected rows but also in different columns. To increase variations and expand the combination we optimize the bit-cell connection without touching the core structure. The two-dimensional sequence-dependent SRAM-based PUF with full SRAM functions generate multiple reliable PUF data sets, the CRPs achieves $rows^{(sequence\ length - 1)} \times columns^{(sequence\ length - 1)}$.

This chapter is organized as follows. An overview of previous work will be

presented in Section II. Section III explains the principle of the proposed design and the operating principle of PUF mode, followed by a discussion of CRPs space expansion in Section IV. Section V presents the measurement results in 65 nm CMOS to validate the proposed technique. Finally, the conclusion is drawn in Section VI.

4.2 Proposed Sequence-Dependent PUF

Conventional PUFs are based on start-up values during power-up reset and they are mainly determined by the mismatches between two cross-coupled inverters. Since each address corresponds to one value, the challenge and response space is equal to the number of bits in the memory.

(a)

(b)

Fig. 4.3 Proposed PUF cells: (a) unit SPUF cell, (b) the layout of the SPUF cell and (c) 2×2 array for on bit SPUF data generation.

Fig. 4.3 (a) shows the schematic of the proposed 6T SPUF bit-cell, consisting of two symmetrical cross-coupled inverters and two access transistors, with regular threshold voltage device. It connects the transistors in a different way compared to conventional 6T SRAM bit-cell, which adopts two split word-lines (WLL and WLR) running orthogonally. Simultaneously, two bit-lines (BL and BLB) are also routed orthogonally with the corresponding word-lines. The layout of the bit-cell is shown in Fig. 4.3 (b), it follows the same arrangement as a conventional SRAM cell, places the BL and WLR lines vertically and BLB and WLL lines horizontally. To avoid performance degradation in PUF mode, the front-end layers need to be symmetric, to reduce the mismatch caused by the layout drawing. The area is similar with convention 6T cell, the size of each bit-cell is $1.77 \mu\text{m} \times 0.735 \mu\text{m}$. Fig. 4.3 (c) depicts a sample of 2×2 array operation with complementary initial values in the PUF mode. To generate one-bit data,

Fig. 4.4 Principle of the PUF mode operation.

we need to turn on two vertical and two horizontal word-lines simultaneously, which connects 4 cells two by two with the bit-lines. In this case, the horizontal word-lines WLL [0] and WLL [1], with the vertical word-lines WLR [0] and WLR [1] are enabled, the voltage levels of BL[0], BL[1], BLB[0], and BLB[1] are determined by the data sets of {Q(a), Q(c)}, {Q(b), Q(d)}, {QB(a), QB(b)} and {QB(c), QB(d)}, respectively. If both QB(a) and QB(c) are '0' or '1', as shown in

Fig. 4.4 (a), the BL [0] will be '1' or '0', the data in two cells won't be rewritten. However, if QB(a) and QB(c) have the opposite state, as shown in Fig. 4.4 (b), the voltage level of BL [0] is decided by the strength of pull up and pull down transistors on Q(a) and Q(c). If pull-down transistor is much stronger than pull up one, Q(a) could be overwritten. If the strengths of mismatches on Q(a) and Q(c) are similar, they could fight on the BL [0], then get back to their previous values. So that, the value of Q(a) is decided by both QB(a) (which is influenced by QB(b)) and Q(c). To sum it up, it is a competition of four cells, the weak cells are overwritten by the strong cells. The response is the result of the mutual effects of four cells.

Unlike the conventional SRAM-based PUFs, the proposed PUF does not directly use the start-up values. To improve the randomness, the proposed technique connects any two nodes from four SPUF bit cells by enabling their word-lines. By doing this, the initial strengths of all the 24 transistors from the selected four PUF cells contribute to the final PUF value. In addition, the generated value depends on the rows and columns we chose. Therefore, the challenge-response pairs are significantly expanded because of the substantially increased combinations of selecting two rows and columns.

The reliability of each SPUF cell depends on the mismatch level of two cross-coupled inverters. A larger mismatch produces a stronger cell. Based on the strength of mismatch, we could define a SPUF cell into four levels. As shown in Fig. 4.5, the strongest cell (I) represents larger mismatches while the weakest cell (IV) indicates smaller mismatches. When we enable both WLL and WLR, it

The strong cell in higher level takes the dominant, finally overwrites the weak cell.

Fig. 4.5 Operation principle of the proposed SPUF data generation.

Fig. 4.6 Dependency of SPUF data on selection sequence order.

can create conflicts in the selected bit-lines. As shown in Fig. 4.5, two strong adjacent cells try to overwrite data in the weakest cell with opposite data. In this situation, the difference between I and IV becomes dominant in this operation, which is larger than that between II and IV (subordinate). Therefore, the data in IV will be overwritten by the data in I.

In the PUF mode, each time enabling four word-lines simultaneously could generate one-bit data. Meanwhile, it could also use multiple steps to generate one-bit data. For example, if adopting 2 steps to generate one-bit data, we need to enable different four word-lines two times. Basically, each time we should choose four cells with complementary states which are interrelated to other steps. Moreover, even if we chose the same step twice, the generated PUF value is also affected by the applied permutation. As depicted in Fig.4.6, the top sequence connects the first two rows and the first two columns, followed by last two rows and last two columns, written as the sequence of $\{(1, 2), (2, 3); (1, 2), (2, 3)\}$. Similarly, the bottom sequence can be written as $\{(2, 3), (1, 2); (2, 3), (1, 2)\}$. The different background color bits represent various cell strengths. Note that a different ordered sequence produces a different final value even if we chose the same rows and columns. In the 2nd and the 3rd row of the final state, the hamming distances are equal to 1. Therefore, both the row and column selection, and the permutation order in the selected rows and columns affect the final PUF value. This allows one PUF chip to generate many PUF data maps, which make the chip much more resilient to attacks by various algorithms including machine learning.

4.3 The Enhancement of Challenge-Response Space

4.3.1 CRPs enhancement in two dimension movement

The number of CRPs is limited in weak PUF, it makes the brute-force attack go through all the possibilities infinite time. However, for the larger CRPs, the attack algorithm needs more effort to break the security. If the CRPs could be polynomially increased by the quantities of basic cells, the security could be improved significantly. So far, most of the SPUFs generate reliable chip ID by aging or utilizing the characteristic of the local transistor, which CRP is 1. These types of SPUF easily become the target of a machine learning attack.

In [35] proposed a sequence-dependent SPUF design, which connects two cells to generate new data. It uses multiple steps to generate one data, which is

Fig. 4.7 Previous technique for changing sequence lengths [34].

(a)

(b)

* 3 X 3 array: the number of combinations is 36

*4 X 4 array: the number of combinations is 576

sequence length $r < \text{Min}(m, n)$

Fig. 4.8 Different cell selection combinations in the proposed PUF. (a) with an adjacent four cells example (b) with a non-adjacent four cells example.

number of basic cells, without changing the conventional 6T bit-cell structure. Fig. 4.7 illustrates the combinations of varying sequence lengths in [35] (only one direction is displayed to simplify the analysis). When 3 rows are used for generating a SPUF value, 2 steps are feasible. In the first step, we have 3 choices and in the second step, we have 2 choices, which provides 6 permutations. Similarly, when 5 rows are employed, 4 steps (Fig. 4.7(Right)) are required and we can obtain 120 permutations. Even though the technique in [34] improves CRPs by utilizing the sequence dependency, the employment of single column limits the number of permutations, which is addressed in this work.

As illustrated in Fig. 4.3 (c), the proposed technique can select 2 rows and 2 columns for PUF data generation. For an m row $\times$ 2 columns array with a sequence length of r , the sequence combination is the same as the movement in [35], which is analyzed above, the number of the challenge to response pairs (CRPs) is written as follows:

$$m \times (m - 1) \times \dots \times (m - r) \quad (3)$$

For the two dimensions movement, any two word-lines in both horizontal and vertical directions could be chosen, Fig. 4.8 exhibits sample combinations of cell selection using 3×3 and 4×4 arrays. We could arrange 2 steps and 3 steps for 3×3 and 4×4 arrays respectively. Fig. 4.8(a) shows how the movement goes in two directions for four adjacent cells. Fig. 4.8(b) presents examples of combinations for non-adjacent four cells for two different sizes of arrays, the arrangement of cells in each step doesn't need to be consistent. The overwritten cells in one step could affect the selected cells in the following steps. Eventually,

the generated data in the final step is influenced by all the movements, and it depends on the sequence and the order.

Comparing to [35], the proposed permutation moves in two directions, which makes the number of combinations of formula in (3) squared. For example, when considering a 3×3 array, the CRPs of the proposed technique is 36, which is square of the CRPs in [34]. The CRPs become 576 when using a 4×4 array. The number of CRPs of the proposed scheme can be estimated by the following expression:

$$[m \times (m - 1) \times \dots \times (m - r)] \times [n \times (n - 1) \times \dots \times (n - r)] \quad (4)$$

Where m is the number of rows, n is the number of columns, and r is the sequence length. Note that r is less than m and n .

4.3.2 Security enhancement analysis

For security, unpredictability is one of the most critical points for SPUF design. According to an arbitrary challenge, the generated responses must be unpredictable, which couldn't be reproduced by the same CRPs from another PUF device or another CRPs from the same PUF device. However, the unpredictability is difficult to quantify, the criteria only could be converted to uniqueness to be evaluated. For the proposed SPUF, because of the improved CRPs, multiple responses are according to the same challenge address based on different sequence and order. Each response should be distinguishable even for the same sequence with a different order in one instance, which makes the

chip configurable. This uniqueness is estimated by the hamming distance of response bit strings from the different permutations or devices [69], which is shown as follow:

$$Uniqueness = \frac{2}{m \times (m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^m \frac{HD(R_i, R_j)}{l} \quad (5)$$

Where m is the number of response bit strings, n is the length of the bit string. $HD(R_i, R_j)$ is the hamming distance between the two-bit strings R_i and R_j . The response strings are from different permutations or devices, which indicates the diversity of the key data generated by various sequences and order.

For the proposed SPUF array, according to the different combinations of operation, the SPUF could generate multiple data maps in one device. Since the proposed design expands the CRPs to the square of the previous one as shown in formula (4), it means a number of security maps could be produced from the same chip. The uniqueness is proofed by the hamming distance from each permutation, to guarantee that the key can't be generated from other permutation with the same challenge. It improves the security for SPUF device significantly.

4.4 SRAM Function Analysis Result

SPUF works as a dual-mode device, which builds the PUF function on the existing hardware environment to minimize the area overhead. It requires the SPUF to work in both modes without degrading the performance of SRAM operations. For the proposed SPUF, the structure of bit-cell keeps the placement of 6 transistors in the conventional SRAM, only changes the connection for control signals, the basic cell unit is explained in section III.

Fig. 4.9(a) displays the schematic of the write operation in SRAM mode. It starts asserting one horizontal word-line WLL [0] and one vertical word-line WLR [0] simultaneously, to write data '0' into the selected cell a. The selected BL [0] and BLB [0] are loaded '0' and '1' respectively, the write ability is the same as conventional 6T SRAM for the selected cell. In this case, since the two word-lines in one cell are running orthogonally, both selected row and column are half-selected cells (b and c). For the selected row, since the WLL [0] is set to VDD, and the unselected bit-line BL [1] is pre-charged to high potential. If the data stored in cell b is '0' the disturbing current will flow from BL [1] to Q, and no current go through QB from BLB [1] because WLR [1] is GND. Comparing to the conventional half-selected cell, since two access transistors are turned on, BLB is charging QB while Q is raised by the disturbing current. Therefore, SNM for the proposed SPUF cell is slightly lower than the conventional SRAM cell (as shown in Fig. 4.9(c)), which is negligible for a write operation in exercisable voltage range. If the data stored in half-selected cell b is '1', both pre-charged BL [1] and Q are logic high, no disturbing current. For the selected column cell

(a)

(b)

(c)

Fig. 4.9 Proposed SPUF cells in SRAM mode: (a) Write operation, (b) Read operation. (c) SNM for both read and write operations comparison between proposed SPUF cell and conventional SRAM cell.

Fig. 4.10 Proposed SRAM PUF architecture.

c, the situation is opposite to cell b. if the data stored in cell c is '0', although WLR [0] is asserted, both QB and BLB [1] are logic '1', no current flow between them. If the data stored in cell c is '1', the disturbing current will drop in QB. It's a mirrored case as cell b, the value of SNM is the same as Fig. 4.9(c).

Fig. 4.9(b) shows the schematic of the read operation for the proposed SPUF. Because the bit-lines are not placed parallel, differential sensing amplifier is not eligible in this case. So that, only WLL [0] is asserted during the read operation, to discharge the BL [0]. For the half-selected cell, cell b suffers the read disturbing current from BL [1], the read SNM is the same as writing one as shown in Fig. 4.9(c), which has negligible degradation in whole voltage comparing to the conventional structure.

Fig.4.10 shows the architecture of the proposed 4k SPUF. To support the proposed configurability, the SRAM has a decoder and a write circuit for both rows and columns. Transmission gate flip flop is used to select multiple word-lines at the same time. The single-ended sensing amplifier is required to detect the value of BLs, which places at the bottom of the macro. The proposed SPUF could switch between PUF mode and SRAM mode without degrading the performance, and the area overhead is minimized.

4.5 Measurement Results and Discussion

A 4kb (64 × 64) SRAM PUF array test chip is fabricated in 65nm CMOS technology. The die photo is shown in Fig. 4.11. The total area is $170 \times 74 \mu\text{m}^2$ and the size is $1.3 \mu\text{m}^2$ per bit cell. A logic analyzer is used to control the supply

Fig. 4.11 Die photo of the proposed SPUF.

Fig. 4.12 Measurement environment.

(a)

(b)

Fig. 4.13 Measured BER sweeping: (a) voltage and (b) temperature.

voltage and operation sequence. A dynamic climate chamber is adopted to adjust the temperature. The experiment platform is shown in Fig. 4.12.

For the common SRAM PUF array, as it is a weak PUF, the challenge to response pairs are limited, especially some of them only use the chip ID (CPR is equal to 1). For the proposed one, the data map is determined by the

sequence length and sequence order we choose. According to the same challenge (address), we could generate multiple responses 5(data maps). For the 64 ×64 PUF, when sequence length equal to 5, the CRPs could be achieved 8.37×10^{17} .

4.5.1 Reliability

For the academic community in PUFs designing, it is necessary to consider reliability as a critical requirement. The proposed SPUF should be reliable to reproduce the response by the same challenge many times, which has different tolerance in the different operating environments. Generally, the SPUF should work well in a wide range of voltage and temperature variations.

First, we have tested the reliability of output over the voltage variation, which is measured as the bit error rate (BER). We sweep the voltage from 0.5 V to 1V at a nominal temperature of 27 °C. As shown in Fig. 4.13 (a), the BER achieved 3% under 0.8 V supply voltage. The maximum bit error ratio is 9% at 0.5 V.

Second, we've evaluated the reliability of output against the temperature variation by sweeping the temperature from -10 °C to 80 °C, under the nominal voltage 0.8 V. We've measured the BER value per 10 °C, as shown in Fig. 4.13 (b). The BER is 3% under 20 °C and the maximum bit error ratio in the temperature range is 14%.

At the nominal operation of 0.8 V/20 °C, the golden data demonstrates the BER of 3%. It can be further reduced by applying temporal majority voting or masking techniques.

4.5.2 Uniqueness

For the common SRAM PUF array, as it is a weak PUF, the challenge to response pairs are limited, especially some of them only use the chip ID (CPR is equal to 1). For the proposed one, the data map is determined by the sequence, order, and length we choose. According to the same challenge (address), we could generate multiple responses (data maps). For the 64×64 PUF, when sequence length equals to 5, the CRPs could be achieved 8.37×10^{17} . To prove the responses are distinguishable for the same challenge, the inter sequence and inter-chip hamming distance are shown in Fig. 4.14 and 15. Fig. 4.14 exhibits the hamming distance with the same sequence #1 and different orders. Under the same temperature and voltage conditions (20°C , 0.5V), using the same chip and sequence #1, we only change the order of the sequence #1, to get the hamming distance with a different number of steps (sequence length). The hamming distance is improved by increasing the sequence length. For the two steps, the hamming distance is 25.62%. The value achieves 42.49% when the sequence length is 5. It demonstrates that for the same initialization and sequence length, the same sequence with different orders could provide unique data maps.

Fig. 4.15 presents the hamming distance for different chips with the same sequence #1. We repeated the same sequence and order pattern in different chips with different sequence length under the same environment, the hamming distance between different chips is higher than 45%. The effect of sequence length is not significant, after 5 steps the average value of the hamming distance is 48.69%.

Fig. 4.14 Measured HD for sequence #1 with order #1 ~ order #38.

Fig. 4.15 Measured HD for sequence #1 with different chips.

Fig. 4.16 shows four data maps from one chip under the same environment. (a) shows the start-up pattern of the chip. This data map could be seen as the chip ID, as it depends on the unique PVT variations. The dark color represent data '1' and grey color represent data '0'. Fig. 4.16 (b) shows the data map after the 5 steps PUF operation with sequence #1 order #1. The hamming distance

All four data maps are generated from one chip

(a) Start-up State

(b) Sequence #1 with order #1

(c) Sequence #1 with order #2

(d) Sequence #2

Fig. 4.16 Measured PUF data maps from one chip.

between (a) and (b) is 0.501. then we start from start-up status again, choose the same sequence #1 with different order #2 (sequence length = 5) to generate a new data map, the pattern is shown in Fig. 4.16 (c). The hamming distance between (b) and (c) is 0.427. In Fig.4.16 (d), it presents the start-up data goes through PUF operations with sequence #2 (sequence length = 5) for the same chip. The inter sequence hamming distance between (b) and (d) is 0.485.

Fig. 4.17 Measured hamming distances.

Fig. 4.17 exhibits the hamming distance between different situations, all chips use the 5 steps sequence length. Sequence #1 with different orders is also shown in Fig. 4.14, the value is 42.49%. Sequence #1 ~ sequence #38 is the hamming distance between the different sequence in one chip, which is 47.32. Sequence #1 & start-up is the difference between the start-up value and the data pattern with sequence #1 and varying orders, the hamming distance value is 49.47%. The last one is the hamming distance with the same sequence and order for different chips, which is shown in Fig. 4.15 (48.69%).

4.5.3 Randomness

```
TEST: monobit_test
  Ones count   = 1756
  Zeroes count = 1892
  PASS
  P=0.024340920043991517
TEST: frequency_within_block_test
  n = 3648
  N = 99
  M = 36
  PASS
  P=0.09958245408651507
TEST: runs_test
  prop 0.48135964912280704
  tau  0.033113308926626096
  vobs 1816.0
  PASS
  P=0.8562002435349985
TEST: longest_run_ones_in_a_block_test
  n = 3648
  K = 3
  M = 8
  N = 16
  chi_sq = 3.5853867852867367
  PASS
  P=0.3098554262482392
TEST: dft_test
  N0 = 1732.800000
  N1 = 1728.000000
  PASS
  P=0.4658276207077625
TEST: non_overlapping_template_matching_test
  PASS
  P=0.9958450385860749
TEST: overlapping_template_matching_test
Insufficient data. 3648 bit provided. 1,028,016 bits required
  FAIL
  P=0.0
TEST: maurers_universal_test
Error. Need at least 387840 bits. Got 3648.
```

Fig. 4.18 NIST test result.

Fig. 4.19 Measured hamming distances.

For a conventional PUF, the generated data is decided by the strengths of mismatch between 6 transistors in a one-bit cell. In the testing chip, the distribution of 0 is 62.45% in raw data when the supply voltage ramps up. For the proposed PUF, the generated data is affected by 24 transistors in a single PUF step, the variation is increased significantly. After PUF operation the sequence length equals to 5, the distribution of 0 in this new map in Fig. 4.16 (b) is 51.12%. In different configurations, the worst distribution we got through testing is 44%.

Fig. 4.18 shows the result of the NIST test, the proposed SPUF passed the frequency, frequency in one block, runs, longest runs, DFT and non-overlap test. The failed ones are caused by insufficient data we provided, we only generated 64x57 bits. The NIST test shows that the SPUF has good randomness.

In the PUF operation, we could use different sequence length to generate one

Table 4.1 Comparison with Previous Works

		This work	2017 VLSI [35]	2015 ISSCC [70]	2014 ISSCC [28]	2004 VLSI [71]	2007 ISSCC [72]
Technology		65nm CMOS	28nm FDSOI	40nm CMOS	22nm CMOS	180nm CMOS	130nm CMOS
Area/PUF bits		5.2 μm^2 (4 cells)	388-970 F^2 (2-5 cells)	—	4.66 μm^2	—	1092 F^2
Energy efficiency (fJ/b)		81 (seq-5, 0.5 V)	97 (seq-5, 0.7 V)	17750	13	—	930
Inter-PUF HD		0.487	0.481 - 0.495	0.5007	0.49	—	0.505
Native BER (%)		3 (0.8V, 20 °C)	3.17 (0.7V, 27 °C)	9 (worst VT)	~8	0.7 - 4.82	3.04
# Possible CRPs		8.37×10^{17} (seq-5)	1.17×10^{11} (seq-5)	$\sim 5.5 \times 10^{28}$	250k	1.4×10^{20}	1 (chip ID)
Tested Condition	Voltage (V)	0.5 - 1	0.5 – 0.9	0.7 – 1.2	0.7 – 0.9	±2%	0.9 – 1.2
	Temp (°C)	-10 - 80	0 - 80	-25 - 125	25 - 50	27 - 67	0 - 80
PUF Type		Bi-Stable	Bi-Stable	Ring Osc.	Bi-Stable	Arbiter	Bi-Stable

data. According to the sequence length, the power consumption is increasing linearly. As shown in Fig. 4.19, if sequence length is 2, the power consumes minimum value, which is 73 pJ at 0.5 V. The value under the same voltage climbs to 186 pJ and 263 pJ when the sequence lengths increase to 3 and 4 respectively. The minimal power consumption is 81 fJ per bit at 0.5 V when the sequence length is 5.

Table 4.1 compares the previous works with the proposed SRAM PUF, like the SRAM based PUF. The area to generate one-bit PUF is around 5.84 μm^2 to 14.6 μm^2 (including 4~10 cells). However, the 4 to 10 cells could be reutilized to generate other bits with a different combination. The actual area overhead is

very small. The energy consumption is 331 pJ at 0.5 V while the sequence length is 5 and minimum bit error rate gets 3% at 0.8 V and 20°C, maintaining the same level as previous work [35]. Ring oscillator and arbiter PUF are strong PUF. This work provides the largest CRPs in the weak PUF type due to the configurable sequence-dependent operation.

4.6 Conclusion

In this chapter, we presented a 2D sequence-dependent PUF, which expands the CRPs by the orders of $\text{rows}^{(\text{sequence length} - 1)} \times \text{columns}^{(\text{sequence length} - 1)}$. Based on the conventional SRAM bit cell, we split the word-lines and run them vertically and horizontally respectively, to connect four cells simultaneously to generate one data. According to various sequence length and order, we could generate multiple data maps in one chip. The chip area is $12580 \mu\text{m}^2$ and $1.3 \mu\text{m}^2$ per bit cell in 65 nm CMOS technology. The minimum energy consumption is 81 fJ/bit at 0.5 V, the bit error rate is 3% at the nominal point (0.8 V/20 °C) and inter hamming distance between chips is 0.4869. For the same sequence length and different order in one chip, the hamming distance achieved 0.4249. In the voltage range (0.5V – 1V) and temperature range (-10 °C – 80 °C), this PUF works well. The measurement results demonstrate that the proposed PUF is configurable and reliable for the physical security keys generation.

Chapter 5 A Sequence Dependent

Configurable PUF Based on 6T SRAM for High Reliability

5.1 Background

Physically unclonable function (PUF) serves to generate and store a secret key for a semiconductor device. The key property of PUF is a unique reproducible response, which makes it difficult to predict or characterize uncontrollable variations from manufacturing. Therefore, PUFs are attractive for security applications such as banking infrastructure, critical communication links, etc [64]. Compared with other existing PUF devices, SRAM based PUFs can be implemented without additional components for generating a response. SRAM based PUFs (SPUFs) operate as weak PUFs. They generate data based upon the physical mismatches produced during fabrication.

An ideal on-chip PUF needs to meet three requisites: uniqueness, reliability, and uniformity [65]. A challenge of PUF design is to keep the unique secure interactions between devices and maintaining the stability to regenerate the code at the same time. Typically, error correction code (ECC), temporal majority voting (TMV) and directed aging have been utilized to improve the reliability [28]. However, both ECC and TMV need additional circuits, which increases silicon area. Directed aging is an effective way to reduce the cost of error correction. Here, selected devices are subject to temperature and voltage stress to

accelerate transistor aging. However, the aged devices degrade the SRAM performance and decrease the lifetime. In [33], they use a native device to sense the threshold voltage difference between two access transistors in conventional 6T SRAM cells to generate highly reliable PUF. However, it is challenging to keep an analog voltage and simultaneously compare slight differences that are generated asynchronously. The SPUFs in [34], [35] generate more variations in addition to the mismatches coming from the fabrication process so that the added variations improve the uniqueness. In [34], they adopted a dual-port SRAM structure to aggravate the conflicts produced during the data generation. In [35], they connect two cells to regenerate one-bit data. This involves 12 transistors and increases the randomness compared with 8 transistors in [34]. Furthermore, the PUF in [35] introduces a sequence-dependent response to make the PUF data programmable. The data is not only affected by initial intrinsic mismatches, but also the sequence length and the order. However, the PUF in [35] only changes the sequence of selected rows, which provides only limited improvement.

This work proposes a new sequence-dependent SRAM-based PUF with high reliability. The proposed novel PUF cell structure with full SRAM functions generate multiple reliable PUF data sets and expand the challenge-response space by the order of $(3 \times \text{rows}^{(\text{sequence length} - 1)} \times \text{columns})$.

5.2 Proposed Reliable PUF

5.2.1 Proposed SPUF cell

Conventional SRAM based PUF generate raw data utilizing a conventional 6T SRAM bit-cell structure, it adopts the start-up data during power-up. The data is affected by noise and mismatch between the cross-coupled inverter pair. The likelihood for the raw data becoming '0' or '1' depends on the PVT variations created during the fabrication.

The proposed SPUF can work under both PUF mode and SRAM mode, Fig. 5.1(a) shows the schematic of the proposed 6T SPUF bit-cell. It consists of two cross-coupled inverters with a minimum size of the transistor, and two access transistors with doubled width. Compared to the conventional 6T SRAM cell, the one being proposed splits the word-line WL to left (WLL) and right (WLR), which controls the two access transistors independently. The bit-lines (BL and BLB) are arranged vertically, each pair of BL and BLB connects to a single differential sensing amplifier. Fig. 5.1(b) displays the layout of the proposed bit-cell. Since the start-up data is generated by the mismatches caused by fabrication, the placement in front-end layers should be symmetric. The layout style is the same as the conventional one with double-sized width of access transistors, only splits the WL to two word-lines. The WLL and WLR are implemented with horizontal metal-3 lines, while the BL and BLB are realized in the vertical metal-2 lines. The cell layout occupies $1.31 \times 0.45 \mu\text{m}^2$, designed using the logic rules of 40 nm CMOS technology.

Fig. 5.1 Proposed SPUF cells: (a) unit SPUF cell and (b) layout of the SPUF cell.

Unlike the conventional SPUF which directly uses raw data as security key, the proposed design utilizes PUF operation to generate one-bit data, improving randomness and reliability. As shown in Fig. 5.2, during the PUF operation, both BL and BLB are pre-charged. The selected row marked with * (asterisk) is turned on for both WLL and WLR, and we choose other two rows, assert either WLL or WLR on each of them. Since both access transistors are turned on in the

Fig. 5.2 3×3 array of the proposed SPUF in PUF mode.

selected row *, the cell on the selected row works as the conventional cell, only the values of BL and BLB are affected by the data stored in the other two selected rows. However, for the other two (unmarked) selected rows, only one access transistor is on. Under this situation, it is single ended writing operation, the write ability is weak. Therefore, the sizes of access transistors are doubled to improve the write margin. We consider the cell with larger mismatch as stronger, smaller mismatch on two sides are weaker. In each BL and BLB, connection between these three selected cells creates conflicts, the stronger cell may overwrite the weaker one in each column. While executing the PUF operation, the final data value is contributed by the initial strength of 18 transistors which are from the selected three cells. The randomness is further improved since more transistors are involved to generate security key compared to 6 transistors in conventional SPUF and 12 transistors in [6].

Cell A	Cell B	Cell C	Prop. C
S	S	S	S
S	S	U	S (A=B) U (A≠B)
S	U	S	S (A=B) U (A≠B)
S	U	U	S
U	S	S	S (A=B) U (A≠B)
U	S	U	S
U	U	S	S
U	U	U	S (A=B) U (A≠B)
1/2	1/2	1/2	3/4

* S: Stable Cell
U: Unstable Cell

Fig. 5.3 Reliability enhancement for PUF operation

5.2.2 Reliability enhancement

Reliability is a crucial criterion for PUF design, the response to the same challenge should be reproducible under different environmental circumstances. The bit error rate (BER) represents the percentage of flipped bits during multiple operations at different time for one device, which evaluates the reliability of a PUF. Fig. 5.3 illustrates the method of reliability enhancement by proposed PUF operation. We simply divide cells into stable cells and unstable cells, and assume the initial stable probability of all cells A, B and C are $\frac{1}{2}$, which means that each cell has 50% chance to be stable. If WLL [C] and WLR [C] are asserted to logic

'1', cell C is marked with *. WLL [A] and WLR [B] are set as high, so that Q [A] is connected to Q [C] and QB [B] is connected to QB [C] through BL and BLB respectively. Fig. 5.4 represents the waveforms under different combinations. In the first situation, initially cells A and C are stable, cell B is unstable. If the initial value of cell A is the same as the initial value of cell B, after the PUF operation the generated cell C is following the value of cells A and B stably. If the value of cell A doesn't equal to cell B, as shown in the first chart in Fig. 5.4, even though the original cell C is stable, the generated cell C becomes unstable. So that under the condition A = S, B = U and C = S (S represents stable cell, U represents unstable cell), the stability of proposed cell C is 50%. In the second scenario, all cells A, B and C are stable. No matter whether the values of cell A and cell B are equal or not, the generated cell C is always stable. It means under the condition A = S, B = S and C = S, the stability of proposed cell C is 100%. Altogether we

have 8 different combinations of cells A, B and C as shown in Fig. 5.3. Since (A = S, B = U and C = S / U) is the same as (A = U, B = S and C = S / U), we show

* S: Stable cell
U: Unstable cell

Fig. 5.4 Waveforms of PUF operation under different situations

waveforms under 6 different scenarios in Fig. 5.4. When $A = B$, the proposed cell C is stable, if $A \neq B$, under ($A = S$, $B = U$ and $C = S$), ($A = U$, $B = S$ and $C = S$), ($A = S$, $B = S$ and $C = U$) and ($A = U$, $B = U$ and $C = U$) these scenarios, the proposed cell C is unstable. To sum up, the stability of the proposed cell C is $\frac{3}{4}$ as shown in Fig. 5.3, which is an improvement compared to $\frac{1}{2}$ in the original cell C.

5.2.3 Sequence Dependent SPUF Analysis

For the proposed SPUF during the PUF mode, each time three cells in one column are connected, to generate one-bit data. Meanwhile, we could compose

Fig. 5.5 Dependency of SPUF data on selection sequence order.

multiple steps to generate one-bit data by choosing different three cells in one column simultaneously. For instance, after the first PUF operation, we could choose different three cells combination as the second step. Basically, we should choose three cells in one column with flipped statuses, and at least one cell interrelated to other steps. Based on the different combinations of cells, for the same challenge, the response should be different. Even if we choose the same sequence and combination with the same location of the cell marked with *, the final states should be influenced by the permutation of steps. Moreover, if we choose the same sequence and combination with the same permutation, the location of the cell marked with * could also decide the final PUF data. Fig. 5.5 displays three examples using two steps with different combinations, the darker bits indicate stronger cells, and lighter bits are overwritten by the darker ones. In the first example, the first sequence connects the top three rows (the second row with an asterisk), and the second step links the bottom three rows together (the fourth row with an asterisk), written as a sequence of (1, 2*, 3) (3, 4*, 5). The second example could be written as a sequence of (3, 4*, 5) (1, 2*, 3). So that the final value of the SPUF is affected by the sequence order even if we choose the same rows with the same location of asterisks. In the 2nd and 3rd rows, the hamming distances are 2 and 1 which are marked with orange dotted circles. Furthermore, as shown in Fig. 5.5, the sequences of the second and third examples are (3, 4*, 5) (1, 2*, 3) and (3, 4, 5*) (1*, 2, 3) respectively. Even with the same rows and sequence permutations, the final states are also influenced by the location of the cell with an asterisk. The hamming distances of final data are 1, 1, 2 and 1 in the 1st, 2nd, 3rd and 5th rows respectively.

Therefore, the PUF key is not only decided by the selected rows and the

sequence order of the selection, but also the location of the asterisk. It means the proposed SPUF could generate different security maps from one device, which makes it become more resilient when facing attacks.

5.2.4 CRPs enhancement

Typically, SRAM based PUF has limited CRPs, in the worst case the CRP is only 1, which is the chip ID of the device. Usually it uses as the golden key, but it also easily becomes a target of an attack. To improve the security, CRP extension is addressed in this work.

Supreet et al. presented a sequence-dependent SPUF with the conventional SRAM 6T cell structure [35], which utilizes two cells to generate one-bit data. They repeat the PUF operation multiple times, and the final data is affected by the length of the sequence and the permutations. Fig. 5.6 explains the way of

Fig. 5.6 Previous technique for changing sequence lengths [35].

Fig. 5.7 Changing sequence length in proposed SPUF.

combination for sequence length in [35]. For three rows, they could adopt two steps (left in Fig. 5.6), in the first step we have three options to choose two rows from three, and in the following step, we get two choices. So totally it is 6 permutations for an array with three rows. Similarly, for 5 rows array, the number of permutations is 120 (shown in Fig. 5.6 right). In this case, if sequence length is r , number of CRPs for an array with m rows and n columns is:

$$m \times (m - 1) \times \dots \times (m - r) \times n \quad (1)$$

For the proposed SPUF, each time we connect three rows together to generate one final data, we can also use multiple operations to increase the CRPs space. As shown in Fig.5.7 left, for an array with 4 rows, two steps are feasible. On the first step, we could choose 3 rows from 4 which has 4 options, and we get 3 choices in the following step. Therefore, for two steps we have 12 choices to combine. In each combination, it could choose one row from three to turn on both sides of bit-lines and sign it with an asterisk, then we could connect them

in 3 different way. Totally, the number of CRPs is 36 in four rows and one column array when the sequence length is 2. Using the analogy, in a 6×4 array, the final CRPs in each column is 1080 with 4 steps PUF operations. So, for an array with m rows and n columns, if the sequence of length is r , the number of CRPs could be written as follow:

$$3 \times [m \times (m - 1) \times \dots \times (m - r)] \times n \quad (2)$$

Here $r \leq (m - 2)$.

To sum up, the responses of the proposed SPUF is decided by the initial start-up data, the sequence length, the applied order and location of cells which are with an asterisk. The expanded CRPs space increases randomness and the security of the PUF.

5.2.5 SRAM Architecture and Operation Flow

Fig.5.8 shows the architecture of the proposed SRAM based PUF. To support the proposed configurability, the SRAM has flip flop and decoder to select multiple word-lines. Each column has one differential sensing amplifier to read data. The bit-lines are controlled by the preb and equ signals. In the SRAM mode, write operation starts by asserting WLL and WLR and loading data in the selected BL and BLB. For the read operation, the WLL and WLR are set simultaneously. Therefore, the selected and half-selected cells have the same condition as the conventional one.

To improve reliability, we need to use bit selection algorithm to generate a reliability map to identify the stable cells in one device [73]. As shown in Fig. 5.9,

Fig. 5.9 The operation flow to generate reliable map.

5.3 Simulation and Measurement Result

A 4kb (64 × 64) SRAM PUF array test chip was fabricated in 40nm CMOS technology. The die photo is shown in Fig. 5.10. the chip area is $156 \times 43 \mu\text{m}^2$, the unit cell size is $0.59 \mu\text{m}^2$. A logic analyzer is used to control the sequence and voltage, and climate chamber is used to adjust the temperature environment.

Fig. 5.10 Die photo of the proposed SPUF.

Fig. 5.11(a) shows the simulation result for the reliability of output over the voltage variation with 1000 samples, which is called a bit error rate (BER). We sweep the voltage from 0.6 V to 1V at nominal temperature of 20 °C. Comparing to the value under 1 V, the BER is always below 2% under the whole voltage range. The maximum bit error ratio is 1.8% at 0.6 V.

Fig. 5.11(b) displays the reliability of output against the temperature variation by sweeping the temperature from 0 °C to 80 °C, under the nominal voltage 1 V.

(a)

(b)

Fig. 5.11 Simulated BER of the proposed SPUF with the golden value of 1V / 20°C. (a) the BER with varying voltage, (b) the BER with varying temperature.

(a)

(b)

Fig. 5.12 Measured BER of SPUF with the golden value (1V / 20 °C). (a) the raw data, (b) the PUF data (sequence length = 1).

Comparing to the value when the temperature is 20 °C, the minimum BER is 1.9% under 40 °C and the maximum BER in the temperature range is 6.1% when the temperature is 80 °C.

Fig. 5.12 shows the measurement result for the bit error rate under 20°C and 1V. We numbered the data as four groups, each group has 64 bits. Both raw data and PUF data with one step sequence (sequence length =1) are generated, each process is repeated 128 times. The bit error rate of raw data is shown in Fig. 5.12(a), for group 1, it has 87.5% BERs are less than 3%. And the distributions of 4%~10%, 10%~20% and >20% of BERs are 3.13%, 3.13% and 6.25% respectively. The worst case is in group 2, the BERs is 84.38% less than 3%, and 6.25% in 20%~50% range. The stable cells (BERs is less than 3%) in group 3 and group 4 are 92.12% and 90.63% respectively.

Fig. 5.12 (b) shows the BERs for the PUF data with a one-step operation (sequence length = 1). Compared to the raw data, the percentages of the stable cells (BERs < 3%) are improved at least 3% in each group. In the worst scenario group 2, the stable cells are increased 6%, and the percentage of ultra-unstable cells (BERs > 10%) is reduced to half. Finally, the stable cells in group 1 to group 4 are 90.63%, 90.63%, 95.31%, and 93.75% respectively. The result can be further improved by increasing the sequence length.

For the common SRAM PUF array, as it is a weak PUF, the challenges to response pairs are limited, especially as some of them use only the chip ID (CRP is equal to 1). For the proposed one, the data map is determined by the sequence length and sequence order we choose. For the same challenge

Fig. 5.13 Inter sequence hamming distance

(address), we could generate multiple responses (data maps). For the 64×64 PUF, when sequence length equals to 5, the CRPs could reach 1.76×10^{11} . To prove the responses are distinguishable for the same challenge, the inter sequence hamming distance is shown in Fig. 5.13.

We observed four groups of data, each group is 64 bits (one row). For each group, the same sequence #1 with different orders are applied, we repeated the operation from one step (sequence length = 1) to five steps (sequence length = 5) under the same supply voltage 1V and room temperature. For the one-step, the average hamming distance in 4 groups is 18.75%, the value increased to 28.12% on the two steps operation. After five steps, the hamming distances are from 42.19% to 46.88%. It demonstrates for the same initialization and the same sequence with different orders, one chip can provide unique data pattern. The CRPs are improved by increasing the number of steps for PUF operation.

5.4 Conclusion

In this chapter, we presented a reliable sequence-dependent PUF, which expands the CRPs by the orders of $(3 \times \text{rows}^{(\text{sequence length} - 1)} \times \text{columns})$ and improves the reliability. Based on the conventional SRAM bit cell, we split the word-lines, to connect three cells simultaneously to generate one data. According to various sequence length and order, we could generate multiple data patterns in one chip. The chip area is $156 \times 43 \mu\text{m}^2$ and $0.59 \mu\text{m}^2$ per bit cell in 40 nm CMOS technology. The bit error rate is improved 3% at the nominal point (1 V/20 °C) by a one-step operation. For the same sequence with a different order in one chip, the hamming distance achieved 0.441 when sequence length is 5. The measurement results demonstrate that the proposed PUF is configurable and reliable for the physical security keys generation.

Chapter 6 Conclusion and Future Work

6.1 Conclusion

This research work starts with the motivation to design an SRAM device for energy efficiency applications. It discusses the restrictions under the ultra-low-voltage condition we are currently facing. The thesis presents several assist circuits to reduce the power consumption from the previous works. Moreover, some schematics for SRAM based PUF to enhance the reliability, randomness, uniqueness, and CRPs are reviewed. This work addresses the enhancement of PUF without degradation of SRAM performance.

Leakage and power efficiency are the primary demanding factors for portable and medical devices. This thesis proposes a vertical read word-line to reduce the dynamic energy consumption for the read operation, through only charge (discharge) of one selected column to evade to switch power on unselected columns and improves the read margin at the same time. Virtual ground is adopted to reduce the leakage of the current, which is dominant in the subthreshold voltage system. For the write operation, this work combines split dual power lines and single-ended writing to improve the write ability and stability under the ultra-low voltage. It achieves the smallest normalized energy per bit when compared to other state-of-the-art works. The SRAM could operate functionally down to 0.26 V.

As the number of CRPs is limited in SPUF, it becomes Achilles' heel to

algorithmic attacks. This thesis proposes a scheme to connect four cells together to generate one-bit data. This technique involves 24 transistors compared to 6 transistors in the conventional solution, which improves the randomness of the SPUF. Furthermore, the final secure data depends on the sequence length and order of the combination, it realizes configurable function in the SPUF device, and the CRPs is improved from 1.17×10^{11} in the previous work to 8.37×10^{17} .

Reliability is one of the most crucial criterions in the PUF design. Compared to other methods, like aging, the proposed SPUF could improve the reliability without degradation of the performance of SRAM. This work connects three cells to generate a one-bit key, to improve the randomness. Meanwhile, this technique could also increase the number of CRPs by programming the combination of sequence length and order, and even the location between three selected cells. The bit error rate is improved 3% at the nominal point (1 V/20 °C) by a one-step operation. For the same sequence with a different order in one chip, the hamming distance achieved 0.441 when sequence length is 5. The measurement results demonstrate that the proposed PUF is configurable and reliable for the physical security keys generation.

6.2 Future Work

The aims of this research work are developing energy-efficient SRAM which operates under near-/subthreshold voltage region and implementing PUF function on the existing SRAM circuit. Based on the techniques presented in this dissertation, still, have many opportunities to explore. For the SRAM based PUF design, there is still more theory analysis to be done [73], especially combining the PUF with an algorithm to further improve the reliability. Currently, we use a hardware device to defend the machine learning attacks [74], and could try to utilize machine learning training to combine with the hardware design [75].

In additional, computing in memory is also in our future work plan. As the reported, $4 + 2T$ SRAM provides searching and logic functions on the existing memory device, which uses N-well to write data in, and adopts read path to process computing [76]. Our future will involve algorithm to assort with the memory design [77], to realize classify function.

List of Author's Publications

- [1] **Lu. Lu**, T. Yoo, V. L. Le, and T. Kim, "A 0.506pJ 16kb 8T SRAM with Vertical Read Word-Lines and Selective Dual Split Power Lines," accepted by *IEEE Trans. Very Large Scale Integr. (VLSI) systems*, 2019.
- [2] **Lu. Lu**, T. Yoo, V. L. Le, and T. Kim, "An Ultra-low Power 8T SRAM with Vertical Read Word Line and Data-Aware Write Assist," in Proc. *IEEE Asian Solid-State Circuits Conference (A-SSCC)*, pp. 143-144, Taiwan, Nov. 2018.
- [3] **Lu Lu**, Tony Tae-Hyoung Kim, "A Sequence Dependent 6T SRAM Based Configurable PUF in 65 nm CMOS", *IEEE International Symposium on Circuits and Systems (ISCAS)*, Sapporo, 2019.
- [4] **Lu Lu**, Ju Eon Kim, Tony Tae-Hyoung Kim, "Low Power ReRAM Device and Circuit Co-Design in Nano-scale CMOS Technology", submitted to *IEEE International Symposium on Circuits and Systems (ISCAS)*, 2020.

Bibliography

- [1] M. Bohr, "The new era of scaling in a SoC world," *IEEE int. Solid-State Circuits conf. Dig. Tech. Papers*, pp. 23-28, 2009.
- [2] S. Damaraju, et al., "A 22nm IA multi-CPU and GPU System-on-Chip, " *IEEE int. Solid-State Circuits conf. Dig. Tech. Papers*, pp. 56-57, 2012.
- [3] R. Islam, et al., "Power Reduction Schemes in Next Generation Intel ATOM Processor based SoC for Handheld Applications," *IEEE Symposium on Very Large Scale Integr. (VLSI) Circuits*, pp. 173-174, 2010.
- [4] H. Lakdawala, et al., "32nm x86 OS-Compliant PC On-Chip with Dual-Core Atom Processor and RF WiFi Transceiver," *IEEE int. Solid-State Circuits conf. Dig. Tech. Papers*, pp. 62-64, 2012.
- [5] The Most Expensive SRAM in the World - 2.0, online: <http://www.monolithic3d.com/blog/the-most-expensive-sram-in-the-world-20>
- [6] Q. Li, et al., "A 5.61 pJ, 16 kb 9T SRAM with Single-ended Equalized Bitlines and Fast Local Write-back for Cell Stability Improvement," *Proc. of the European Solid-State Device Research Conference*, pp. 201-204, 2012.
- [7] T. W. Oh, et al., "Power-Gated 9T SRAM Cell for Low-Energy Operation", *IEEE Trans. Very Large Scale Integr. (VLSI) systems*, vol. 25, no. 3, pp. 1183–1187, Nov, 2017.
- [8] L. Atias, et al., "A Low-Voltage Radiation-Hardened 13T SRAM Bitcell for Ultralow Power Space Applications", *IEEE Trans. Very Large Scale Integr. (VLSI) systems*, vol. 24, no. 8, pp. 2622–2633, Aug, 2016.
- [9] Q. Li, et al., "A 5.61 pJ, 16 kb 9T SRAM with Single-ended Equalized Bitlines and Fast Local Write-back for Cell Stability Improvement," *Proc. of the European Solid-State Device Research Conference*, pp. 201-204, 2012.
- [10] W. Choi, et al., "A Refresh-Less eDRAM Macro with Embedded Voltage Reference and Selective Read for an Area and Power Efficient Viterbi Decoder," *IEEE J. Solid-State Circuits*, pp. 2451-2462. July, 2015,
- [11] M. H. Tu, et al., "A Single-Ended Disturb-Free 9T Subthreshold SRAM With Cross-Point Data-Aware Write Word-Line Structure, Negative Bit-Line, and Adaptive Read Operation Timing Tracing", *IEEE J. Solid-State Circuits*, vol. 47, no. 6, pp. 1469–1482, Apr., 2012.

- [12] L. Chang, et al., "Stable SRAM Cell Design for the 32 nm Node and Beyond" in *IEEE Symposium on Very Large Scale Integr. (VLSI) Systems*, pp. 128-129, 2005.
- [13] Y. H. Chen, et al., "A 16 nm 128 Mb SRAM in High-κMetal-Gate FinFET Technology with Write-Assist Circuitry for Low-VMIN Applications", *IEEE J. Solid-State Circuits*, vol. 50, no. 1, pp. 170–177, Sept. 2015.
- [14] O. Hirabayashi, et al., "A process-variation-tolerant dual-power-supply SRAM with 0.179μm cell in 40nm CMOS using level- programmable wordline driver", *IEEE int. Solid-State Circuits conf. Dig. Tech. papers*, pp. 458–459, 2009.
- [15] Y. Sinangil ; et al.," A 128 Kbit SRAM With an Embedded Energy Monitoring Circuit and Sense-Amplifier Offset Compensation Using Body Biasing", *IEEE J. Solid-State Circuits*, vol. 49, no. 11, pp. 2730–2739, Sept. 2014.
- [16] B. Zhai, et al., "A sub-200mV 6T SRAM in 0.13μm CMOS", *IEEE int. Solid-State Circuits conf. Dig. Tech. papers*, pp. 332–606, 2007.
- [17] T. H. Kim; et al., " A 0.2 V, 480 kb Subthreshold SRAM With 1 k Cells Per Bitline for Ultra-Low-Voltage Computing", *IEEE J. Solid-State Circuits*, vol. 43, no. 2, pp. 518–529, Jan. 2008.
- [18] M. F. Chang, et al., "A Compact-Area Low-VDDmin 6T SRAM With Improvement in Cell Stability, Read Speed, and Write Margin Using a Dual-Split-Control-Assist Scheme", *IEEE J. Solid-State Circuits*, pp. 2498–2514, May, 2017.
- [19] T. Song, et al., "A 14 nm FinFET 128 Mb SRAM With VMINEnhancement Techniques for Low-Power Applications", *IEEE J. Solid-State Circuits*, vol. 50, no. 15, pp. 158–169, Oct. 2015.
- [20] M. Nabavi, et al.," A 290-mV, 3.34-MHz, 6T SRAM With pMOS Access Transistors and Boosted Wordline in 65-nm CMOS Technology," *IEEE J. Solid-State Circuits*, vol. 53, no. 2, pp. 656–667, Feb. 2018.
- [21] Y. Yang, et al., "Single-Ended 9T SRAM Cell for Near-Threshold Voltage Operation With Enhanced Read Performance in 22-nm FinFET Technology", *IEEE Trans. Very Large Scale Integr. (VLSI) systems*, vol. 23, no. 11, pp. 2748–2752, Nov. 2015.
- [22] W. Choi, et al., " A Refresh-Less eDRAM Macro with Embedded Voltage Reference and Selective Read for an Area and Power Efficient Viterbi Decoder" *IEEE J. Solid-State Circuits*, pp. 2451-2462, July, 2015.
- [23] M. D. Yu; et al., "Performance metrics and empirical results of a PUF cryptographic key generation ASIC". *IEEE International Symposium on Hardware-Oriented Security and Trust (HOST)*, pp. 108–115, 2012.

- [24] L. Zhang ; et al., "Exploiting Process Variations and Programming Sensitivity of Phase Change Memory for Reconfigurable Physical Unclonable Functions", *IEEE Trans inform. Forensics and Security*, vol. 9, no. 6, pp. 921–932, Apr. 2014.
- [25] L. Zhang ; et al., " PCKGen: A Phase Change Memory based cryptographic key generator", in *Proc. IEEE int. Symp. Circuit and Syst.*, pp. 1444–1447, 2013.
- [26] M. Cortez, et al., " Intelligent Voltage Ramp-Up Time Adaptation for Temperature Noise Reduction on Memory-Based PUF Systems", *IEEE Trans. On Comp. design of Integrated circuits and syst.*, VOL. 34, NO. 7, pp. 1162–1175, Apr. 2015
- [27] V. v. der Leest, et al., "On Improving Reliability of SRAM-Based Physically Unclonable Functions". *IEEE, Journal of Low Power Electronic and Applications*. pp. 1–15, 2017.
- [28] S. K. Mathew, et al., " A 0.19pJ/b PVT-Variation-Tolerant Hybrid Physically Unclonable Function Circuit for 100% Stable Secure Key Generation in 22nm CMOS". *IEEE int. Solid-State Circuits conf.*, pp. 278–280, 2014.
- [29] N. Sklavos, et al., " Hardware Security and Trust: Design and Deployment of Integrated Circuits in a Threatened Environment." *Book*, 2017.
- [30] J. B. Velamala; et al., "Compact modeling of statistical BTI under trapping/detrapping," *IEEE Trans. Electron Devices*, vol. 60, no. 11, pp. 3645–3654, Sept. 2013.
- [31] X. Guo, W. Burleson, M. Stan, "Modeling and experimental demonstration of accelerated self-healing techniques," in *Proc. 51st Annu. Design Autom. Conf.*, p. 171-1–171-6, 2014.
- [32] Md N. Islam, et al., " A Guide to Graceful Aging: How Not to Overindulge in Post-Silicon Burn-in for Enhancing Reliability of Weak PUF", *IEEE International Symposium on Circuits and Systems (ISCAS)*, p. 1–4, 2017.
- [33] J. Y. Li ; et al., "A technique to transform 6T-SRAM arrays into robust analog PUF with minimal overhead," *IEEE Intl. Symp. on Circuits and Systems*, pp. 1–4, 2017.
- [34] C. Q. Liu, et al., " A new write-contention based dual-port SRAM PUF with multiple response bits per cell," *IEEE Intl. Symp. on Circuits and Systems*, pp. 1–4, 2017.
- [35] S. Jeloka, et al., " A sequence dependent challenge-response PUF using 28nm SRAM 6T bit cell," *IEEE, Symp. on Very Large Scale Integr. (VLSI) Circuits*, pp.1 C270-271, 2017.
- [36] E. J. Nowak, et al., "Turning Silicon on its Edge, Overcoming Silicon Scaling Barriers with Double-gate and FinFET Technology," *IEEE Circuit Device Mag.*, vol. 20, no. 1, pp. 20–31, Jan. 2004.

- [37] C. T. Chuang ; et al., "High-performance SRAM in nanoscale CMOS: Design challenges an techniques," in *IEEE Int. Workshop on Memory Technology, Design and Testing*, Dec. 3-5, 2007, pp. 4–12.
- [38] H. Jeong et al., "Offset-Compensated Cross-Coupled PEET Bit-Line Conditioning and Selective Negative Bit-Line Write Assist for High-Density Low-Power SRAM," *IEEE Transactions on Circuits and Systems-I*, vol. 62, no. 4, pp. 1062–1070, April 2015.
- [39] W. Choi ; et al., "A Refresh-Less eDRAM Macro With Embedded Voltage reference and Selective Read for an Area and Power Efficient Viterbi Decoder," *IEEE J. Solid-State Circuits*, pp. 2451-2462.Oct. 2015.
- [40] I. J. Chang ; et al., "A 32 kb 10T sub-threshold SRAM array with bit-interleaving and differential read scheme in 90 nm CMOS," *IEEE J. Solid-State Circuits*, vol. 44, no. 2, pp. 650–658, Feb. 2009
- [41] B. Wang, et al., "Design of an Ultra-low Voltage 9T SRAM with Equalized Bitline Leakage and CAM-assisted Energy Efficiency Improvement," *IEEE Transactions on Circuits and Systems-I*, Vol. 62, pp. 441-448, Nov. 2015.
- [42] M. H. Tu ; et al., "A Single-Ended Disturb-Free 9T Subthreshold SRAM with Cross-Point Data-Aware Write Word-Line Structure, Negative Bit-Line, and Adaptive Read Operation Timing Tracing," *IEEE J. Solid-State Circuits*, pp. 1469-1482. June. 2012.
- [43] L. Wen ; et al., "Bit-Interleaving-Enabled 8T SRAM With Shared Data-Aware Write and Reference-Based Sense Amplifier," *IEEE Transaction on Circuit and Systems-II*, pp. 643-647. July. 2016.
- [44] Y. Wang, et al., "A 1.1 GHz 12 uA/Mb-Leakage SRAM Design in 65 nm Ultra-Low-Power CMOS Technology With Integrated Leakage Reduction for Mobile Applications", *IEEE J. Solid-State Circuits*, pp. 172-179, Jan. 2008.
- [45] S. L. Wu, et al., "A 0.5-V 28-nm 256-kb Mini-Array Based 6T SRAM With Vtrip-Tracking Write-Assist", *IEEE J. Solid-State Circuits*, pp. 1791–1802, Mar. 2017.
- [46] T. Song, et al., "A 14nm FinFET 128 Mb SRAM With VMIN Enhancement Techniques for Low-Power Applications," *IEEE J.Solid-State Circuits*, vol. 50, no.1, pp. 158–169, Jan. 2015.
- [47] E. Karl, et al., "A 0.6 V, 1.5 GHz 84 Mb SRAM in 14 nm FinFET CMOS Technology with Capacitive Charge-Sharing Write Assist Circuitry", *IEEE J. Solid-State Circuits*, vol. 51, no. 1, pp. 222–229, Oct. 2016.

- [48] Y. H. Chen, et al., "A 16nm 128 Mb SRAM in High-k Metal-Gate FinFET Technology With Write-Assist Circuitry for Low-VMIN Applications," *IEEE J. Solid State Circuits*, vol. 50, no. 1, pp. 170–177, Jan. 2015.
- [49] A. Grover, et al., "A 32 kb 0.35–1.2 V, 50 MHz–2.5 GHz Bit-Interleaved SRAM With 8 T SRAM Cell and Data Dependent Write Assist in 28-nm UTBB-FDSOI CMOS," *IEEE J. Solid-State Circuits*, vol. 64, no. 9, pp. 2438–2447, Sep. 2017.
- [50] G. Pasandi; S. M. Fakhraie, "A 256-kb 9T Near-Threshold SRAM With 1k Cells per Bitline and Enhanced Write and Read Operations", *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 23, no. 11, pp. 319–332, Nov. 2015.
- [51] S. Lutkemeier ; et al., "A 65 nm 32 b subthreshold processor with 9T multi-Vt SRAM and adaptive supply voltage control," *IEEE J. Solid-State Circuits*, vol. 48, no. 1, pp. 2438–2446, Jan. 2013.
- [52] T. Song, et al., "A 10 nm FinFET 128 Mb SRAM With Assist Adjustment System for Power, Performance, and Area Optimization," *IEEE J. Solid-State Circuits*, vol. 52, no. 1, pp. 240–249, Jan. 2017.
- [53] L. Lu; et al., "An Ultra-low Power 8T SRAM with Vertical Read Word Line and Data Aware Write Assist", in *Proc. Asian Solid-State Circuits Conf. (A-SSCC)*, pp. 143–144. 2018.
- [54] L. Chang, et al., "CMOS circuit performance enhancement by surface orientation optimization," *IEEE Transactions on Electron Devices*. Volume: 51, Issue: 10, Page s: 1621 – 1627, Sept. 2004.
- [55] F. Hamzaoglu ; et al., "A 3.8 GHz 153 Mb SRAM design with dynamic stability enhancement and leakage reduction in 45 nm high k metal gate CMOS technology," *IEEE J. Solid State Circuit*, vol.44, pp. 148–154, Jan. 2009.
- [56] I. Fukushi, et al., "A system level memory power optimization technique using multiple supply and threshold voltages", *IEEE Symposium on VLSI Circuits. Digest of Technical Papers* (Cat. No.98CH36215), pp:142 – 145. 1998.
- [57] R. Boumchedda, et al., "Energy-Efficient 4T SRAM Bitcell with 2T Read-Port for Ultra-Low-Voltage Operations in 28 nm 3D Monolithic CoolCube™ Technology," *IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH)*, page:1-7. 2018.
- [58] Y. W. Chiu, et al., "8T Single-ended sub-threshold SRAM with cross-point data-aware write operation," *IEEE/ACM International Symposium on Low Power Electronics and Design*, page: 169-174. 2011.

- [59] M. Nabavi, et al., "A 290-mV, 3.34-MHz, 6T SRAM With pMOS Access Transistors and Boosted Wordline in 65-nm CMOS Technology," *IEEE J. Solid-State Circuits*, vol. 53, no. 2, pp. 656–667, Feb. 2018.
- [60] C. H. Chang; et al., "A Retrospective and A Look Forward: Fifteen Years of Physical Unclonable Function Advancement," *IEEE Circuit and System Magazine*, pp: 32–62, 2017.
- [61] S. P. Skorobogatov. (2011, Aug.). Illegitimi non Carborundum Santa Barbara, CA. Invited keynote talk given at CRYPTO. [online] Available:<http://people.csail.mit.edu/rivest/2011-08-15-rivest-CRYPTO-keynote.pdf>
- [62] B. Gassend, et al., "Silicon Physical Random Functions," in *Proc. 9th ACM Conf. Computer and Commun. Security* Washington, DC, pp: 148–160. 2002.
- [63] R. Rappu, "Physical One-Way Functions," *Science*, vol. 297, no. 5589, pp: 2026–2030, Sept. 2002.
- [64] L. Kusters ; et al., "Security of Helper Data Schemes for SRAM-PUF in Multiple Enrollment Scenarios," *IEEE Intl. Symp. on Information Theory.*, pp. 1803–1807, 2017.
- [65] A. Maiti, et al., "A Systematic Method to Evaluate and Compare the Performance of Physical Unclonable Functions," in *IACR Cryptology ePrint Archive*. 2011.
- [66] P. Koeberl, et al., "Evaluation of a PUF Device Authentication Scheme on a Discrete 0.13 μ m SRAM," in *Proc. Int. Conf. Trusted Syst.* Berlin, Germany: Springer, 2011, pp. 271–288.
- [67] L. Zhang ; et al., "Statistical Analysis and Design of 6T SRAM Cell For Physical Unclonable Function with Dual Application Modes," *IEEE Intl. Symp. on Circuits and Systems*, pp. 1410–1413, 2015.
- [68] M. Cortez; et al., "Modeling SRAM start-up behavior for physical unclonable functions," in *Proc. IEEE Int. Symp. Defect Fault Tolerance VLSI Nanotechnol. Syst. (DFT)*, Oct. 2012, pp. 1–6.
- [69] A. Maiti ; et al., "A Large Scale Characterization of RO-PUF," *IEEE Intl. Symp. Hardware-Oriented Security and Trust*, Anaheim, CA, pp. 94–99, June 2017.
- [70] K. Y. Yang; et al., "A Physically Unclonable Function with BER $<10^8$ for Robust Chip Authentication Using Oscillator Collapse in 40nm CMOS," *IEEE int. Solid-State Circuits. Dig. Tech.*, pp. 1–3, 2015.
- [71] J. W. Lee ; et al., "A Technique to Build a Secret Key in Integrated Circuits for Identification and Authentication Applications," *IEEE, Symp. on Very Large Scale Integr. (VLSI) Circuits*, pp. 176–179, 2004.

- [72] Y. Su; et al., "A 1.6pJ/bit 96% Stable Chip-ID Generating Circuit using Process Variations", *IEEE int. Solid-State Circuits conf. Dig. Tech.* pp. 406–407, 2007.
- [73] K. Xiao, et al., "Bit Selection Algorithm Suitable for High-Volume Production of SRAM-PUF." *IEEE International Symposium on Hardware-Oriented Security and Trust (HOST)*, pp. 101–106, 2014.
- [74] X. D. Xi ; et al., " Strong subthreshold current array PUF with 265challenge-response pairs resilient to machine learning attacks in 130nm CMOS", *IEEE Symposium on Very Large Scale Integr. (VLSI) Circuits*, pp. C268–C269, 2017.
- [75] Y. J. Wen ; Y. Lao, " Enhancing PUF reliability by machine learning", *IEEE Intl. Symp. on Circuits and Systems*, pp. 1–4, 2017.
- [76] Q. Dong ; et al., " A 4 + 2T SRAM for Searching and In-Memory Computing With 0.3-V". *IEEE J. Solid-State Circuits*, vol. 53, no. 4, pp.1006-1015, Dec. 2018.
- [77] Y. H. Chen, et al., " Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks", *IEEE J. Solid-State Circuits*, vol. 52, no. 1, pp.127-138, Nov. 2017.