Novel Non-Volatile Memory Devices and Applications

Tsegereda Esatu

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2023-97
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2023/EECS-2023-97.html

May 11, 2023

Copyright © 2023, by the author(s).
All rights reserved.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission.

Novel Non-Volatile Memory Devices and Applications

By

Tsegereda Esatu

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering - Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Tsu-Jae King Liu, Chair
Professor Ming C. Wu
Professor Junqiao Wu

Spring 2023

Novel Non-Volatile Memory Devices and Applications

Copyright ©2023

by

Tsegereda Esatu

Abstract

Novel Non-Volatile Memory Devices and Applications

By

Tsegereda K. Esatu

Doctor of Philosophy in Engineering Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Tsu-Jae King Liu, Chair

The performance of semiconductor logic and memory devices has improved significantly with advancements in complementary metal-oxide-semiconductor (CMOS) manufacturing technology to miniaturize transistors, resulting in increased integrated circuit ("chip") processing speed, energy efficiency, and cost per function. The proliferation of the Internet of Things (IoT) and the generation and processing of large data sets, commonly known as "big data," have increased demand for non-volatile (NV) information storage that can be embedded with energy-efficient logic switches so that certain computational tasks can be performed in the memory itself. To achieve ultra-low-power electronics, alternative switching devices that can be operated with much smaller voltages than CMOS transistors are required. Micro-electro-mechanical (MEM) switches have been shown to be promising for ultra-low-power digital computing applications due to their negligible I_{OFF} and abrupt switching behavior across a wide range of operating temperatures. Therefore, they have attracted growing interest for embedded energy-efficient logic and NVM applications. In addition to energy efficiency, it is also crucial to maintain the security and confidentiality of information collected and shared by IoT devices. To address this issue, memory devices that can store data and function as hardware security key generator are desirable. Emerging Resistive Random Access Memory (ReRAM) is a promising technology for hardware security applications due to its inherent variability.

This dissertation focuses on novel non-volatile memory devices and their applications. First, logic MEM switches are demonstrated to be operable as NV memory devices using controlled welding and unwelding of the contacting electrodes. Reprogrammability with consistently low programmed state resistance, and excellent (essentially infinite) retention time at elevated temperature, are experimentally demonstrated. These results indicate that MEM switches are promising for low-cost implementation of ultra-low-power integrated systems.

Next, this dissertation presents a prototype MEM switch design incorporating a floating gate (FG) for non-volatile charge storage. The FG-MEM NV switch potentially can achieve much longer data retention time than a conventional floating-gate MOS transistor, since there exists an air gap between conductive electrodes when the switch is in the OFF-state. Initial experimental results and FG-MEM NV switch design improvements are discussed.

Finally, this dissertation proposes a new method of generating Physically Unclonable Function (PUF) encryption keys that leverage the inherent random variability of ReRAM device

programming time, for hardware security applications. The design and operation of a ReRAM device is presented, followed by a detailed discussion of the PUF key generation scheme. The randomness and reliability of the generated keys are then evaluated. The randomness of the ReRAM-based PUF is found to be of high quality compared to previous PUF implementations. Therefore, ReRAM technology enables the incorporation of both NVM and PUF functions within the same chip.

To my family,

for their endless love and support

Contents

Contents ii

List of Figures iv

List of Tables ix

1 Introduction 1

1.1 Brief History and Evolution of Integrated Circuits 1
1.2 Memory Requirements for IoT 3
1.3 Energy-Efficient Computing for IoT 5
1.4 Ensuring IoT Security 8
1.5 Dissertation Objectives and Overview 9
1.6 References 10

2 A Reprogrammable MEM Switch Utilizing Controlled Contact Welding 13

2.1 Introduction 13
2.2 MEM Switch Operation 14
2.3 MEM Switch Design and Fabrication Process 16
2.4 Non-Volatile MEM Switch Operation 19
2.5 Demonstration of NV-MEM Switch Reprogrammability 28
2.6 Simulation Study 30
2.7 Discussion 32
2.8 Summary 34
2.9 References 34

3 Investigation of Floating-Gate MEM Switch for Non-Volatile Memory Application 36

3.1 Introduction 36
3.2 Proposed FG-MEM NV Switch Design 37
3.3 Experimental Investigation of FG-MEM NV Switches 40
3.4 FG-MEM NV Switch Design Improvements 45
3.5 Summary 45
3.6 References 45

4 Highly Reliable and Secure PUF Using Resistive Memory Integrated into a 28nm CMOS Process 48

4.1 Introduction 48

4.2 ReRAM Structure and Operation 50
4.3 Inherent Stochastic Behavior of ReRAM for PUF 55
4.4 PUF Generation Scheme Using Switching Time Variation 58
4.5 PUF Randomness Evaluation 63
4.6 PUF Performance Benchmarking 66
4.7 Conclusion 68
4.8 References 68

5 Conclusion **72**

5.1 Contributions of This Work 72
5.2 Suggestions for Future Work 73
5.3 References 74

List of Figures

1.1 Evolution of chip component density scaling trends. Transistor density is defined as the number of transistors on a chip divided by the chip area. SRAM bit density and DRAM bit density have advanced at the same pace as the transistor density (reproduced from [8]). 2

1.2 Exponential growth of (a) connected devices and (b) data creation (adapted from [11] 3

1.3 Memory hierarchy diagram [19]. 4

1.4 Conceptual illustrations of (a) the switching I-V characteristics of a high-Vth n-channel MOSFET, a low-Vth n-channel MOSFET, and an ideal switch; (b) dynamic, static, and total energy consumed per operation of a CMOS digital logic circuit. The lower limit for CMOS energy efficiency exists due to MOSFET OFF-state leakage (reproduced from [26]). 6

1.5 Cross-section illustrations of a 4-terminal MEM switch in a) OFF–state, and b) ON–state. 7

1.6 A typical I_{DS}-*vs.*-V_G characteristic of a MEM switch for bidirectional voltage sweep, showing abrupt switching behavior. 8

1.7 Simple cross-section schematic illustrating the metal-insulator-metal structure of a ReRAM device. 9

2.1 Schematic diagram and mechanical spring model of a four-terminal MEM switch designed for digital logic applications. Electrostatic actuation and spring restoring forces are illustrated. 14

2.2 Plan-view scanning electron micrograph (SEM) image of a fabricated MEM switch. 17

2.3 Schematic cross-sectional views along A-A' cutline in Fig. 2.2: (a) OFF–state (b) ON–state. 17

2.4 Schematic device cross-sections along B-B', C-C', and D-D' (cf. Figure 2.2) illustrating key MEM switch fabrication steps. 19

2.5 Pulsed cold switching oxide breakdown procedure: (a) circuit schematic and (b) voltage timing waveforms (not to scale). Adapted from [13]. 20

2.6 MEM switch program operation: (a) circuit schematic and (b) voltage timing waveforms (not to scale). 22

2.7 Contact welding in a MEM switch, induced by Joule heating: (a) Due to surface roughness, physical contact and current flow occur only at one or more asperities. (b) During "Program" operation, Joule heating softens the contact, resulting in contacting asperity growth and increased area of physical contact, effectively welding the electrodes together. 23

2.8 Illustration of erase operation (a) circuit schematic and (b) voltage timing waveform. 24

2.9 Contact un-welding in a programmed MEM switch: During "Erase" operation, Joule heating weakens the bonding strength so that the spring restoring force causes the contact to be broken. 24

2.10 Circuit schematic diagram of read operation. 25

2.11 Measured I_{DS}-V_{GB} curves for a MEM switch before programming (blue) and after one P/E cycle (red). 26

2.12 AFM analyses of MEM switch source electrode topography: Scans (1.6µm x 1.6µm) of electrode surfaces for (a) an unprogrammed contact, and (b) a P/E-cycled contact; height distributions within the contact dimple region for (c) the unprogrammed contact and (d) the P/E-cycled contact. (e) Height *vs*. distance for an asperity, showing that it is ~13 nm tall. 27

2.13 SEM images of MEM switch source electrodes: (a) for a fresh switch and (b) for a P/E-cycled switch. More prominent features (asperities) can be seen in the contact dimple for the cycled switch. 27

2.14 Operating windows for MEM switch (a) program operation and (b) erase operation. The shaded regions indicate the combinations of V_{DS} pulse duration and voltage for successful re-program/erase operation. Damaged switches are not functional afterwards, *i.e.*, they cease to conduct any current. 28

2.15 Evolution of MEM switch I_{DS}-V_{GB} characteristic through multiple program/erase cycles. The change in VON from cycle to cycle is non-deterministic because it depends on the shape and height of the contacting asperities. 29

2.16 Evolution of the measured resistance of a MEM switch in the programmed state ($R_{program}$) and in the erased state (R_{erase}), over multiple P/E cycles. $R_{program}$ is consistently below 1kΩ. R_{erase} essentially indicates an open circuit, as the measurement limit due to noise is ~10^9 Ω. 29

2.17 Data retention testing of multiple MEM switches (1 erased, 6 programmed) at high temperature (200˚C) in vacuum (~1 µTorr). The current noise floor is higher at elevated temperature, resulting in smaller apparent R_{erase}. 30

2.18 COMSOL simulation of Joule heating in a MEM switch during erase operation (cf. Table 2.1). 31

2.19 Measured EDX spectrum of source electrode showing the presence of W and O at the surface. 32

2.20 Evolution of peak contact temperature (cf. Fig. 2.18) during erase operation. 32

2.21 Top view SEM image of a MEM switch that failed after 6 program/erase cycles. 33

3.1 (a) Schematic isometric view of the floating-gate MEM switch studied in this work and cross-sectional views along the ‘B’ cutline in (b) OFF-state (c) ON-state. 37

3.2 Schematic illustrating capacitances within a FG-MEM switch in the ON-state. 39

3.3 Schematic cross-sectional views along A, B and C cut-lines in Figure 3.1(a) illustrating the fabrication process steps for the FG-MEM switches studied in this work. 41

3.4 Measured I-V characteristics of a FG-MEM switch showing symmetric values of VON. (Negative VGB sweep is shown in Black while positive VGB sweep is shown in Red.) 42

3.5 Measured I-V characteristics of a FG-MEM switch before (red) and after (blue) programming for negative and positive VGB sweeps. VON is no longer symmetrical due to injected charges and is shifted by +0.79 V to the right, after programming. 43

3.6 Change in FG-MEM switch turn-ON voltage with (a) program and (b) erase time, for different program/erase voltages. 43

3.7 Zoomed-in cross-sectional schematic of the initial FG-MEM NV-switch design (cf. Figure 3.3(g)) showing that the floating gate (FG) contacts the gate electrode in the anchor region, allowing charge on the FG to leak away. 44

3.8 Data retention characteristics of three FG-MEM devices at 200˚C. Most of the charge stored on the FG in each of the three devices leaked away within ~1 minute. 44

4.1 Typical PUF based authentication scheme for IoTs. 49

4.2 (a) Cross-sectional transmission electron micrograph of a fabricated 1 transistor (1T) 1 ReRAM (1R) cell (b) higher-magnification view showing ReRAM layer stack: inert bottom electrode (W), switching layer (AlOx), and Al-based top electrode (AlNx). 51

4.3 Schematic illustration of the switching mechanisms of a Conductive Bridge ReRAM device: (a) program operation and (b) erase operation. 52

4.4 (a) I-V measurement circuit schematic. (b) Typical measured I-V curves of a ReRAM device. ‘Cycle 1’ is the 1st cycle from pristine device (‘form’ and ‘erase’) and ‘Cycle 2’ shows ‘program’ and ‘erase’ operation. Voltage sweep directions are indicated by arrows. 52

4.5 Distributions of measured ReRAM cell current after verified program/erase cycles. Excellent endurance (more than 10,000 cycles) is seen for the two distinct states (ON and OFF). 53

4.6 ReRAM retention performance. Clearly differentiated ‘ON’ (‘1’) and ‘OFF’ (‘0’) states are retained after baking at 225°C for 1 hour. The projected retention time at 85°C is greater than 98 years. 54

4.7 ReRAM forming current distributions before and after baking at 265°C for 1 hour, confirming solder reflow process compatibility. 55

4.8 Characterization of DC forming voltage randomness of 74 ReRAM cells........................ 56

4.9 (a) Cycle-to-cycle program voltage distribution for 18 ReRAM devices. Each device is cycled 100 times. The edges of the box indicate the 25th and 75th percentiles and the median value is indicated inside the box. (b) Program/erase parameters used. More than 1 erase pulse was used if needed. 56

4.10 Cell array architecture used for measuring ReRAM program/erase times. ‘BL’, ‘WL’ and ‘SL’ denote bit line, word line, and source line, respectively. 57

4.11 Measured distributions of ReRAM program voltage for various program pulse widths. The edges of the red box indicate the 25th and 75th percentiles and the median program voltage is indicated inside the box. 58

4.12 Measured distributions of ReRAM switching time for various values of program voltage. (The switching time is the product of program pulse width and pulse count.) The edges of the black box indicate the 25th and 75th percentiles and the median value of switching time is indicated inside the box. 58

4.13 Programming voltage pulse (black) and measured voltage (red) across 5 kΩ series resistor. (Once the ReRAM cell is programmed, the voltage across the resistor increases.) A large difference in switching time is seen for the two different ReRAM cells in (a) and (b).. 59

4.14 Circuit diagram illustrating voltage-differential based PUF generation. A pair of 1T-1R cells (‘Cell 1’ and ‘Cell 2’) forms one PUF bit-cell. ‘BL’, ‘WL’ and ‘SL’ denote bit line, word line, and source line, respectively. 59

4.15 Timing diagram for PUF bit generation during Gentle Program step. Once one cell is programmed, the voltage on the bit line (VBL) drops, preventing the programming of the other cell in the pair. Rtran is the ON-state resistance of the programmed cell transistor. 60

4.16 Simulation of bit-line voltages (for program voltages of 3.2V, 3.3V and 3.4V) when 100uA current compliance is set. It takes ~18ns for the bit-line voltage to drop after a cell is programmed. 61

4.17 Measured ReRAM current before and after soaking step. ON-state current is higher after soaking due to enhanced filament strength. 62

4.18 Percentages (%) of ‘0’s (OFF-state cells) and ‘1’s (ON-state cells) in 102,400,000 generated PUF bits. The percentages are very close to 50%, the ideal value for randomness. 62

4.19 Inter- and Intra- hamming distance (HD) of generated PUF bits (102,400,000 bits). PUF key length is 256 bits and total PUF count is 400,000. Inter-HD follows an ideal Gaussian distribution. There is no overlap between inter- and intra- HD distributions...... 64

4.20 Stability of ‘1’ and ‘0’ bits after 150°C, 50 hours (PostBake) compared to before bake (PreBake). 100% retention is achieved for both ‘0’ and ‘1’ states. 64

4.21 Auto correlation test results of 400,000 PUF keys of length 256 bits each, showing no correlation between PUF key bits. 65

4.22 ReRAM read endurance (measured I_{Cell} vs. Read #) showing that distinction between ON and OFF devices is well maintained after 5E13 read operations. 65

4.23 Randomness comparison of Intel TRNG (which uses non-deterministic hardware sources for RDRAND), Intel RDRAND (Digital RNG), and the ReRAM-based PUF demonstrated using 28nm-generation CMOS technology in this work. Entropy of 0.990 indicates $2^{-0.990}$ = 50.3% chance for 0 or 1 in any sequence/circumstance. (1.0 corresponds to perfect entropy of binary bits.) 67

List of Tables

2.1 Default parameters for MEM switch program, erase and read operations 25
2.2 Key material properties and values used for simulation 31
2.3 Mechanical and electrical properties of alternative contacting electrode materials 33
3.1 FG-MEM design parameter values used in this study 41
4.1 Parameters for ReRAM program and erase cycling 53
4.2 PUF bit generation conditions used in gentle program and soaking steps 60
4.3 Detailed results for the standard NIST SP800-22 randomness check test suite 66
4.4 Comparison of advanced PUF implementations 67

Acknowledgments

As I reflect upon this momentous occasion, I cannot help but feel overwhelmed with a deep sense of gratitude and accomplishment. It has been a long and arduous journey to reach this point, filled with moments of doubt and uncertainty. However, as Philippians 4:13 states, “I can do all things through Christ who strengthens me.” With the help of God Almighty, who has blessed me with countless opportunities and guidance, I have finally reached this day. This has indeed been an incredible journey and the accomplishment cannot be attributed solely to my efforts, but rather to all those who have played a role in this journey.

First and foremost, I would like to express my sincere gratitude to my advisor, Professor Tsu-Jae King Liu, for her exceptional mentorship, leadership, thoughtfulness, support, and encouragement. Her technical innovation, research expertise, and visionary leadership have been a tremendous inspiration to me as a scholar. In addition, her guidance and mentorship have helped me to grow and develop significantly in my field. I am deeply grateful for her unwavering support even in times when I doubted myself. As I embark on the next phase of my life, I aspire to follow in her footsteps, both in achieving technical excellence and in shaping the next generation. Thank you, Prof. King Liu!

I would also like to thank Prof. Ming Wu and Prof. Junqiao Wu for serving on my qualifying examination committee and dissertation committee. I am very grateful for their invaluable feedback and guidance on my research projects. Additionally, I’m grateful to Prof. Ana Claudia Arias, who not only served on my preliminary examination and qualifying exam committee but also shared invaluable life advice from her own academic journey. I would like to thank Prof. Jeffrey Bokor for dedicating his time to helping me understand advanced engineering concepts beyond what was taught in class. I also want to express my appreciation to Dr. Hei Kam, alumnae of King Liu group, for generously providing me with insightful suggestions and constructive feedback throughout the course of our reprogrammable MEM switches project. The constructive feedback and guidance were integral to the success of this project, and I am truly thankful for the collaboration.

Thank you to Prof. Gireeja Ranade and Dr. Jaeseok Jeon for allowing me to serve as a Graduate Student Instructor for EECS16A and EE130/230A. I have found the experience thoroughly enjoyable and rewarding as well. I thank Profs. Ali Javey, Vivek Subramanian, Clark Nguyen, Ali Niknejad, and Eli Yablonovitch for their expert guidance and outstanding courses in the field of electrical engineering.

Throughout my research journey, I have received steadfast support from various organizations, especially from the CrossBar Inc. team. I want to express my sincere gratitude to Dr. Sung Hyun Jo, Dr. Zhi Li, Dr. Amit Prakash, and Dr. Derek Lau, for their valuable contributions to the ReRAM project. Their insightful suggestions, brainstorming discussions, and constructive feedback have been immensely helpful. I am truly grateful for their guidance and support.

My sincere thanks also goes to the King Liu research group who made my experience at Berkeley as smooth and stress-free as possible. I truly cherished the meaningful interactions we had as both colleagues and friends. I am especially thankful to Dr. Sergio Almeida, Dr. Benjamin Osoba, Dr. Zhixin Alice Ye, Dr. Urmita Sikder, Dr. Xaioer Hu, Lars P. Tatum, Dasom Lee, Dr. Xi (Robin) Zhang, Dr. Fei Ding, Dr. Thomas Rembert, Dr. Rebecca Mih, and Dr. Min-Wu Kim. I am deeply grateful for the expertise and the frequent discussions we have had, which have played a

crucial role in my academic growth and development. I also would like to thank my friends and colleagues from the Black Graduate Engineering and Science Students (BGESS) for being by my side encouraging me and allowing me to serve as a board member. Thank you Dr. Liya Weldegebriel, Dr. Carlos Biaou, and Dr. Juan Pablo Llinas for providing a support system within and outside of academics.

I am fortunate to have had the opportunity to interact with the staff at UC Berkeley's Marvell Nanofabrication Laboratory, and I am deeply grateful to Bill Flounders, Dave Taosaka, Ryan Rivers, Allison Dove, Joanna Bettinger, Richelieu Hemphill, Sam Tsitrin, Jay Morford, and Jason Chukes for their invaluable assistance. They not only taught me how to use the fabrication equipment but also helped me address process issues and ensured that the equipment remained operational. Their support has been immensely valuable.

I would also like to thank all the faculty and staff members of the EECS department and College of Engineering including Audrey Sillers, Susanne Kauer, Tiffany Reardon, Meltem Erol, Lea Marlor, Dr. Michael Bartl, Charlotte Jones, Dr. Kedrick Perry, and many others for their support and guidance in administrative matters. In particular, I am sincerely grateful to Shirley Salanio for her continuous support, prompt replies, and encouragement during challenging times. Additionally, I would also like to thank Dr. Sheila Humphrey for her kindness and willingness to share her knowledge.

I also would like to express my gratitude to Dr. Satish Naidu, Dr. Jose Romero, and Dr. Atul Shah from Intel Inc., for the valuable knowledge they imparted to me during my internship.

Thank you to the University of California Office of the President and the National GEM Consortium for their generous fellowship support and funding, which made my graduate studies possible. Additionally, I would like to thank the UC-HBCU initiative for the opportunity to conduct research at Berkeley during my undergraduate studies, which played a pivotal role in my decision to pursue a Ph.D. at UC Berkeley.

I am also deeply grateful to all the young adult fellowship, choir, and my entire church family at Abenezer Evangelical Church. Their love, encouragement, spiritual guidance, and hospitality have meant the world to me, and I truly feel like I am home. Thank you all!

Finally, I am tremendously grateful to my family for their unconditional love and support. I am deeply indebted to them for their sacrifices, which have helped shape me into the person I am today. Their constant guidance, support, and endless caring have helped me navigate through life's many challenges. Special thanks to my husband, Bekalu Aneme, for taking care of me and sharing all my laughs and tears. I can never get through my Ph.D. journey without his love and patience.

Chapter 1

Introduction

1.1 Brief History and Evolution of Integrated Circuits

Integrated circuits (ICs), also known as microchips, are tiny electronic devices that integrate a large number of electronic components such as transistors, resistors, and capacitors on a single chip of silicon. The advent of the IC chip was a revolutionary innovation that transformed the electronics industry and paved the way for advanced information processing and communication systems [1]. This breakthrough has led to the development of personal computers, smartphones, and other portable electronic devices that have become vital in daily life. Furthermore, the proliferation of computing has profoundly advanced many areas of technology, including the Internet of Things (IoT), cloud computing, artificial intelligence, and other domains. By enabling the integration of numerous components onto a single chip, microchips have made it possible to design smaller, more powerful, and more energy-efficient devices that have enhanced people's productivity and convenience.

The history of ICs can be traced back to the invention of the transistor in 1947 by John Bardeen, Walter Brattain, and William Shockley at Bell Labs. The transistor was a replacement for the bulky and inefficient vacuum tubes that were used in electronics at the time. It was smaller, used less power, and was more reliable than vacuum tubes, making it ideal for use in electronic devices. In 1958, Jack Kilby of Texas Instruments invented the first integrated circuit. Kilby's circuit comprised a tiny piece of germanium with a few transistors and other components etched onto its surface [2]. In 1965, Gordon Moore, the co-founder of Intel Corporation, made an observation that the number of Complementary Metal Oxide Semiconductor (CMOS) transistors on a single chip doubled about every year and later in 1967 adjusted to every two years [3][4]. This prediction, which became known as Moore's Law, has been remarkably accurate over the past several decades and has driven the rapid advancement of computer technology. Nowadays, modern microprocessor chips contain many billions of transistors enabled by the steady miniaturization of the transistor, which allows exponential pace of improvement in computational speed while lowering the cost and energy consumed per function [5].

Along with transistor (switches used to implement digital logic and for signal amplification), the advancement of memory devices has also been a key factor in the development of modern information technology. In the early days of computing, data was stored on magnetic tape, punch cards, and other physical media which were large, heavy, and expensive. However, the invention of ICs led to the creation of the first solid-state memory device, the metal-oxide-semiconductor (MOS) memory [6][7]. Today, solid-state memory devices are ubiquitous and used in a wide range of applications, including personal computers, smartphones, wearable devices, servers, and data centers.

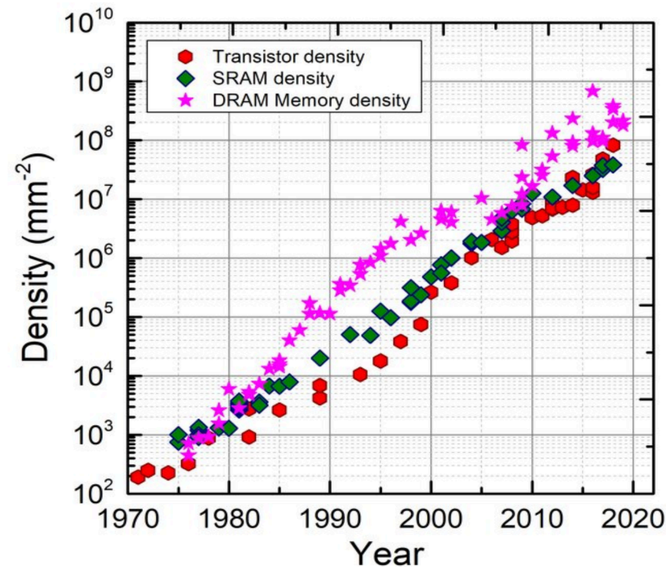
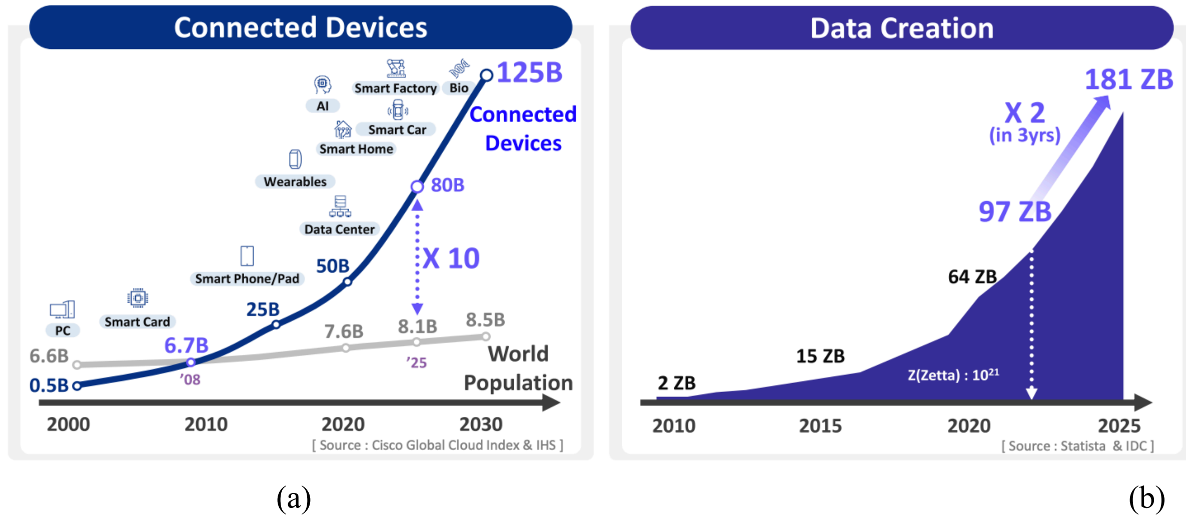


Figure 1.1: Evolution of chip component density scaling trends. Transistor density is defined as the number of transistors on a chip divided by the chip area. SRAM bit density and DRAM bit density have advanced at the same pace as the transistor density (reproduced from [8]).

Despite rumors of the end of Moore's law in recent years [9], the semiconductor industry has continued to steadily advance IC manufacturing. This progress is driven by growing demand for easy access to information and mobile connectivity, resulting in the widespread acceptance and rapid advancement of internet and wireless communication technologies. Today we live in a globally connected society with universal access to information, which is increasingly collected by machines with the emergence of the IoT.

The Internet of Things (IoT) refers to objects with embedded electronic devices that are used to collect, process and communicate information via a wired or wireless network. Over the past decade, there has been an exponential growth in the number of connected devices and data generated worldwide. This trend is primarily driven by the proliferation of IoT devices [10]. As shown in Figure 1.2, the number of connected devices is expected to reach 125 billion by 2030. The exponential growth in the number of connected devices has also led to an exponential growth in the amount of digital data generated. It is estimated that the total amount of digital data generated worldwide will reach 181 zettabytes by 2025, up from just 2 zettabytes in 2010 [11][12]. This

growth trend is expected to continue in the years ahead. However, this growth also brings challenges, such as data privacy and security concerns, as well as the need for more energy-efficient and more reliable data storage and processing solutions.



(a) (b)

Figure 1.2: Exponential growth of (a) connected devices and (b) data creation (adapted from [11]).

1.2 Memory Requirements for IoT

The widespread availability of electronic devices and the generation and processing of large data sets ("big data") necessitate the need for a storage medium that is both non-volatile and energy-efficient, with high endurance and retention, low manufacturing cost, and most importantly, embedded with logic circuitry so that certain computational tasks can be performed in the memory itself for faster processing speed [13]. Meanwhile, it is also crucial to maintain the security and confidentiality of information collected and shared by IoT devices. A recommended solution to achieve this is to utilize memory devices that can store data and function as hardware security keys as well.

1.1.1 Traditional Memory Devices

Modern computer architectures have a hierarchical memory system, as illustrated in Figure 1.3, comprising different types of conventional memory devices ranging from high-speed, small-capacity storage devices to slower, large-capacity devices [14–16].

Conventionally the memory pyramid from top to bottom (small capacity to large capacity) comprises central processor unit (CPU) registers and cache memory on the same chip as the CPU, and main working memory and long-term storage memory on separate chips. CPU registers at the top of the pyramid are the fastest type of memory used in a computer. Registers are built into the processor and are used to store the instructions and data that are frequently accessed or currently

being executed by the processor. However, registers are the most expensive type of memory and have very limited capacity, typically only a few bytes [17].

Next in the hierarchy is cache memory. Caches are designed to reduce the time it takes to access data from the main memory by temporarily storing copies of data frequently used. Typically, caches are slower than the CPU registers and mostly comprise Static Random Access Memory (SRAM) cell arrays due to their higher access speed compared to other types of memory [14][18]. A typical SRAM cell consists of six transistors (two cross-coupled CMOS inverters, forming a latch, plus two pass-gate transistors for accessing the two storage nodes of the latch) to store one bit of information while power is supplied. Yet, it has a limitation in terms of storage density (meaning that it stores less data per unit area on a chip compared to other types of memory). This makes it more expensive and impractical for large capacity and long-term storage.

Main memory comprises Dynamic Random Access Memory (DRAM) chips. A DRAM cell stores one bit of information using a capacitor and an access transistor. One of the main benefits of DRAM is its high storage density, *i.e.*, it can store more data per unit area compared to SRAM. This makes it more cost-effective for applications where large amounts of data need to be temporarily stored. However, the simpler cell design results in slower data access speeds [15]. A DRAM cell requires regular refreshing to restore charge on its capacitor (because the access transistor has non-zero leakage current). For this reason it consumes more power than non-volatile memory.

Although both SRAM and DRAM are useful for temporary data storage, they are volatile, meaning data will be lost when power to the chip is turned off. For long-term data storage, hard-disk drives (HDDs) have been used for over 50 years. HDDs have several benefits, including nonvolatility, large storage capacity, and low cost per bit. Their main downside is their slow data transfer rate for both reading and writing operations.

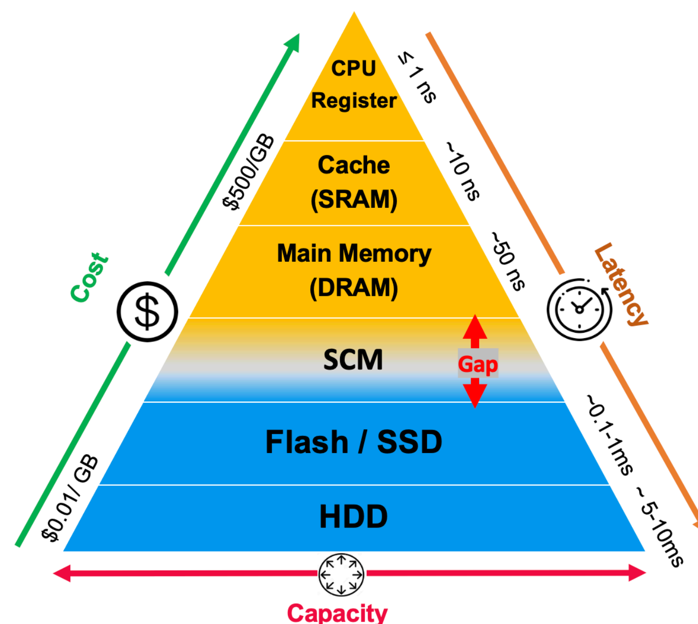


Figure 1.3: Memory hierarchy diagram [19].

In recent years, due to the performance (latency) gap between DRAM and HDD, flash memory devices have emerged as a solution. Flash memory stores information by injecting (erasing) electrons to (from) a charge-storage layer through an electrically insulating dielectric material. These devices have become widely used in solid-state drives (SSD), displacing traditional HDD for some applications because of their faster speed and lower power consumption. Flash memories have also found extensive usage in portable electronic devices, such as mobile phones, and USB flash drives [20]. Vertical (three-dimensional, or 3-D) stacking of flash memory cells has been adopted to increase storage density and increase storage capacity [21].

Among the well-established types of memory devices there exists a sizable gap in performance (latency) between volatile memory (DRAM) and non-volatile storage (Flash) devices. Emerging Non-Volatile Memory (eNVM) device technologies aim to bridge this gap.

1.1.2 Emerging Non-Volatile Memory Technologies

Increasing requirements for reduced IoT power consumption have driven efforts to develop new devices for storage-class memory (SCM), filling the gap in performance and storage density that currently exists between volatile memory and non-volatile storage devices [22]. Among several emerging memory device technologies, phase change memory (PCM), spin-transfer torque random access memory (STT-RAM), ferroelectric RAM (FeRAM), and resistive random-access memory (RRAM) are the most promising.

These emerging technologies employ novel materials (metal oxides, ferroelectric oxides, ferromagnetic metals, chalcogenides, carbon materials, *etc.* [23]) and mechanisms such as quantum mechanical phenomena, redox reaction, phase transition, spin-state, molecular reconfiguration, *etc.* to change the state of the memory device and thereby store information. Some have a simple two-terminal cell structure that is most amenable for high-density storage.

Although the dream of a "universal memory" (that offers high speed, large capacity and low cost) has not yet been achieved, SCM devices have the potential to bridge the performance gap between storage and memory, to enable new computer architectures such as brain-inspired computing systems and new applications such as hardware security systems. Practical manufacturing challenges exist, such as device-to-device variability and reliability. Therefore, there are still room for innovation in SCM devices to meet future computing needs.

1.3 Energy-Efficient Computing for IoT

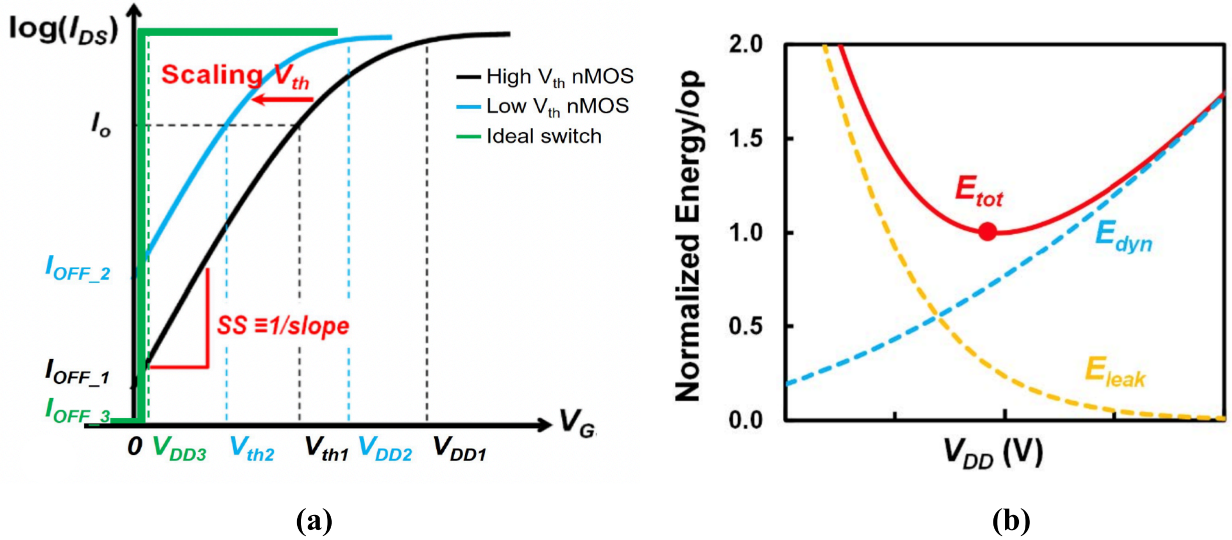
1.1.3 Energy-Efficiency Limit for CMOS Technology

Continual improvements in digital IC performance and reductions in cost per function have been enabled by the steady miniaturization of CMOS transistors. Ideally, as the dimensions of a transistor shrink, the operating voltage (V_{DD}) should decrease to avoid increasing the peak electric field [24]. However, since the 90 nm technology node, voltage scaling slowed down because the threshold voltage (V_{th}) of a MOS field-effect transistor (MOSFET) cannot be too close to 0 V; otherwise, the off-state leakage current (I_{OFF}), which increases exponentially with a linear reduction in V_{th}, will result in unacceptably high static power dissipation.

To explain this, Figure 1.4(a) shows a standard drain current *vs.* gate voltage semi-log plot, for two n-channel MOSFETs with different values of V_{th} as well as for an ideal switch. I_{OFF} is defined as the drain current when $V_{GS} = 0V$ and $V_{DS} = V_{DD}$, where V_{DD} is the power supply voltage.

$$I_{OFF} \propto \exp(-\frac{V_{th}}{SS}) \tag{1.1}$$

where the subthreshold swing (SS) is defined as the inverse slope of the log(I_D)-V_G curve. The lower limit of SS is 60 mV/dec at room temperature due to the Boltzmann energy distribution of electrons in the source region of a MOSFET [25].



(a) **(b)**

Figure 1.4: Conceptual illustrations of (a) the switching *I-V* characteristics of a high-V_{th} n-channel MOSFET, a low-V_{th} n-channel MOSFET, and an ideal switch; (b) dynamic, static, and total energy consumed per operation of a CMOS digital logic circuit. The lower limit for CMOS energy efficiency exists due to MOSFET OFF-state leakage (reproduced from [26]).

The total energy dissipated (E_{tot}) per digital operation is composed of a dynamic energy dissipation component (E_{dyn}) and a static energy dissipation component (E_{leak}):

$$E_{tot} = E_{dyn} + E_{leak} \tag{1.2}$$

If V_{DD} is lowered to decrease E_{dyn} then the transistor on-state current (I_{ON}) will be lowered, resulting in slower digital circuit operation. The more time (t_{delay}) it takes for a digital circuit to complete its function, the more energy is wasted due to transistor off-state leakage since $E_{leak} \propto I_{off}V_{DD}t_{delay}$. If V_{th} is lowered together with V_{DD} to maintain I_{ON} for fast circuit operation, I_{OFF} is exponentially higher, again resulting in higher E_{leak}. Therefore as V_{DD} is lowered, E_{leak} increases while E_{dyn} decreases, so that a minimum value of E_{tot} exists (at $V_{DD} = V_{th}$).

As illustrated by the green curve in Figure 1.4 (a), an ideal switch would have an abrupt switching characteristic with $I_{off} \approx 0$ and $V_{th} \approx 0$ to enable ultra-low energy per operation. Thus, in order to achieve dramatically improved energy efficiency, an alternative switching device is needed.

1.1.4 MEM Switches for Energy-Efficient Computing

Microelectromechanical (MEM) switches operate by making and breaking physical contact between two conductive electrodes (drain and source electrodes) so they have essentially zero I_{OFF}, enabling zero static power consumption, and abrupt switching characteristics across a wide range of temperatures [27], enabling lower supply voltage V_{DD}. Therefore, MEM switches have been investigated for energy-efficient logic switches and NVM device applications [28-30].

Figure 1.5 shows the schematic cross-section of body-biased MEM switch consisting of four terminals: gate, body, drain, and source electrodes. The movable structure is referred to as the gate and has an attached conductive metal strip underneath, called the drain, that is electrically insulated from the gate by a thin dielectric layer. A fixed actuation electrode underneath the gate is referred to as the body; the source electrode is coplanar with the body and runs underneath the drain in the dimpled contact region. In the OFF state, an air gap separates the conducting source and drain electrodes, hence no current flows between them as shown in Figure 1.5(a). When a sufficiently large voltage is applied across the gate and body such that $V_{GB} > V_{ON}$, the attractive electrostatic force (F_{elec}) actuates the movable gate electrode downwards towards the fixed body electrode, bringing the drain electrode into physical contact with the source electrode so that the device turns ON abruptly, allowing current to flow (Figure 1.5(b)). To turn OFF the switch, the voltage across gate and body is reduced below V_{OFF}, so that the spring restoring force (F_{sp}) of the deformed movable electrode actuates is upward to break contact between the drain and source. As shown in Figure 1.6, V_{OFF} is smaller than V_{ON} due to contact adhesive force F_{adh} between the drain and source electrodes in the ON state. The hysteresis voltage (V_H) is the difference between V_{ON} and V_{OFF}, and can be engineered to be very small (on the order of 10 mV) [30].

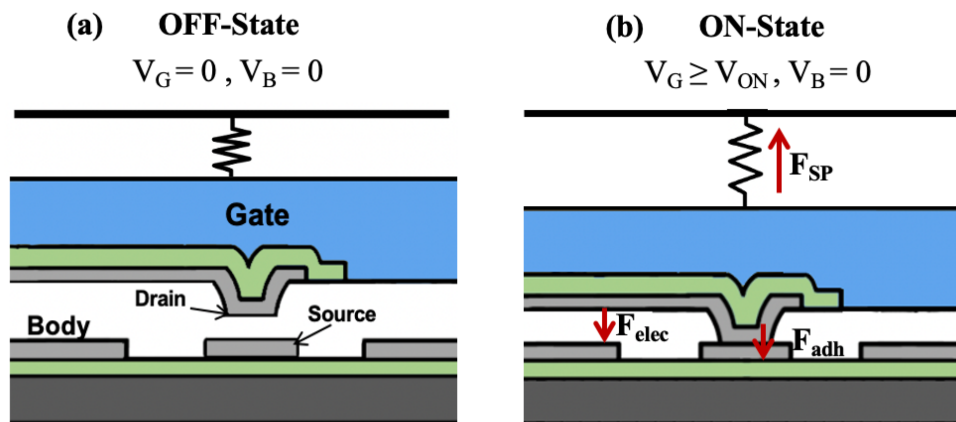


Figure 1.5: Cross-section illustrations of a 4-terminal MEM switch in a) OFF–state, and b) ON–state.

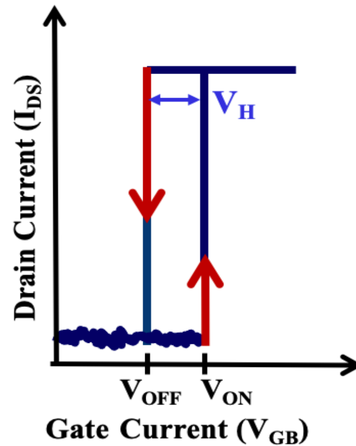


Figure 1.6: A typical I_{DS}-*vs.*-V_G characteristic of a MEM switch for bidirectional voltage sweep, showing abrupt switching behavior.

Due to their negligible OFF-state leakage and abrupt switching characteristics, MEM switches are attractive for energy-efficient logic and NVM applications.

1.4 Ensuring IoT Security

IoT devices collect, process, and exchange a significant amount of potentially security-risking and/or private information. They are vulnerable to various malicious attacks, particularly because they communicate wirelessly and operate in an energy-constrained environment with very limited hardware resources [32][33], *e.g.*, their memory capacity may be insufficient for storing encryption keys.

Physically unclonable functions (PUFs) are promising for authentication and secure encryption key generation/storage without expensive hardware [34]. Rather than storing keys in digital memory, PUFs leverage inherent variability in the integrated circuit manufacturing process that results in random variations in physical properties and hence electrical characteristics from device to device. Although device variability is not desirable for circuit performance, it is crucial for PUF implementation.

Resistive Random Access Memory (ReRAM) is an eNVM technology of keen interest for PUF key generation due to its low-power switching characteristics and relatively large device-to-device variability. As illustrated in Figure 1.7, a ReRAM device comprises two terminals with an oxide "switching layer" sandwiched between the two metallic electrodes. The ReRAM switching mechanism is based on the movement of ions under the influence of an electric field, to form or to remove an electrically conductive filament within the switching layer [35]. To program the ReRAM device, a positive voltage pulse is applied to the top electrode to cause a conductive filament to form, resulting in a low resistance state. To erase the ReRAM device, a negative voltage pulse is applied to the top electrode to break the filament, resulting in a high resistance state.

Depending on the composition of the conductive filament, a ReRAM device can be classified as one of two types: metal ion based, also referred to as conductive-bridge random access memory (CBRAM); or oxygen vacancy based random access memory (OxRAM). In a CBRAM device,

filament formation/breakage occurs via migration of metal ions and subsequent reduction/oxidation (redox) reactions. In an OxRAM device, filament formation occurs via migration of oxygen ions or vacancies [36].

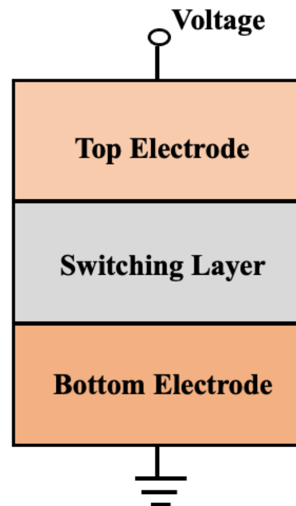


Figure 1.7: Simple cross-section schematic illustrating the metal-insulator-metal structure of a ReRAM device.

Random device-to-device resistance variability can be leveraged to generate a unique PUF key based on the resistance of each cell within an ReRAM cell array. ReRAM-based PUFs can be implemented within a standard ReRAM cell array, enabling information storage together with PUF storage. Moreover, ReRAM-based PUFs have relatively low power consumption and a small footprint, making them ideal for use in resource-limited settings like embedded systems [37][38].

1.5 Dissertation Objectives and Overview

The main objective of this dissertation is to investigate novel/emerging NVM devices for ultra-low-power electronics applications. A significant portion of this work focuses on adapting micro-electro-mechanical (MEM) switches for eNVM applications. Another area of focus is the application of Resistive Random Access Memory (ReRAM) devices for efficient implementation of physical unclonable functions (PUFs).

In chapter 2, the concept of a reprogrammable NV-MEM switch is presented. After an introductory overview of the MEM switch design and fabrication process, program and erase operations are discussed. It is experimentally demonstrated that MEM switches can be programmed and erased with relatively small voltage (< 3V) and that they have excellent retention characteristics at elevated temperature (200°C). This allows non-volatile information storage at zero incremental fabrication cost.

In chapter 3, a floating-gate MEM switch design is investigated. The device fabrication process and initial experimental results are presented. Challenges their possible solutions are discussed.

In chapter 4, a novel PUF key architecture and generation scheme that utilizes the inherent program-time variation of resistive random access memory (ReRAM) cells as an entropy source is presented. ReRAM design and robust nonvolatile memory operation is first discussed, followed by the PUF generation scheme and randomness evaluation using a standard NIST test suite. Further verifications to confirm the high reliability of the generated PUF keys are described.

Chapter 5 summarizes the key findings and contributions of this dissertation. Suggestions for future research are also offered.

1.6 References

[1] 60 Years of Integrated Circuits. *Nat Electron,* vol. 1, no. 483, September 2018, doi: 10.1038/s41928-018-0145-6.

[2] W. Brinkman, D. Haggan, and W. Troutman, “A History of the Invention of the Transistor and Where It Will Lead Us,” *IEEE Journal of Solid-State Circuits*, vol. 32, no. 12, pp. 1858-1865, Dec. 1997, doi: 10.1109/4.643644.

[3] Gordon E. Moore, “Cramming more components onto integrated circuits,” *Electronics Magazine*, vol. 38, no. 8, April 1965, doi:10.1109/N-SSC.2006.4785860.

[4] G.E. Moore, “Progress in Digital Integrated Electronics,” *Proc. Technical Digest Int’l Electron Devices Meeting*, vol. 21, pp. 11–13, 1975, doi: 10.1109/N-SSC.2006.4804410.

[5] F. Faggin, “The Birth of the Microprocessor,” *IEEE Micro*, vol. 41, no. 6, pp. 16-19, 2021, doi: 10.1109/MM.2021.3112302.

[6] D. Klein, “The History of Semiconductor Memory: From Magnetic Tape to NAND Flash Memory,” *IEEE Solid-State Circuits Magazine*, vol. 8, no. 2, pp. 16-22, Spring 2016, doi: 10.1109/MSSC.2016.2548422.

[7] S. Asai, “Semiconductor Memory Trends,” in Proceedings of the IEEE, vol. 74, no. 12, pp. 1623-1635, Dec. 1986, doi: 10.1109/PROC.1986.13681.

[8] H. -S. P. Wong et al., “A Density Metric for Semiconductor Technology [Point of View],” *Proceedings of the IEEE*, vol. 108, no. 4, pp. 478-482, April 2020, doi: 10.1109/JPROC.2020.2981715.

[9] T. N. Theis and H.-S. P. Wong, “The End of Moore's Law: A New Beginning for Information Technology,” *Computing in Science & Engineering*, vol. 19, no. 2, pp. 41-50, Mar.-Apr. 2017, doi: 10.1109/MCSE.2017.29.

[10] M. Bali, K.Gupta, D. Koundal, A. Zaguia, S. MahajanA. Pandit, “Smart Architectural Framework for Symmetrical Data Offloading in IOT,” *Symmetry* 13, 1889, October 2021, doi:10.3390/sym13101889.

[11] D. Ha, “Energy Efficient CMOS Scaling for 1nm and Beyond,” *2022 IEDM Short Course 1-2*, 2022.

[12] A. Weissberger. Cisco’s Annual Internet Report (2018–2023) forecasts huge growth for IoT and M2M; tepid growth for Mobile. *IEEE Com. Soc*. February 20, 2020

[13] J. Lipman, S. Corp, "NVM Memory: A Critical Design Consideration for IoT Applications," *Design and Reuse*, 2017.

[14] S. Manegold, "Memory Hierarchy" *Encyclopedia of Database Systems*, Springer, pp 2222–2229: 2018, doi:10.1007/978-1-4614-8265-9_657.

[15] J.L. Hennessy and D.A. Patterson, "*Computer Architecture – A Quantitative Approach*," 3rd ed. San Mateo: Morgan Kaufmann; 2003.

[16] D.R. Parthasarathi, "Computer Architecture, Memory Hierarchy Design – Basics – Computer Architecture" *INFLIBNET Centre.*

[17] S. Kumar and P. K. Singh, "An overview of modern cache memory and performance analysis of replacement policies," *IEEE International Conference on Engineering and Technology (ICETECH)*, pp. 210-214, 2016 doi: 10.1109/ICETECH.2016.7569243.

[18] K. Osada, "Fundamentals of SRAM Memory Cell," *Low Power and Reliable SRAM Memory Cell and Array Design*. *Springer Series in Advanced Microelectronics*, vol 31. 2011, doi:10.1007/978-3-642-19568-6_2.

[19] Chuang Qian. Electro-Mechanical Devices for Ultra-Low-Power Electronics. PhD thesis, EECS Department, University of California, Berkeley, May 2017. URL: http://www2.eecs.berkeley.edu/Pubs/TechRpts/2017/EECS-2017-21.html.

[20] V. . -Y. Aaron and J. . -P. Leburton, "Flash memory: towards single-electronics," *IEEE Potentials*, vol. 21, no. 4, pp. 35-41, Oct.-Nov. 2002, doi: 10.1109/MP.2002.1044216.

[21] Y. Li, "3D NAND Memory and Its Application in Solid-State Drives: Architecture, Reliability, Flash Management Techniques, and Current Trends," *IEEE Solid-State Circuits Magazine*, vol. 12, no. 4, pp. 56-65, 2020, doi: 10.1109/MSSC.2020.3021841.

[22] C. H. Lam, "Storage Class Memory," *IEEE International Conference on Solid-State and Integrated Circuit Technology*, pp. 1080-1083, 2010, doi: 10.1109/ICSICT.2010.5667551.

[23] A. Chen, "Emerging nonvolatile memory (NVM) technologies," *45th European Solid State Device Research Conference (ESSDERC)*, pp. 109-113, 2015, doi: 10.1109/ESSDERC.2015.7324725.

[24] R.H. Dennard, F.H. Gaensslen, H.-N. Yu, V.L. Rideout, E. Bassous, A.R. Leblanc, "Design of ion-implanted MOSFETs with very small physical dimensions," IEEE Journal of Solid-State Circuits, Vol. 9, 1974.

[25] K. K. Ng and S. Sze, "Physics and properties of semiconductors - A review," *Physics of Semiconductor Devices. John Wiley Sons*, pp. 5–75, 2006, doi: 10.1002/9780470068328.ch1.

[26] Xiaoer Hu. Micro-Electro-Mechanical Relay Technology for Beyond-Von-Neumann Computer. PhD thesis, EECS Department, University of California, Berkeley, May 2017. URL: https://www2.eecs.berkeley.edu/Pubs/TechRpts/2022/EECS-2022-251.html

[27] X. Hu, S. F. Almeida, Z. Alice Ye, and T. -J. K. Liu, "Ultra-Low-Voltage Operation of MEM Relays for Cryogenic Logic Applications," *IEEE International Electron Devices Meeting (IEDM)*, pp. 34.2.1-34.2.4, 2019, doi: 10.1109/IEDM19573.2019.8993629.

[28] O. Y. Loh and H. D. Espinosa, “Nanoelectromechanical Contact Switches,” *Nature Nanotechnology*, vol. 7, no. 5, p. 283, 2012, doi: 10.1038/nnano.2012.40.

[29] J. E. Jang, S. N. Cha, Y. J. Choi, D. J. Kang, T. P. Butler, D. G. Hasko, J. E. Jung, J. M. Kim, and G. A. Amaratunga, “Nanoscale Memory Cell Based on a Nanoelectromechanical Switched Capacitor,” *Nature Nanotechnology*, vol. 3, no. 1, p. 26, 2008, doi: 10.1038/nnano.2007.417.

[30] B. Osoba et al., “Sub-50 mV NEM Relay Operation Enabled by Self-Assembled Molecular Coating,” *IEEE International Electron Devices Meeting*, pp. 26.8.1-26.8.4, 2016, doi: 10.1109/IEDM.2016.7838489.

[31] S. W. Lee, S. J. Park, E. E. Campbell, and Y. W. Park, “A Fast and Low-Power Microelectromechanical System-Based Non-Volatile Memory Device,” *Nature Communications*, vol. 2, p. 220, 2011, doi: 10.1038/ncomms1227.

[32] B. Halak, M. Zwolinski, and M. S. Mispan, “Overview of PUF-Based Hardware Security Solutions for the Internet of Things,” *IEEE 59th International Midwest Symposium on Circuits and Systems (MWSCAS)*, pp. 1-4, 2016, doi: 10.1109/MWSCAS.2016.7870046.

[33] W. Trappe, R. Howard, and R. S. Moore, “Low-Energy Security: Limits and Opportunities in the Internet of Things,” *IEEE Security & Privacy*, vol. 13, no. 1, pp. 14-21, Jan.-Feb. 2015, doi: 10.1109/MSP.2015.7.

[34] R. Maes, “Physically Unclonable Functions: Properties. In: Physically Unclonable Functions,” Berlin, Heidelberg: Springer, pp. 49-80, 2013, doi: 10.1007/978-3-642-41395-7_3.

[35] F. Zahoor, T.Z. Azni Zulkifli, and F.A. Khanday, “Resistive Random Access Memory (RRAM): an Overview of Materials, Switching Mechanism, Performance, Multilevel Cell (mlc) Storage, Modeling, and Applications,” *Nanoscale Res Lett,* vol. 15, no. 90, 2020, doi: 10.1186/s11671-020-03299-9.

[36] A. Calderoni, S. Sills, and N. Ramaswamy, “Performance comparison of O-based and Cu-based ReRAM for high-density applications,” *IEEE 6th International Memory Workshop (IMW)*, pp. 1-4, 2014, doi: 10.1109/IMW.2014.6849351.

[37] F. Pan, C. Chen, Z. Wang, Y. Yang, J. Yang, F. Zeng, “Nonvolatile Resistive Switching Memories-Characteristics, Mechanisms and Challenges,” *Progress in Natural Science: Materials International,* vol 20, pp. 1-15, 2010, doi:10.1016/S1002-0071(12)60001-X.

[38] ReRAM Memory Overview: *Crossbar Inc*. URL: https://www.crossbar-inc.com/technology/reram-overview/.

Chapter 2

A Reprogrammable MEM Switch Utilizing Controlled Contact Welding

2.1 Introduction

With the emergence of the Internet of Things (IoT), the number of electronic devices has rapidly increased, and the volume of generated data has grown exponentially. The IoT refers to a network of objects that are wirelessly connected together. This is made possible by embedding into the objects ("things") electronic information processing devices, sensors, actuators, *etc*. for the purpose of communicating data and information with other objects. The number of IoT devices worldwide is forecast to almost triple from 9.7 billion in 2020 to more than 29 billion IoT devices in 2030 [1], when the total amount of data generated would be approximately 10 times more than exists today [2].

With the proliferation of IoT devices, as well as "edge" and cloud computing, electricity consumption of electronics is growing at an exponential pace. Therefore, new computing and memory device technologies are needed to enable much more energy-efficient digital computing and non-volatile (NV) data storage [3]. For IoT applications, low manufacturing cost is also an important requirement.

Micrometer-scale electro-mechanical (MEM) switches are considered an attractive option for IoT applications and wearable/disposable electronics due to their negligible OFF-state power consumption and abrupt switching characteristics enabling zero standby power consumption and milli-Volt operation across a wide range of operating temperatures [4–6]. MEM switches can be fabricated using standard CMOS integrated circuit manufacturing processes and can be designed to be non-volatile, *e.g.*, serving as reconfigurable interconnects [7], [8].

In this chapter, MEM switches designed for digital logic applications are demonstrated to be multi-time programmable via controlled contact welding and un-welding. This newfound

capability provides for greater versatility of device operation, enabling non-volatile information storage to be embedded with digital logic circuitry with no incremental fabrication cost.

2.2 MEM Switch Operation

MEM switches operate by making and breaking physical contact between two conductive electrodes, separated physically by an air gap in the OFF-state. As shown in Figure 2.1, the movable top structure is referred to as the gate, which is mechanically suspended by a spring with effective spring constant k_{eff} while the underlying fixed electrode is referred as the body. g_d and g_o are the as-fabricated contact gap size and the actuation gap size, respectively.

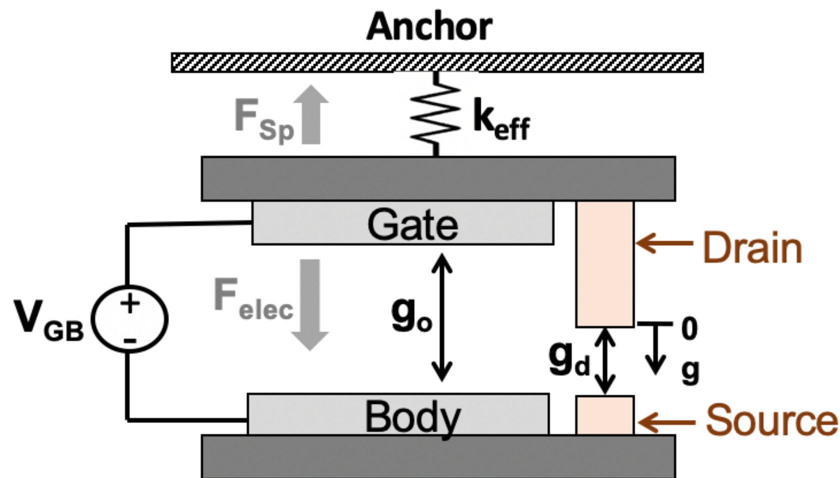


Figure 2.1: Schematic diagram and mechanical spring model of a four-terminal MEM switch designed for digital logic applications. Electrostatic actuation and spring restoring forces are illustrated.

In the OFF-state, an air gap exists between the drain and source conductive electrodes, preventing any current flow. When a voltage is applied across the gate and body (V_{GB}), an attractive electrostatic force F_{elec} is exerted on the top movable gate electrode toward the fixed body electrode. Displacement of the gate electrode from its equilibrium position produces a spring restoring force (F_{sp}) opposing the electrostatic force (F_{elec}):

$$F_{sp} = -k_{eff} \times g \quad (2.1)$$

where g is the displacement of the top electrode from its original position towards the bottom fixed electrode. (The negative sign of the spring restoring force indicates that the force is in the upward direction.)

$$F_{elec} = \frac{1}{2}\frac{\varepsilon_o A_{ACT} V_{GB}^2}{(g_o - g)^2} \tag{2.2}$$

where ε_o is the vacuum permittivity, A_{ACT} is the effective actuation (overlap) area, and V_{GB} is the potential difference between the gate and body electrodes. The net force F_{net} on the gate electrode is given by:

$$F_{net} = F_{sp} + F_{elec}$$

$$= -k_{eff} \times g + \frac{1}{2}\frac{\varepsilon_o A_{ACT} V_{GB}^2}{(g_o - g)^2} \tag{2.3}$$

Taking a derivative with respect to g, one obtains the following equation:

$$\frac{dF_{net}}{dg} = -k_{eff} + \frac{\varepsilon_o A_{ACT} V_{GB}^2}{(g_o - g)^3} \tag{2.4}$$

As the applied voltage V_{GB} is increased, the movable structure is physically displaced ($g > 0$) to maintain force balance ($F_{net} = 0$). However, since the electrostatic force (F_{elec}) increases super-linearly while $|F_{sp}|$ increases linearly with decreasing actuation gap size ($g - g_o$), eventually g reaches a critical value beyond which force balance is no longer possible and the movable structure is "pulled-in" to contact. The value of V_{GB} at which this phenomenon occurs is referred to as the pull-in voltage V_{PI}. Both V_{PI} and the critical displacement g can be calculated by setting F_{net} and $\frac{dF_{net}}{dg}$ equal to 0 in Equations 2.3 and 2.4:

$$V_{PI} = \sqrt{\frac{8k_{eff} g_o^3}{27\varepsilon_o A_{ACT}}} \tag{2.5}$$

$$g = \frac{g_o}{3} \tag{2.6}$$

For the MEM switch to operate in pull-in (PI) mode, the as-fabricated contact gap size g_d should be greater than 1/3 of the as-fabricated actuation gap size g_o [3][4][6]. If the as-fabricated contact gap size g_d is less than $g_o/3$, (*i.e.*, $g_o > 3g_d$), however, then the switch turns ON before it enters into the PI region of operation, *i.e.*, it operates in non-pull-in (NPI) mode. The applied voltage required to turn on a NPI MEM switch is

$$V_{ON} = \sqrt{\frac{2k_{eff}(g_o - g_d)^2}{\varepsilon_o A_{ACT}}} \tag{2.5}$$

In the ON-state there exists an attractive contact adhesive force F_{adh} between the contacting source and drain electrode surfaces, so the net force equation becomes

$$F_{net} = F_{sp} + F_{elec} + F_{adh} \tag{2.6}$$

If the contacting surfaces comprise the same material, then the adhesive force is due primarily to Van der Waals force which is proportional to the apparent contact surface area [9,10].

Note that F_{adh} is additive to F_{elec}, *i.e.*, it helps to keep the MEMS switch in the ON–state. For this reason, the minimum value of applied voltage required to keep the MEMS switch ON is less than V_{ON}. That is, a MEM switch exhibits hysteretic switching behavior. The spring restoring force of the movable plate must overcome both the electrostatic force and the contact adhesive force in order to turn off the MEMS switch:

$$k_{eff} g = \frac{1}{2} \frac{\varepsilon_o A_{ACT} V_{OFF}^2}{(g_o - g)^2} + F_{adh} \tag{2.7}$$

$$V_{OFF} = \sqrt{\frac{2(k_{eff} g_d - F_{adh})(g_o - g_d)^2}{\varepsilon_o A_{ACT}}} \tag{2.8}$$

The hysteresis voltage is defined as

$$V_H = V_{ON} - V_{OFF} \tag{2.9}$$

and is the minimum actuation voltage-swing required to operate the MEMS switch.

Solving for V_H for a NPI-mode switch with $g_d \leq \frac{g_o}{3}$:

$$V_H \approx F_{adh} \sqrt{\frac{2 g_d}{\varepsilon_o A_{ACT} k_{eff}}} \tag{2.10}$$

Previous studies indicate that PI-mode switch designs can provide for the lowest switching energy due to a relatively compliant (lower k_{eff}) structure. However, this makes them prone to stuck-ON failure (*i.e.,* $F_{SP} < F_{adh}$) [11]. Therefore, to avoid stuck-ON failures, body-biased (V_B > 0) NPI-mode MEM switch designs with stiffer structures are more practical for reliable operation. Therefore NPI-mode switches are used for the work of this dissertation.

2.3 MEM Switch Design and Fabrication Process

Figure 2.2 is a plan-view scanning electron micrograph (SEM) image of a fabricated MEM switch; it is designed for digital logic applications and comprises two electrical switches and hence has two pairs of source/drain contact electrodes [6].

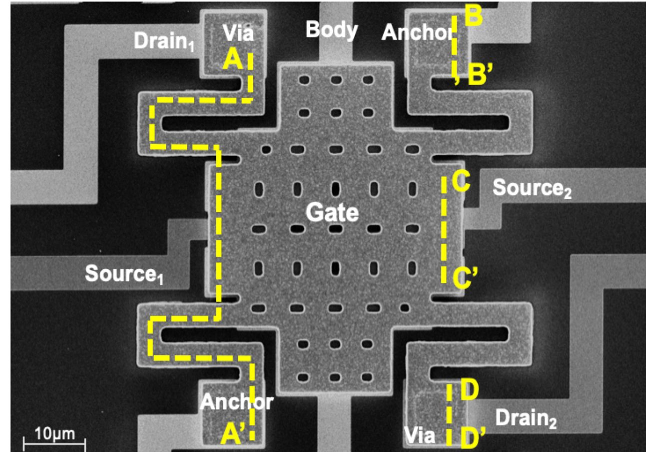


Figure 2.2: Plan-view scanning electron micrograph (SEM) image of a fabricated MEM switch.

As shown in the schematic cross-sectional views in Figure 2.3, the 2-contact (2C) device comprises a movable gate electrode suspended by four folded-flexure beams over a fixed body electrode. The drain electrodes are attached to and routed underneath the gate electrode and electrically insulated from it by an Al_2O_3 gate-dielectric layer.

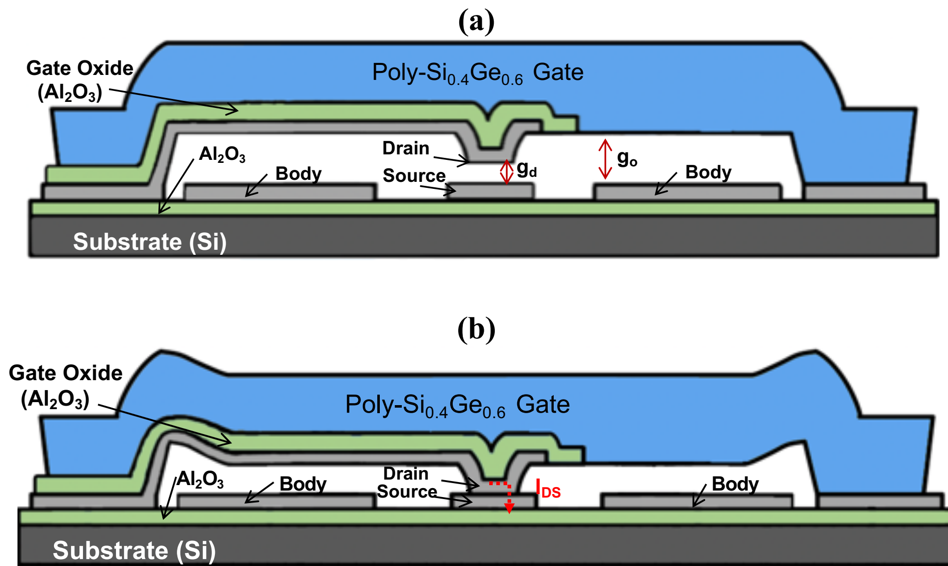


Figure 2.3: Schematic cross-sectional views along A-A' cutline in Fig. 2.2: (a) OFF–state (b) ON–state.

In the OFF–state (Fig. 2.3 (a)), the as-fabricated air gap separates the conductive drain and source electrodes, so that no current can flow between them. The movable gate is electrostatically actuated downward toward the body electrode when a voltage is applied between the gate and body (V_{GB}). When V_{GB} is larger than the turn-on voltage (V_{ON}), the electrostatic force overcomes the opposing spring restoring force of the folded flexure beams. This causes each of the drain

electrodes comes into physical contact with its underlying source electrode, allowing current (I_{DS}) to flow under the influence of an applied voltage between the drain and source electrodes (V_{DS}). This state is referred to as the ON–state (Fig. 2.3 (b)).

To turn off the switch, V_{GB} is reduced toward 0V so that the spring restoring force (F_{sp}) of the folded flexure beams actuates the movable structure upward, causing the drain and source electrodes to break contact. The voltage at which I_{DS} drops back to zero is referred to as the turn-off voltage (V_{OFF}). Due to contact adhesive force between the contacting surfaces, V_{OFF} is always smaller than V_{ON}. The hysteresis voltage (V_H) is defined as $V_{ON} - V_{OFF}$.

Figure 2.4 illustrates key steps of the MEM switch fabrication process along cutlines B-B', C-C', and D-D' of Figure 2.2. The MEM switches are fabricated using conventional planar processing techniques with a maximum substrate temperature below 450 °C for compatibility with post-CMOS integration [4]. Initially, an 80 nm-thick Al_2O_3 layer is deposited on a silicon substrate using an atomic layer deposition (ALD) system (Fig. 2.4 (a)). Next, a 60 nm-thick tungsten (W) layer is deposited by sputtering (Figure 2.4 (b)) and patterned to form the fixed body electrodes, drain electrodes, and source electrodes as shown in Figure 2.4 (c). Then, a 160 nm-thick sacrificial low-temperature deposited SiO_2 layer (LTO_1) is deposited using low pressure chemical vapor deposition (LPCVD), followed by contact "dimple" region definition (Fig. 2.4 (d)). A second sacrificial layer of 60 nm-thick SiO_2 layer (LTO_2) is then deposited and routing via regions are defined (Fig. 2.4 (e)). Note that the thickness of the second layer determines the as-fabricated air gap thickness between the drain electrode and the source electrode in the contact regions (g_d) while the combined thickness of LTO_1 and LTO_2 layers determines the as-fabricated actuation air-gap thickness (g_o) between the movable structure and the body electrode.

Afterwards, a second 60 nm-thick tungsten is deposited and patterned to form drain electrodes (Fig. 2.4 (f)). Next, a 55 nm-thick gate-insulating Al_2O_3 is deposited and patterned to define anchor regions (Fig. 2.4 (g)). A 1.9 µm-thick p-type heavily *in-situ* doped polycrystalline- $Si_{0.4}Ge_{0.6}$ (poly-SiGe) structural layer is then deposited using LPCVD and patterned to form the movable gate electrode (Fig. 2.4 (h)). Finally, the structural layer is released by selectively removing the sacrificial LTO layers using vapor hydrofluoric acid (HF). In this work, g_d is designed to be less than one third of g_o so that the MEM switch operates in non-pull-in mode, which is beneficial for minimizing V_H [12].

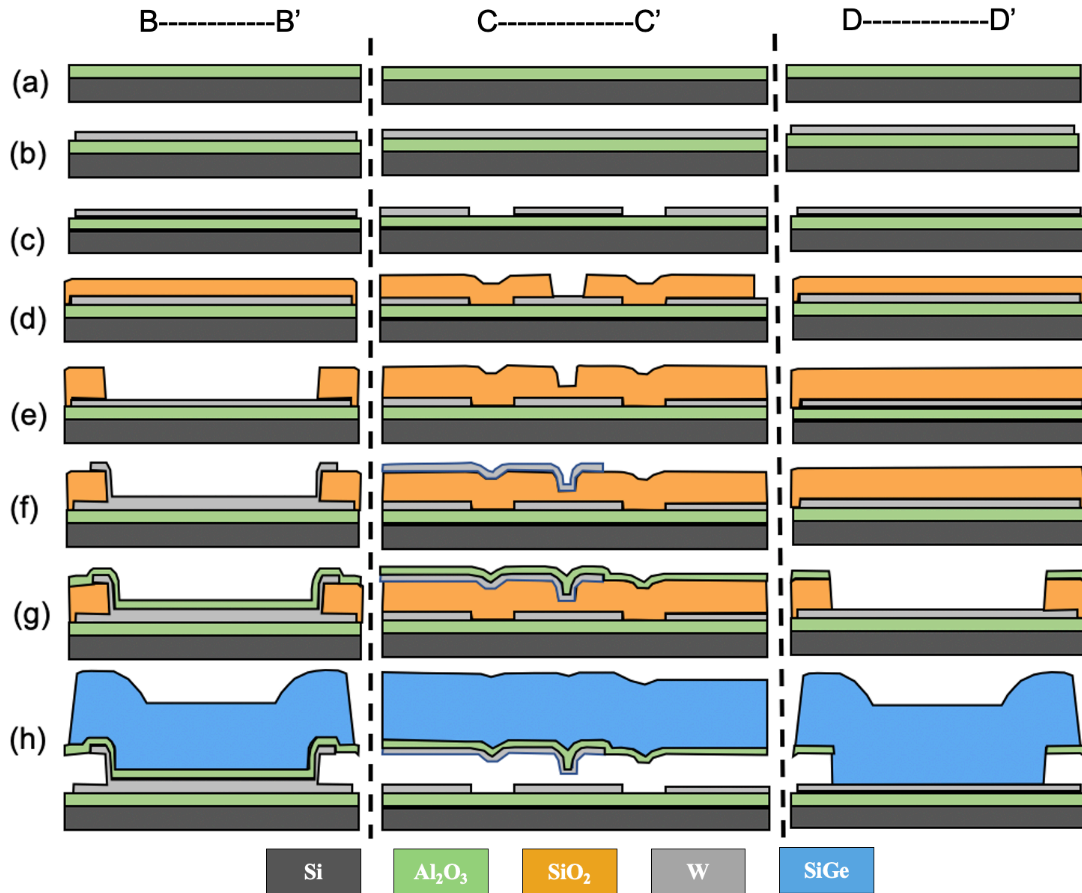


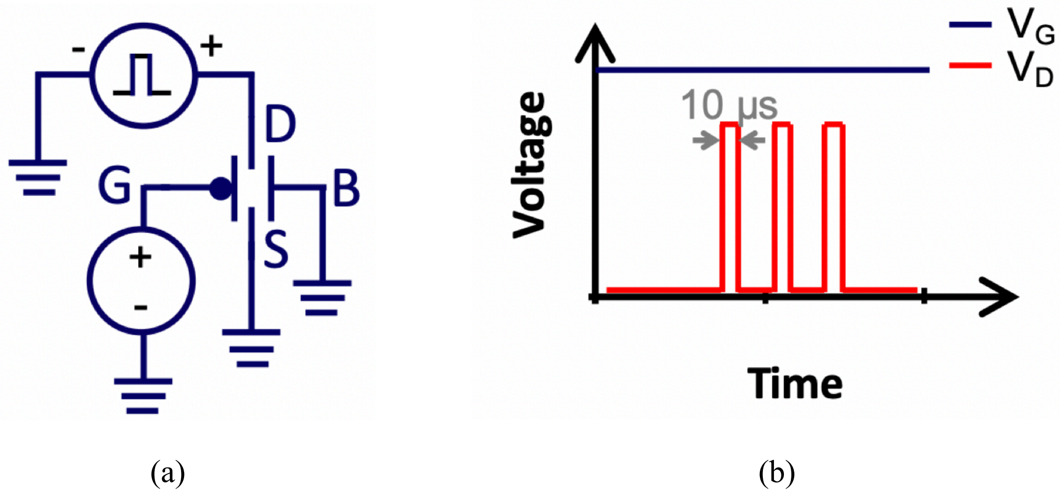
Figure 2.4: Schematic device cross-sections along B-B', C-C', and D-D' (cf. Figure 2.2) illustrating key MEM switch fabrication steps.

2.4 Non-Volatile MEM Switch Operation

A MEM switch can be operated as a non-volatile (NV) memory device if it remains in the ON-state after the applied voltage is removed, and if it can be reprogrammed (reset) to the OFF-state. The feasibility of controllably welding and un-welding the contacting electrodes of a MEM switch is proposed herein to enable a MEM switch to be used not only for digital logic applications but also for embedded non-volatile information storage.

In this work the MEM switch conductive electrodes are formed with tungsten, which oxidizes upon exposure to air. Prior to programming a MEM switch, the native oxide layers on the surfaces of the conductive electrodes must be electrically broken down to promote current conduction and thereby achieve low ON-state resistance (R_{ON}). Therefore, a pulsed cold-switching oxide breakdown procedure [13] was used in this work. As illustrated in Figure 2.5, the switch is first turned on by applying a voltage $V_{GB} = V_{ON} + 2V$; subsequently 3 voltage pulses (10 µs each) are applied across the drain-source contact. The initial magnitude of the drain-to-source voltage (V_{DS}) is 6V. Afterwards R_{ON} is checked by measuring the drain current (I_{DS}) with applied $V_{GB} > V_{ON}$ and

$V_{DS} = 0.5V$. If R_{ON} is larger than 1 kΩ then additional voltage pulses are applied with larger $V_{DS} =$ 8V. This native oxide breakdown procedure was found to be effective for consistently achieving sub-1kΩ R_{ON} [13].



(a) (b)

Figure 2.5: Pulsed cold switching oxide breakdown procedure: (a) circuit schematic and (b) voltage timing waveforms (not to scale). Adapted from [13].

2.4.1 Analytical Modeling

An analytical model for contact welding was derived in [14]; only the key results are presented herein. When a voltage pulse with amplitude V_{DS} and duration Δt is applied between the drain and source in the ON–state, the temperature of the contacting asperities on the contact electrode surfaces increases due to Joule heating. Assuming all contacting asperities are similar in size and that the size of an asperity is small compared to that of the switch, the rate of heating ($\dot{Q}$) at the asperity is:

$$\dot{Q} \approx \frac{{V_{DS}}^2}{R_{Total}^2} \cdot R_C \tag{2.11}$$

$$R_C = \rho_E/(2a_0) \tag{2.12}$$

where R_C is the contact resistance, ρ_E is the electrical conductivity, a_o is the radius of the contacting asperity, and R_{Total} is the total drain-to-source resistance.

From solving the heat equation (Equation (2.11)), the peak temperature at the contact (T_C) is:

$$T_C = V_{DS} \cdot \frac{1}{4l} \frac{R_C}{R_{Total}} \sqrt{f(\tau)} \tag{2.13}$$

where $l = 2.44\text{E-}8$ W·Ω /K^2 is the Lorenz number, $f(\tau) = 1 - e^{\tau}\left(1 - erf\left(\sqrt{\tau}\right)\right)$, $\tau = \Delta t/t_{TH}$, $t_{TH} = \rho_E c a_o^2/\lambda$ is the thermal time constant, λ is the thermal conductivity, and c is the specific heat capacity.

As T_C approaches the melting temperature of the contacting material, atomic movement occurs at the contacting asperities and the area of physical contact increases. If atomic movement is diffusive, the rate of asperity growth should have an Arrhenius dependence on temperature [15][16],

$$\frac{da}{dt} \propto \frac{D}{a^n} \Longrightarrow \frac{da}{dt} \propto \frac{D_o \exp\left(-E_A/k_B T\right)}{a^n} \tag{2.14}$$

Integrating Equation 2.14 over time *t*, the increase in the radius of the contacting asperity can be estimated as:

$$\Delta a_o \sim [D_o exp(-E_A/k_B T_C)]\Delta t \tag{2.15}$$

where D is the atomic diffusivity, D_o is the diffusion coefficient, a is the average radius of the contacting asperity, E_A is the activation energy, k_B is the Boltzmann constant and n is a factor between 2 and 4 that depends on the exact diffusion mechanism [14].

The drain and source electrodes become welded together when the metal-to-metal bonding energy is greater than the spring restoring energy – *i.e.,* when Δa_o reaches a critical value. Substituting Equation (2.13) into Equation (2.15), we obtain the following relationship between programming time and programming voltage [14]:

$$\Delta t \propto exp\left(\frac{B}{V_{DS}}\right) \tag{2.16}$$

where B is a constant that depends on the contacting material properties. In words, the time required to weld the contacting surfaces together decreases exponentially with increasing V_{DS}.

2.4.2 Program and Erase Operation

Guided by the analytical modeling, a range of program and erase conditions were explored in this work. To program a MEM switch, a voltage ($V_{GB} > V_{ON}$) is applied between the gate and body to actuate the switch into the ON–state; then two voltage pulses are applied between the drain and source electrodes, as illustrated in Figure 2.6 (b). In order to avoid permanent contact damage of the MEM switch due to the Joule heating, programming pulses (V_{DS}) are applied as two 1ms pulses with 50% duty cycle instead of a one 2ms pulse. This provides a time for the contacting asperities to cool back down to ambient temperature before the second pulse is applied.

The time required for the contact to cool down after the first programming voltage pulse is solved using heat dissipation analysis. In the heat dissipation process, thermal diffusivity (α) is the characteristic affecting the rate of heat transfer due to a temperature gradient, and it is calculated as follows:

$$\alpha = \frac{k}{\rho c_p} \tag{2.17}$$

where ρ is the material density and c_p is specific heat capacity of the material.

The heating time constant (τ) is given by

$$\tau = \frac{L^2}{\alpha} \tag{2.18}$$

where L is the length of the material through which heat flows to reach the heat sink from the heat generation point. In this case, L is the distance from the contacting drain/source electrode asperity to the wafer surface, which is approximately 12μm.

Using Equation (2.18) and properties of tungsten material, the estimated time it takes for the heat to dissipate is ~2μs.

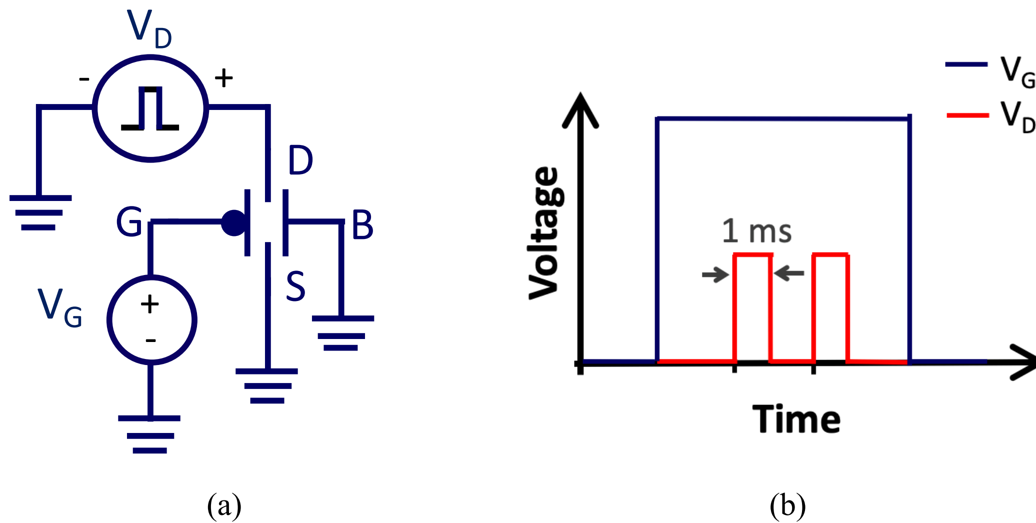


Figure 2.6: MEM switch program operation: (a) circuit schematic and (b) voltage timing waveforms (not to scale).

As shown in Figure 2.7 (a), the bottom conductive electrode (source) has larger surface roughness than the top conductive electrode (drain) because the top electrode is formed from a tungsten layer deposited onto the LTO_2 sacrificial layer conformally deposited over the bottom electrode resulting in a smoother surface. Due to surface roughness, physical contact between the conductive electrodes in the ON–state conductance occurs only at asperities. Current flow (in response to non-zero applied voltage V_{DS}) causes Joule heating, which raises the local temperature at the contact point(s), T_C. If T_C is sufficiently high to soften the contacting material resulting in atomic movement, the area of physical contact increases and new bonds can form between the contacting electrode surfaces. Thusly, the contacts can be effectively welded together as illustrated

in Figure 2.7(b). In the programmed state, the strength of the weld is greater than F_{sp} so that the switch remains ON when the gate voltage is removed, *i.e.,* V_{GB} is reduced to 0V.

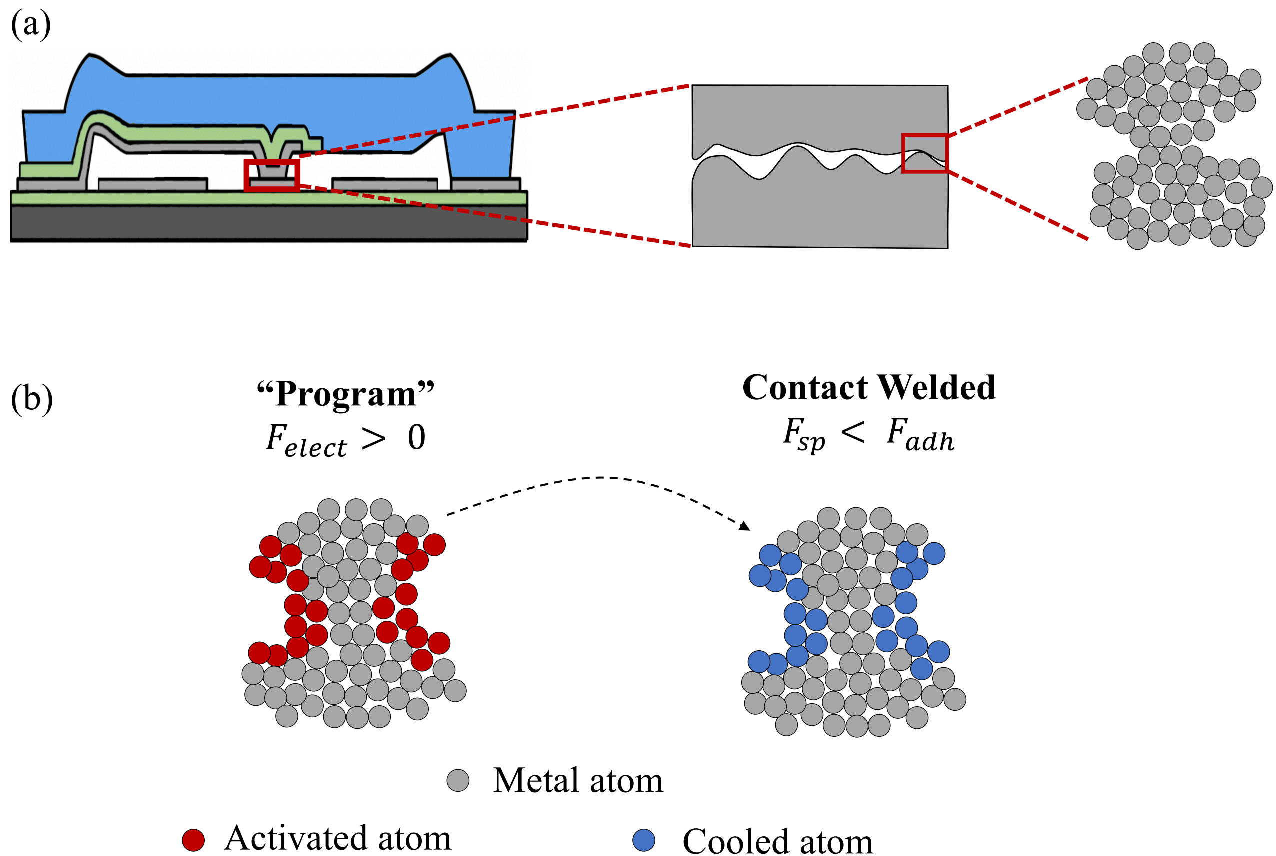


Figure 2.7: Contact welding in a MEM switch, induced by Joule heating: (a) Due to surface roughness, physical contact and current flow occur only at one or more asperities. (b) During "Program" operation, Joule heating softens the contact, resulting in contacting asperity growth and increased area of physical contact, effectively welding the electrodes together.

To erase the MEM switch, a voltage pulse is simply applied between the drain and source electrodes, as illustrated in Figure 2.8. If the Joule-heating induced temperature rise at the welded asperity is sufficient to soften the contacting material, *i.e.,* weaken the bonding strength between atoms, the spring restoring force can cause the contact to be broken as illustrated in Figure 2.9 so that the switch turns OFF.

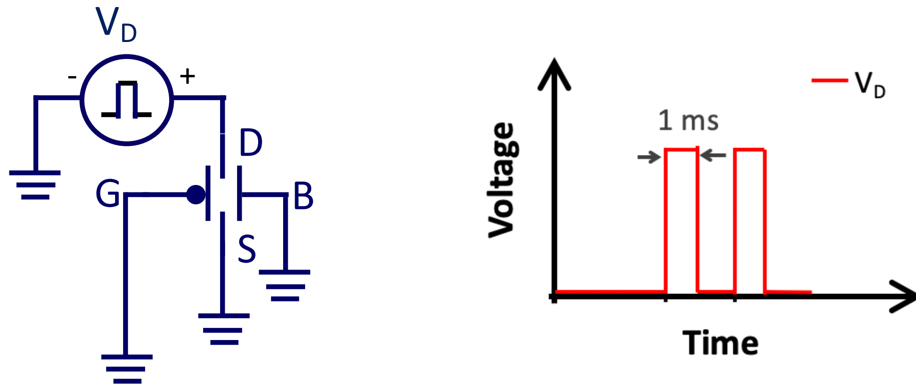


Figure 2.8: Illustration of erase operation (a) circuit schematic and (b) voltage timing waveform.

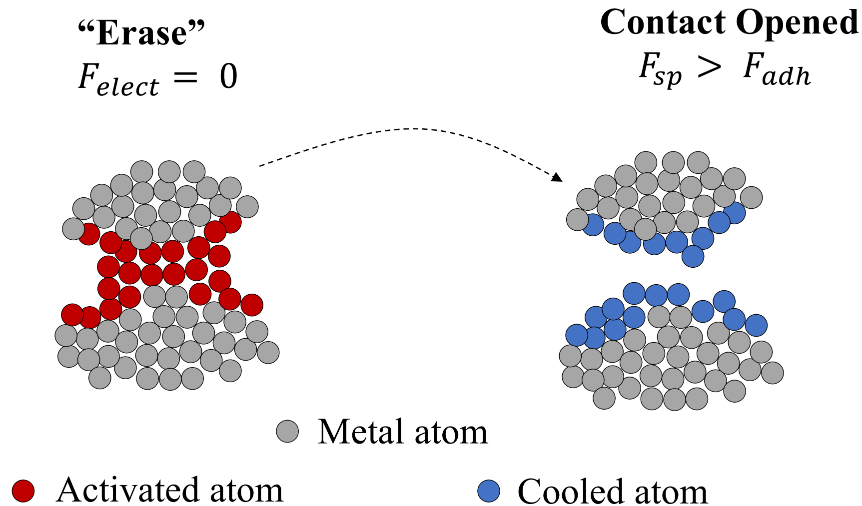


Figure 2.9: Contact un-welding in a programmed MEM switch: During "Erase" operation, Joule heating weakens the bonding strength so that the spring restoring force causes the contact to be broken.

The state of the MEM switch can be easily read by applying a small V_{DS} (*e.g.* 100 mV) and measuring I_{DS}, as illustrated in Figure 2.10. If any current flows (*i.e.,* $I_{DS} > 0$) then the switch is programmed. Program, erase and read conditions used for subsequent study are listed in Table 2.1.

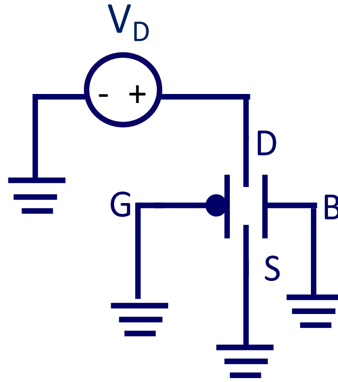


Figure 2.10: Circuit schematic diagram of read operation.

Table 2.1: Default parameters for MEM switch program, erase and read operations.

	Program	Erase	Read
$\mathbf{V_{GB}}$	V_{ON} + 2V	0 V	0 V
$\mathbf{V_{DS}}$	2.5 V	2.7 V	100 mV
Pulse Width	1ms	1ms	-
Pulse count	2	2	-

Figure 2.11 shows measured $I_D - V_{GB}$ characteristics for a MEM switch before and after the first program/erase (P/E) cycle. It is observed that V_{ON} is reduced by approximately 2.5 V. This can be explained by contact asperity growth (in height as well as in diameter) resulting from the programming and erase processes, which effectively reduces the contact air-gap thickness (g_d). The hysteresis voltage (V_H) is also reduced from ~0.3V to 0V. This is consistent with the explanation of larger contacting asperity height, resulting in larger average separation between the source and drain electrode surfaces in the ON-state, which results in reduced Van der Waals force.

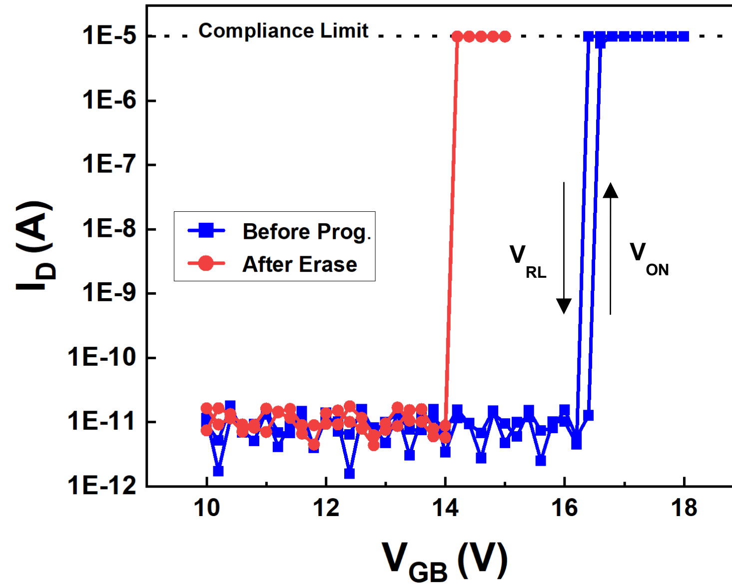


Figure 2.11: Measured I_{DS}-V_{GB} curves for a MEM switch before programming (blue) and after one P/E cycle (red).

This explanation is supported by atomic force microscopy (AFM) analyses of the source electrodes for a fresh (never-programmed) switch and a cycled (programmed & erased, P/E) switch, in Figure. 2.12 (a) and (b), respectively. (The movable electrode layer stack – including the drain electrodes – was physically removed using scotch tape to expose the fixed electrodes.) Figure 2.13 shows SEM images of (a) fresh and (b) P/E-cycled source electrode surfaces.

Both the AFM and SEM analyses show that the cycled source electrode has larger asperities within the contact dimple region than does the fresh source electrode. The surface height distribution for the fresh electrode is Gaussian whereas it is skewed for the cycled electrode, indicative of atomic movement to grow the size of the contacting asperities (Figures 2.12 (c) and (d)). The AFM line scan in Figure 2.12 (e) shows that a large asperity is approximately ~13 nm in height. Based on the measured reduction in V_{ON} and using Equation 2.5, the reduction in g_d is estimated to be 30 nm; therefore, similarly sized opposing asperities are expected on the surface of the cycled drain electrode. Since the average spacing between the contacting electrode surfaces is increased by ~30 nm due to asperity growth, Van der Waals force (which is the dominant component of the contact adhesive force [9]) in the ON–state is greatly reduced and hence V_H is greatly reduced.

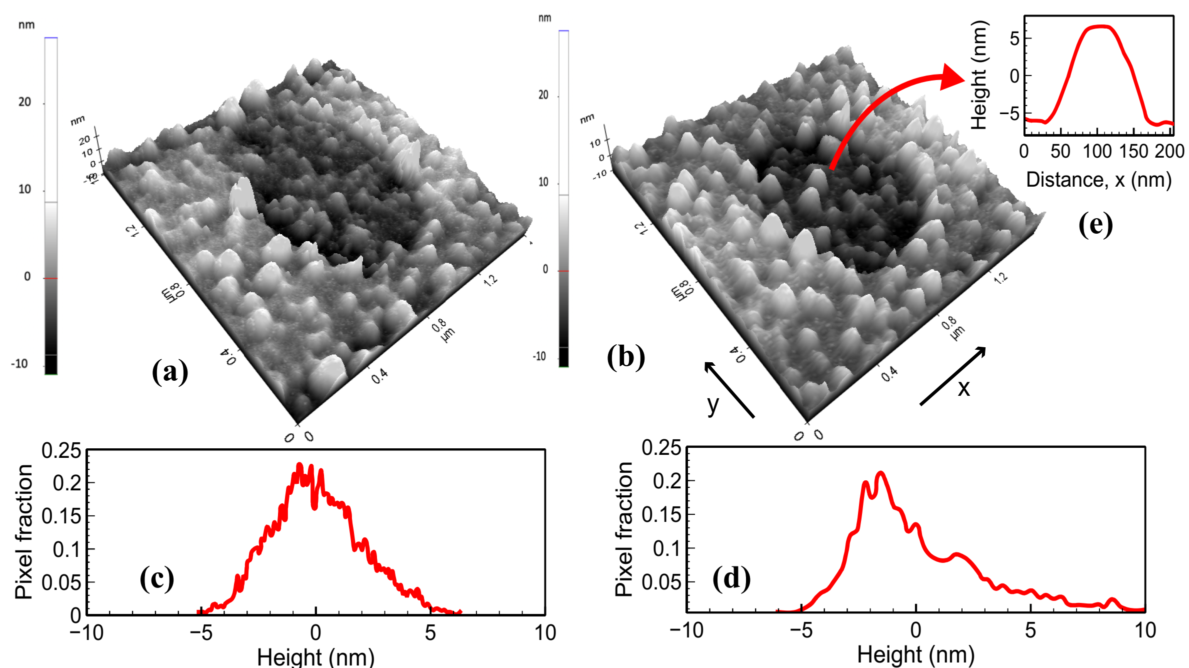


Figure 2.12: AFM analyses of MEM switch source electrode topography: Scans (1.6μm x 1.6μm) of electrode surfaces for (a) an unprogrammed contact, and (b) a P/E-cycled contact; height distributions within the contact dimple region for (c) the unprogrammed contact and (d) the P/E-cycled contact. (e) Height *vs*. distance for an asperity, showing that it is ~13 nm tall.

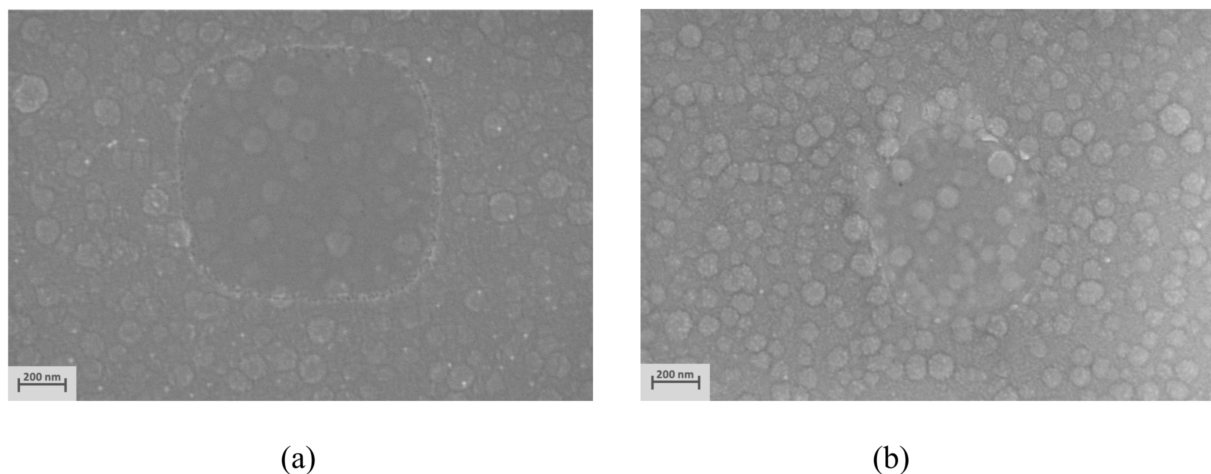
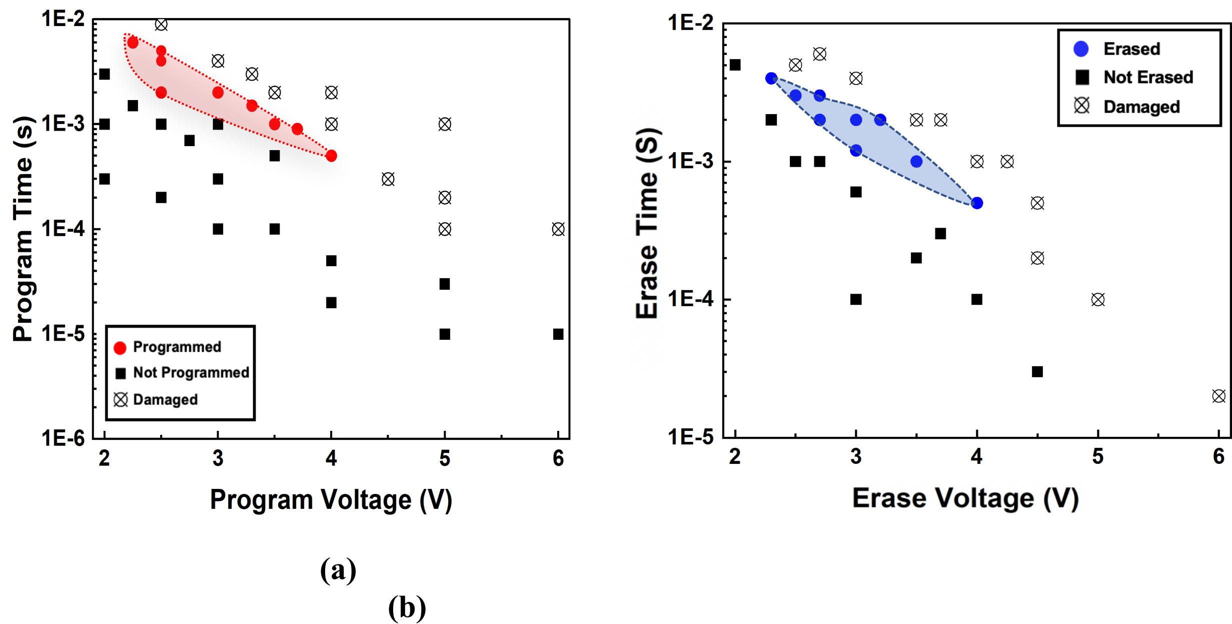


Figure 2.13: SEM images of MEM switch source electrodes: (a) for a fresh switch and (b) for a P/E-cycled switch. More prominent features (asperities) can be seen in the contact dimple for the cycled switch.

2.5 Demonstration of NV-MEM Switch Reprogrammability

A functionality that is desirable for embedded memory devices is re-programmability, *i.e.* the ability for the device to be reset to its original state and programmed again afterwards. The number of P/E cycles before failure (*i.e.* endurance) of NV-MEM switches is dependent on the program and erase conditions, as well as the electrical and mechanical properties of the structural and contact electrode materials [17]. Therefore, a range of program and erase conditions were investigated for tungsten electrode MEM switches in this work.

Figures 2.14 (a) and 2.14 (b) show the experimental results for program and erase operating windows, respectively. If the voltage and/or pulse width are too small, then the switch does not change state. If the voltage and/or pulse width are too large, then irreparable damage is caused to the contact so that the switch no longer functions properly. The shaded regions in Figure 2.14 indicate the voltage-time windows for reprogrammable operation.



(a)

(b)

Figure 2.14: Operating windows for MEM switch (a) program operation and (b) erase operation. The shaded regions indicate the combinations of V_{DS} pulse duration and voltage for successful re-program/erase operation. Damaged switches are not functional afterwards, *i.e.*, they cease to conduct any current.

Figure 2.15 shows measured $I_D - V_{GB}$ characteristics for a MEM switch before and after multiple program/erase cycles. It can be seen that V_{ON} varies substantially from cycle to cycle, indicating that the contact asperity height and shape change with each cycle. Nevertheless, low programmed-state resistance (< 1kΩ) is maintained with each cycle, as can be seen from the endurance testing results in Figure 2.16.

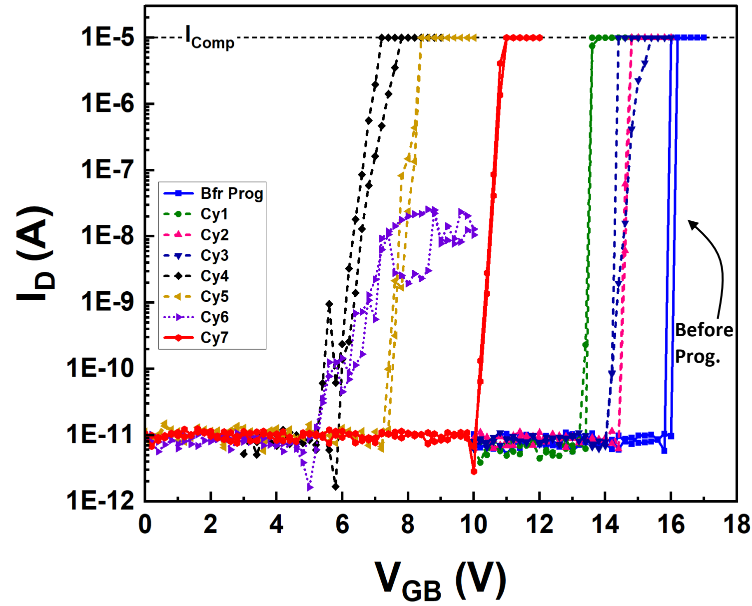


Figure 2.15: Evolution of MEM switch I_{DS}-V_{GB} characteristic through multiple program/erase cycles. The change in V_{ON} from cycle to cycle is non-deterministic because it depends on the shape and height of the contacting asperities.

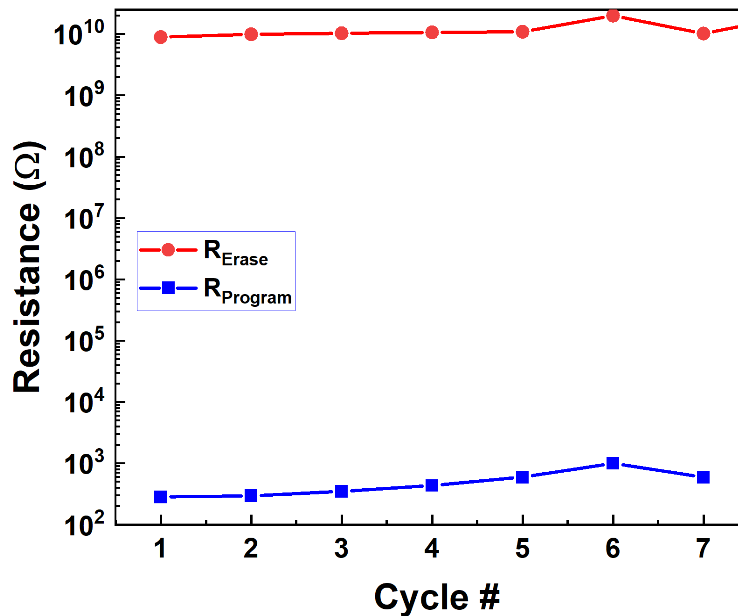


Figure 2.16: Evolution of the measured resistance of a MEM switch in the programmed state ($R_{program}$) and in the erased state (R_{erase}), over multiple P/E cycles. $R_{program}$ is consistently below 1kΩ. R_{erase} essentially indicates an open circuit, as the measurement limit due to noise is ~10^9 Ω.

It is also important to consider the basic data-retention requirement for non-volatile memory devices. Retention refers to the capability to retain a state (programmed or erased) over time with no voltages applied, across a wide temperature range. Figure 2.17 shows the results of data retention testing of multiple MEM switches (1 erased, 6 programmed) at high temperature (200˚C) in vacuum (~1 µTorr) over 5 hours. The programmed-state resistance is slightly larger at elevated temperature due to degraded electron mobility, but it is very stable over time; this indicates that the welded asperities are very stable, providing for excellent data retention.

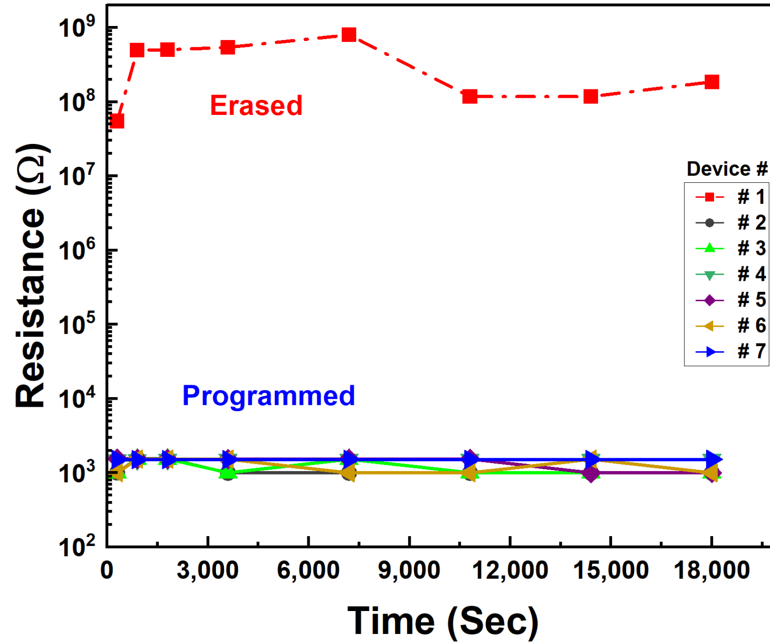


Figure 2.17: Data retention testing of multiple MEM switches (1 erased, 6 programmed) at high temperature (200˚C) in vacuum (~1 µTorr). The current noise floor is higher at elevated temperature, resulting in smaller apparent R_{erase}.

2.6 Simulation Study

To facilitate the understanding of MEM switch program and erase processes, COMSOL simulation of Joule heating in a MEM switch in the ON-state was performed. As shown in Figure 2.18, a welded-asperity structure with double napped conical shape – 100 nm wide at the base, 50 nm wide at the middle, and 13 nm tall for one side asperity – is simulated. Structural dimensions are selected to be consistent with AFM measurements (cf. Figure 2.12) and with measured electrical resistance values (cf. Figure 2.16). As indicated by the energy-dispersive X-ray (EDX) analysis results shown in Figure 2.19 and as discussed in [12], a thin native oxide forms on the surface of the tungsten electrode. Therefore, the simulated structure comprises 1.5 nm-thick tungsten trioxide (WO_3)-coated drain/source contact asperities that are welded together. Table 2.2 summarizes material properties and values used in this simulation.

Figure 2.20 shows how the peak contact temperature changes with time when a 2.7V, 1ms erase pulse is applied at time $t = 0$ with a ramp time of 0.1 μs. The simulation results show that the peak temperature at the contacting asperity (T_C) reaches 1366˚C, which is close to the melting temperature of WO_3 and certainly is sufficient to soften the contact material and cause atomic movement. Note that the heat is largely confined to the WO_3 contacting layers because the thermal conductivity of WO_3 is two orders of magnitude lower than that of tungsten.

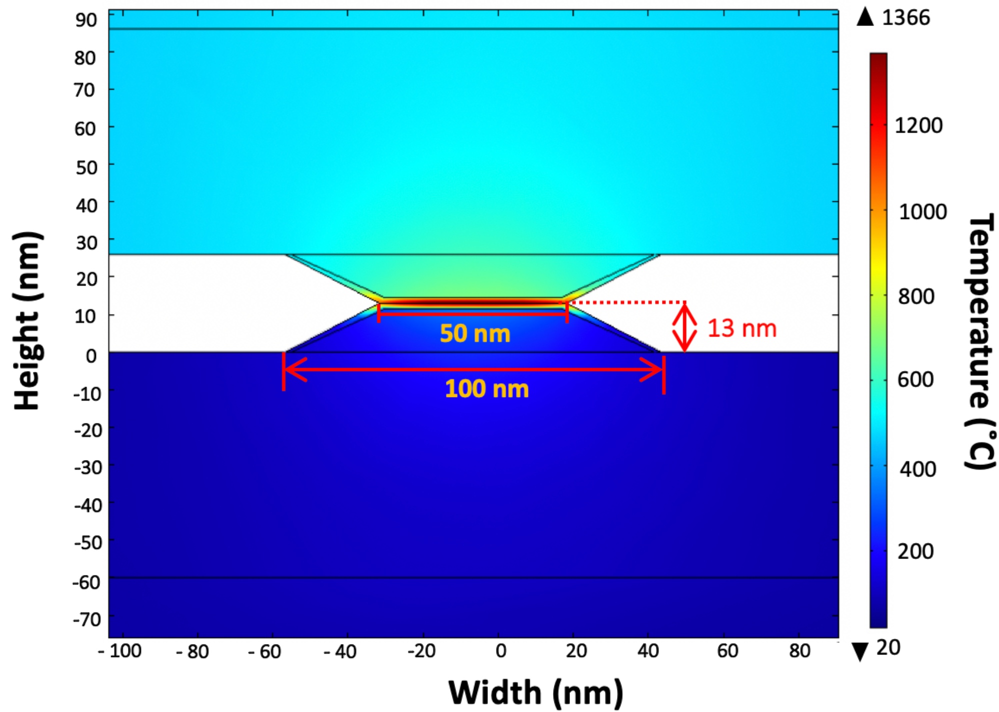


Figure 2.18: COMSOL simulation of Joule heating in a MEM switch during erase operation (cf. Table 2.1).

Table 2.2: Key material properties and values used for simulation.

	Electrical Conductivity σ (S/m)	**Thermal Conductivity** κ (W/(m K))	**Heat Capacity** Cp (J/(Kg K))	**Density** ρ (kg/m^3)
W	2x10^7	174	132	19,350
WO_3	1.6x10^3	1.63	170	7,160
Al_2O_3	0	35	730	3,965
Poly-SiGe	1.67x10^5	11	464	4,740
Si Substrate	4	1.31x10^2	700	2,329

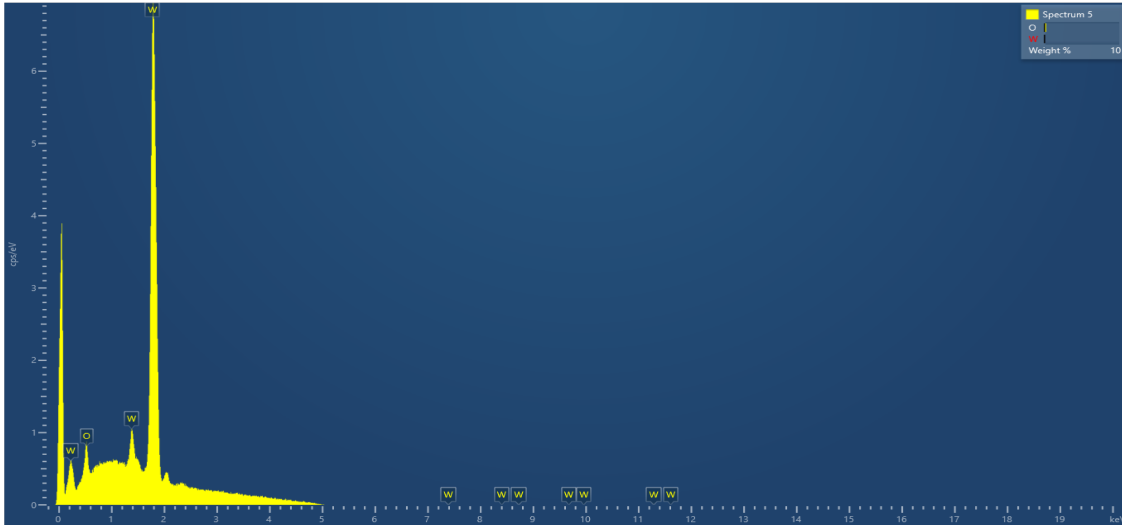


Figure 2.19: Measured EDX spectrum of source electrode showing the presence of W and O at the surface.

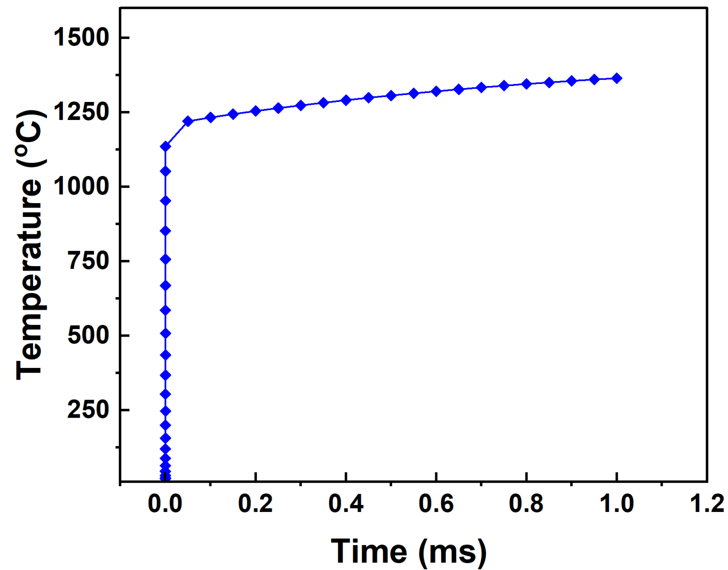


Figure 2.20: Evolution of peak contact temperature (cf. Fig. 2.18) during erase operation.

2.7 Discussion

Simulations show that the properties of the structural and conductive/contacting electrode materials play an important role in determining the endurance of a NV-MEM switch. The SEM image in Figure 2.21 shows how repeated P/E-cycling can cause permanent damage to the heated

portions of the switch (top left corner of the movable structure along with the folded-flexure beam and drain electrode). Alternative contacting electrode materials with lower melting temperature than tungsten can be explored in the future to reduce the thermal stress sustained by the other layers of the MEM switch during program/erase operation, to alleviate this issue. Table 2.3 lists the relevant properties of alternative contacting materials.

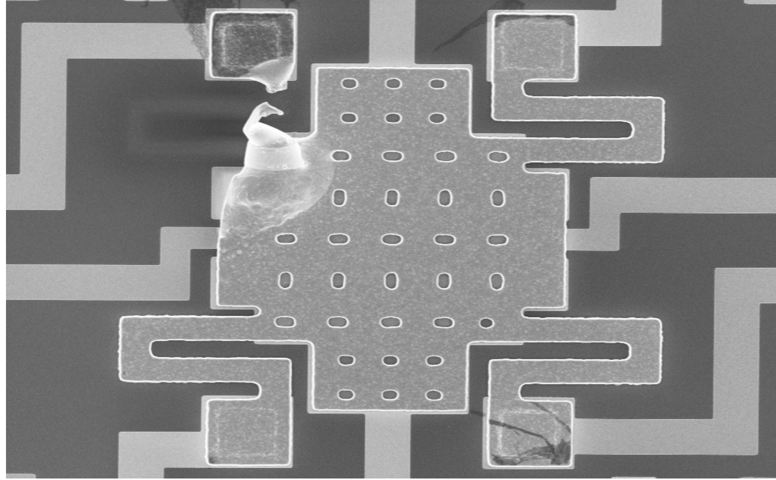


Figure 2.21: Top view SEM image of a MEM switch that failed after 6 program/erase cycles.

Table 2.3: Mechanical and electrical properties of alternative contacting electrode materials [18]

Contact Material	**Tungsten (W)**	**Aluminum (Al)**	**Niobium (Nb)**	**Nickle (Ni)**	**Ruthenium (Ru)**	**Gold (Au)**
Melting Point (K)	3695	933.5	2750	1728	2607	1337
Specific heat capacity (J/kg.K)	132	904	265	445	238	129
Resistivity (Ωm)	5.6×10^{-8}	2.6×10^{-8}	1.5×10^{-7}	7×10^{-8}	7×10^{-8}	2.2×10^{-8}
Thermal conductivity (W/(mK))	170	235	54	91	120	320

2.8 Summary

In this chapter MEM switches designed for digital logic application are demonstrated to be able to function as non-volatile memory devices through controlled contact welding and un-welding processes. It is demonstrated that MEM switches can be programmed and erased with relatively small voltage (<3V). Re-programmability with consistently low programmed-state resistance, and excellent (essentially infinite) retention time at elevated temperature, are experimentally demonstrated. The experimental findings in this work indicate that MEM switches are promising for low-cost implementation of ultra-low-power integrated microsystems. Further work is needed to optimize the contact and structural materials to achieve greater program/erase cycling endurance.

2.9 References

[1] Vailshery, Lionel Sujay. "IOT Connected Devices Worldwide 2019-2030." *Statista*, 22 Aug. 2022. URL: https://www.statista.com/statistics/1183457/iot-connected-devices-worldwide/.

[2] "Big Data for Sustainable Development." *United Nations*, United Nations. URL: https://www.un.org/en/global-issues/big-data-for-sustainable-development.

[3] Chuang Qian. "Electro-Mechanical Devices for Ultra-Low-Power Electronics". PhD thesis, EECS Department, University of California, Berkeley, May 2017. URL: http://www2.eecs.berkeley.edu/Pubs/TechRpts/2017/EECS-2017-21.html.

[4] X. Hu, S. F. Almeida, Z. Alice Ye and Tsu-Jae King Liu, "Ultra-Low-Voltage Operation of MEM Relays for Cryogenic Logic Applications," *IEEE International Electron Devices Meeting (IEDM),* 2019, pp. 34.2.1-34.2.4, doi: 10.1109/IEDM19573.2019.8993629.

[5] Benjamin Osoba, Bivas Saha, Liam Dougherty, Jane Edgington, Chuang Qian, Farnaz Niroui, Jeffrey H. Lang, Vladimir Bulovic, Junqiao Wu, and Tsu-Jae King Liu, "Sub-50 mV NEM relay operation enabled by self-assembled molecular coating," *IEEE International Electron Devices Meeting (IEDM)*, 2016, pp. 26.8.1-26.8.4, doi: 10.1109/IEDM.2016.7838489.

[6] Z. A. Ye, S. Almeida, M. Rusch, A. perlas, W. Zhang, U. Sikder, J. Jeon, V. Stojanović, and Tsu-Jae King Liu, "Demonstration of 50-mV Digital Integrated Circuits with Microelectromechanical Relays," *IEEE International Electron Devices Meeting (IEDM)*, 2018, pp. 4.1.1-4.1.4, doi: 10.1109/IEDM.2018.8614663.

[7] K. Kato, V. Stojanović, and Tsu-Jae King Liu, "Embedded Nano-Electro-Mechanical Memory for Energy-Efficient Reconfigurable Logic," *IEEE Electron Device Letters*, vol. 37, no. 12, pp. 1563-1565, Dec. 2016, doi: 10.1109/LED.2016.2621187.

[8] U. Sikder, L. P. Tatum, T.-T. Yen, and Tsu-Jae King Liu, "Vertical NV-NEM Switches in CMOS Back-End-of-Line: First Experimental Demonstration and Array Programming Scheme," *IEEE International Electron Devices Meeting (IEDM)*, 2020, pp. 21.2.1-21.2.4, doi: 10.1109/IEDM13553.2020.9372116.

[9] R. Maboudian and R. T. Howe, "Critical review: Adhesion in surface micromechanical structures," *Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures Processing, Measurement, and Phenomena*, vol. 15, no. 1, pp. 1–20, 1997; doi: 10.1116/1.589247.

[10] J. Yaung, L. Hutin, J. Jeon, and Tsu-Jae King Liu, "Adhesive force characterization for mem logic relays with sub-micron contacting regions," *Journal of Microelectromechanical Systems*, 23(1):198-203, 2014. doi:10.1109/JMEMS.2013.2269995.

[11] S. Saha, M. S. Baghini, M. Goel, and V. R. Rao, "Sub-50-mV nanoelectromechanical switch without body bias," *IEEE Transactions on Electron Devices*, vol. 67, no. 9, pp. 3894–3897, 2020. doi: 10.1109/TED.2019.2951615.

[12] E. Falicov, J. Marvin, Z. A. Ye, S. F. Almeida, D. Contreras, Tsu-Jae King Liu, and M. Spencer, " Breakdown and Healing of Tungsten-Oxide Films on Microelectromechanical Relay Contacts," *Journal of Microelectromechanical Systems*, vol. 31, no. 2, pp. 265-274, April 2022, doi: 10.1109/JMEMS.2021.3135259.

[13] Alice Ye. "Millivolt Micro-Electro-Mechanical Relay Devices Circuits". PhD thesis, EECS Department, University of California, Berkeley, Dec 2021. URL: http://www2.eecs.berkeley.edu/Pubs/TechRpts/2021/EECS-2021-239.html.

[14] H. Kam, E. Alon, and Tsu-Jae King Liu, "A predictive contact reliability model for MEM logic switches," *International Electron Devices Meeting (IEDM)*, 2010, pp. 16.4.1-16.4.4, doi: 10.1109/IEDM.2010.5703375.

[15] J. R. Black, "Electromigration—A brief survey and some recent results," *IEEE Transactions on Electron Devices*, vol. 16, no. 4, pp. 338-347, April 1969, doi: 10.1109/T-ED.1969.16754.

[16] M. Shatzkes and J. R. Lioyd, "A model for conductor failure considering diffusion concurrently with electromigration resulting in a current exponent of 2," *Journal of Applied Physics.* 59, 3890-38931986. URL: https://doi.org/10.1063/1.336731.

[17] Urmita Sikder. "Nanoelectromechanical Switch Design and Implementation in Back-End-of-Line Technology". PhD thesis, EECS Department, University of California, Berkeley, May 2022. URL: https://www2.eecs.berkeley.edu/Pubs/TechRpts/2022/EECS-2022-20.pdf.

[18] *The Photographic Periodic Table of the Elements*, URL: https://periodictable.com/.

Chapter 3

Investigation of Floating-Gate MEM Switch for Non-Volatile Memory Application

3.1 Introduction

With the rise of cloud computing and the Internet of Things, the traditional von Neumann computing architecture – in which digital computation and information storage functions are physically separated – is facing new challenges related to the large amount of data and the increasing burden of communication between the memory and the processing unit [1]. Therefore, embedded memory technology is of interest for faster and more energy-efficient computing. Ideally, both the switches used for digital computation and the non-volatile memory (NVM) devices should be operated with very low energy [2]. Electrostatically actuated mechanical switches can achieve immeasurably low OFF-state leakage current (I_{OFF}) and abrupt switching behavior enabling sub-50 mV operation at room temperature [3, 4] and sub-25 mV operation at 77 K [5]; hence they are promising for ultra-low-power digital computing.

In Chapter 2, operation of MEM switches as NVM devices using controlled contact welding and unwelding was demonstrated. In this chapter, a different approach is investigated, by fabricating MEM switches with a floating gate electrode embedded within the gate dielectric to store electronic charge. Unlike flash memory devices, MEM switches have an air gap in the off state that prevents charge leakage, which could provide for long data retention time both at room temperature and in a harsh environment (such as space). In Section 3.2 the structure and operating principle of a floating gate (FG) MEM switch for embedded memory application is presented. Section 3.3 presents a device fabrication process and initial experimental results. Improvements to the device design and fabrication process are then proposed in Section 3.4. Section 3.5 summarizes this chapter.

3.2 Proposed FG-MEM NV Switch Design

Figure 3.1 illustrates the FG-MEM switch design used in this work, comprising 2 pairs of source/drain electrodes (therefore 2 separate contacts), as well as corresponding schematic cross-sectional views in the ON-state and in the OFF-state. The FG-MEM switch comprises a movable gate stack consisting of the "control" gate and the charge-storing floating gate which is electrically insulated by dielectric material. The movable gate stack is suspended by four folded-flexure beams over a fixed body electrode. Each drain electrode extends underneath the gate electrode and is attached to it via an insulating dielectric layer. Herein the thinner dielectric material layer between the control gate and the floating gate is called the tunnel oxide while the thicker dielectric material layer between the floating gate and the drain electrode is called the gate oxide. The fixed source electrodes are co-planar with the body electrode.

This switch operates similarly as the MEM switches discussed in Chapter 2: in the OFF–state, an air gap physically separates each drain electrode from its underlying source electrode, so that no current (I_{DS}) flows; in the ON–state, the movable gate stack is actuated downward by electrostatic force between the gate and body such that each drain electrode comes into contact with its underlying drain electrode to allow current flow with applied drain/source voltage difference (*i.e.*, $V_{DS} \neq 0V$).

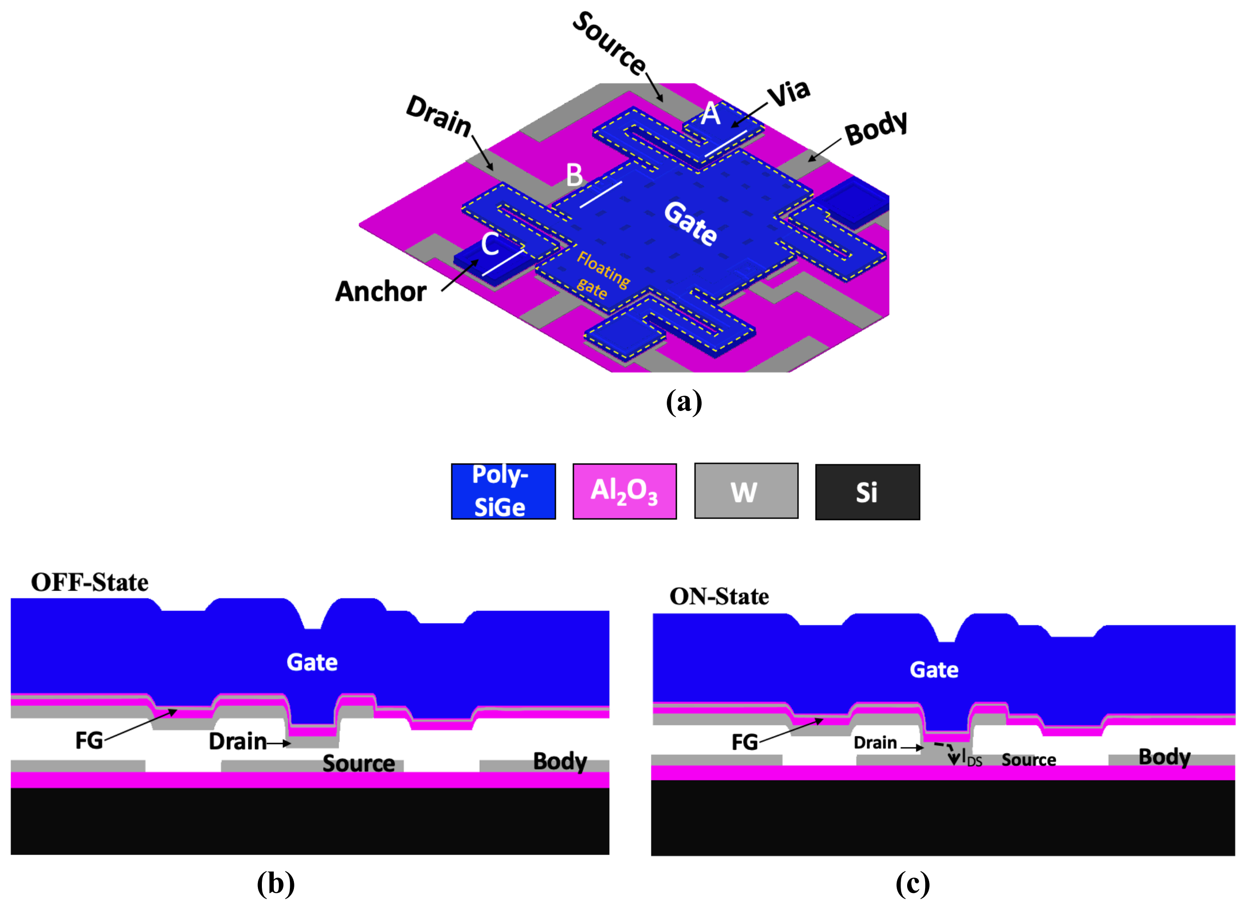


Figure 3.1: (a) Schematic isometric view of the floating-gate MEM switch studied in this work and cross-sectional views along the 'B' cutline in (b) OFF-state (c) ON-state.

Similarly as in a metal-oxide-semiconductor (MOS) transistor based flash memory cell [6–8], the control gate is used to turn the device ON and OFF, while the floating gate is used to store electronic charge to change the turn-ON voltage (V_{ON}). Fowler-Nordheim (F-N) tunneling is the process used to transfer charges to or from the floating gate though a thin tunneling oxide [9–12]. F-N tunneling occurs when a large electric field is induced across a thin dielectric layer (approximately 10 nm or less in thickness) [12] resulting in a triangular (*vs*. rectangular) potential barrier which increases the probability of electrons tunneling through the barrier [10][13][14][15].

The generalized FN tunneling current density (J_{FN}) is given by the following equation:

$$J_{FN} = \frac{q^3 {V_{ox}}^2}{8\pi h \phi_B {t_{ox}}^2} \exp\left(-\frac{4\sqrt{2m_e^*}\,\phi_B^{3/2} t_{ox}}{3hqV_{ox}}\right) \tag{3.1}$$

where q is the electronic charge, V_{ox} is the voltage dropped across the dielectric layer, t_{ox} is the thickness of the dielectric layer, h is Planck's constant, ϕ_B is the barrier height, and m_e^* is the electron effective mass. A detailed derivation of this equation can be found in [10].

A. *Program, Erase, Read Operation*

When electronic charge is injected to or removed from the floating gate, V_{ON} changes. Therefore, information can be encoded in the turn-ON voltage of a FG-MEM switch. Ideally, the total amount of charge on the floating gate does not change after the applied voltage is removed, to achieve non-volatile information storage.

If a large programming voltage (V_{Prog}) pulse were to be simply applied between the control gate and body electrodes, as in a conventional floating-gate flash memory cell, the MEM switch would turn on and off, physically making and breaking contact. To avoid unnecessary contact wear, the switch should be maintained in the OFF–state during program and erase operations. Therefore, in this work, programming and erasing voltage pulses were applied to the drain electrode while the other electrodes (control gate, source, and body) were biased at ground (0 V).

The application of a positive voltage to the drain electrode results in an electric field in the gate oxide and also an electric field in the tunnel oxide, attracting electrons toward the drain. Because the tunnel oxide is thin, electrons can tunnel from the control gate into the floating gate. (The gate oxide is much thicker, so electron tunneling into the drain electrode is negligible.) Similar to the situation of a floating gate MOS transistor [6][11], the electric field within the tunnel oxide is

$$E = \frac{V_{FG}}{t_{ox}} \tag{3.2}$$

where t_{ox} is tunnel oxide thickness and V_{FG} is the floating gate voltage, which can be calculated using the relationship

$$Q = CV \tag{3.3}$$

where Q is the charge stored, C is the capacitance of the storage node, and V is the voltage at the storage node. Considering the general case in which there is net charge stored on the floating gate (FG), *i.e.,* $Q_{FG} \neq 0$,

$$Q_{FG} = C_{FG_D}(V_{FG} - V_D) + C_{FG_CG}\,(V_{FG} - V_{CG}) + C_{FG_B}\,(V_{FG} - V_B) \tag{3.4}$$

where Q_{FG} is the amount of stored charge, C_{FG_D} , C_{FG_CG}, C_{FG_B} are the capacitances between the FG and drain, between the FG and control gate, and between the FG and body electrode, respectively, as illustrated in Figure 3.2. V_{FG} is the FG voltage and V_D, V_{CG}, V_B are the voltages at the drain, control gate, and body electrode, respectively.

While the programming voltage (V_{Prog}) pulse is applied the drain electrode (*i.e.,* $V_{Prog} = V_D$) the other electrodes are grounded ($V_{CG} = V_B = 0V$), so

$$Q_{FG} = C_{FG_D}\left(V_{FG} - V_{Prog}\right) + C_{FG_CG}\,(V_{FG}) + C_{FG_B}\,(V_{FG}) \tag{3.5}$$

$$Q_{FG} = \left[C_{FG_D} + C_{FG_CG} + C_{FG_B}\right] V_{FG} - C_{FG_D}\,V_{Prog} \tag{3.6}$$

Solving for V_{FG},

$$V_{FG} = V_{Prog} \cdot \left[\frac{C_{FG_D}}{C_{FG_D}+C_{FG_CG}+C_{FG_B}}\right] + \left[\frac{Q_{FG}}{C_{FG_D}+C_{FG_CG}+C_{FG_B}}\right] \tag{3.7}$$

Herein the capacitive coupling ratio α is defined as follows:

$$\alpha = \frac{C_{FG_D}}{C_{FG_D}+C_{FG_CG}+C_{FG_B}} = \frac{C_{FG_D}}{C_T} \tag{3.8}$$

where $C_T = C_{FG_D} + C_{FG_CG} + C_{FG_B}$

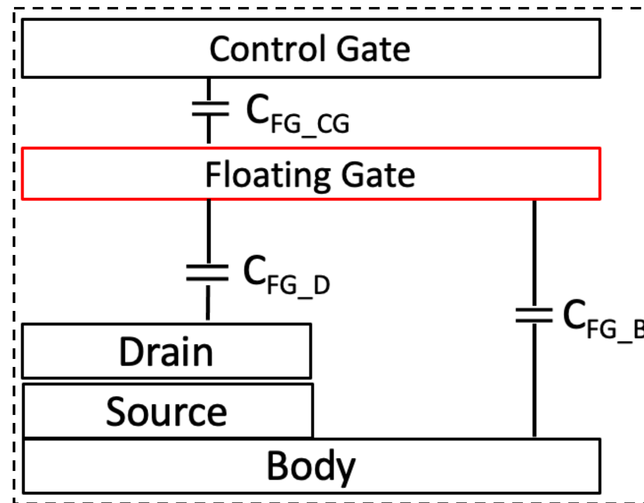


Figure 3.2: Schematic illustrating capacitances within a FG-MEM switch in the ON-state.

If electrons are stored on the floating gate, then a larger positive voltage or a smaller negative voltage must be applied to the gate electrode relative to the body electrode to induce the amount of electrostatic force needed to actuate the movable gate stack downward such that the drain electrodes come into contact with their respective source electrodes, *i.e.*, to turn on the FG-MEM switch. When a positive shift in turn-ON voltage ($\Delta V_{\mathrm{ON}} > 0$) is observed, the switch is referred as programmed or in the '1' state. To erase the device, a negative voltage pulse is applied to the drain electrode to cause electrons to tunnel back to the gate. After programming and erase operations, if V_{ON} returns to its initial value then the switch is referred as erased or in the '0' state.

The shift in V_{ON} due to (negative) charge stored on the floating gate is:

$$\Delta V_{\mathrm{ON}} = -\frac{Q_{FG}}{C_{FG_CG}} \tag{3.9}$$

$C_{FG_CG} = A\varepsilon_{Al_2O_3}/t_{ox}$ where A is the area of the floating gate (which is roughly the gate actuation area) and $\varepsilon_{Al_2O_3}$ is the dielectric permittivity of the tunnel oxide.

3.3 Experimental Investigation of FG-MEM NV Switches

The fabrication process and initial experimental results for FG-MEM switches are presented in this section.

3.3.1 Device Fabrication

The device fabrication process used in this work is very similar to that used to fabricate conventional MEM switches (cf. Figure 2.4). The main difference is a thin tungsten metal layer added between gate-insulating layer formation steps, as described in detail below.

An 80 nm-thick electrically insulating Al_2O_3 layer is deposited over the silicon wafer substrate by atomic layer deposition (ALD) (Figure 3.3 (a)). Next, a 60 nm-thick tungsten (W) layer is deposited by sputtering. Lithography and reactive ion etching (RIE) processes are performed to pattern the W layer to form fixed electrodes (body electrodes, drain electrodes, and source electrodes) as shown in Figure 3.3(b). Subsequently, a 160 nm-thick sacrificial SiO_2 layer (LTO_1) is deposited using low pressure chemical vapor deposition (LPCVD), followed by contact "dimple" region definition (Figure 3.3 (c)). A second 60 nm-thick sacrificial SiO_2 layer (LTO_2) is deposited and patterned to form via regions (Figure 3.3 (d)).

To form the movable gate stack, 60 nm-thick W drain electrodes are formed (Figure 3.3 (e)). Afterwards, instead of a 55 nm-thick Al_2O_3 gate insulator layer, a 45 nm-thick Al_2O_3 gate insulator layer is deposited. Then a 5 nm-thick W layer is deposited to form the floating gate. Next, a 10 nm-thick Al_2O_3 tunnel oxide layer is deposited and etched together with the LTO layers to define the structural anchor regions (Figure 3.3 (f)).

The control gate electrode is formed by depositing a 1.9 μm-thick p-type heavily *in-situ* doped polycrystalline-$Si_{0.4}Ge_{0.6}$ (poly-SiGe) layer by LPCVD and patterning it. Finally, the gate stack is released by selectively removing the sacrificial LTO layers using vapor hydrofluoric acid (HF) (Figure 3.3 (g)).

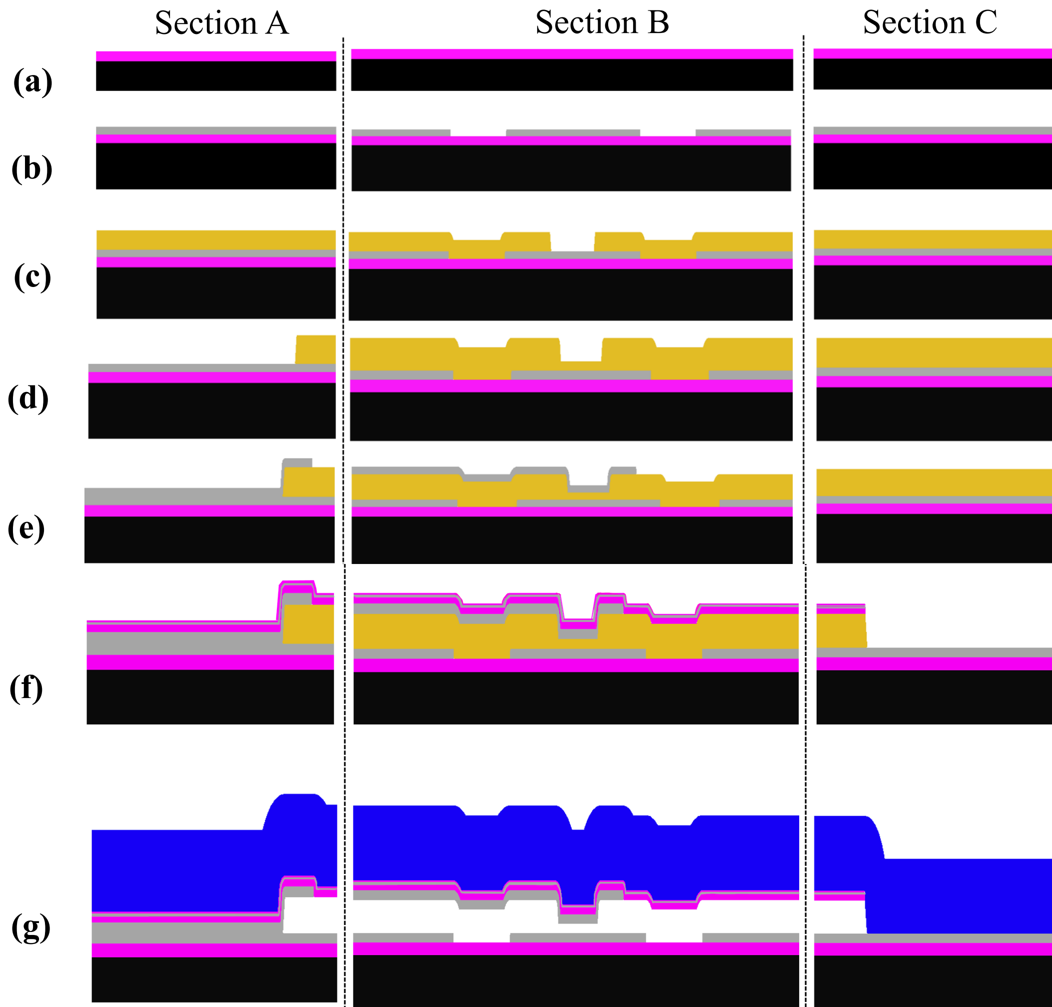


Figure 3.3: Schematic cross-sectional views along A, B and C cut-lines in Figure 3.1(a) illustrating the fabrication process steps for the FG-MEM switches studied in this work.

Table 3. 1: FG-MEM design parameter values used in this study.

Design parameter	Symbol	Values
Poly-$Si_{0.4}Ge_{0.6}$ Thickness	t	1.9 μm
Suspension Beam Width	W	2 μm
Suspension Beam Length	L	12 μm
Actuation Gap Thickness	g_0	220 nm
Contact Gap Thickness	g_d	60 nm
Gate Actuation Area	A	1062 μm^2
Total Drain Electrode Area (Drains on left and right)	A_{Drain}	~216 μm^2
Total Contact Dimple Area	A_{CONT}	2 μm^2

3.3.2 Electrical Characterization

Electrical characteristics of fabricated FG-MEM switches were measured at room temperature in a Lakeshore TTPX cryogenic vacuum probe station at ~1.5 μTorr to minimize oxidation of the W electrode surfaces. Prior to collecting data, a native-oxide breakdown process was conducted by turning on the switch by applying $V_{GB} = V_{ON} + 2V$ and then applying a millisecond voltage pulse ($V_{DS} = 3V, \sim 5ms$) across the contact using the Agilent B1500A Semiconductor Device Parameter Analyzer. Afterwards, ON-state resistance R_{ON} is measured, and if $R_{ON} > 1k\Omega$ then an additional voltage pulse V_{DS} is applied. This pulsed oxide-breakdown procedure is discussed in [16].

Figure 3.4 shows measured current-*vs.*-voltage (*I*-*V*) characteristics for a fresh FG-MEM switch, for positive and negative sweeps of the gate voltage with the body biased at ground and $V_{DS} = 0.5\,V$. It can be seen that the device turns on with positive gate voltage at $V_{ON} = +14.3\,V$ and with negative gate voltage at $V_{ON} = -14.2\,V$, which is approximately symmetrical. This indicates that there is no charge stored on the floating gate.

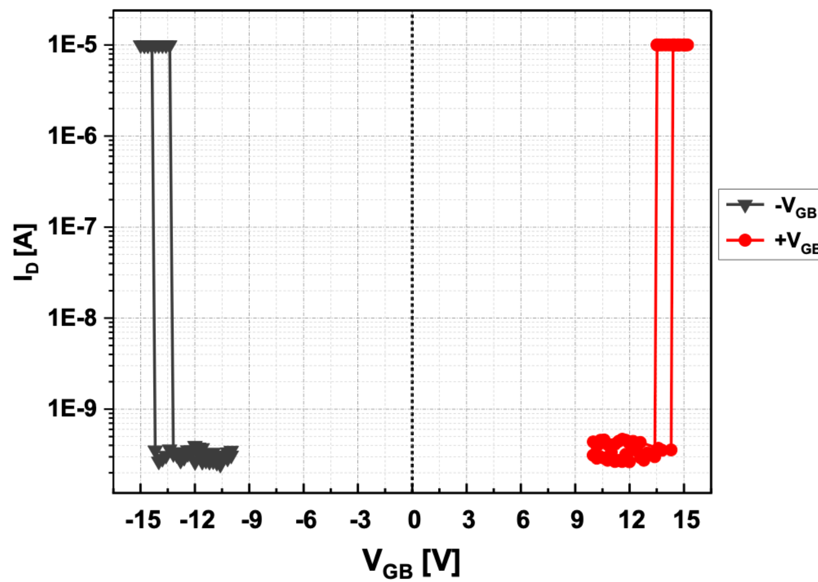


Figure 3.4: Measured *I*-*V* characteristics of a FG-MEM switch showing symmetric values of V_{ON}. (Negative V_{GB} sweep is shown in Black while positive V_{GB} sweep is shown in Red.)

Figure 3.5 shows measured *I*-*V* characteristics for a FG-MEM switch before (red) and after (blue) it is programmed with a single voltage pulse of magnitude $V_{Prog} = 30V$ and 10 μs width. As can be seen from the figure, there is an asymmetric shift in switching voltage (ΔV_{ON} is approximately +0.79V) indicating that some charge was successfully injected onto the FG.

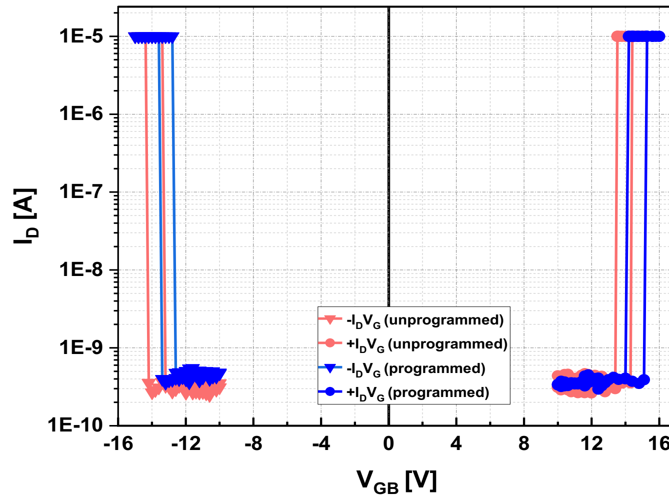


Figure 3.5: Measured *I-V* characteristics of a FG-MEM switch before (red) and after (blue) programming for negative and positive V_{GB} sweeps. V_{ON} is no longer symmetrical due to injected charges and is shifted by +0.79 V to the right, after programming.

The programming/erasing characteristics of the FG-MEM switch device are shown in Figure 3.6 for different program/erase voltages. ΔV_{ON} saturates as the program time increases because the electric field within the tunneling oxide decreases as the FG is charged (cf. Eqn. (3.7)). The much faster erase operation is due to a leakage path between the floating gate and the structural anchor. Figure 3.7 (cf. Figure 3.3 (g) Section C) shows that the fabricated floating gate is in direct contact with the anchor, providing a pathway for charges to leak away.

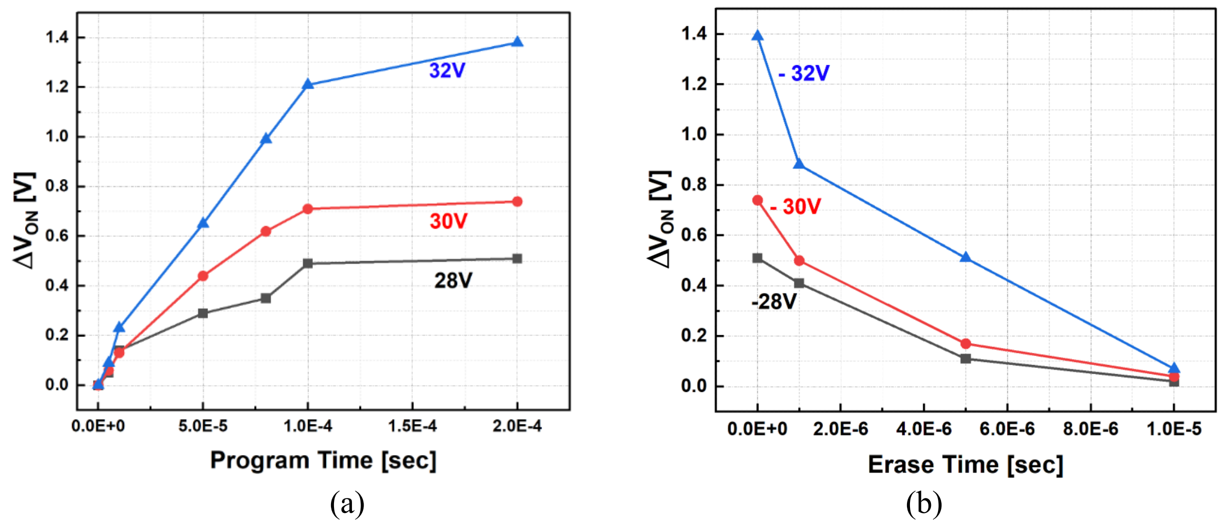


Figure 3.6: Change in FG-MEM switch turn-ON voltage with (a) program and (b) erase time, for different program/erase voltages.

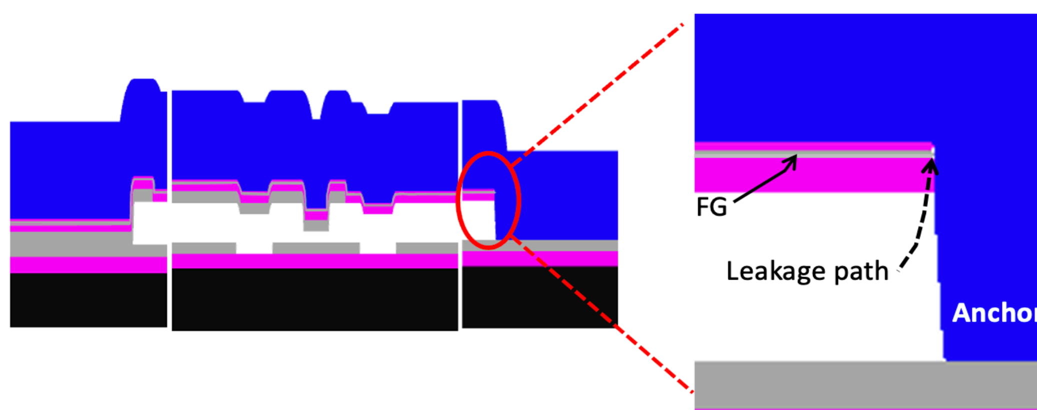


Figure 3.7: Zoomed-in cross-sectional schematic of the initial FG-MEM NV-switch design (cf. Figure 3.3(g)) showing that the floating gate (FG) contacts the gate electrode in the anchor region, allowing charge on the FG to leak away.

3.3.3 Data Retention

Figure 3.8 shows data-retention characteristics for three different FG-MEM devices maintained at 200˚C in vacuum. First the initial V_{ON} values were recorded for the three fresh devices at room temperature (Device #1: 16.04V, Device #2: 16.85V, Device #3: 16.0V). Subsequently, a programming voltage pulse (30V, 10μs) was applied to each device, followed by V_{ON} read to verify each device was programmed. Then the devices were heated to 200˚C for accelerated data-retention testing. Unfortunately, all devices' V_{ON} returned to their initial values (or ΔV_{ON} went back to zero) within ~1 minute, as can be seen from Figure 3.8. This means that all the stored charges leaked out of the floating gate. This issue is also likely due to a leakage path between the floating gate and the structural anchor, as shown above in Figure 3.7.

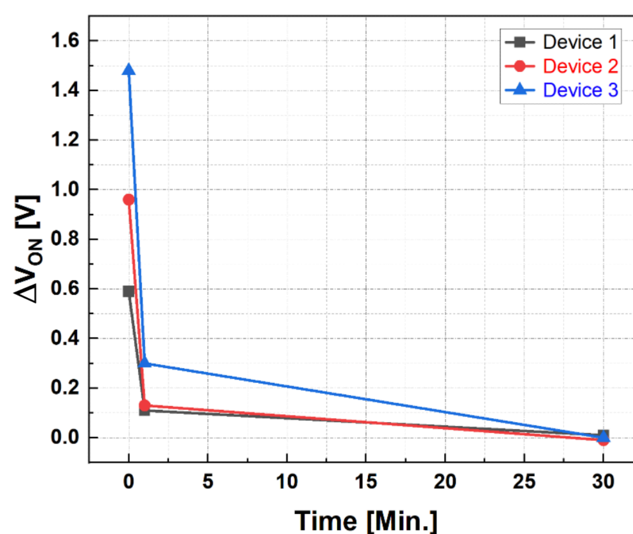


Figure 3.8: Data retention characteristics of three FG-MEM devices at 200˚C. Most of the charge stored on the FG in each of the three devices leaked away within ~1 minute.

3.4 FG-MEM NV Switch Design Improvements

The initial experimental results presented above reveal some challenges for practical implementation of FG-MEM NV switches: (i) relatively small values of ΔV_{ON}, (ii) slow programming speed, and (iii) poor retention time. Causes and proposed solutions for these are discussed in this section.

Since ΔV_{ON} is inversely proportional to C_{FG_CG} (cf. Eqn. (3.9), it is desirable to reduce C_{FG_CG} in order to increase ΔV_{ON} or to reduce the amount of stored charge (Q_{FG}) required to achieve the desired value of ΔV_{ON}. This means that the thickness of the dielectric material layer between the control gate and the floating gate should be increased.

To increase programming speed, the electric field in the tunnel oxide layer during the programming process should be increased. This means that the floating gate voltage (V_{FG}) should be larger. The capacitive coupling ratio α for the initial FG-MEM NV devices in this work is less than 0.05, which means that V_{FG} is more than 20 times smaller than V_{prog} (cf. Eqn. (3.7). Therefore, the thickness of the dielectric material layer between the floating gate and the drain should be decreased and/or the thickness of the dielectric material layer between the floating gate and the control gate should be increased.

To achieve both larger values of ΔV_{ON} and to improve programming speed, the positions of the tunnel oxide and gate oxide should be interchanged; that is, the gate oxide between the control gate and floating gate should be thick (45 nm) while the tunnel oxide between the floating gate and drain electrode should be thin (10 nm).

The final challenge to address for the FG-MEM NV switch is the poor data retention time. Because the floating gate is not patterned with a separate mask, its edge is in physical contact with the anchor region of the gate electrode (Figure 3.7); this provides a path for charge to leak away from the floating gate, resulting in faster erase speed (cf. Figure 3.6(b)) and short retention time. To eliminate this leakage path, a separate mask must be used to pattern the FG; also, dielectric sidewall spacers (formed by conformally depositing a layer of Al_2O_3 after the FG is patterned, and then anisotropically etching the Al_2O_3 layer) can be used to completely insulate the FG.

3.5 Summary

A floating-gate-based non-volatile MEM switch is proposed and experimentally investigated in this work. Design improvements are proposed to achieve faster programming speed and long retention time (typically 10 years) to make this device technology practical for NVM applications.

3.6 References

[1] M. Goudarzi, S. Ilager, R. Buyya, "Cloud Computing and Internet of Things: Recent Trends and Directions," *In New Frontiers in Cloud Computing and Internet of Things*, pp. 3–29., 2022, doi: 10.1007/978-3-031-05528-7_1.

[2] L. Benini, A. Macii, and M. Poncino, Massimo, “Energy-Aware Design of Embedded Memories: A Survey of Technologies, Architectures, and Optimization Techniques,” *Association for Computing Machinery*, vol. 2, no. 1, February 2003, doi: 10.1145/605459.605461.

[3] Z. A. Ye, S. Almeida, M. Rusch, A. Perlas, W. Zhang, U. Sikder, J. Jeon, V. Stojanovic, and T.-J. K. Liu, “Demonstration of 50-mv Digital Integrated Circuits with Microelectromechanical Relays,” *IEEE International Electron Devices Meeting (IEDM)*, pp. 4.1.1-4.1.4, 2018, doi:10.1109/IEDM.2018.8614663.

[4] C. Qian, A. Peschot, B. Osoba, Z. A. Ye and T. -J. K. Liu, “Sub-100 mV Computing with Electro-Mechanical Relays,” *IEEE Transactions on Electron Devices*, vol. 64, no. 3, pp. 1323-1329, March 2017, doi: 10.1109/TED.2017.2657554.

[5] X. Hu, S. F. Almeida, Z. Alice Ye and Tsu-Jae King Liu, “Ultra-Low-Voltage Operation of MEM Relays for Cryogenic Logic Applications,” *IEEE International Electron Devices Meeting (IEDM),* 2019, pp. 34.2.1-34.2.4, doi: 10.1109/IEDM19573.2019.8993629.

[6] M. R. Zakaria, et al., “An Overview and Simulation Study of Conventional Flash Memory Floating Gate Device Using Concept FN Tunnelling Mechanism.” *Proc. of V-th Int. Conf. on Intelligent Systems, Modeling and Simulation*, pp. 775-780, 2014.

[7] S.-H Lim and K.-H Park, “An efficient NAND flash file system for flash memory storage,” *in IEEE Transactions on Computers*, vol. 55, no. 7, pp. 906-912, July 2006, doi: 10.1109/TC.2006.96.

[8] Z.A.K. Durrani and H. Ahmed, “Nanosilicon Single-Electron Transistors and Memory,” *in Nanosilicon*, pp. 335-359, 2008, doi: 10.1016/B978-008044528-1.50011-7.

[9] R. Bez, E. Camerlenghi, A. Modelli, and A. Visconti, “Introduction to Flash Memory,” *Proceedings of the IEEE*, vol 4, pp 489-502, 2003, doi: 10.1109/JPROC.2003.811702.

[10] C.K. Perkins, et al., “Demonstration of Fowler–Nordheim Tunneling in Simple Solution-Processed Thin Films,” *ACS Applied Materials & Interfaces,* 10 (42), 36082-36087, 2018, doi: 10.1021/acsami.8b08986.

[11] S Maikap, *et al*., “Charge Trapping Characteristics of Atomic-Layer-Deposited HfO_2 Films with Al_2O_3 as a Blocking Oxide for High-Density Non-Volatile Memory Device Applications,” *Semiconductor Science and Technology*, vol. 22, no. 8, June 2007, dio: 10.1088/0268-1242/22/8/010.

[12] L. Chong, K. Mallik, and C.H. de Groot, “The Vertical Metal Insulator Semiconductor Tunnel Transistor: A proposed Fowler–Nordheim Tunneling Device”, *Microelectronic Engineering*, vol. 81, pp. 171-180, 2005, doi: 10.1016/j.mee.2005.03.003.

[13] M.Y. Ghannam, and R.P. Mertens, “Polycrystalline Silicon in ULSI,” *Encyclopedia of Materials: Science and Technology (Second Edition)*, pp 7152-7158, 2001, doi: 10.1016/B0-08-043152-6/01268-7.

[14] A. Kumar, M. Das, and S. Mukherjee, “Oxide Based Memristors: Fabrication, Mechanism, and Application,” *Materials Science and Materials Engineering*, 2018, doi: 10.1016/B978-0-12-803581-8.10384-4.

[15] P. Pavan, R. Bez, P. Olivo and E. Zanoni, “Flash memory cells-an overview,” in Proceedings of the IEEE, vol. 85, no. 8, pp. 1248-1271, Aug. 1997, doi: 10.1109/5.622505.

[16] Alice Ye. Millivolt Micro-Electro-Mechanical Relay Devices Circuits. PhD thesis, EECS Department, University of California, Berkeley, Dec 2021. URL: https://www2.eecs.berkeley.edu/Pubs/TechRpts/2021/EECS-2021-239.pdf.

Chapter 4

Highly Reliable and Secure Physical Unclonable Function Using Resistive Memory Integrated into a 28nm CMOS Process

4.1 Introduction

With the proliferation of mobile computing and communication devices and the emergence of the Internet of Things (IoT), hardware security is becoming a key concern [1]. The connectivity of IoT devices is mandatory but also makes them vulnerable to malicious attacks and data leaks. Therefore, security and privacy are considered as key challenges to future growth of the IoT [2]. The requirements for encryption key generation and storage include a source of randomness to guarantee unpredictability and uniqueness of the key, and a protected memory to reliably store the key [3].

Software algorithms are commonly used to generate encryption keys to secure information and communication channels. While software-generated encryption keys are cost-effective and relatively easy to update and maintain, they can be less secure than their hardware equivalents. Also, continual encryption and decryption of data can significantly slow system performance [4].

Hardware-based encryption is considered to be more secure and faster because circuitry within the hardware is responsible for encryption and authentication so only authorized users can access the data in the hardware [4] [5]. In the past, secured electrically erasable programmable read-only memory (EEPROM) or other battery-backed static random-access memory (SRAM) was used for secret key storage. Physical Unclonable Functions (PUFs) embedded in integrated circuits (“chips”) have attracted attention in recent years because they offer a lower-cost, lower-power and

tamper-resistant solution for authentication, as well as key generation for cryptography applications [6]–[8].

PUFs exploit inherent hardware variations to produce an unclonable, unique device response to a given input without the need for expensive hardware or secured memory modules [1]. Unlike other hardware security primitives, PUFs do not store keys in memory [9]. Rather, a PUF generates device-specific output (the device's digital fingerprint) when queried with a certain input. Typically, the generated output is a string of binary digits (bitstream) determined by inherent physical variations that occurred during the chip's manufacturing process [10]–[12].These process-induced variations result in variability in transistor threshold voltage, transistor drive current, and parasitic resistances and parasitic capacitances across the chip. Although process variation may be an unwanted effect from the circuit designer's viewpoint, it is vital for building PUF circuits because unpredictability increases security, as the device-specific output cannot be predicted based on some deterministic or quasi-deterministic algorithm. [1].

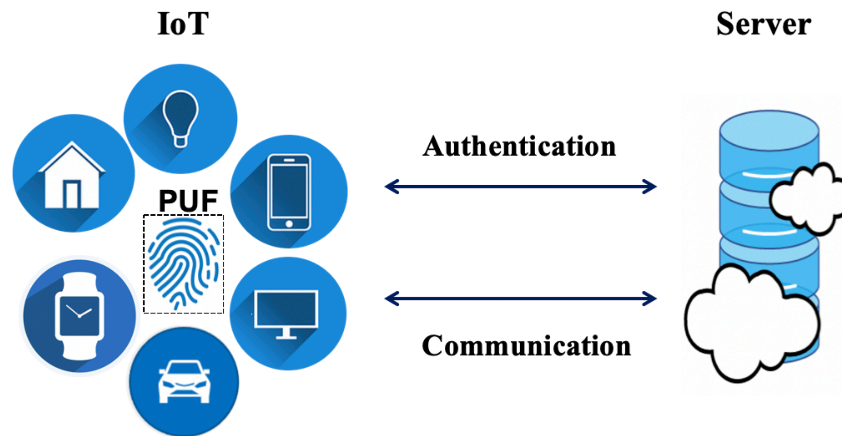


Figure 4.1: Typical PUF based authentication scheme for IoTs.

Various PUF architectures based on ring oscillators, arbiter circuits, and various types of memory cells (*e.g.*, SRAM) have been reported [13]–[18]. Among these, filamentary resistive random access memory (ReRAM) based PUFs stand out due to their high degree of randomness, low bit error rate (BER), resilience to variations in environmental conditions and resistance to invasive attack [9],[19]. Therefore, ReRAM-based PUF is a promising candidate for applications requiring both hardware security and embedded non-volatile memory (NVM). In advanced security protocols, PUFs are required to cooperatively fulfill requirements for device authentication and encrypted communication between IoT devices and a cloud-based server, as illustrated in Figure 4.1.

In this chapter, a highly reliable and resilient ReRAM-based PUF architecture and generation scheme that is fully compatible with 28 nm-generation CMOS technology is presented. Both NVM and PUF functions are realized in the same chip. Generated PUF bitstreams satisfy all NIST SP800-22 randomness assessments and demonstrate 100% retention after 50 hours at 150˚C.

4.2 ReRAM Structure and Operation

Resistive random-access memory (ReRAM or RRAM) is a type of NVM that relies on the electrical resistance modification of a switching medium in order to store information. It is of particular interest for emerging high-density NVM applications, owing to its characteristics of simple cell structure, fast program and erase (P/E) speed, excellent scalability, low power consumption, and compatibility with standard complementary metal-oxide-semiconductor (CMOS) process technology [20-26].

A ReRAM device is a metal/insulator/metal (MIM) structure comprising the switching layer (SL) sandwiched between an inert bottom electrode (BE) and a top electrode (TE) [26][27]. The application of a voltage across the ReRAM device enables a transition from a high-resistance state (HRS) or OFF–state to a low-resistance state (LRS) or ON–state. The physical mechanism of resistance change depends on the materials used.

A wide variety of materials have been investigated as ReRAM electrodes. Among these, the most abundant and commonly used are elementary materials such as W, Al, Ti, Cu, Ag, and Pt or nitride-based compound materials such as TaN and TiN [28]. A variety of materials have been investigated for the ReRAM SL, but metal oxides and nitrides have been most extensively studied and are preferred primarily due to their compatibility with CMOS back-end-of-line (BEOL) processing [28]. Examples of insulating materials that exhibit non-volatile resistance switching are metal oxides such as TiO_x, HfO_x, AlO_x, SiO_x and nitrides such as AlN_x, SiN_x [28].

In this work, ReRAM devices are integrated into the BEOL process of a standard 28nm-generation CMOS technology. Figure 4.2(a) shows the one transistor (1T) - one ReRAM (1R) memory cell configuration, with ~120nm ReRAM lateral dimension. The ReRAM device comprises a 30nm tungsten (W) bottom electrode, a 3nm AlOx switching layer, and a 90nm AlNx top electrode (TE) as shown in Figure 4.2(b).

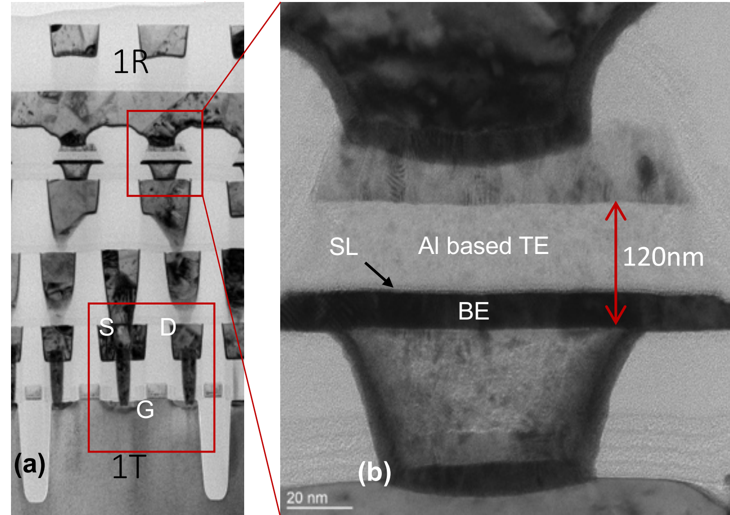


Figure 4.2: (a) Cross-sectional transmission electron micrograph of a fabricated 1 transistor (1T) 1 ReRAM (1R) cell (b) higher-magnification view showing ReRAM layer stack: inert bottom electrode (W), switching layer (AlOx), and Al-based top electrode (AlNx).

The switching mechanism of the AlOx-based ReRAM device used in this work is the formation of a conductive metallic filament within the dielectric layer to electrically connect the two electrodes. Such ReRAM devices are called Conductive Bridge ReRAMs (CB-ReRAMs). Figure 4.3 illustrates the metallic filamentary switching mechanism. The application of a positive voltage on the TE relative to the BE ground potential causes metal atoms from the TE to become ionized through the process of oxidation [29]. The metal ions then drift through the SL toward the BE, forming a conductive filament. The metal ions are reduced when they come into electrical contact with the BE (Figure 4.3(a)). This process is called forming and is considered to be the initial soft breakdown of the MIM structure. The applied voltage used for the forming process is referred to as the forming voltage (V_f).

Upon application of a negative voltage to the TE (or a positive voltage to the BE with the TE maintained at ground potential), Joule heating together with the induced electric field cause oxidation (ionization) of the metal filament [30][31]. The metal ions then drift back toward the TE and are reduced when they come into electrical contact with the TE (Figure 4.3(b)). The ruptured filament in this OFF-state results in relatively large resistance (R_{OFF}) between the TE and BE. The voltage at which this rupturing process occurs is called the erase voltage (V_{erase}).

Subsequent application of a voltage to switch the device to the LRS is referred to as a program operation, and the applied voltage is referred to as the program voltage (V_{pgm}). The resistance between the TE and BE in the ON-state (R_{ON}) is relatively small.

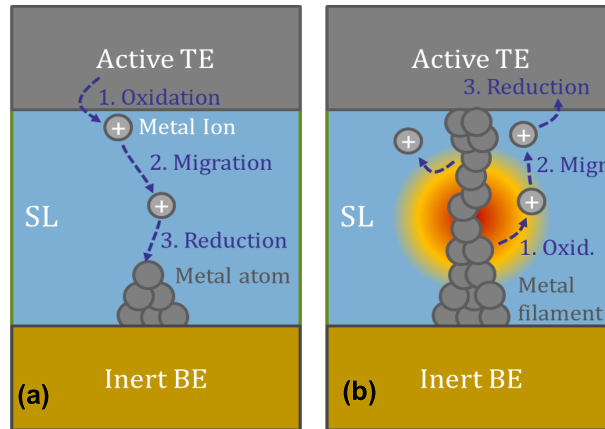
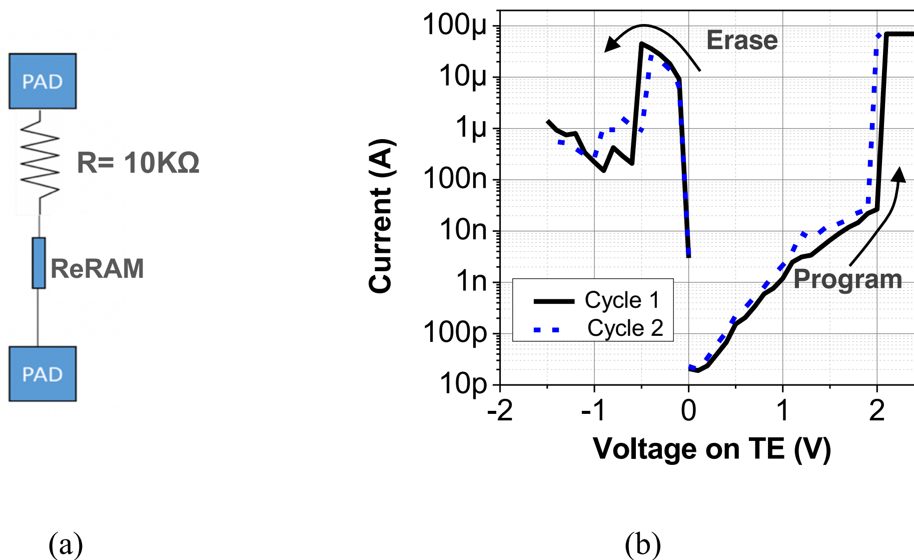


Figure 4.3: Schematic illustration of the switching mechanisms of a Conductive Bridge ReRAM device: (a) program operation and (b) erase operation.

Figure 4.4 shows a circuit schematic used to measure DC current and a typical measured DC current *vs*. voltage (*I-V*) curves for a ReRAM device after a 400°C alloying process. Unlike oxygen-vacancy-based resistive memory which usually requires a high-voltage forming step before program/erase cycling [32] [33], our ReRAM shows relatively small (less than 2V) forming and program voltages. This is because the switching-layer thickness and TE interface are engineered to achieve a low voltage in the first cycle, or "forming process." In addition, cycle 1 and 2 program voltages are almost identical, indicating that the filament formed within the switching layer during programming is retracted almost completely back to the top electrode after erase operation.



(a) (b)

Figure 4.4: (a) I-V measurement circuit schematic. (b) Typical measured I-V curves of a ReRAM device. 'Cycle 1' is the 1st cycle from pristine device ('form' and 'erase') and 'Cycle 2' shows 'program' and 'erase' operation. Voltage sweep directions are indicated by arrows.

ReRAM operation can involve frequent switching between HRS and LRS; each transition between states can cause defects in the insulating switching layer, consequently resulting in lower OFF-state resistance (R_{OFF}) and degrading ReRAM performance. Thus, it is important to perform endurance testing to determine the maximum number of program/erase cycles before the device fails (with indistinguishable HRS and LRS). This is done by performing multiple program/erase cycles and intermittently applying a read voltage (V_{Read} = 0.3 V) and measuring the current (to monitor R_{OFF} and R_{ON}). Table 4.1 shows the program/erase conditions used to obtain the endurance testing results for 1Mb of ReRAM cells shown in Figure 4.5. Due to ReRAM switching voltage variability, programming/erasing pulses were applied repeatedly until the device was verified to be programmed/erased.

In Figure 4.5, The measured ON-state (‘1’ bit) and OFF-state (‘0’ bit) current distributions indicate excellent endurance, exceeding 10,000 program/erase cycles, making this ReRAM technology suitable for NVM applications. The increase in ON-state current with cycling is due to degradation of the electrically insulating property of the switching layer.

Table 4. 1: Parameters for ReRAM program and erase cycling.

Cycling Parameters	Value
Form/ Program Voltage	3.2V
Form/ Program Pulse Width	200µs / 5µs
Form/ Program Current	300µA
Erase Voltage	-2.2V
Erase Pulse Width	20µs

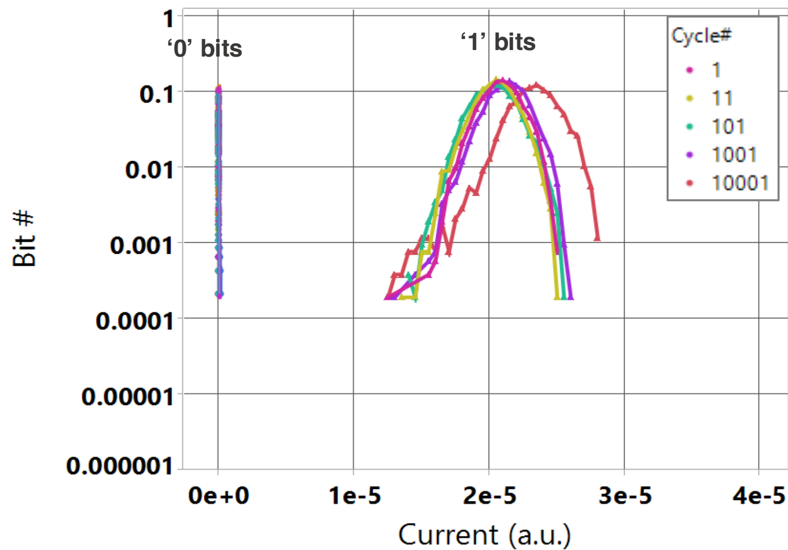


Figure 4.5: Distributions of measured ReRAM cell current after verified program/erase cycles. Excellent endurance (more than 10,000 cycles) is seen for the two distinct states (ON and OFF).

Data retention (stability of LRS and HRS after undergoing 10,000 program/erase cycles) is another NVM device performance requirement. As shown in Figure 4.6, excellent post-cycling data retention characteristics are seen. ON–state and OFF–state conductance remains well separated after a 1-hour 225˚C bake; the projected retention time exceeds 98 years at 85°C, assuming an activation energy (E_a) of 1.5 eV [34]. Figure 4.7 shows the current distribution of ReRAM devices after forming (pre-cycling). As can be seen from this figure, the devices well maintain their state after a 265°C 1-hour bake (which extrapolates to 10 years at 125°C), with good margin between ON and OFF states. This confirms that the ReRAM devices are compatible with a reflow soldering process, which is important for low-cost chip packaging.

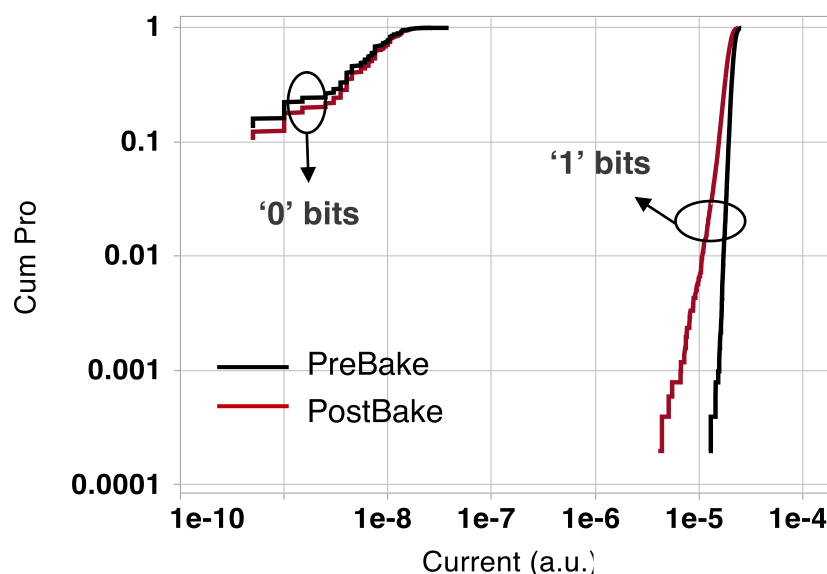


Figure 4.6: ReRAM retention performance. Clearly differentiated 'ON' ('1') and 'OFF' ('0') states are retained after baking at 225°C for 1 hour. The projected retention time at 85°C is greater than 98 years.

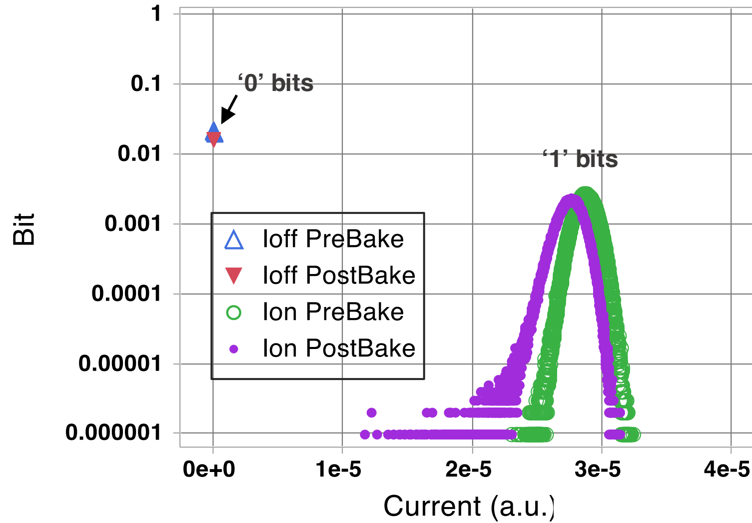


Figure 4.7: ReRAM forming current distributions before and after baking at 265°C for 1 hour, confirming solder reflow process compatibility.

4.3 Inherent Stochastic Behavior of ReRAM for PUF

Although ReRAM can be used for NVM applications, the inherently stochastic nature of the device is a significant limiting factor. The switching voltage, switching current, HRS/LRS resistances, and switching delay all exhibit a high degree of variation from device-to-device and from cycle-to-cycle [35][36]. Such randomness arises from the formation and rupture of the conductive filament, manufacturing process variations, and/or defects in the switching layer. This stochastic behavior can be leveraged as a source of randomness for PUFs, however [37, 38]. In this section, different aspects of ReRAM randomness are investigated.

The measured forming *I-V* characteristics for 74 ReRAM cells from a single wafer, plotted in Figure 4.8, show that the forming voltage varies randomly from device to device across a ~1V window from 2.4V to 3.4V. (The voltage step size was 50 mV, which is why many curves overlap in their transition regions.) The program voltage for each device also randomly varies from cycle to cycle, as can be seen from Figure 4.9 which plots V_{pgm} for 18 devices over 100 cycles.

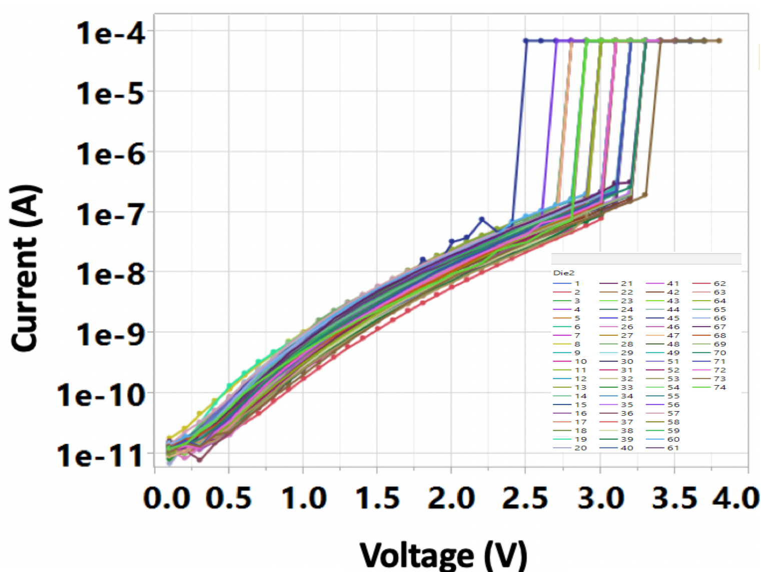
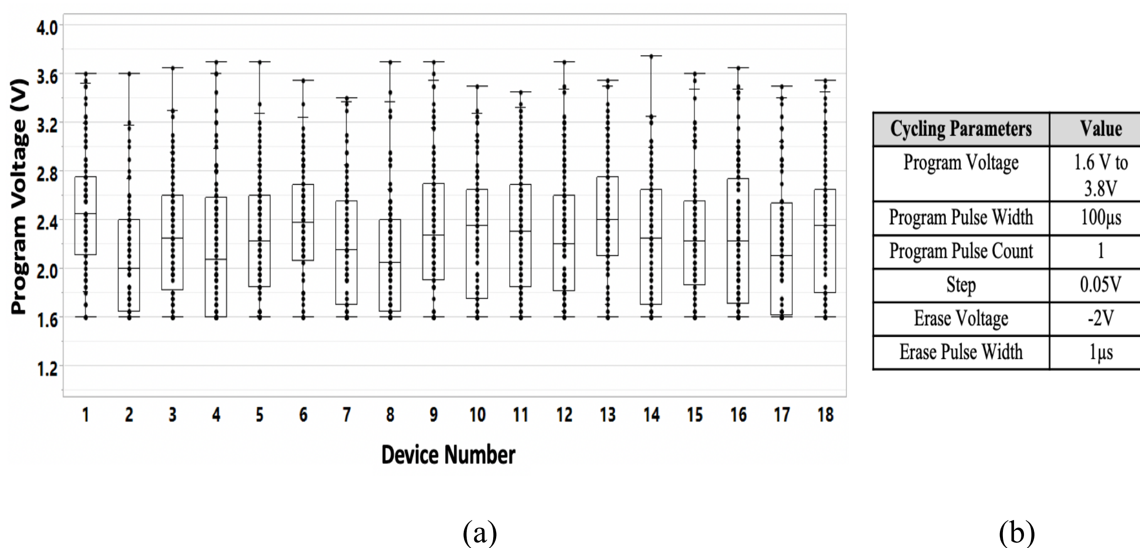


Figure 4.8: Characterization of DC forming voltage randomness of 74 ReRAM cells.



Cycling Parameters	Value
Program Voltage	1.6 V to 3.8V
Program Pulse Width	100µs
Program Pulse Count	1
Step	0.05V
Erase Voltage	-2V
Erase Pulse Width	1µs

(a) (b)

Figure 4.9: (a) Cycle-to-cycle program voltage distribution for 18 ReRAM devices. Each device is cycled 100 times. The edges of the box indicate the 25th and 75th percentiles and the median value is indicated inside the box. (b) Program/erase parameters used. More than 1 erase pulse was used if needed.

The effects of program voltage pulse width (PW) and pulse count were investigated using an array of 163,840 1Transistor-1ReRAM (1T-1R) cells as illustrated in Figure 4.10. $V_{Read} = 0.3$V was used for verification of successful switching. For each value of PW (*e.g.*, 10µs) the program

voltage applied to the TE was successively incremented by 0.1V until the ReRAM device switched to the ON-state. The results shown in Figure 4.11 indicate that as the PW is increased, the range of program voltage required to switch ON the ReRAM generally diminishes. The effects of program voltage and PW on ReRAM switching time (*i.e.*, product of PW and pulse count required to switch ON the ReRAM) are shown in Figure 4.12. Randomness in switching time is seen and is greater for lower program voltage and shorter PW.

The stochastic nature of ReRAM switching time can be utilized as the source of randomness for PUFs. From the data in Figure 4.12, program voltage parameters to achieve 50% probability of switching can be determined, for PUF generation discussed in the next section.

Figure 4.10: Cell array architecture used for measuring ReRAM program/erase times. ‘BL’, ‘WL’ and ‘SL’ denote bit line, word line, and source line, respectively.

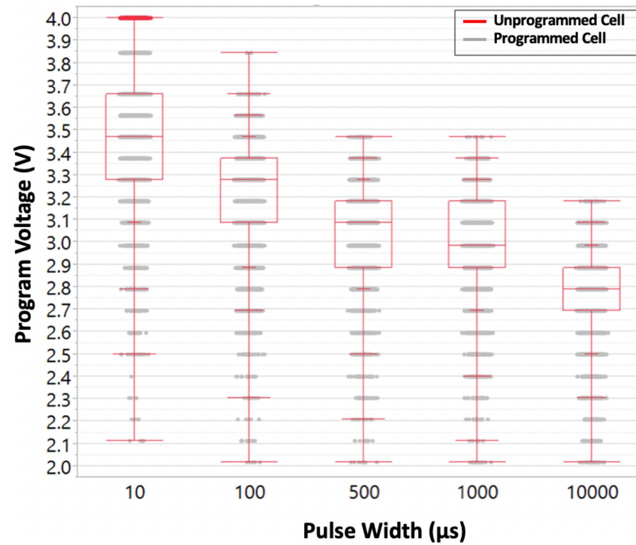


Figure 4.11: Measured distributions of ReRAM program voltage for various program pulse widths. The edges of the red box indicate the 25th and 75th percentiles and the median program voltage is indicated inside the box.

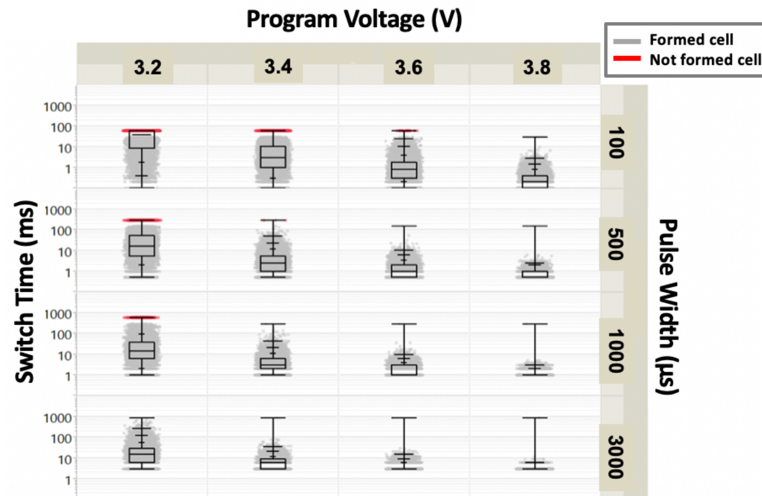


Figure 4.12: Measured distributions of ReRAM switching time for various values of program voltage. (The switching time is the product of program pulse width and pulse count.) The edges of the black box indicate the 25th and 75th percentiles and the median value of switching time is indicated inside the box.

4.4 PUF Generation Scheme Using Switching Time Variation

The scheme proposed herein exploits the stochastic nature of ReRAM switching time as an entropy source. Figure 4.13 shows measured voltage waveforms for two different ReRAM devices during

program/switching operation. The programming voltage pulse (black) is applied to the ReRAM device in series with a 5 kΩ resistor, while the output voltage (red) across the 5 kΩ resistor is monitored. The voltage across the resistor increases when the ReRAM cell switches to the ON–state. For the same program voltage (V_{pgm}), the two ReRAM devices exhibit different switching times of ~140ns and ~35ns, confirming stochasticity. Therefore, it is proposed to exploit the mismatch in program time between two ReRAM devices for PUF generation. Figure 4.14 is a circuit diagram of an array of 1T-1R cells illustrating voltage-differential based PUF generation, where the TEs for a pair of ReRAM devices along a selected row are connected together to generate one bit of the PUF.

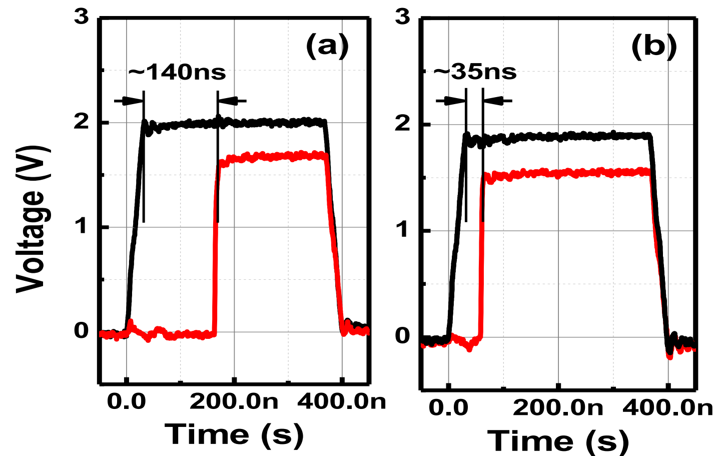


Figure 4.13: Programming voltage pulse (black) and measured voltage (red) across 5 kΩ series resistor. (Once the ReRAM cell is programmed, the voltage across the resistor increases.) A large difference in switching time is seen for the two different ReRAM cells in (a) and (b).

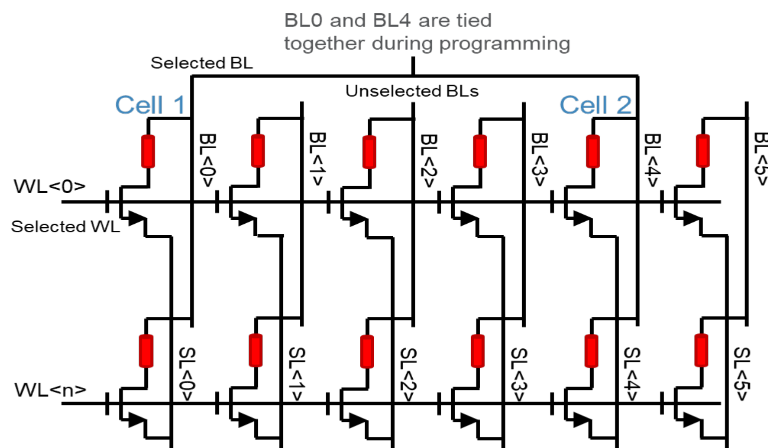


Figure 4.14: Circuit diagram illustrating voltage-differential based PUF generation. A pair of 1T-1R cells (‘Cell 1’ and ‘Cell 2’) forms one PUF bit-cell. ‘BL’, ‘WL’ and ‘SL’ denote bit line, word line, and source line, respectively.

Reliable PUF generation requires two steps for each bit-cell: A gentle program step followed by a “soaking” step. In-between these steps, a read operation is performed to verify the state of the bit-cell. In the gentle program step, a program voltage pulse is applied to the two connected bit-lines while the selected word line is pulsed high and source lines are grounded. Table 4.2 lists the parameter values used in this work, and Figure 4.15 shows a timing diagram for the PUF bit generation process.

Table 4.2: PUF bit generation conditions used in gentle program and soaking steps.

PUF Generation Parameters	Value
Gentle Program Voltage	3.2V
Gentle Program Pulse Width	200µs
Gentle Program Current Compliance	100µA
Soaking Voltage	3.2V
Soaking Pulse Width	200µs
Soaking Pulse Count	1
Soaking Current Compliance	450µA
Read/Verify Voltage	0.3V

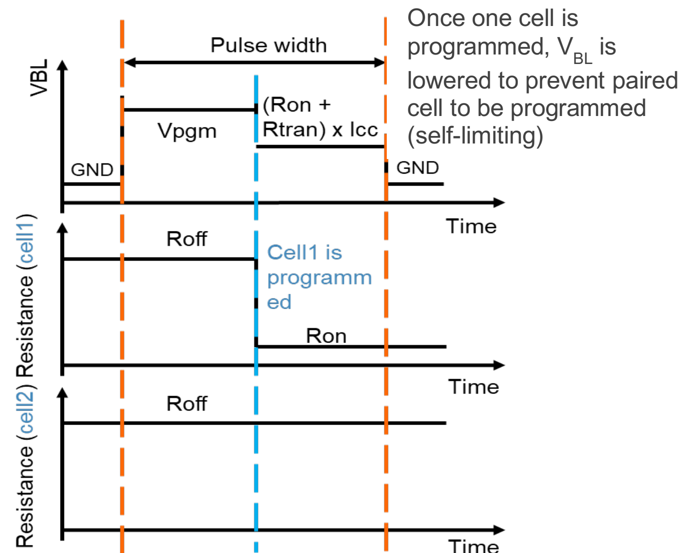


Figure 4.15: Timing diagram for PUF bit generation during Gentle Program step. Once one cell is programmed, the voltage on the bit line (VBL) drops, preventing the programming of the other cell in the pair. Rtran is the ON-state resistance of the programmed cell transistor.

Note that this PUF bit generation scheme requires peripheral circuitry to limit the maximum current flowing through a cell to ~100uA, to protect the ReRAM devices from being damaged and to ensure the switching-lowered bit-line voltage is insufficient to program the paired ReRAM cell connected in parallel. To verify this, a circuit simulation was performed and the results are shown in Figure 4.16. The simulated bit-line voltage is seen to drop from >3V to 0.4V within ~18ns during the gentle program step; this time window is much too short for the paired cell to be programmed.

A soaking step is applied to all of the cells programmed (switched ON) by the gentle program step, to enhance the filament thickness and strength by applying a larger voltage pulse, ensuring a highly stable PUF key. Figure 4.17 illustrates the difference between measured cell current before and after soaking, showing the importance of this step.

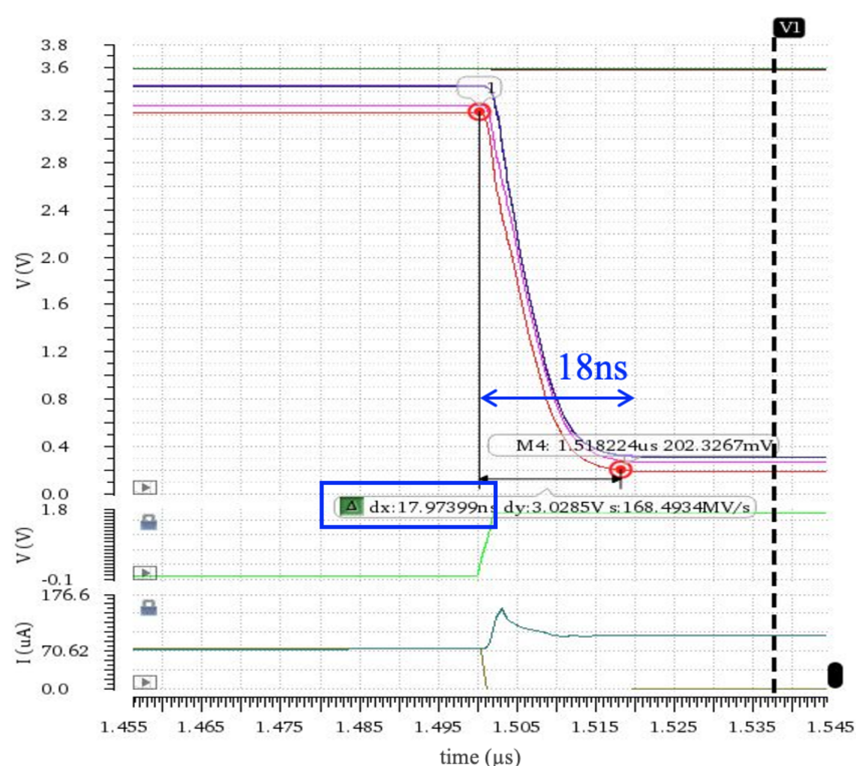


Figure 4.16: Simulation of bit-line voltages (for program voltages of 3.2V, 3.3V and 3.4V) when 100uA current compliance is set. It takes ~18ns for the bit-line voltage to drop after a cell is programmed.

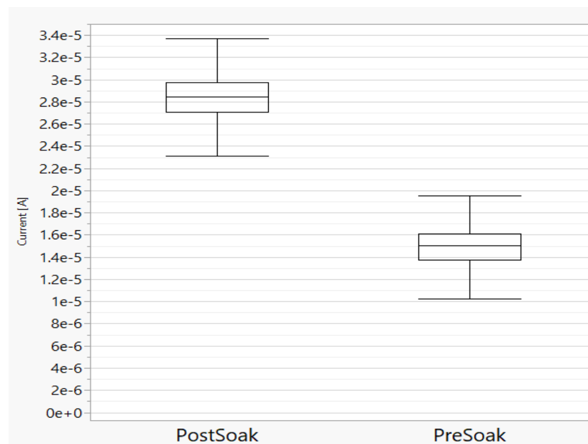


Figure 4.17: Measured ReRAM current before and after soaking step. ON-state current is higher after soaking due to enhanced filament strength.

The two-step PUF bit generation process described above is used to randomly produce complementary bit pairs. For higher throughput generation, multiple pairs of ReRAM cells can be programmed in parallel; the number is limited by the maximum current that can be supplied by the charge pump circuit. The PUF bitstream is then generated by reading/concatenating only the first or only the second bit of every pair.

Due to the stochastic nature of ReRAM switching time, there is extremely low probability that both cells in a ReRAM pair are programmed simultaneously, detrimentally impacting the PUF quality of randomness. As shown in Figure 4.18, the ratio of generated '0's to generated '1's in our experimental dataset is very close to 50%, confirming that such cases are outliers and hence do not significantly diminish the effectiveness of the PUF key generation algorithm.

It is also worthwhile to point out here that any portions of a manufactured ReRAM array can be used to implement PUF bit-cells; the regular NVM and PUF portions of the array are physically indistinguishable, which is advantageous for making the PUF more secure against attacks. Furthermore, our PUF scheme does not need additional circuitry beyond the standard NVM ReRAM circuitry.

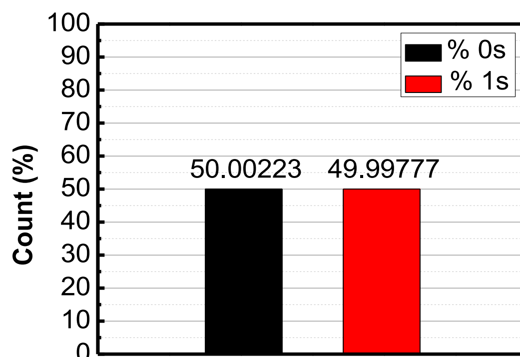


Figure 4.18: Percentages (%) of '0's (OFF-state cells) and '1's (ON-state cells) in 102,400,000 generated PUF bits. The percentages are very close to 50%, the ideal value for randomness.

4.5 PUF Randomness Evaluation

For assessing and quantifying randomness, 204,800,800 ReRAM cells are used to generate 102,400,400 PUF bits. The PUF bits are divided to make 400,000 PUF keys of 256-bit length. Note that in practice the length of a PUF key varies depending on the application.

Randomness quality is evaluated by assessing multiple criteria. The two most standard metrics for PUF evaluation are the intra-Hamming distance (HD) and the inter-HD. The intra-HD is a measure of stability of the PUF response under nominal conditions and characterizes a bit by the average number of varying output bits compared to a predefined reference output. As multiple readouts should not result in different PUF responses, the ideal intra-HD is 0%. Although the ideal value is 0%, in reality PUF readouts are nonzero due to electronic noise resulting in supply voltage variations, variations in environmental conditions such as temperature, and aging. [39] [40]. The inter-HD is a measure of uniqueness between different keys, indicating the average number of bits that differ from each other in two randomly selected PUFs. Ideally, this metric should have a Gaussian distribution with a mean of 50% [39-41].

To estimate the Intra and Inter-HD values, the fractional Hamming distance (FHD) is used as defined in [40] [42]. For k devices and the challenges (inputs) C_1 and C_2, the intra-HD is estimated as:

$$\mathrm{HD}_{\mathrm{Intra}} = \frac{1}{\mathrm{k}} \sum_{i=1}^{k} FHD\ (R_{i,1} R_{i,2})\ \times 100\% \tag{4.1}$$

where $R_{i,1}$ is the response from device i for challenge C_1 and $R_{i,2}$ is the response from device i for challenge C_2.

For k devices the inter-HD between distinct PUFs for a challenge C is defined as

$$\mathrm{HD}_{Inter} = \frac{2}{\mathrm{k(k-1)}} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} FHD\ (R_i R_j)\ \times 100\% \tag{4.2}$$

where i and j correspond to two different devices. R_i is the response from device i for the challenge C. R_j is the response from device j for the same challenge C.

Due to inherent stochastic properties of the ReRAM, there is an equal probability of either cell in a pair being programmed first, resulting in a bitstream with high entropy, characterized by inter-HD and intra-HD. Figure 4.19 shows inter-HD and intra-HD of 400,000 PUF keys generated from ReRAM integrated into a 28nm-generation CMOS technology. No overlap between intra-HD and inter-HD (false rejection rate, false acceptance rate $\ll 10^{-12}$) was observed. Inter-HD has the ideal Gaussian distribution with $\mu = 0.50000$, $\sigma = 0.03128$.

Stability over time is another PUF requirement. As shown in Figure 4.20, excellent (100%) retention of the generated PUF bits is seen after baking at 150°C for 50 hours, with minimal reduction in the conductance of the programmed (‘1’) bits. Accordingly, the inter- and intra-HD distributions are expected to be resilient to thermal stress and lifecycle decline.

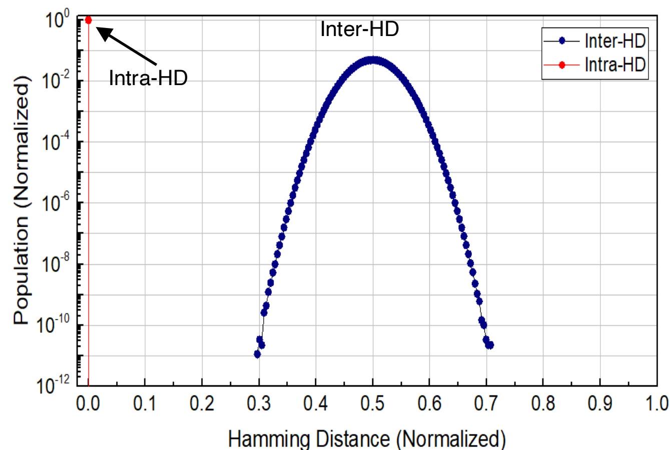


Figure 4.19: Inter- and Intra- hamming distance (HD) of generated PUF bits (102,400,000 bits). PUF key length is 256 bits and total PUF count is 400,000. Inter-HD follows an ideal Gaussian distribution. There is no overlap between inter- and intra- HD distributions.

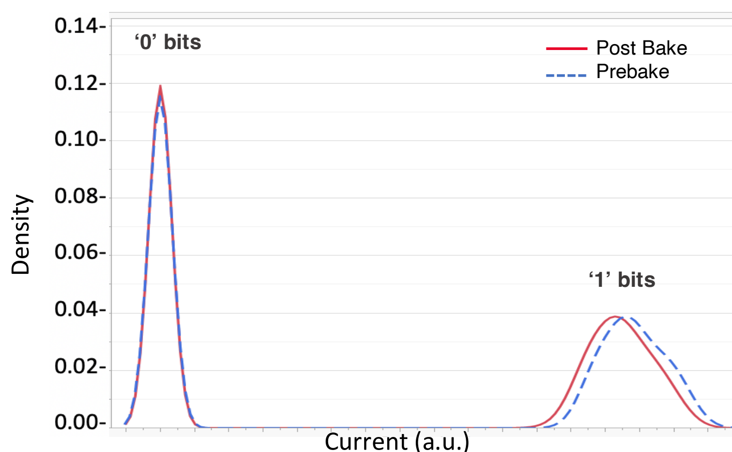


Figure 4.20: Stability of ‘1’ and ‘0’ bits after 150°C, 50 hours (PostBake) compared to before bake (PreBake). 100% retention is achieved for both ‘0’ and ‘1’ states.

Another important PUF metric is autocorrelation, that is, any patterns or trends over time which make PUF keys susceptible to modelling and machine learning attacks [43]. No correlation is required to prevent the possibility of predicting the PUF key using such patterns. Thus, an autocorrelation test is performed on the generated 400,000 PUF keys of length 256 bits each. The test results in Figure 4.21 show no correlation between the bits, ensuring the keys are safe from any attackers that reads patterns. Note that ReRAM cells are also immune to read disturbance, as cell states remain distinct after 5×10^{13} read operations (Figure 4.22).

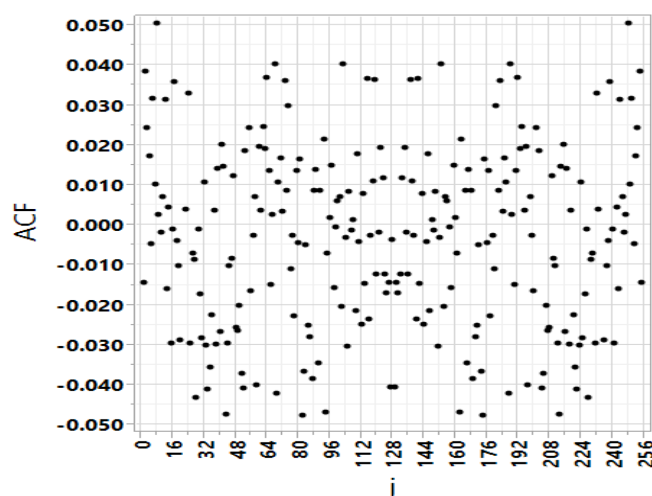


Figure 4.21: Auto correlation test results of 400,000 PUF keys of length 256 bits each, showing no correlation between PUF key bits.

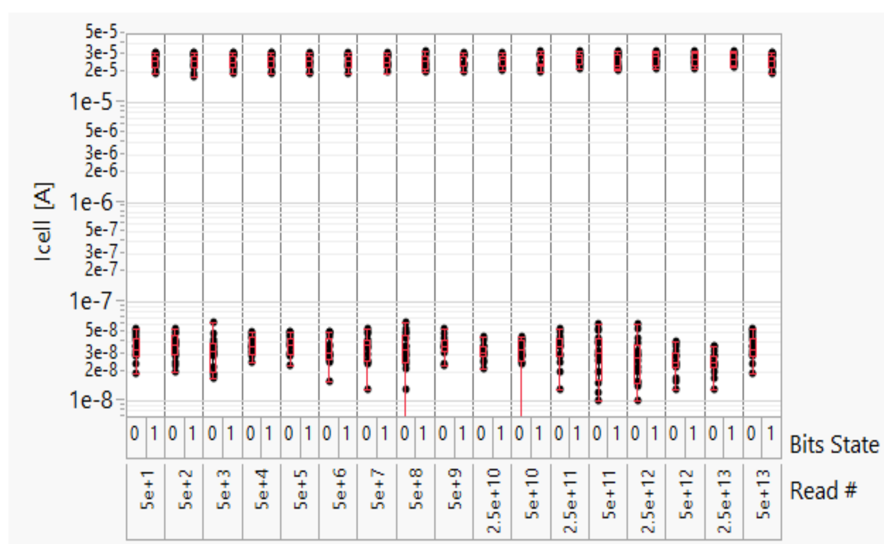


Figure 4.22: ReRAM read endurance (measured I_{Cell} vs. Read #) showing that distinction between ON and OFF devices is well maintained after 5E13 read operations.

The final evaluation of the ReRAM PUF key uses the National Institute of Standards and Technology (NIST) platform, which provides a statistical test suite for random and pseudorandom number generators for cryptographic applications [44]. The NIST suite provides about 15 statistical tests that can be applied to a sequence to compare and evaluate the sequence to a truly random sequence. Thus, the generated 102,400,000 PUF bit stream is evaluated using the NIST SP800-22 test suite [44], and the results are summarized in Table 4.3.

There are two methods to determine whether a bit stream is random. The first method is based on examining whether the distribution of p-values is uniform; the minimum value recommended by NIST for a bit stream to pass this test is 0.0001. The second method is based on computing the proportion of sequences that passed (*i.e.*, with p-value ≥ 0.01); the minimum required values to conclude that a bit stream under test is random are indicated in Table 4.3. As can be seen from the table, each of the 15 different randomness tests conclude that the generated PUF bitstreams satisfy both criteria.

Table 4.3: Detailed results for the standard NIST SP800-22 randomness check test suite.

NIST SP 800-22		METHOD 1			METHOD 2		
STATISTICAL TEST		P-VALUE	NIST SPEC	CONCLUSION	PROPORTION	NIST SPEC	CONCLUSION
1	Frequency	0.181557	> 0.0001	**RANDOM**	99/100	≥ 96/100	**RANDOM**
2	Block Frequency	0.834308		**RANDOM**	98/100		**RANDOM**
3	Cumulative Sums	0.437274		**RANDOM**	100/100		**RANDOM**
4	Runs	0.075719		**RANDOM**	99/100		**RANDOM**
5	Longest Run	0.816537		**RANDOM**	99/100		**RANDOM**
6	Rank	0.153763		**RANDOM**	98/100		**RANDOM**
7	FFT	0.137282		**RANDOM**	99/100		**RANDOM**
8	Non-Overlapping Template	0.851383		**RANDOM**	99/100		**RANDOM**
9	Serial	0.816537		**RANDOM**	97/100		**RANDOM**
10	Overlapping Template	0.474986		**RANDOM**	100/100		**RANDOM**
11	Universal	0.816537		**RANDOM**	97/100		**RANDOM**
12	Approximate Entropy	0.005762		**RANDOM**	100/100		**RANDOM**
13	Linear Complexity	0.834308		**RANDOM**	99/100		**RANDOM**
14	Random Excursions	0.654467		**RANDOM**	66/67	≥ 63/67	**RANDOM**
15	Random Excursions Variant	0.337162		**RANDOM**	66/67		**RANDOM**

4.6 PUF Performance Benchmarking

Herein the PUF generation scheme proposed in this work is compared against industry benchmarks for randomness, specifically the Intel True Random Number Generator (TRNG) and Intel Digital Random Number Generator (RDRAND). RDRAND [45] is a high-quality cryptographic randomness source available in Intel's Ivy Bridge processors [46]. The randomness components of Intel's TRNG (which uses non-deterministic hardware sources for RDRAND), Intel RDRAND, and the ReRAM-based PUF demonstrated in this work are tested with 100 1Mb bitstreams (totaling 100 Mb) using the NIST SP800-90B test suite [47]. As can be seen from Figure 4.23, the ReRAM-generated PUF demonstrates randomness quality comparable to Intel's RDRAND.

PUFs can be vulnerable to power analysis attacks that detect differences in read power consumption of bit-cells in different states [48]. An advantage of the ReRAM-based PUF proposed in this work is that it can provide a safeguard against power analysis attack [49] if the read operation is performed on bit-cell pairs, since the power consumption would be the same regardless of the PUF bit-cell state as there is always one ReRAM cell in the '1' state and the other in the '0' state.

Furthermore, if needed (*e.g.*, when an attack is detected), the PUF bitstream can be erased by programming the ReRAM devices all to the '0' state or all to the '1' state; this "key zeroization" ability is an additional benefit of ReRAM-based PUF for increased security. Table 4.4 provides a summary comparison of previously reported PUFs with the ReRAM-based PUF in this work.

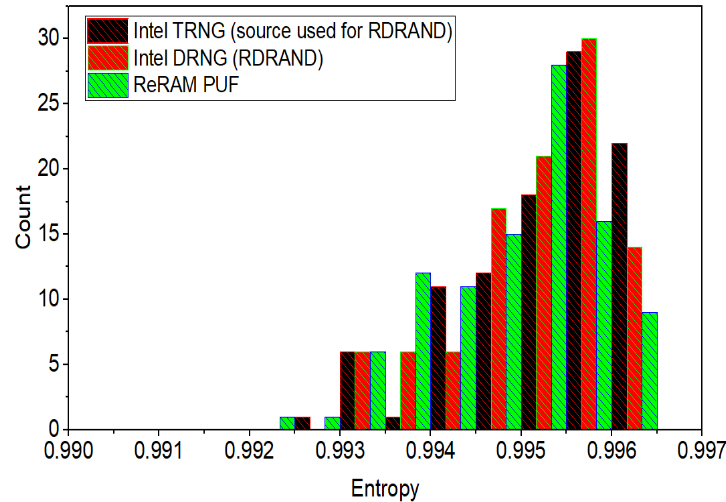


Figure 4.23: Randomness comparison of Intel TRNG (which uses non-deterministic hardware sources for RDRAND), Intel RDRAND (Digital RNG), and the ReRAM-based PUF demonstrated using 28nm-generation CMOS technology in this work. Entropy of 0.990 indicates $2^{-0.990} = 50.3\%$ chance for 0 or 1 in any sequence/circumstance. (1.0 corresponds to perfect entropy of binary bits.)

Table 4.4: Comparison of advanced PUF implementations

	28nm ReRAM PUF	**IEDM'20 [14]**	**IEDM'20 [15]**	**IEDM'21 [16]**	**VLSI'19 [17]**
Technology	28nm	0.13µm	N/A	N/A	14nm
Entropy Source	ReRAM program time	RRAM retention time	VRAM current	Leakage mismatch	dFuse
Bit error rate in typical case	0	0	5.10% @ 85˚C	1.66% @ 85˚C	0.78%
Hamming distance	0.50000	0.49990	0.50004	0.49971	0.49758
NIST test (SP800-22, SP800-90B)	SP800-22: all SP800-90B: all	SP800-22: not all SP800-90B: N/A	SP800-22: not all SP800-90B: all	SP800-22: not all SP800-90B: all	N/A

4.7 Conclusion

A highly reliable and secure PUF architecture and generation scheme that exploits the inherent randomness of ReRAM device program time as an entropy source is experimentally demonstrated using a 28nm-generation CMOS manufacturing process. The fabricated ReRAM devices show excellent non-volatile memory performance, with endurance exceeding 10^4 program/erase cycles and data retention projected to exceed 10 years at 125°C. The generated ReRAM PUF bitstream passed all of the NIST SP 800-22 randomness tests, showed 100% data retention at 150°C, and no overlap between inter- and intra-hamming distance (HD) with inter-HD following an ideal Gaussian distribution. The quality of randomness of ReRAM PUF is found to be comparable to that of Intel RDRAND.

4.8 References

[1] B. Halak, M. Zwolinski and M. S. Mispan, "Overview of PUF-based hardware security solutions for the internet of things," *IEEE 59th International Midwest Symposium on Circuits and Systems (MWSCAS),* pp. 1-4, 2016, doi: 10.1109/MWSCAS.2016.7870046.

[2] R. Roman, P. Najera and J. Lopez, "Securing the Internet of Things," *in Computer*, vol. 44, no. 9, pp. 51-58, Sept. 2011, doi: 10.1109/MC.2011.291.

[3] R. Maes, "Physically Unclonable Functions: Properties. In: Physically Unclonable Functions," *Berlin, Heidelberg: Springer,* pp. 49-80, 2013, doi: 10.1007/978-3-642-41395-7_3.

[4] T. Team, "Software vs. Hardware Encryption: The Pros and Cons," *Dynamic Solutions Group*, (2021). URL:https://www.dsolutionsgroup.com/software-vs-hardware-encryptio.

[5] "Hardware vs Software Encryption" *Data Recovery Specialists*. URL: http://www.datarecoveryspecialists.co.uk/blog/hardware-vs-software-encryption.

[6] Herder, M. -D. Yu, F. Koushanfar, and S. Devadas, "Physical Unclonable Functions and Applications: A Tutorial," in *Proceedings of the IEEE*, vol. 102, no. 8, pp. 1126-1141, Aug. 2014, doi: 10.1109/JPROC.2014.2320516.

[7] R. Maes and I. Verbauwhede, "Physically unclonable functions: A study on the State of the Art and Future Research Directions", *Towards Hardware-Intrinsic Security*, Berlin, Heidelberg: Springer, pp. 3-37, 2010. doi: 10.1007/978-3-642-14452-3_1.

[8] G. E. Suh and S. Devadas, "Physical Unclonable Functions for Device Authentication and Secret Key Generation," *44th ACM/IEEE Design Automation Conference,* pp. 9-14, 2007.

[9] S. Jain, T. Wilson, S. Assiri, and B. Cambou, "Bit Error Rate Analysis of Pre-formed ReRAM-Based PUF," *Intelligent Computing. SAI 2022. Lecture Notes in Networks and Systems*, vol 508. Springer, Cham, doi:10.1007/978-3-031-10467-1_54.

[10] Felicetti, M. Lanuzza, A. Rullo, D. Saccà, and F. Crupi, "Exploiting Silicon Fingerprint for Device Authentication Using CMOS-PUF and ECC," *IEEE International Conference on Smart Internet of Things (SmartIoT)*, pp. 229-236, 2021, doi: 10.1109/SmartIoT52359.2021.00043.

[11] B. Gassend, D. Clarke, M. van Dijk, and S. Devadas, “Controlled Physical Random Functions,” *18th Annual Computer Security Applications Conference, Proceedings.*, pp. 149-160, 2002, doi: 10.1109/CSAC.2002.1176287.

[12] V. P. Yanambaka, S. P. Mohanty, and E. Kougianos, “Making Use of Manufacturing Process Variations: A Dopingless Transistor Based-PUF for Hardware-Assisted Security,” in *IEEE Transactions on Semiconductor Manufacturing*, vol. 31, no. 2, pp. 285-294, May 2018, doi: 10.1109/TSM.2018.2818180.

[13] K. Yang, Q. Dong, D. Blaauw, and D. Sylvester, “14.2 A Physically Unclonable Function with BER $<10^{-8}$ for Robust Chip Authentication Using Oscillator Collapse in 40nm CMOS,” *IEEE International Solid-State Circuits Conference - (ISSCC) Digest of Technical Papers*, pp. 1-3, 2015, doi: 10.1109/ISSCC.2015.7063022.

[14] J. Yang, D. Chen, Q. Ding, J. Fang, X. Xue, H. Lv, X. Zeng, and M. Liu, “A Novel PUF Using Stochastic Short-Term Memory Time of Oxide-based RRAM for Embedded Applications,” in *IEEE International Electron Devices Meeting (IEDM)*, pp. 39.2.1-39.2.4, 2020, doi: 10.1109/IEDM13553.2020.9372050.

[15] J. Yang D. Lei, D. Chen, J. Li, H. Jiang, Q. Ding, Q. Luo, X. Xue, H. Lv, X. Zeng, and M. Liu, “A Machine-Learning-Resistant 3D PUF with 8-layer Stacking Vertical RRAM and 0.014% Bit Error Rate Using In-Cell Stabilization Scheme for IoT Security Applications,” in *IEEE International Electron Devices Meeting (IEDM)*, pp. 28.6.1-28.6.4, 2020, doi: 10.1109/IEDM13553.2020.9372107.

[16] Q. Ding, H. Jiang, J. Li, C. Liu, J. Yu, P. Chen, Y. Zhao, Y. Ding, T. Gong, Q. Luo, J. Yang, Q. Liu, H. Lv, and M. Liu, “Unified 0.75pJ/Bit TRNG and Attack Resilient 2F2/Bit PUF for Robust Hardware Security Solutions with 4-layer Stacking 3D NbOx Threshold Switching Array,” *IEEE International Electron Devices Meeting (IEDM)*, pp. 39.2.1-39.2.4, 2021, doi: 10.1109/IEDM19574.2021.9720641.

[17] R. Hsieh, H. W. Wang, C. H. Liu, Steve S. Chung, T. P. Chen, S. A. Huang, T. J. Chen, and Osbert Cheng, “Embedded PUF on 14nm HKMG FinFET Platform: A Novel 2-bit-per-cell OTP-based Memory Feasible for IoT Security Solution in 5G Era,” *Symposium on VLSI Technology*, pp. T118-T119, 2019, doi: 10.23919/VLSIT.2019.8776515.

[18] C. Böhm, M. Hofer, and W. Pribyl, “A Microcontroller SRAM-PUF,” *5th International Conference on Network and System Security*, pp. 269-273, 2011, doi: 10.1109/ICNSS.2011.6060013.

[19] J. Kim, H. Nili, G. C. Adam, N. D. Truong, D. B. Strukov and O. Kavehei, “Predictive Analysis of 3D ReRAM-based PUF for Securing the Internet of Things,” *IEEE Region Ten Symposium (Tensymp)*, pp. 91-94, 2018, doi: 10.1109/TENCONSpring.2018.8692038.

[20] R. Waser, M. Aono, “Nanoionics-based Resistive Switching Memories,” *Nanosci. Technology*, pp. 158–165, 2009, doi, 10.1142/9789814287005_0016.

[21] J. Yang, D. Strukov, and D. Stewart, “Memristive Devices for Computing,” *Nature Nanotech*, vol. 8, pp. 13–24, 2013, doi: 10.1038/nnano.2012.240.

[22] Pan, S. Gao, C. Chen, F. Zeng, “ Recent Progress in Resistive Random Access Memories: Materials, Switching Sechanisms, and Performance,” *Mater Sci. and Eng. Reports*, vol. 83, pp. 1-59, Sept 2014, doi: 10.1016/j.mser.2014.06.002.

[23] H-S. P. Wong *et al.*, “Metal–Oxide RRAM,” *Proceedings of the IEEE*, vol. 100, no. 6, pp. 1951-1970, June 2012, doi: 10.1109/JPROC.2012.2190369.

[24] R. Waser, R. Dittmann, G. Staikov, and K Szot, “Redox-based Resistive Switching Memories-Nanoionic Mechanisms, Prospects, and Challenges,” Advanced Materials, vol. 21, pp. 2632–63. July 2009, doi: 10.1002/adma.200900375.

[25] M. J. Kim et al., “Low Power Operating Bipolar TMO ReRAM for Sub-10 nm Era,” International Electron Devices Meeting (IEDM), pp. 19.3.1-19.3.4, 2010, doi: 10.1109/IEDM.2010.5703391.

[26] J. Hertz, “Crossbar Reimagines ReRAM Technology for Physically Unclonable FunctionsAll About Circuits, July 2021. URL: https://www.allaboutcircuits.com/news/crossbar-reimagines-reram-technology-for-pufs/.

[27] H. Akinaga and H. Shima, “Resistive Random Access Memory (ReRAM) Based on Metal Oxides,” Proceedings of the IEEE, vol. 98, no. 12, pp. 2237-2251, Dec. 2010, doi: 10.1109/JPROC.2010.2070830.

[28] F. Zahoor, et al, “Resistive Random Access Memory (RRAM): An Overview of Materials, Switching Mechanism, Performance, Multilevel Cell (mlc) Storage, Modeling, and Applications,” Nanoscale Res. Lett., vol 15, 90, 2020, doi: 10.1186/s11671-020-03299-9

[29] M. Kozicki, H. Barnaby, “Conductive Bridging Random Access Memory–Materials, Devices and Applications,” Semicond. Sci. Technol., vol. 31(11), 2016, doi: 10.1088/0268-1242/31/11/113001

[30] T. Tsuruoka, K Terabe, T Hasegawa, and M Aono, “Forming and Switching Mechanisms of Cation-Migration-Based Oxide Resistive Memory,” Nanotechnology, vol. 21, 425205, 2010.

[31] D. Kumar, R. Aluguri, U. Chand, and T.Y. Tseng, “Metal Oxide Resistive Switching Memory: Materials, Properties and Switching mechanisms,” Ceramics International, vol. 43, pp. S547-S556, Aug. 2017.

[32] A. Calderoni, S. Sills and N. Ramaswamy, “Performance Comparison of O-based and Cu-based ReRAM for High-Density Applications,” IEEE 6th International Memory Workshop (IMW), pp. 1-4, 2014, doi: 10.1109/IMW.2014.6849351.

[33] F. Pebay-Peyroula, T. Dalgaty, and E. Vianello, “Entropy Source Characterization in HfO2 RRAM for TRNG Applications,” 15th Design & Technology of Integrated Systems in Nanoscale Era (DTIS), pp. 1-2, 2020, doi: 10.1109/DTIS48698.2020.9081294.

[34] D. Ielmini, “Resistive Switching Memories Based on Metal Oxides: Mechanisms, Reliability and Scaling,” Semicond. Sci. Technol., vol. 31, no. 6, May 2016, doi: 0.1088/0268-1242/31/6/063002.

[35] A. Kalantarian et al., “Controlling Uniformity of RRAM Characteristics Through the Forming Process,” IEEE International Reliability Physics Symposium (IRPS), pp. 6C.4.1-6C.4.5, 2012, doi: 10.1109/IRPS.2012.6241874.

[36] N. Vasileiadis, P. Dimitrakis, V. Ntinas and G. C. Sirakoulis, “True Random Number Generator Based on Multi-State Silicon Nitride Memristor Entropy Sources Combination,” International Conference on Electronics, Information, and Communication (ICEIC), pp. 1-4, 2021, doi: 10.1109/ICEIC51217.2021.9369817.

[37] Y. Pang, et al., "A novel PUF Against Machine Learning Attack: Implementation on a 16 Mb RRAM Chip," IEEE International Electron Devices Meeting (IEDM), pp. 12.2.1-12.2.4, 2017, doi: 10.1109/IEDM.2017.8268376.

[38] M. R. Mahmoodi, et al., "Ultra-Low Power Physical Unclonable Function with Nonlinear Fixed-Resistance Crossbar Circuits," IEEE International Electron Devices Meeting (IEDM), pp. 30.1.1-30.1.4, 2019, doi: 10.1109/IEDM19573.2019.8993618.

[39] A. Herkle, et al., "Extracting Weak PUFs from Differential Nonlinearity of Digital-to-Analog Converters," IEEE International Symposium on Circuits and Systems (ISCAS), pp. 1-5, 2020, doi: 10.1109/ISCAS45731.2020.9181190.

[40] A. Herkle, J. Becker and M. Ortmanns, "Exploiting Weak PUFs From Data Converter Nonlinearity—E.g., A Multibit CT \Delta\Sigma\Delta\Sigma Modulator," IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 63, no. 7, pp. 994-1004, July 2016, doi: 10.1109/TCSI.2016.2555238.

[41] A. Sadr and M. Zolfaghari-Nejad, "Weighted Hamming distance for PUF performance evaluation", Electron. Lett., vol. 49, no. 22, pp. 1376-1378, Oct. 2013, doi: 10.1049/el.2013.2326.

[42] R. Maes, "Physically unclonable functions: Constructions, properties and applications," Ph.D. dissertation, Katholieke Universiteit Leuven, Leuven, Belgium, 2012. URL: https://www.esat.kuleuven.be/cosic/publications/thesis-211.pdf.

[43] T. Arul, et al., "A Study of the Spatial Auto-Correlation of Memory-Based Physical Unclonable Functions," European Conference on Circuit Theory and Design (ECCTD), pp. 1-4, 2020, doi: 10.1109/ECCTD49232.2020.9218302.

[44] A. Rukhin, et al., "A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications," SP 800-22, Rev. 1a, April 2010. URL: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=906762.

[45] Intel, "What is Intel® Secure Key Technology?", 2022, URL: https://www.intel.com/content/www/us/en/developer/articles/technical/what-is-secure-key-technology.html.

[46] D. James, "Intel Ivy Bridge unveiled — The First Commercial Tri-gate, High-k, Metal-Gate CPU," Proceedings of the IEEE 2012 Custom Integrated Circuits Conference, pp. 1-4, 2012, doi: 10.1109/CICC.2012.6330644.

[47] M. Turan, E. Barker, J. Kelsey, K. McKay, M. Baish, and M. Boyle, "Recommendation for the Entropy Sources Used for Random Bit Generation," SP 800-90B, Jan. 2018. URL: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-90B.pdf.

[48] R. Govindaraj, S. Ghosh and S. Katkoori, "Design, Analysis and Application of Embedded Resistive RAM Based Strong Arbiter PUF," IEEE Transactions on Dependable and Secure Computing, vol. 17, no. 6, pp. 1232-1242, 1 Nov.-Dec. 2020, doi: 10.1109/TDSC.2018.2866425.

[49] H. J. Mahanta, A. K. Azad and A. K. Khan, "Power analysis attack: A vulnerability to smart card security," International Conference on Signal Processing and Communication Engineering Systems, pp. 506-510, 2015, doi: 10.1109/SPACES.2015.7058206.

Chapter 5

Conclusion

5.1 Contributions of This Work

Advancements in integrated circuit (IC) manufacturing technology over the past several decades have significantly improved the performance of semiconductor logic and memory devices. Miniaturization of the transistor (the basic building block of an IC) has enabled the production of ever more complex integrated circuit, resulting in increased chip functionality and energy efficiency. In recent years, the proliferation of Internet of Things (IoT) devices and the generation and processing of "big data" have highlighted the need for non-volatile (NV) information storage that can be embedded with energy-efficient logic switches for NV memory (NVM) and hardware security applications.

The objective of this dissertation is to investigate new NVM technologies and applications to address the need for embedded memory and hardware security in IoT devices. Micro-electro-mechanical (MEM) switch designs which previously were developed for ultra-low-power computing are investigated herein for embedded NVM application. Emerging Resistive Random-Access Memory (ReRAM) technology inherently has high device-to-device variability, which is leveraged in this work for hardware security application.

MEM switches are promising for ultra-low-power digital computing because they have the ideal properties of abrupt turn-ON/turn-OFF behavior and zero OFF-state leakage current, enabling sub-50 mV operating voltage across a wide range of operating temperatures [1-2]. MEM switches can be designed to operate as non-volatile (NV) switches, *e.g.*, for reconfigurable interconnects [3-4]. These properties make them attractive for IoT applications such as wearable or disposable electronics that require ultra-low power consumption and relatively low

manufacturing cost. In Chapter 2 the design, fabrication, and operation of logic MEM switches as multi-time programmable NV cells is presented. Controlled contact welding and unwelding is demonstrated to be a viable approach for programming and erasing, with relatively small voltage pulses (< 3V) and excellent retention at elevated temperature (200°C). This allows NV information storage to be embedded with digital logic circuitry with zero incremental fabrication cost.

By inserting a floating gate (FG) layer in the MEM switch fabrication process, electronic charge storage within the gate stack is possible as an alternative means for data storage. In principle, the data retention time of a FG-MEM NV switch should be much longer than that of a conventional floating-gate transistor (used for flash memory devices), because an air gap exists in the FG-MEM NV switch in the OFF-state. In Chapter 3 the structure and operating principle, and the fabrication process flow for the first prototype FG-MEM NV switches are described. Measured electrical characteristics are presented and improvements to the switch design are discussed.

As mobile computing and communication devices become more pervasive and the IoT continues to grow, hardware security is increasingly becoming a concern. Therefore, Resistive Random-Access Memory (ReRAM), an emerging NV memory technology that can be integrated into a conventional CMOS process, is investigated for hardware security application in Chapter 4. ReRAM is presented as a promising technology for implementation of Physically Unclonable Functions (PUFs). A highly reliable and secure ReRAM-based PUF architecture and generation scheme that exploits the inherent randomness of ReRAM device program time as an entropy source is experimentally demonstrated. Generated PUF bitstreams satisfy all the randomness assessments and demonstrate 100% retention after 50 hours at 150˚C. The quality of randomness of ReRAM PUF compares well against that of previously reported PUF implementation schemes. ReRAM technology enables both NVM and PUF functions to be incorporated within the same chip.

5.2 Suggestions for Future Work

While this research has investigated new applications for emerging switch and memory device technologies, there is still room for further enhancements.

5.2.1 NV MEM Switches

The NV-MEM switch design used to demonstrate controlled contact welding and unwelding in this work can be optimized to achieve greater program/erase cycling endurance. In this regard, alternative contacting electrode and structural materials should be explored. The use of contact materials with lower melting temperature than tungsten, such as Niobium, Ruthenium, Gold, *etc.*, should reduce the energy required for program and erase operations, as well as thermal stress sustained by the other layers of the MEM switch during program/erase operation. Switch miniaturization is another pathway for reducing the voltage and time needed to weld and unweld a contacting electrode, as well as the cell area for lower cost and/or higher NV storage capacity.

FG-based MEM switches are potentially advantageous for energy-efficient embedded logic and NVM applications due to their negligible OFF-state leakage and abrupt switching characteristics. The design improvements described in Section 3.4 should be pursued in order for

FG-MEM switches to achieve program/erase speed and energy, endurance and retention performance specifications for a wide range of IoT applications.

5.2.2 ReRAM

Future work for ReRAM-based PUFs should focus on real-world applications, particularly in the field of hardware security. These include device authentication, random number generation, memory protection, *etc*. It would be beneficial to conduct experimental demonstrations of these applications to assess the practicality of ReRAM-based PUFs.

5.3 References

[1] X. Hu, S. F. Almeida, Z. Alice Ye, and Tsu-Jae King Liu, "Ultra-Low-Voltage Operation of MEM Relays for Cryogenic Logic Applications," *IEEE International Electron Devices Meeting,* pp. 34.2.1-34.2.4, 2019, doi: 10.1109/IEDM19573.2019.8993629.

[2] B. Osoba, B. Saha, L. Dougherty, J. Edgington, C. Qian, F. Niroui, J. H. Lang, V. Bulovic, J. Wu, and T.-J. K. Liu, "Sub-50 mV NEM Relay Operation Enabled by Self-Assembled Molecular Coating," *IEEE International Electron Devices Meeting*, pp. 26.8.1-26.8.4, 2016, doi: 10.1109/IEDM.2016.7838489.

[3] K. Kato, V. Stojanović, and Tsu-Jae King Liu, "Embedded Nano-Electro-Mechanical Memory for Energy-Efficient Reconfigurable Logic," *IEEE Electron Device Letters*, vol. 37, no. 12, pp. 1563-1565, Dec. 2016, doi: 10.1109/LED.2016.2621187.

[4] U. Sikder, L. P. Tatum, T.-T. Yen, and Tsu-Jae King Liu, "Vertical NV-NEM Switches in CMOS Back-End-of-Line: First Experimental Demonstration and Array Programming Scheme," *IEEE International Electron Devices Meeting*, pp. 21.2.1-21.2.4, 2020, doi: 10.1109/IEDM13553.2020.9372116.