Semiconductor Memory Applications in Radiation Environment, Hardware Security and Machine Learning System

by

Rui Liu

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved October 2018 by the
Graduate Supervisory Committee:

Shimeng Yu, Chair
Yu Cao
Hugh Barnaby
Jae-sun Seo

ARIZONA STATE UNIVERSITY

December 2018

ABSTRACT

Semiconductor memory is a key component of the computing systems. Beyond the conventional memory and data storage applications, in this dissertation, both mainstream and eNVM memory technologies are explored for radiation environment, hardware security system and machine learning applications.

In the radiation environment, e.g. aerospace, the memory devices face different energetic particles. The strike of these energetic particles can generate electron-hole pairs (directly or indirectly) as they pass through the semiconductor device, resulting in photo-induced current, and may change the memory state. First, the trend of radiation effects of the mainstream memory technologies with technology node scaling is reviewed. Then, single event effects of the oxide based resistive switching random memory (RRAM), one of eNVM technologies, is investigated from the circuit-level to the system level.

Physical Unclonable Function (PUF) has been widely investigated as a promising hardware security primitive, which employs the inherent randomness in a physical system (e.g. the intrinsic semiconductor manufacturing variability). In the dissertation, two RRAM-based PUF implementations are proposed for cryptographic key generation (weak PUF) and device authentication (strong PUF), respectively. The performance of the RRAM PUFs are evaluated with experiment and simulation. The impact of non-ideal circuit effects on the performance of the PUFs is also investigated and optimization strategies are proposed to solve the non-ideal effects. Besides, the security resistance against modeling and machine learning attacks is analyzed as well.

Deep neural networks (DNNs) have shown remarkable improvements in various intelligent applications such as image classification, speech classification and object

localization and detection. Increasing efforts have been devoted to develop hardware accelerators. In this dissertation, two types of compute-in-memory (CIM) based hardware accelerator designs with SRAM and eNVM technologies are proposed for two binary neural networks, i.e. hybrid BNN (HBNN) and XNOR-BNN, respectively, which are explored for the hardware resource-limited platforms, e.g. edge devices.. These designs feature with high the throughput, scalability, low latency and high energy efficiency. Finally, we have successfully taped-out and validated the proposed designs with SRAM technology in TSMC 65 nm.

Overall, this dissertation paves the paths for memory technologies' new applications towards the secure and energy-efficient artificial intelligence system.

ACKNOWLEDGMENTS

My PhD journey would not have been possible without the support of my family, professors, colleagues, collaborators and friends.

First, I would like to give special thanks to my PhD advisor Dr. Shimeng Yu, for supporting me during these past four and half years. I am very grateful for his scientific advice and knowledge and many insightful discussions and suggestions. His guidance helped me in all the time of research, and I have learned a lot from him in the way of writing, communication and presentation. I would also like to thank the members of my PhD committee, Dr. Yu Cao, Dr. Hugh Banarby and Dr. Jae-sun Seo, for their insightful comments and encouragement.

I greatly appreciate and acknowledge the support received Tsinghua University and National Tsing Hua University. I am equally thankful to Dr. Huaqiang Wu, Dr. He Qian and Dr. Bin Gao from Tsinghua University and Dr. Meng-Fan Chang from National Tsing Hua University and other team members from each research group, Yachuan Pang, Wei Wu, Dr. Win-San Khwa, Dr. Wei-Hao Chen, Xin Si, for providing the chips and the tape-out opportunities to evaluate or verify most of my designs described in this dissertation.

Big thanks to all my fellow colleagues and friends at ASU, particularly, Dr. Ligang Gao, Dr. Jiyong Woo, Dr. Pai-Yu Chen, Dr. Zhiwei Li, Xiaoyu Sun, Xiaochen Peng, Panni Wang, Bin Dong, Manqing Mao, Dr. Zihan Xu, Dr. Abinash Mohanty, Xiaocong Du, Deepak Kadetotad, Minkyu Kim, Xiaoyang Mi, Shihui Yin, Dr. Wenhao Chen, for their excellent cooperation, stimulating discussions, encouragement, hardship and happiness we experienced together, in the last four and half years. Besides, I would like to extend my

thanks to Mr. James Laux for helping me with the requisite research tools for my research, and Ms. Delilah Alirez for helping me with the institutional facilities.

I would also like to thank my colleagues from my internship at HP labs for their wonderful collaboration. I would like to particularly single out my supervisors, Dr. Richard Stanley Williams, Dr. John-Paul Strachan and my tutor, Dr. Suhas Kumar for their excellent cooperation, their valuable guidance and for all the opportunities I was given to conduct my research at HP labs.

Finally, I owe thanks to the people who mean a lot to me, my parents and my parents in law, my husband and my sisters, for showing their faith in me and giving me spiritual and financial support and supporting any decision I made, for their selfless love, care, sacrifice they did to shape my life. I consider myself as the luckiest person in the world to have such a lovely and caring family, standing beside me with their love and unconditional support.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

Figure Page

LIST OF TABLES

Table Page

Table Page

1 INTRODUCTION

Semiconductor memory represents a significant portion of the semiconductor market. It is a key component of the computing system and widely used in various applications (for smart phone, tablet, laptop, server, data center, network, communication, gaming, consumer, etc.). The requirements for different applications are different in memory bandwidth, latency, system performance, power, capacity size, and so on. To meet the growing needs for semiconductor market with diverse requirements, there are different types of memory technologies that are being used. As the demands grow, new memory technologies are being introduced and the existing technologies are being further developed.

1.1 Overview of MOS Memories

From a system viewpoint, semiconductor memories are divided into two major categories: volatile memory and non-volatile memory (NVM). Volatile memory loses its storing connect when the power is removed. In contrast, non-volatile memory is able to retain its stored data virtually forever when the power is turned off. Within the volatile memory, there are two major types: dynamic random access memory (DRAM) and static random access memory (SRAM). Within the NVMs, there are a few different major types of memory technology, including Mask ROM, programmable read only memory (PROM) and Flash memory. For Mask ROM, data are written during chip fabrication by the use of a photo mask that contains the write data. For the PROM, however, data are written after chip fabrication. Depending on its erasable or non-erasable characteristics, PROM is further classified into two categories: EPROM (Erasable PROM, typically by ultra-violet light source) and EEPROM (Electrically Erasable PROM). In compared to the other NVMs,

the Flash memory is a random access memory that can be written to, as well as read from. On the other hand, the FLASH memory has many advantages over the other NVMs.

1.2 Mainstream MOS Memory Technologies

Fig. 1.1 shows schematic circuits of mainstream MOS memory cells, which store binary information, "1" or "0", on the above-described memory arrays.

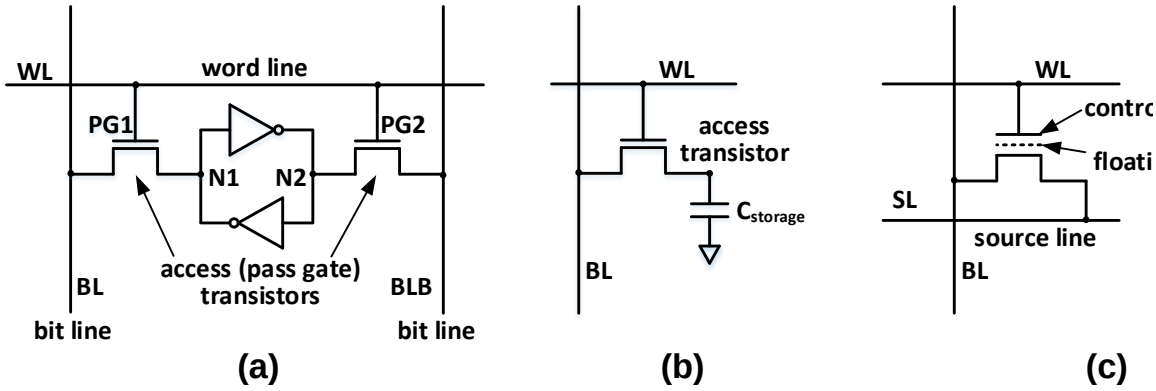


Fig. 1.1 Mainstream memory cell schematics (a) SRAM; (b) DRAM; (c) FLASH.

1.1.1 SRAM

A conventional SRAM cell usually consists of six transistors (Fig. 1.1(a)), referred to as 6T SRAM. Two cross-coupled inverters are used to store the bi-stable information like in a latch. Two pass gate transistors are needed for the access control during the write and read operations. To write the value to the cross-coupled inverters, a differential voltage between a high voltage (i.e. VDD) and a low voltage (i.e. ground) is first applied to a pair of bit lines (i.e. BL and BLB) and then the value can be written to storage nodes (i.e. N1 and N2) by turning on the two access transistors, PG1 and PG2. For example, logic "1" is written to storage node N1 and logic "0" is written to storage node N2. First, BL and BLB are pre-charged to VDD and ground, respectively. Then the data is written the N1 and N2 by turning on PG1 and PG2. The read operation is performed by detecting the polarity of differential signal voltage developed on the bit lines through a sense amplifier. More

specifically, both BL and BLB are first pre-charged to VDD. With turning on the access transistors, the voltages on BL and BLB could be discharged depending on the data stored in the storage nodes, thus resulting a voltage difference between bit lines. On the other hand, the data in SRAM cell is held statically as long as the power is applied.

1.1.2 DRAM

A DRAM cell consists of an access transistor and a capacitor for storing charges, as shown in Fig. 1.1(b). The level of charge on the storage capacitor determines the data stored in the DRAM: the presence of charge in the capacitor indicates a logic "1" and the absence of charge indicates a logical "0". To write a value into the DRAM cell, switch on the access transistor by applying a high voltage (Vpp) to the WL; then apply a voltage (VDD or ground) to the BL depending on the data to be written into DRAM cell, thus the storage capacitor (Cs) being charged or discharged accordingly. For example, logic "1" is written to storage capacitor. VDD is applied to BL and Cs is charged to a high voltage. In the read operation, charge sharing takes place between bit line and storage capacitance. BL is pre-charged to Half VDD in the read operation. Then with asserting WL, charge redistribution occurs between BL and Cs. Depending on the stored data at the Cs, we will see different resultant voltage developed on the BL, which is sensed by a sense amplifier connected to the bit line. Since the charge redistribution destroys the stored information in the read operation read operation is destructive in DRAM and a simultaneous write-back must be contained, which is conducted by the sense amplifier during the sensing phase. Besides, the leakage current (e.g. through off-state current of the access transistor, or the reversed biased the drain-body PN junction in the storage node) degrades an initial high stored voltage, finally causing the loss of information. Therefore, a "refresh" operation is

necessary to retain the value in the storage capacitor before the stored voltage becomes excessively decayed. The “refresh” operation is usually performed by a read operation since it has a write-back operation in DRAM. A succession of refresh operation at a given time interval retains the data. The time interval, which is determined by the leakage current, is about 64ms. The name DRAM is derived from the fact that data is dynamically retained by refresh operation, which differs from SRAM.

1.1.3 FLASH

Each FLASH memory cell consists of the transistor structure with the source and drain electrodes separated by the channel. Above the channel in the FLASH memory cell there is a floating gate which is separated from the channel by a tunneling oxide layer as shown in Fig. 1.1(c). Above the floating gate is the control gate isolated with an inter-poly oxide layer. The presence of charge in the floating gate will then determine the threshold voltage thereby whether the channel will conduct or not. During the read operation, a “1” at the output corresponds to a low threshold voltage or the channel being in its low resistance state. The program/erase cycle for FLASH memory uses a process called Fowler-Nordheim tunneling. The process is performed by applying sufficient large the voltage accross the control gate and the substrate. If the energy barrier is like a triangular shape with thin enough tunneling thickness, electrons could quantum tunnel through the tunneling oxide layer. Generally the program/erase process is needs hundreds of microseconds.

There are two basic types of FLASH memory: NAND and NOR FLASH memory. Although they use the same basic technology, the way they are addressed for reading and writing is slightly different with different array architectures. NAND FLASH memory has a string of floating gate transistors, and is accessed sequentially and must be erased by

block. When NAND FLASH memory is to be read, the contents must first be paged into memory-mapped RAM. This makes the presence of a memory management unit essential. NOR FLASH memory is able to read individual transistor cell randomly with similar array architecture like DRAM. NAND FLASH memory has higher integration density while NOR FLASH memory has much faster read speed, thus they are suitable for different applications, e.g. massive data storage for NAND while code storage for NOR.

1.3 Advantages and Disadvantages of Memory Technologies

Table 1.1 Device Characteristics of Mainstream Memories

	SRAM	DRAM	FLASH	
			NOR	NAND
Cell Area	> 100 F^2	6 F^2	10 F^2	< 4 F^2 (3D)
Multi-bit	1	1	2	3
Voltage	< 1V	< 1V	> 10V	> 10V
Access Time	~ 1ns	~ 10ns	~ 50ns	~ 10μs
Write Energy/bit	~fJ	~10fJ	~100pJ	~10fJ
Endurance	> 10^{16}	> 10^{16}	> 10^5	> 10^4
Retention	Volatile	~64ms	> 10years	> 10years

There are various advantages and disadvantages between different memory technologies. Table 1.1 presents to performance metrics of those memory technologies discussed above. It is necessary to consider all of these when determining the optimum type of memory to be used. If we compare the cell are and access time among different memory technologies, we will see that the SRAM shows the fastest access speed but the largest cell area, however, NAND Flash shows the smallest cell area but the slowest access speed. For a given capacity size, the memory technology with larger cell area is more expensive. However, the programmers desire fast memory with large capacity size, which

is challenging to be satisfied with only one of the memory technologies we discussed. In practical, most programs do not access all code or data uniformly, which is referred as to the principle of locality. Locality occurs in time (temporal locality) and in space (spatial locality). With taking advantages of locality and trade-offs between cost and access speed of memory technologies, an economical solution was led to hierarchies based on memories of different speeds and cell areas. Due to the advantage of low cost, despite medium speed, DRAM is widely and extensively used for the main memory, while SRAM, which features high speed, despite high cost, is used for the cache memory in computers and mobile devices.

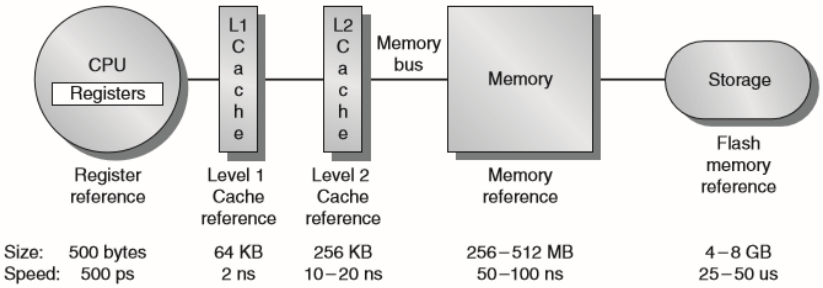


Fig. 1.2 Typical memory hierarchies in a personal mobile devices.[1]

1.4 Memory Hierarchy

Fig. 1.2 shows a multilevel memory hierarchy used in mobile device, including typical sizes and speeds of access. From top layers to bottom layers, SRAM, DRAM, and FLASH are the mainstream memory technologies serving as cache, main memory, and solid-state-drive (SSD), respectively. Moving left the hierarchy, the memory write/read latency decreases. Moving right the hierarchy, the memory capacity increases. Since fast memory is expensive, a memory hierarchy is organized into several levels. The one closer to processor is smaller, faster, and more expensive than the next lower level, which is farther from the processor.

1.5 Emerging Non-Volatile Memory (eNVM) Technologies

1.5.1 Overview

All these mainstream memory technologies discussed above are charged-based memories, which means the charges are used to represent the storing information. They are facing challenges with technology node scaling to 10 nm node or beyond, due to reduced amount of the charges stored in the storing node and the easy loss of them at nanoscale. As a result, it causes degradation of the performance, reliability, and noise margin, etc. In this context, non-charge based emerging memory technologies proposed by the research community and are under active research and development in the industry, but none of them is mature for high-volume production yet. The emerging NVM (eNVM) candidates are spin-transfer-torque magnetic random access memory (STT-MRAM) [2], phase change random access memory (PCRAM) [3], and resistive random access memory (RRAM) [4].

All these emerging NVM technologies are non-volatile two-terminal devices, and they differentiate their states by the switching between two resistance states: a high resistance state (HRS) and a low resistance state (LRS). The switching between the two states can be enabled by electrical inputs. However, each of the eNVMs has its unique switching physics as shown in Fig. 1.3. The resistance of STT-MRAM is determined by the relative direction of two ferromagnetic layers separated by a thin tunneling insulator layer, which corresponds to LRS if they are in parallel configuration, HRS otherwise. The resistance of PCRAM is determined by the phase of chalcogenide material, which corresponds to LRS if it is in crystalline phase and HRS if it is in amorphous phase. RRAM relies on the formation (corresponding to LRS) and the rupture (corresponding to HRS) of conductive filaments in the insulator between two electrodes.

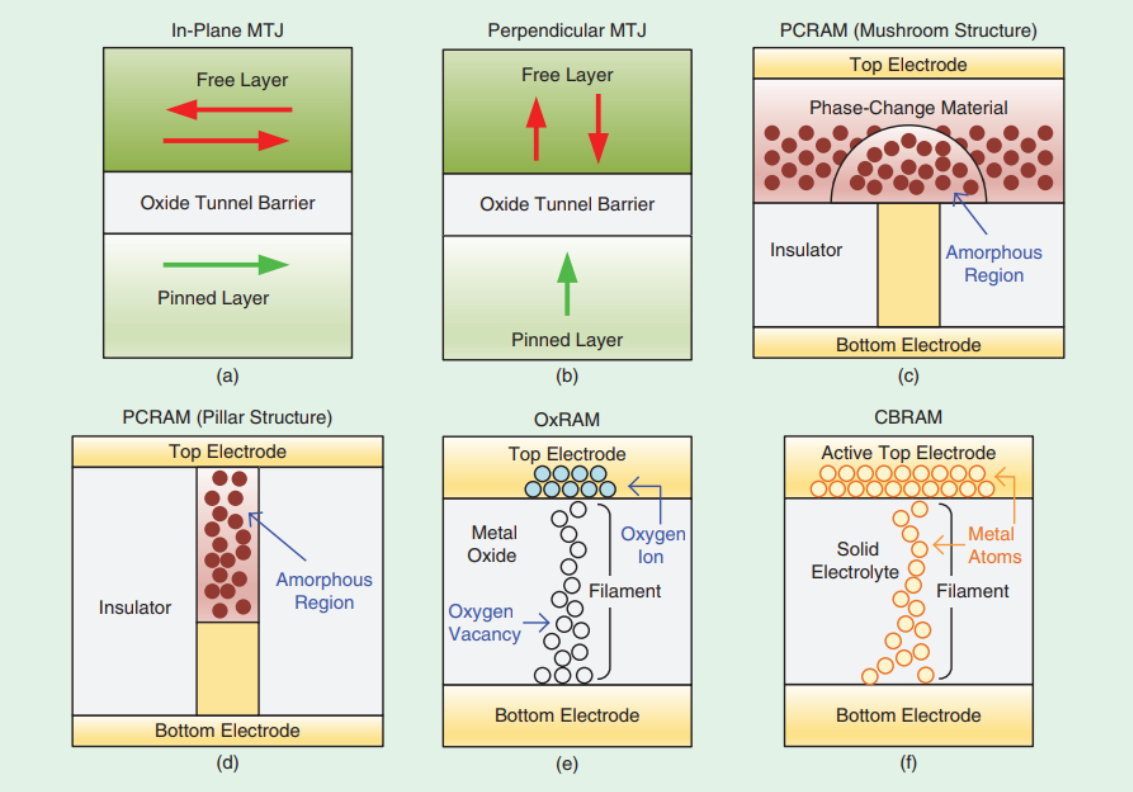


Fig. 1.3 Diagrams of switching mechanism of emerging memory technologies[5]. © 2016 IEEE

Due to the different underlying physics, the device characteristics are also different among different emerging NVM technologies. Table 1.2 compares the typical device characteristics between the eNVMs and the mainstream memory technologies. For different eNVM technology could be used for different applications with considering their unique characteristics. For example, the STT-MRAM was proposed for the SRAM replacement in the last-level cache in [6]. The RRAM was proposed to replace the NOR FLASH for code storage and more ambitiously to replace NAND FLASH as data storage in [7].

Table 1.2 Device Characteristics of Mainstream and Emerging Memories[5].

	MAINSTREAM MEMORIES				EMERGING MEMORIES		
			FLASH				
	SRAM	DRAM	NOR	NAND	STT-MRAM	PCRAM	RRAM
Cell area	>100 F^2	6 F^2	10 F^2	<4F^2 (3D)	6~50F^2	4~30F^2	4~12F^2
Multibit	1	1	2	3	1	2	2
Voltage	<1 V	<1 V	>10 V	>10 V	<1.5 V	<3 V	<3 V
Read time	~1 ns	~10 ns	~50 ns	~10 µs	<10 ns	<10 ns	<10 ns
Write time	~1 ns	~10 ns	10 µs–1 ms	100 µs–1 ms	<10 ns	~50 ns	<10 ns
Retention	N/A	~64 ms	>10 y	>10 y	>10 y	>10 y	>10 y
Endurance	>1E16	>1E16	>1E5	>1E4	>1E15	>1E9	>1E6~1E12
Write energy (J/bit)	~fJ	~10fJ	~100pJ	~10fJ	~0.1pJ	~10pJ	~0.1 pJ

Notes: F: feature size of the lithography. The energy estimation is on the cell-level (not on the array-level). PCRAM and RRAM can achieve less than 4F^2 through 3D integration. The numbers of this table are representative (not the best or the worst cases).

1.5.2 Emerging NVM Array Architectures

Array Architectures

In general, there are two types of RRAM array architectures for integration. The first integration architecture is the one-transistor and one-resistor (1T1R) structure (see Fig. 1.4(a)), where each RRAM cell is in series with a cell selection transistor, as shown. The addition of a selection transistor is able to isolate the selected cell from other unselected cells. The word line (WL) controls the gate of the transistor, thus voltage amplitude applied to the WL can determine the compliance current of the RRAM cell. The bit line (BL) connects to the RRAM anode (top electrode) and contact via of the drain of the transistor connects to the RRAM cathode (bottom electrode). The source line (SL) connects to the source of the transistor. The minimum cell area for the 1T1R architecture is six F^2 (F is the lithography feature size) if a minimum size transistor is used with aggressive borderless DRAM design rules. The cell area increases as the size of the selection transistor is increased when a minimum sized transistor cannot provide sufficient programming current. The second integration architecture is crossbar structure, where rows and columns are perpendicular to each other with RRAM cells sandwiched in between, as shown in Fig. 1.4(b). The crossbar architecture can achieve a smaller footprint of four F^2, thus it can

support a higher integration density than the 1T1R architecture. Crossbar architecture consists of only RRAM devices without selection transistors in the array. At the edge of the array, RRAM devices on the same word line (WL) or bit line (BL) share a common driver (e.g. CMOS inverter). The driver should provide sufficient current for the programming current of the selected cell in addition to the sneak path current of the unselected cells. A simple crossbar array suffers from a problem known as "sneak paths" that limits the array size, increases power consumption, and degrades write/read margins. Therefore, today's crossbar designs typically use stacked cell structures composed of 1-selector and 1-resistor (1S1R) in series. The use of two-terminal selector devices with strong I-V nonlinearity has been shown to significantly suppress sneak currents

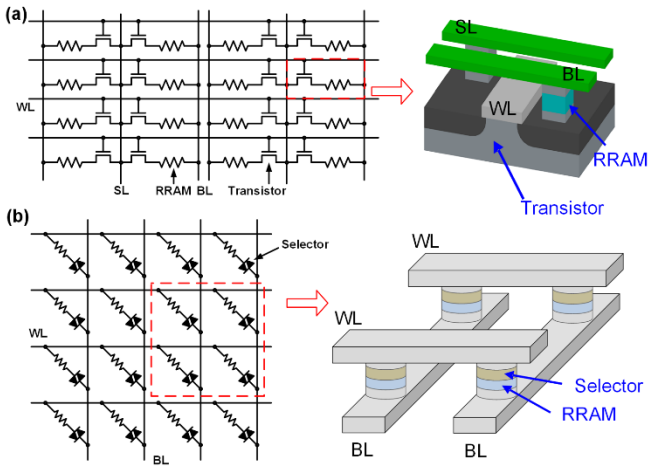


Fig. 1.4 Schematic and layout diagram of (a) 1T1R and (b) crossbar array architectures[8]. © 2015 IEEE.

Write and Read Schemes

To write the RRAM device, it includes two operations, SET and RESET.

For the 1T1R architecture, a positive voltage (i.e. V_{WL}) is applied to WL to turn on the transistor of the selected cell and the WLs of the other unselected cells are grounded. During the SET operation a write voltage (i.e. V_{BL}) is applied to the BL of the selected cell

while SL is grounded. During the SET operation, in contrast, a write voltage (i.e. V_{SL}) is applied to the SL of the selected cell while BL is grounded to reverse the current, as the typical RRAM operation needs bipolar switching. During the both SET and RESET operation, all the BLs and SLs of the unselected cells are grounded. In the read operation, the bias configuration is similar to the SET operation. The only difference is that a smaller V_{BL}, typically less than 0.5 V, is used in the read operation to suppress the read disturbance. The sense amplifier are generally used in the read operation to compare the read-out current along the BL with a reference to determine if the RRAM cell is in HRS or LRS. Because the transistors are turned off for the unselected cells, there is no cross-talk or interference issues in the 1T1R array, and each cell can be independently and randomly accessed. Multiple-bits can be written (or read) in parallel into (or from) the same row by activating multiple columns. It is necessary to point out that different WL voltages are used for SET, RESET and read, respectively. Typically, the WL voltage in RESET operation is larger than the ones in both SET and read operations because part of the SL voltage is dropped on the RRAM cell; thus a larger WL voltage is needed to turn on the transistor. Due to different WL voltages, the SET and RESET operations cannot be performed simultaneously on the same selected row. Therefore, a two-step programming process is needed if we have both "1"(s) and "0"(s) to be rewritten into the 1T1R cells on the same row.

For the crossbar architecture, cross-talk or interference exists between cells in the crossbar array doe to no selection transistors in the array. In order to write the RRAM cells successfully, the V/2 scheme is usually employed. During the SET operation, a write voltage (i.e. V_W) is applied to WL of the selected cell while the BL of the selected cell is

grounded. During the RESET operation, in contrast, a write voltage (i.e. V_W) is applied to BL of the selected cell while the WL of the selected cell is grounded. The WLs and BLs of the unselected RRAM cells in both SET and RESET operations are biased to $V_W/2$. During the read operation, the WL of the selected cell is grounded and the rest of WLs and BLs are applied by a read voltage (i.e. V_R).

1.6 Overview of Contribution

In this dissertation, both mainstream memory and eNVM technologies are investigated and explored in radiation environment, hardware security system and machine learning applications.

For the use in radiation environment, single event effects are investigated on the oxide based RRAM with two different architectures (1T1R and crossbar array architectures). A physics-based SPICE model calibrated with HfOx RRAM is employed for circuit and array-level simulations. The model captures the resistance switching dynamic responses to ion-induced voltage transients. RRAM state-flipping is attributed to transient photocurrents at neighboring transistors. SBU caused by either single-event upset (SEU) or multiple-event upset (MEU) is studied in the 1T1R array. The simulation results correlate with experimentally observed phenomena in HfOx RRAM under heavy ion irradiation. In addition, circuit simulation is performed to investigate the impact of transient induced soft errors in a crossbar array. The crossbar array itself is transistor-free, thus the sensitive locations are the peripheral circuitry (i.e. drivers) only at the edge of the array. The simulations show that the crossbar array with HfOx based RRAM is highly radiation tolerant thanks to the V/2 bias scheme. However, multiple-bits upset (MBU) occurs if other oxides are used that lower activation energies in pursuit of low operation voltage. The

voltage spikes generated at the edge of the array may propagate along the rows or the columns, causing MBUs since there is no isolation between cells in the crossbar array. Thus a trade-off between low power operation and radiation hardness, has to be considered when the crossbar array is designed for use in radiation environments. Besides, the SEU sensitivity is investigated from the system-level. From a circuit-level perspective, the 1T1R is only susceptible to SBU due to the isolation of cells, while in the crossbar, MBU may occur because ion-induced voltage spikes generated on drivers may propagate along rows or columns. Three factors are considered to evaluate system-level susceptibility: the upset rate, the sensitive area, and the vulnerable time window. Our analysis indicates that the crossbar architecture has a smaller maximum bit-error-rate (BER) per day as compared to the 1T1R architecture for a given sub-array size, I/O width and susceptible time window. The result will be presented in Chapter 2.

For the hardware security system application, two RRAM-based Physical Unclonable Function (PUF) implementations are proposed for cryptographic key generation (weak PUF) and device authentication (strong PUF), respectively. In the weak PUF implementation, the entropy source or randomness comes from stochastic switching mechanism and intrinsic variability of the RRAM devices, which is unlike conventional manufacturing process variation based silicon PUFs. The RRAM weak PUF properties such as uniqueness and reliability are experimentally evaluated with 1 kb HfO_2 based RRAM arrays. Firstly, our experimental results show that selection of the split reference and offset of the split sense amplifier (S/A) significantly affect the uniqueness. More dummy cells are able to generate a more accurate split reference, and relaxing transistor's sizes of the split S/A can reduce the offset, thus achieving better uniqueness. The average

inter-Hamming distance (HD) of 40 RRAM weak PUF instances is ~42%. Secondly, we propose to use the sum of the read-out currents of multiple RRAM cells for generating one response bit, which statistically minimizes the risk of early retention failure of a single cell. The measurement results show that with 8 cells per bit, 0% intra-HD can maintain more than 50 hours at 150 °C or equivalently 10 years at 69 °C by 1/kT extrapolation. Finally, we propose a layout obfuscation scheme where the entire S/A are randomly embedded into the RRAM array to improve the RRAM weak PUF's resistance against invasive tampering. The RRAM cells are uniformly placed between M4 and M5 across the array. If the adversary attempts to invasively probe the output of the S/A, he has to remove the top-level interconnect and destroy the RRAM cells between the interconnect layers. Therefore, the RRAM weak PUF has the "self-destructive" feature. In the strong PUF implementation, the sneak paths in the resistive X-point or cross-point array are exploited as the entropy source. The entanglement of the sneak paths in the X-point array greatly enhances the entropy of the physical system, thereby increasing the space of challenge-response pairs (CRPs). To eliminate the undesired collision or diffuseness in X-point PUF with "analog" resistance distribution, "digital" resistance distribution is employed. The effect of design parameters and non-ideal properties in X-point array on the performance of X-point PUF is comprehensively investigated by SPICE simulation. The simulation results show that: 1) the PUF's performance presents strong dependence on the percent of cells in the on-state, thus should be carefully optimized for the robustness against the reference current variation of the S/A; 2) the interconnect resistance decreases the column current thus the reference current should scale down with the scaling of technology node; 3) larger on/off ratio is desired to achieve low power consumption and high robustness against reference current

variation; 4) the device-to-device variation might degrade the performance of X-point PUF, which can be mitigated with write-verify programming scheme in the PUF construction phase. In addition, the proposed X-point PUF presents no correlation between challenges and responses, and strong security against the possible SPICE modeling attack and machine learning attack. Compared to the conventional Arbiter PUF, the X-point PUF has benefits in smaller area, lower energy and enhanced security. The design methodologies and results will be presented in Chapter 3.

For the application in machine learning, increasing efforts have been devoted to develop hardware accelerators. Various hardware platforms (i.e. GPU and FPGA) are used to process DNNs and the various optimizations (i.e. ASIC chips) are used to improve throughput and energy efficiency without impacting application accuracy. However, all of them are based on Von Neumann architecture and memory bottleneck still exists. In the context, in-memory computing was proposed by embedding computing within memory and reducing intermediate data transfer in order to achieve area, throughput and energy efficiency improvement. In our work, two binary neural networks, i.e. hybrid BNN (HBNN) and XNOR-BNN, are explored for the hardware resource-limited platforms, e.g. edge devices. In the HBNN, the weights are binarized to +1/-1 while the neuron activations are binarized to 1/0. In contrast, both the weights and neuron activations are binarized to +1/-1 in XNOR-BNN. With SRAM technology, 6T SRAM and custom 8T SRAM are proposed as bit cells for HBNN and XNOR-BNN implementations, respectively. With RRAM technology, 2 1T1R cells and 4 1T1R cells are proposed as bit cells for HBNN and XNOR-BNN implementations, respectively. In our design, the high-precision multiply-and-accumulate (MAC) is replaced by bitwise multiplication for HBNN or XNOR operation

for XNOR-BNN plus bit-counting operations. To parallelize the weighted sum operation, we activate multiple word lines in the SRAM or RRAM array simultaneously and digitize the analog voltage developed along the bit line by a multi-level sense amplifier (MLSA). To enable the design scalability for arbitrary size of weight matrices in DNNs, we digitized the analog weighed sums collected from each memory array to digital ones with higher precision rather than binary with MLSAs. In order to maintain the sufficient partial sum precision, we propose to use the learned nonlinear quantization technique rather than linear quantization to mitigate the accuracy degradation due to quantization. With 64×64 sub-array size and 3-bit MLSA, HBNN and XNOR-BNN architectures can minimize the accuracy degradation to 2.37% and 0.88%, respectively, for an VGG-like network on the CIFAR-10 dataset. Design space exploration of the proposed architectures with the conventional row-by-row access scheme and our proposed parallel access scheme are also performed, showing significant benefits in the area, latency and energy-efficiency. Finally, we have successfully taped-out and validated the proposed HBNN and XNOR-BNN designs with SRAM technology in TSMC 65 nm process with measured silicon data, achieving access time ~2.3ns, and energy-efficiency ~111 TOPS/W for HBNN and ~65 TOPS/W for XNOR-BNN, respectively. The design methodologies and result will be presented in Chapter 4.

2 MEMORIES USED IN RADIATION ENVIRONMENT

2.1 Overview

Radiation-induced transient faults arise from energetic particles, such as alpha particles from packaging material and neutrons from the atmosphere, generating electron-hole pairs (directly or indirectly) and transient photocurrent as they pass through a semiconductor device. A sufficient amount of accumulated charge or transient current spike may change the state of a logic device, such as a latch, or gate, thereby introducing a logical fault into the circuit's operation. Because this type of fault does not reflect a permanent malfunction of the device, it is referred to as soft error.

2.1.1 Major Radiation Resources

Typical sources of ionizing radiation are the cosmic rays, the Van Allen radiation belts, Solar flares, nuclear reactors in power plants, particle accelerators, residual radiation from isotopes in chip packaging materials, and nuclear explosions.

Cosmic rays consist of approximately 90% protons (i.e hydrogen), 9% alpha particles (i.e. helium), and 1% heavy ions, together with x-ray and gamma ray radiation [9]. Earth's magnetic field and atmosphere shields the planet from 99.9 percent of the radiation from space. Therefore, they are primarily a concern for spacecraft and high-altitude aircraft. Van Allen radiation belts contain electrons (up to 10 MeV) and protons (up to 100 MeV) trapped in the geomagnetic field [10]. The particle flux density and the location of the peak flux can vary wildly depending primarily on solar activity and the magnetosphere. Due to their position, they endanger satellites. Solar particle events come from the sun and consist of a large flux of high-energy (several GeV) protons and heavy ions, and together with x-ray radiation, which will present a significant radiation hazard to spacecraft. Nuclear reactors

produce gamma radiation and neutron radiation, which can affect sensor and control circuits in nuclear power plants. Particle accelerators produce high energy protons and electrons, and the secondary particles produced by their interactions produce significant radiation damage on sensitive control and particle detector components, of the order of magnitude of 10 MRad[Si]/year for systems such as the Large Hadron Collider (LHC). Chip packaging materials were an insidious source of radiation that was found to be causing soft errors in DRAM chips in the 1970s [11]. Traces of radioactive elements in the packaging of the chips were producing alpha particles, which were then occasionally discharging some of the capacitors used to store the DRAM data bits. These effects have been reduced today by using purer packaging materials, and employing error-correcting codes (ECC) to detect and often correct DRAM errors. Nuclear explosions produce a short and extremely intense surge through a wide spectrum of electromagnetic radiation, an electromagnetic pulse (EMP), neutron radiation, and a flux of both primary and secondary charged particles.

2.1.2 Types of Single-Event Effects

In recent years, the dominant radiation effect in space electronic systems has become the family of single-event effects (SEEs). SEEs arise through the action of a single ionizing particle as it penetrates sensitive nodes within electronic devices. There are a variety of possible single event effects and they are very important as they can cause malfunctions in microelectronics devices operating in the space environment. The basic effects are as follows:

(1) SET—single-event transient, which means the transient introduce by single event.

(2) SEU—single-event upset, which means the temporary state change of memory or register. This do not cause lasting damage to the device. A single ion can cause single-bit upset (SBU) or multiple-bit upset (MBU) in several adjacent memory cell in advance technology nodes. SEU will be the primary effect in our study.

(3) SEFI—single-event functional interrupt, which means the control circuits (e.g. state machines) are corrupted into an undefined state by an single-even upset

(4) SEL—single event latchup, which means device with a parasitic PNPN structure is latched in high current state or "shorted".

(5) SES—single-event snapback, which means regenerative current mode in NMOS

(6) SEB— single-event burnout, which means device draws high current and burns out. This is hard error.

(7) SEGR— single-event gate rupture, which means gate is destroyed in power MOSFETs.

Considering these and other problems in the space systems, it becomes important to understand the single particle errors in the integrated circuit (IC) systems for the space-based use. This problem becomes even more important becomes even more important as device dimensions scale, and denser integrated systems are placed in space or satellite applications. With the device dimensions keeping scaling, a single bit information in the memory devices is represented by an extremely small value of charge and noise margins are very tight as well. For example, if a typical dynamic random access memory (DRAM) can tolerate approximately 100mV of noise on the bit storage node with 10's fF of storage capacitance, then this value of noise corresponds to a charge of only 10's thousands electrons. Any disturbance of this delicate balance by an incident particle is intolerable.

Therefore, it is essential to recognize and get familiar with the effects of space radiation on the electronics in that hostile environment.

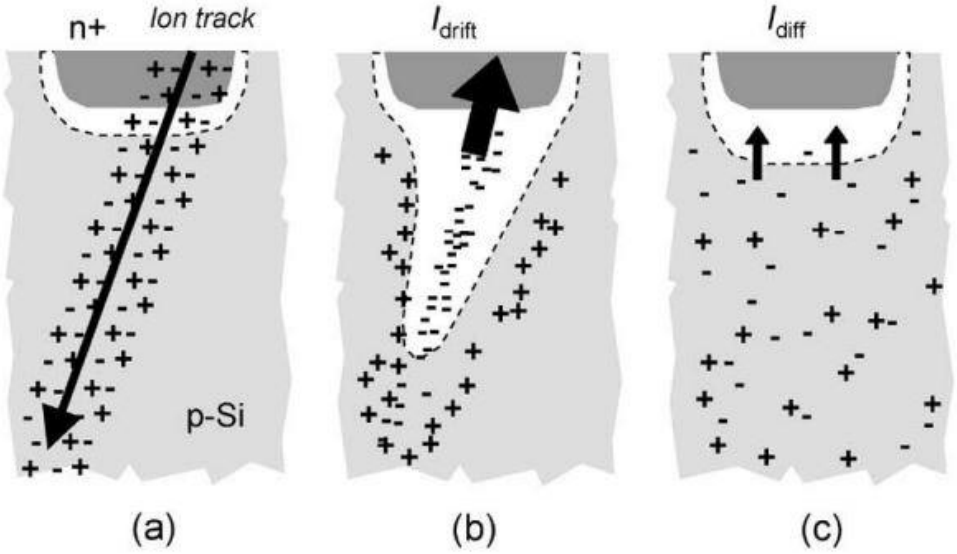


Fig. 2.1 Charge generation and collection phases in a reverse-biased PN junction [12].© 2005 IEEE.

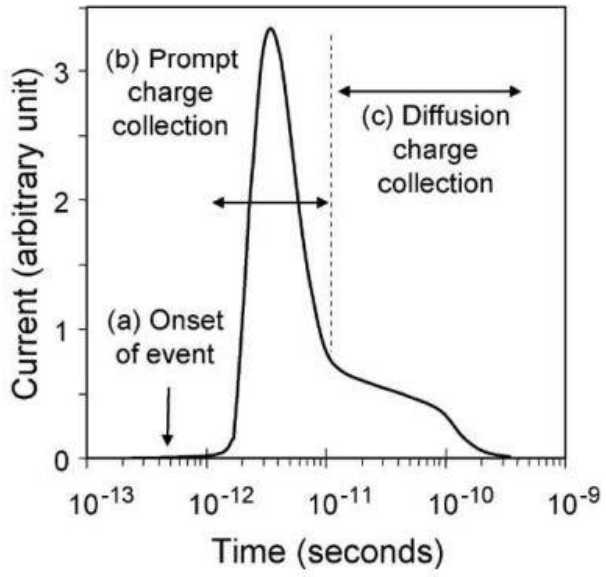


Fig. 2.2 The resultant current transient caused by an ion strike on silicon device with a reverse-biased PN junction in Fig. 2.1[12]. © 2005 IEEE.

2.2 Physical Mechanism of Radiation Effects in MOS Device.

When an ion strikes a silicon device, the ion track leaves a dense plasma of electron-hole pairs along its path as shown in Fig. 2.1(a). If the electron-hole plasma is generated in a region with an electric field, e.g. at a PN junction, electrons and holes are rapidly split by the electric field creating a large current or voltage transient at the circuit node [12]. The charge collection has two phases. First, a prompt collection phase that lasts for the order of hundreds of picoseconds and the carries are collected in the original depletion region and the funnel region [13], as shown in Fig. 2.1(b). Second, a delayed phase that lasts for hundreds of nanoseconds and the carries diffuse up to the depletion region and are quickly collected by the junction electric field (Fig. 2.1(b)). Fig. 2.2 shows the resultant current transient in the different phases shown in Fig. 2.1. The reverse-biased PN junction is the most charge-sensitive part of circuits.

2.3 Radiation Effects in Mainstream Semiconductor Memories.

Extensive investigation on the radiation effects on the mainstream semiconductor memories has been reported in past a few decades since they have reached the industrial maturity. In this section, only the trend of soft errors with technology node scaling is summarized for different memory technologies.

2.3.1 DRAM

The radiation-induced soft error was first reported in 1970s, which mainly caused by the alpha particles emitted by uranium and thorium impurities in packaging materials [11]. With technology node scaling, The volume of charges on the storing node decreases as with the decreasing area of the junction (junction/well doping also plays a role) while cell capacitance remains relatively constant with scaling since it is dominated by the external

3-D capacitor cell. Such DRAM device scaling results in DRAM cell voltage scaling. Voltage reduction reduces the critical charge (Q_{crit}), but with concurrent aggressive junction volume scaling, a much more significant reduction in collected charge is observed. The net result to DRAM soft error rate (SER) is shrinking about 4~5× per generation. While DRAM bit SER has been reduced by more than 1000 times over seven generations, the DRAM system SER has remained essentially unchanged [12]. System requirements have increased the memory density (bits/system) almost as fast as the SER reduction provided by technology scaling. Thus, DRAM system reliability has remained roughly constant over many generations.

2.3.2 SRAM

In contrast, early SRAM was more robust against SER because of high operating voltages and the fact that data in an SRAM are stored as an active bi-stable state (made up of two cross-coupled inverters), each one strongly driving the other to keep the SRAM bit in its programmed state. The Q_{crit} for the SRAM cell is largely defined by the charges on the node capacitance which keep the node voltage at the proper value. With technology scaling, the SRAM junction area has been deliberately minimized to cell area. On the other hand, the SRAM operating voltage has been scaled down to minimize power as well. With each successive SRAM generation, reductions in the amount of charges in the storing node due to cell area shrinking have been cancelled out by big reductions in operating voltage and reductions in node capacitance. SRAM single bit SER was initially increasing with each successive generation. Most recently, as the technology nodes have been reduced into the deep submicron regime (< 0.25 μm), the SRAM bit SER has saturated and may even be decreasing. This saturation is primarily due to the saturation in voltage scaling,

reductions in amount of junction charges, and increased charge sharing due to short-channel effects with neighboring nodes. Although the SRAM bit SER has saturated in the deep submicron region, the SRAM system SER does not saturate as scaling also implies increased memory density. The exponential growth in the amount of SRAM in processors has led the SRAM system SER to increase with each generation with no end in sight. This trend is of great concern to chip manufacturers since SRAM constitutes a large part of all advanced integrated circuits today.

2.4 Radiation Effects of Emerging NVM Technology

Emerging NVM technologies include STT-MRAM, PCRAM, and RRAM. The oxide-based resistive random access memory (OxRAM) is one of the most promising emerging NVM technologies due to its attractive attributes including excellent scalability (<10 nm), low programming voltage (<3 V), fast switching speed (<10 ns), high on/off ratio (>10), high endurance (up to 10^{12} cycles) and compatibility with the silicon CMOS technology [4]. In general, there are two types of RRAM array architectures for integration, 1T1R and crossbar array architectures. A simple crossbar array suffers from "sneak path" that limits the array size, increases the power consumption, and degrades the write/read margin [14]. However crossbar designs now use stacked cell structures composed of one-selector and one-resistor (1S1R) in series [15, 16]. The use of 1S1R structures has been shown to significantly suppress sneak current [17, 18]. In the recent years, industrial development efforts focused on oxide RRAM has led to the production of various prototype chips, e.g. ITRI's 4Mb HfOx 1T1R array [19], Panasonic's 8Mb TaOx crossbar array [20], SanDisk/Toshiba's 32Gb MeOx crossbar array [21].

For the space applications, the lack of low-cost high-density radiation-hardened NVM is one of the key challenges in the design of systems for the hostile space environment. Today's FLASH technology can only sustain a total ionizing dose (TID) up to 75 krad(Si) and FLASH suffers functional failures during write due to the radiation-induced charge pump degradation [22]. Oxide RRAM devices have already been demonstrated to be robust to TID > 1 Mrad (Si) in individual cell [23-26]. However, there are limited studies on the SEE-induced soft errors in RRAM at the array-level. There are a few paper reporting on the single event effect (SEE) susceptibility of HfOx-based [27], TaOx-based [28] and chalcogenide-based 1T1R structures [29].

Single event modeling plays a key role in the understanding of the observed-error mechanisms in existing systems, as well as the prediction of errors in newly designed systems. There are two different aspects of interest. First is the analysis of various types of single event experiments to help understand the phenomena. Second is the modeling of the various aspects of the phenomena that allow prediction of SEE rates in space. SEU plays a key role in memory devices. In this section, we mainly focus on the investigation of SEU in RRAM design with both 1T1R and crossbar architectures at the circuit level.

2.4.1 Modeling the RRAM Device

We employed a physics-based RRAM SPICE model [30] with adding new feature of resistance retention failure mechanism that is essential for RRAM reliability evaluation. The heat conduction process in the transient form is also reformulated, which is necessary for applications that explore ultrafast I-V responses below nanoseconds, e.g., in assessing the susceptibility of RRAM devices to a heavy ion strike. The model parameters are validated with the experimental data of IMEC HfOx-based RRAM devices [31-33]. It is

known that the switching characteristics of oxide based RRAM are strongly dependent on the distribution of oxygen vacancies in the oxide layer, where vacancies act as electron hopping sites for current conduction. In this model, a single dominant conductive filament in one dimension is assumed. The primary internal variable that determines the resistance is the gap distance, which is defined as the average distance between the electrode and the tip of the conductive filament. Gap distance is modulated under the sufficient electric field, thereby changing the RRAM resistance through an electron tunneling conduction mechanism.

We first use the RRAM SPICE model in a 1T1R cell structure to determine its programming conditions for SET and RESET. If not specified, the PTM model of the 45nm bulk transistors [34] is used in the HSPICE simulations in this work. In the 1T1R array architecture, the SET operation is performed by applying positive voltage pulse on the BL and WL, while grounding the SL and body of the transistor. RESET is performed by applying positive pulses on WL and SL, while grounding BL and the body. The WL pulse has a longer width than the BL or SL pulses as shown in Fig. 2.3(a) and (b). Fig. 2.3(c) shows the current transients of BL and Fig. 2.3(d) plots the resistance transients of RRAM during SET and RESET operations. In SET operation, a relatively lower voltage pulse is applied on WL (1.15 V for a transistor with W/L ~ 3) to provide a compliance current which sets a lower-bound for the LRS resistance (~100 kΩ). In RESET, a larger positive bias is applied on the WL (3 V for transistor with W/L ~ 3) so that the resistance of transistor is low enough that most of the voltage on the SL and BL drops across the RRAM. If different sized transistors are used in 1T1R, bias conditions are varied in order to switch RRAM between the same HRS (~1 MΩ) and LRS (~100 kΩ). Table 2.1 lists bias

conditions for transistor with W/L ratios of 1, 2 and 3, respectively. Higher voltage biases are needed to provide the same amount of switching current for smaller W/L transistors.

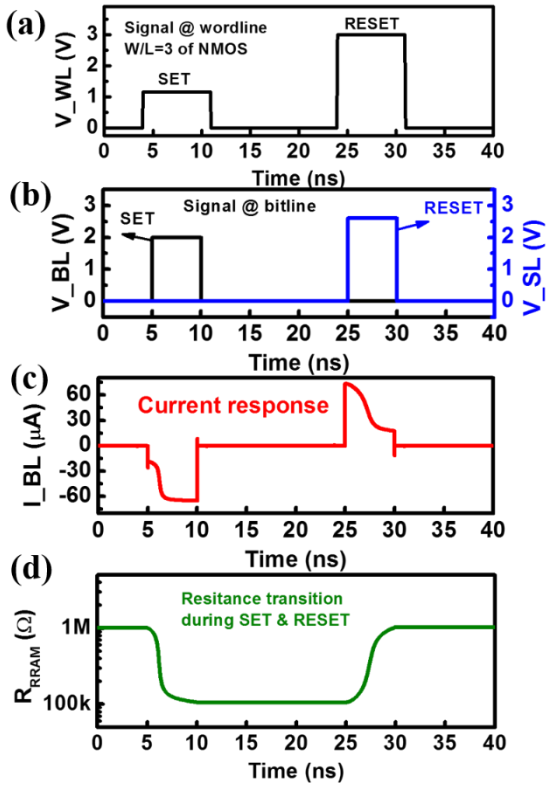


Fig. 2.3 Transient waveforms of 1T1R in SET and RESET operation using RRAM SPICE model. © 2015 IEEE.

Table 2.1 Three Sizes of Selection Transistor and Associated Programming Conditions

W/L	V_WL (V) SET/RESET	V_BL (V)	V_SL (V)
1	2.2/3.9	2.4	3.28
2	1.4/3.2	2.2	2.8
3	1.15/3.0	2.0	2.6

2.4.2 Modeling the Photocurrent at the Device

When a heavy ion strikes the reverse biased PN junction formed by the drain and body of the transistor, it will create electron-hole pairs, resulting in photocurrent flowing

between the drain and the bulk. We employ a photocurrent model as a function of LET following prior work [18]. The total energy lost per unit path length when the particle strikes the drain of a transistor can be expressed in terms of LET as,

$$E = LET \times \rho_0 \tag{2.1}$$

where ρ0 is density of the irradiated material, in this case Si.

When the ionizing particle penetrates through the drain-body junction, electron-hole pairs can be rapidly collected in drain (electrons) and body (holes) due to the reverse bias field. The electron-hole pair generation rate (G) can be estimated as,

$$G = \frac{E}{w \times \pi \times r^2 \times \tau} \tag{2.2}$$

where w is a material property that specifies the energy required to generate a single electron-hole pair, τ is the ion transit time, and r is the path track radius. Here we assumed that the radius is a linear function of LET with α being the constant of proportionality between r and LET [35],

$$r = r_0 + \alpha \times LET \tag{2.3}$$

The transport of electron-hole pairs result in a reverse drain-body diode current, namely a transient photocurrent (I_{ph}) [36] that can be approximated as,

$$I_{ph} = qAG(x_j + L_e) \tag{2.4}$$

where q is the electronic charge, A is the drain area, x_j is the junction depth of the drain and L_e is electron diffusion length. Table 2.2 summarizes the parameters in the photocurrent model.

In order to validate the model described above, we simulated 10-CMOS-inverter delay chain as in [37]. The particle strike occurs on the drain of the off-NMOS of the first inverter in the chain. The transistor model used here is the 0.13μm bulk transistors in the PTM

model. In addition, a dynamic capacitor, in parallel with photocurrent, is modeled as the changing capacitance due to shrinking of the depletion region of PN junction and electron and hole pairs generated by the particle strike [38]. The parameters used for simulation are the same as those used in [37]. The peak current through the NMOS transistor during the strike as a function of LET is shown in Fig. 2.4. The result shows that our analytical photocurrent model fits well to the TCAD simulation results in [37].

Table 2.2 Variable Declaration

Variables	Value
ρ0	2.33e3 mg/cm-3
w	3.6e-6 Mev
r0	1e-5 cm
τ	10e-12 s
α	1.05e-7
xj	1.4e-6 cm
Le	1e-4 cm

In this work, the photocurrent model as described above is used to simulate the effects of ion strikes with various LETs. Fig. 2.5 shows peak photocurrent as a function of LET and transistor size. The results indicate that the photocurrent increases monotonically with the increase in LET and W/L ratio.

2.5 RRAM Sensitivity for Upset

As discussed in session 1.5.2, 1T1R and crossbar are the two typical architectures for integration. In the 1T1R design, each RRAM cell is integrated with a selection transistor. Crossbar architecture consists of only RRAM devices without selection transistors in the

array. At the edge of the array, RRAM devices on the same WL or BL share a common driver (e.g. CMOS inverter).

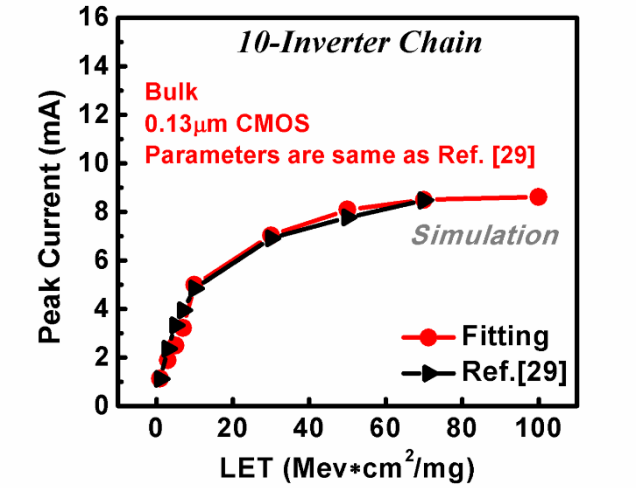


Fig. 2.4 Peak transient current as a function of strike LET for 0.13μm bulk technology [37]. © 2015 IEEE.

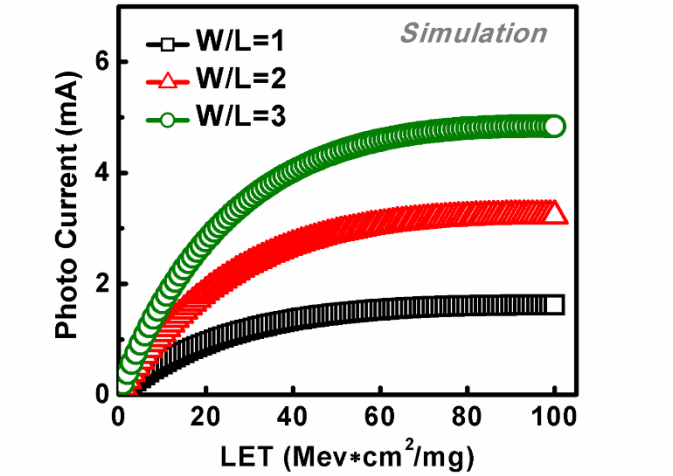


Fig. 2.5 Photocurrent transient peak as a function of strike LET and transistor size [8]. © 2015 IEEE.

The write operation involves two processes, SET and RESET. In the 1T1R architecture, the WL of the selected cells are set to a positive voltage (i.e. V_WL) to make sure the selection transistor is "ON" when a cell is being written (SET) or erased (RESET). In the SET operation, the selected BL is fixed to a positive voltage of V_BL and source line (SL)

is grounded (Fig. 2.6(a)). In the RESET operation, the selected BL is grounded and the selected SL are biased by a positive voltage of V_SL. The other unselected WL, SL and BL are all grounded. In the crossbar architecture, the 1/2 V bias scheme is usually employed [14]. In this scheme, unselected WLs and BLs are all biased to half write voltage Vw/2, while the selected WL and BL are biased to full write voltage Vw and 0 V for the SET operation, 0 V and Vw for the RESET operation, respectively. Therefore, during either SET or RESET, some cells share either WL or BL signals with the selected cell. These "half-selected" cells are colored as purple (WL shared) or blue (BL shared) in (Fig. 2.6b). As these "half-selected" cells are biased, they suffer certain degree of undesired write disturbance. All the unselected cells sharing neither WL nor BL with the selected cell will have 0 V applied across the electrodes if neglecting the interconnect wire resistance. We focus on the 1/2 V bias scheme for the crossbar architecture.

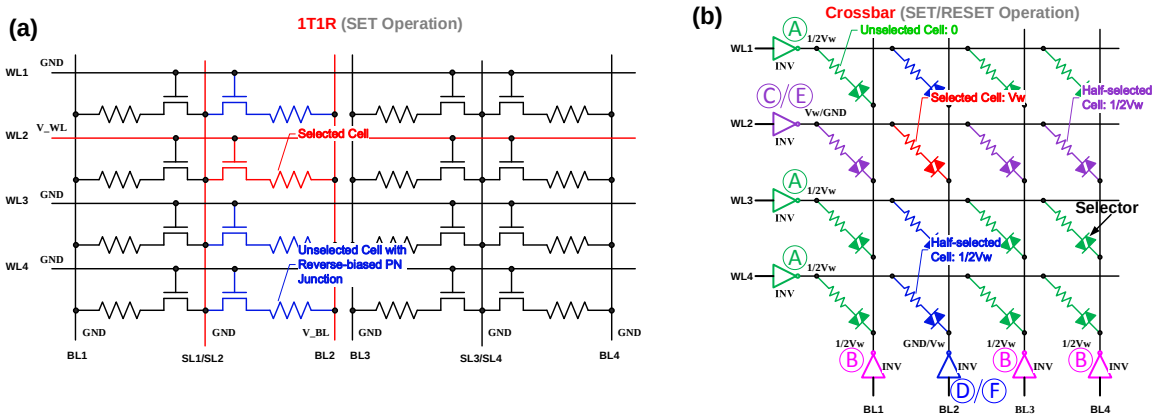


Fig. 2.6 Schematic of (a) the 1T1R architecture and (b) crossbar architecture with 1/2 V bias scheme.

2.5.1 Sensitivity for Upset in 1T1R Architecture

In the 1T1R architecture, each RRAM device has a selection transistor in series. The susceptible RRAM are the cells in blue in the same column as the selected cells being

written (see Fig. 2.6(a)). In the SET operation, the sensitive locations are the drain-body junctions tied through unselected RRAM devices to the selected BL. The generated holes will sink to the body of the transistor, while the electrons will be collected to the drain of the transistor. Therefore, ion-induced positive photocurrent will flow through the RRAM cell from anode to cathode. Depending on the magnitude of current, voltage on the cathode might be pulled down even to a negative potential. This could cause a suitably high positive voltage differential between anode and cathode thereby making an RRAM cell susceptible to HRS to LRS flipping. However, in the RESET operation, the sensitive locations are source-body junctions tied to the selected SL. A negative photocurrent is generated only for the angular strike. This is because there is a parasitic bipolar effect caused by the forward biasing of body-drain junction [39]. This may cause a spurious transition from LRS to HRS under extreme circumstances. In work [40], no LRS to HRS upset was shown in their experiments. This suggests the flipping from LRS to HRS has a very small possibility in the 1T1R structure. Therefore, only HRS to LRS transition with normal incidence strike is considered for the 1T1R architecture at the circuit level modeling and system level analysis.

SEU

An SEU occurs when a single incident ion changes the RRAM's state. For memory application with binary states, we define that it is an HRS to LRS upset when the resistance falls below 100kΩ from an initial high state (i.e. HRS). In contrast, it is an LRS to HRS upset when resistance increases above 1MΩ from an initial low state (i.e. LRS). Given a sufficiently high photocurrent, the drain voltage, which initially is 2V as in the case of IMEC HfOx RRAM device and when the transistor W/L ratio is three, drops to a negative

value. This results in a corresponding net positive voltage transient across the RRAM, which triggers the HRS to LRS upset in the 1T1R.

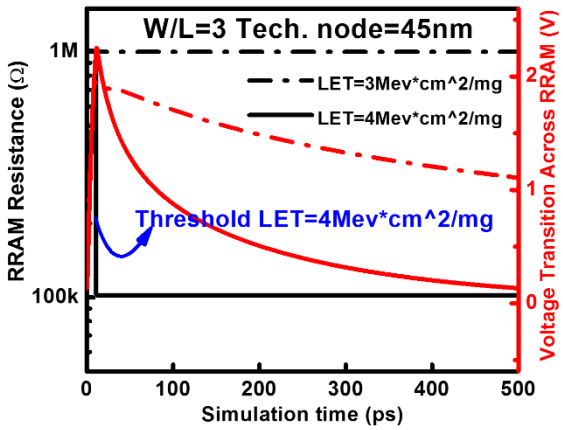


Fig. 2.7 Simulated voltage transients generated across 1T1R and their corresponding resistance change. The threshold LET is in range of 3 to 4 MeV·cm^2/mg [8]. © 2015 IEEE.

Fig. 2.7 presents the voltage transient waveforms and the corresponding resistance transition, respectively, of a RRAM with 1T1R structure under the particle strike. The simulation result indicates the threshold LET of SEU is ranging from three to four MeV*cm^2/mg. Since the RRAM resistance change is more strongly dependent on voltage than current [41, 42], the likelihood of an upset increases substantially when a larger positive bias is applied on the BL during the particle strike [40]. In fact, our simulation result shows that the threshold LET decreases less than three MeV·cm^2/mg when the access transistor W/L ratio is two. This is because a larger BL voltage (i.e. 2.2 V) is required to SET RRAM (see Table 2.1). It is noted that the threshold LET obtained here is smaller than what was reported in [40] because the transistor's drivability in our study is smaller than the transistors used in [40], thus a higher BL voltage is needed, increasing the susceptibility.

MEU

RRAM is stable in a continuum of resistance levels. If multiple ions strike the same cell before a new data is written into the cell, the cumulative change in resistance by each particle strike can eventually upset the cell, even though each strike itself is not able to change state of the cell. Multiple-event upset (MEU) was experimentally observed in the heavy ion irradiation test on 1T1R [40]. Fig. 2.8 shows that a MEU occurs when several ions with the same LET of three MeV·cm^2/mg (which is smaller than the threshold LET) strike the drain of the transistor consecutively.

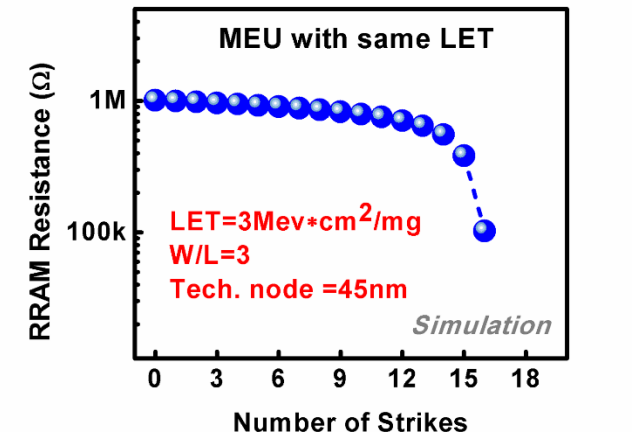


Fig. 2.8 Resistance change in accumulative fashion, causing MEU. The incident ion LET is 3MeVcm^2/mg, which is smaller than the threshold LET for SEU [8]. © 2015 IEEE.

2.5.2 Sensitivity for Upset in Crossbar Architecture

Crossbar array is an architecture attractive for ultra-high density applications and holds the potential for the 3D integration. In order to support a large-scale crossbar array, a selector is generally stacked with the RRAM cell to suppress parasitic sneak currents from unselected cells. In our simulation, a state-of-the-art selector, FAST selector is employed [18], which shows excellent I-V nonlinearity, steep turn on slope and high endurance. With the help of the selector, the resistance of the half selected cells is assumed to be 10^9 Ω, which is able to support a 1024 × 1024 array.

In the crossbar array, there are drivers (i.e. CMOS inverters) at the edge of the array to drive WLs and BLs. Any "OFF" transistor of the driver has a reversed biased PN junction, which is sensitive to the incident ion strike. If an incident particle strikes the drain of the "OFF" transistor (PMOS or NMOS), a transient voltage spike occurs at the output of the driver. The spike generated on the driver at the edge propagates along the WL or BL since there is no isolation between cells in the crossbar array. This is a particular problem in the crossbar architecture, which, however, does not exist in 1T1R design. This can lead to a MBU.

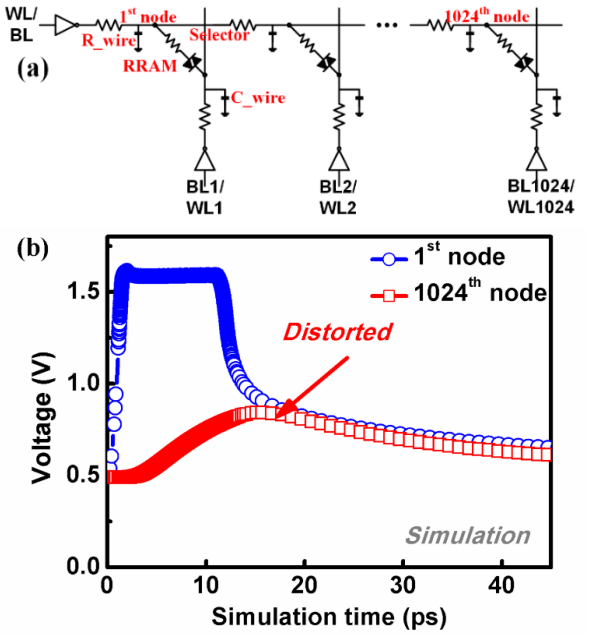


Fig. 2.9 (a) Schematic of simulation circuit for analyzing the crossbar array taking into account the parasitic wire capacitance and resistance, (b) Single event transient spike propagation in 1024×1024 RRAM array. The transient spike is significantly attenuated when it propagates across the entire array [8]. © 2015 IEEE.

Fig. 2.9(a) shows a schematic used to examine the MBU mechanism in the RRAM crossbar array. We consider a 1024×1024 array where one WL drives 1024 RRAM cells. Each cell unit consists of a wire resistor and a wire capacitor and a RRAM cell with a

FAST selector. The wire resistance and capacitance are calculated in 45 nm using the ITRS table [43]. Once a single event transient has been produced in the struck device, e.g., PMOS of an inverter at the edge of WL, its transient evolution along the WL can be simulated in HSPICE. To deliver programming current to the crossbar array, the driver's NMOS W/L is set to be three, and PMOS W/L is set to be six. Fig. 2.9(b) illustrates the voltage spike transient's propagation at the extremes of the WL in 1024 ×1024 crossbar array. For a given LET, the voltage transient, although substantial at the stuck node at the edge of the array, is significantly attenuated as it travels along the WL. Therefore, only a certain number of cells that are closer to the edge of the array may see sufficiently large net voltage transient to flip the states. In addition, we also investigate the numbers of cells which will be susceptible to the heavy ion strike on the transistors at the near edge of the array.

The net voltage difference across RRAM cells can be either positive or negative depending on whether the incident ion strikes the drain of the driver's PMOS or the NMOS. This implies both HRS to LRS and LRS to HRS upsets can occur with roughly the same probability. Table 2.3 summarizes all possible sensitive transistors ("OFF" transistors) and the potential upset types during SET or RESET operation in the crossbar architecture as shown in Fig. 2.6(b). The drivers at edge are classified into six categories depending on their bias voltages. Drivers A and B are the drivers with outputs of Vw/2 on the rows (WL) and columns (BL), respectively, in both SET and RESET operations. Drivers C and E are WL drivers with outputs of Vw in SET and ground in RESET, respectively. Drivers D and F are BL drivers with output of ground in SET and Vw in RESET, respectively. The descriptions of the driver categories can be referred to Fig. 2.6(b). Due to parasitic capacitances, wire resistances and the loading effects of other cells, the generated spike

amplitude is degrading while traveling. Thus only the cells that are close to the driver at edge are susceptible to the upset. Depending on the LET of ion, different numbers of cells may experience upsets, namely, from P_A (LET) for Driver A to P_F (LET) for Driver F.

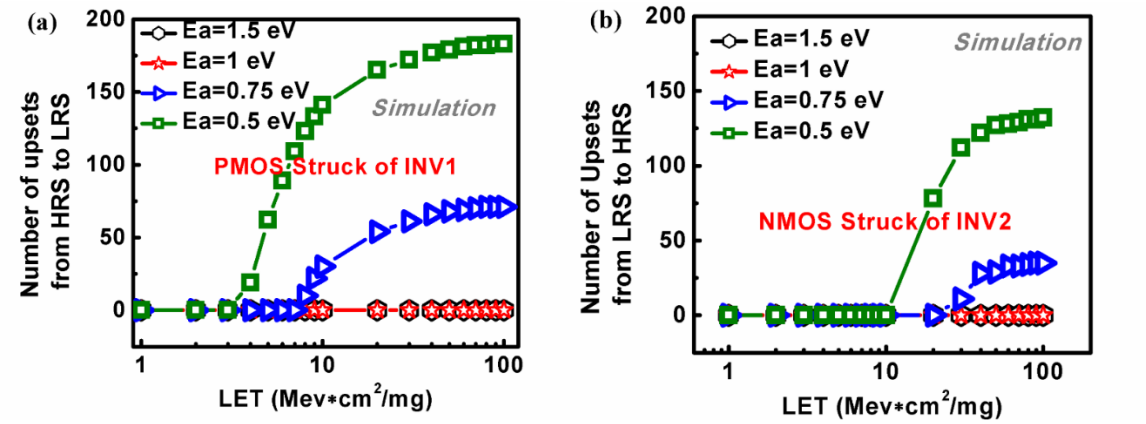


Fig. 2.10 Multiple-bit upset in crossbar array induced by single event transient with RRAM having different switching voltage or Ea. (a) HRS to LRS flipping and (b) LRS to HRS flipping [44]. © 2016 IOP.

It should also be noted that the susceptibility of the RRAM crossbar array strongly depends on bias voltage. If Vw is 2V as in the case of IMEC HfOx RRAM, then half-selected RRAM cell may see -1.7V spike on it in the worst case. As -1.7V is less (in absolute value) than the RESET voltage (-2V), half-selected RRAM may not be flipped from LRS to HRS. If Vw requirement for switching is smaller, e.g. 1V, then half-selected RRAM may see net -1.2V voltage spike in the worst case. As -1.2 V is larger (in absolute value) than the RESET voltage (-1V), half-selected RRAM cell could be flipped from LRS to HRS. On the other hand, the switching voltage of RRAM strongly is also determined by the activation energy (Ea) of oxygen vacancy migration, which is strongly dependent on the material set of the RRAM.

Next, we will reduce the Ea in the RRAM device model to lower the Vw required for the write operation. In practice, lowering Ea can be achieved by changing RRAM oxide

materials. Table 2.4 lists the Vw required for different RRAMs with different Ea. The Vw is reduced to be 0.6 V when Ea is decreased to 0.5 eV. Fig. 2.10 illustrates the simulation results of MBU for HRS to LRS flipping and LRS to HRS flipping, respectively. It is seen that the reduced switching voltage of RRAM (i.e., lower Ea) significantly increases the number of bit upsets in the crossbar array. We interpret this phenomena as a certain particle strike is more easily to cause more number of RRAM cells with lower Ea to see net voltages exceeding their switching voltages. It is also noted that the threshold LET of a PMOS struck is smaller than that of NMOS struck because the area of PMOS is twice the size of NMOS in an inverter. It is also observed that HRS to LRS flipping has a lower threshold LET than LRS to HRS flipping as the RESET needs a larger voltage than SET as predicated by the RRAM model.

Table 2.3 Sensitive Transistors and Potential Upset Types in Fig. 2.6(b)

Process	Driver (Inverter)	Bias	Off Transistor	Upset Type	Upset Bits
SET	A	1/2Vw	NMOS	LRS -> HRS	P_A(LET)
	B	1/2Vw	NMOS	HRS -> LRS	P_B(LET)
	C	Vw	NMOS	LRS -> HRS	P_C(LET)
	D	GND	PMOS	LRS -> HRS	P_D(LET)
RESET	A	1/2Vw	NMOS	LRS -> HRS	P_A (LET)
	B	1/2Vw	NMOS	HRS -> LRS	P_B (LET)
	E	GND	PMOS	HRS -> LRS	P_E(LET)
	F	Vw	NMOS	HRS -> LRS	P_F(LET)

Table 2.4 Required Vw of RRAM with Different Ea

Ea (eV)	Vw (V)
1.5	2
1	1.3
0.75	0.95
0.5	0.6

2.6 Sensitivity Comparison between 1T1R and Crossbar Architectures

As the MBU effect only occurs in the low-voltage operation, to compare the sensitivity for upset between 1T1R and crossbar, Vw is designed to be 1V by changing the activation energy of an oxide material in the RRAM SPICE model. Table 2.5 shows the voltage biases with same RRAM oxide material during SET and RESET operations in the 1T1R and crossbar architectures.

Table 2.5 Voltage Bias for 1T1R and Crossbar Architectures

	1T1R			Crossbar
Process	V_WL (V)	V_BL (V)	V_SL (V)	V_W (V)
SET	0.7	1.0	0	1.0
RESET	1.6	0	1.1	1.0

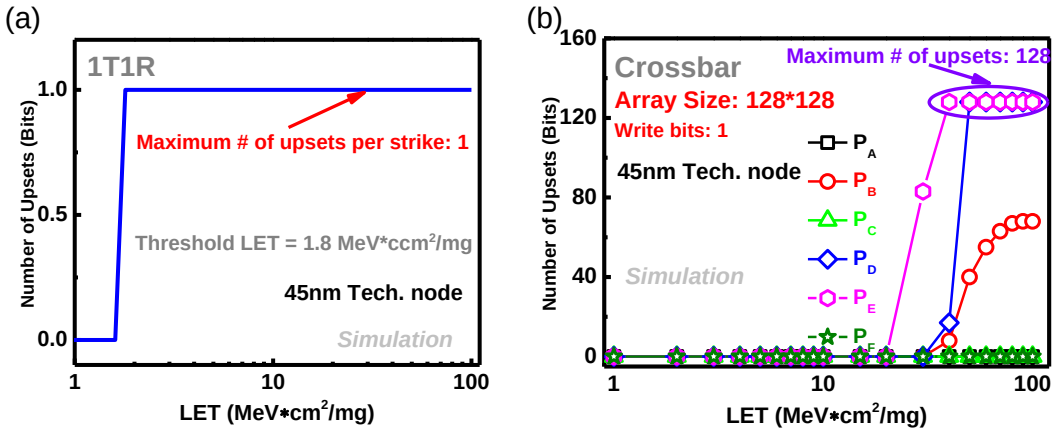


Fig. 2.11 (a) SBU from HRS to LRS due to SEE in 1T1R array. (b) MBU from HRS to LRS or from LRS to HRS induced by SET [44]. © 2016 IOP.

As mentioned above, only SBUs are assumed to occur when a heavy ion strikes the cathode node of RRAM during the SET operation in a 1T1R structure.

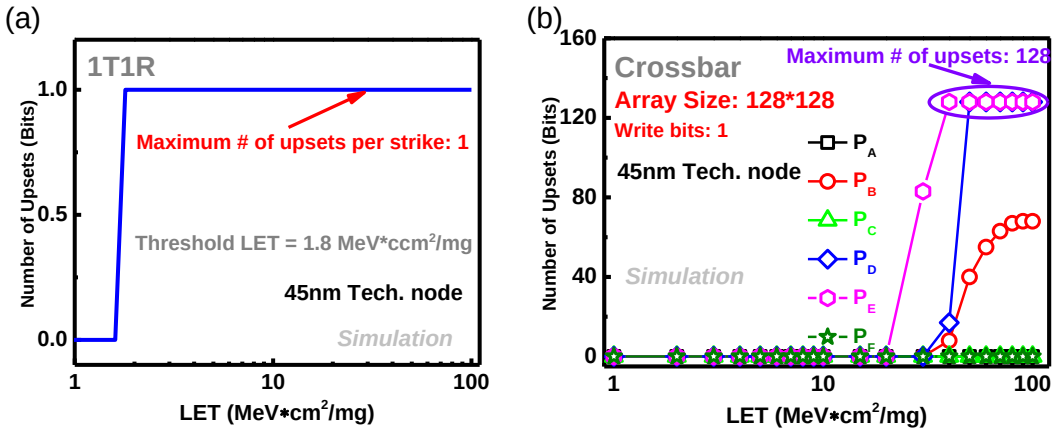


Fig. 2.11(a) presents the simulated single cell upset probability for HRS to LRS flipping as a function of the incident ion LET. The threshold LET is found to be 1.8 MeV·cm^2/mg from the HSPICE simulations. The threshold LET is generally affected by a few factors. The first factor is the sensitive area, which decides the magnitude of ion-induced photo-current and the net voltage differential across RRAM. The second factor is the intrinsic switching voltage of RRAM, which determines how much net voltage across RRAM can flip its desired state. The third factor is the applied voltage on BL, which will determine how much voltage transient is required to cause enough voltage difference on the RRAM. Compared to work [28], our simulation results presented a larger LET at a given applied voltage on BL. For example, the threshold LET of 1T1R was around 4 MeV·cm^2/mg and it might be less than 1 MeV·cm^2/mg at a given BL voltage of 2 V (extrapolated in work [28]). This is because the sensitive area or size of access transistor used in our work (i.e. W = 135 nm, L = 45 nm) is much smaller than that in work [28] (i.e. W = 1 µm, L = 100 nm). The threshold LET in this work (1.8 MeV·cm^2/mg) is decreased from simulation results shown in previous section (i.e. 3-4 MeV·cm^2/mg). The reason is that the RRAM

switching voltage is reduced by reducing the activation energy of an oxide material in the RRAM SPICE model.

For the crossbar architecture, both HRS to LRS and LRS to HRS flipping can occur in the crossbar architecture as discussed earlier. In order to quantify the number of upsets of either type that may occur, HSPICE simulations are performed on arrays programed with two data patterns. One array is preprogrammed with all the cells in LRS (~100 kΩ) to get the number of upsets from LRS to HRS. The other one is preprogrammed with all the cells in HRS (~1 MΩ) to obtain the number of upsets from HRS to LRS. In the crossbar array simulation, the parasitic capacitances and wire resistances (for 45nm node) are considered.

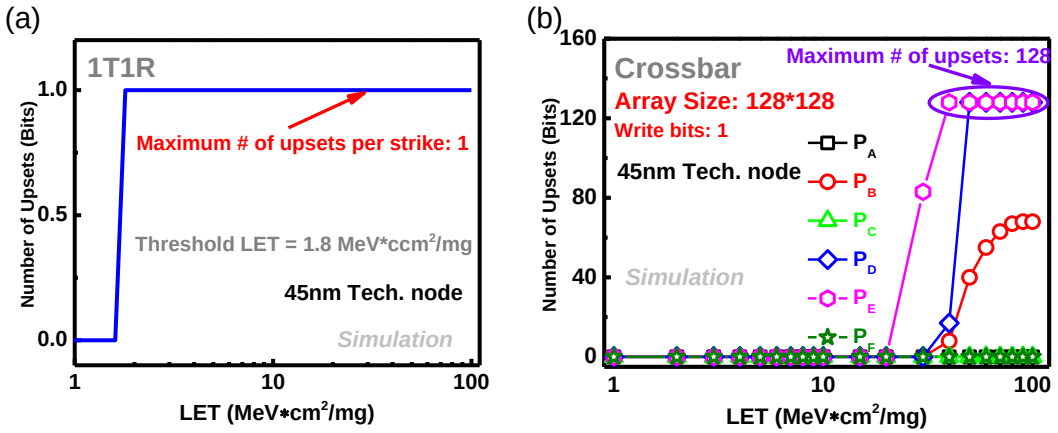


Fig. 2.11(b) shows the MBU effect simulation result for a 128 × 128 array when only one bit is written into the array. It indicates that the number of upsets increases as the incident ion LET increases. This can be attributed to the larger voltage spike generated at the output of the driver when the incident ion LET is larger. It also suggests that the number of upsets strongly varies with the types of drivers being struck. First, Drivers A and B have the same output voltage of 1/2 Vw and sensitive transistor (i.e. NMOS). A strike on the NMOS drain of driver A on the row can cause LRS to HRS upset. However, a strike on the NMOS drain of driver B on the column can cause HRS to LRS upset. The number of upsets,

P_A, is smaller than PB for a given LET due to a lower switching voltage from HRS to LRS in the RRAM model. The difference between P_D and PE has a similar explanation. Compared to P_B, P_E presents more upsets because the drain area of PMOS is larger than that of NMOS, which results in a larger photocurrent and larger voltage spike. Finally, P_F has a smaller number of upsets compared to P_B for a given LET because higher WL voltage (i.e. Vw) requires higher photocurrent to pull it down to a certain voltage.

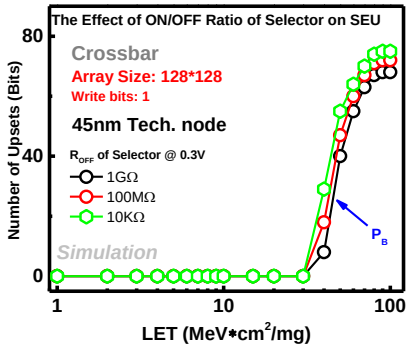


Fig. 2.12 The number of upset bits increases as the lower quality threshold selector with lower OFF resistance is employed [44]. © 2016 IOP.

In addition, we found that the quality of selector also affects the number of upsets in the crossbar architecture. Fig. 2.12 shows the number of upsets under the P_B condition for different threshold selector with different R_{OFF} resistances. The result suggests that a selector with higher off-resistance is desired in the design of crossbar memory for stronger radiation hardness. If we compare the 1T1R and the crossbar architectures, the crossbar architecture generally presents a higher threshold LET than the 1T1R architecture.

2.6.1 Single Event Upset Rate

Since the number of upsets varies significantly with the types and locations of transistors being struck and the states of RRAM cells, especially in the crossbar architecture, we have to consider all the strike scenarios. In order to obtain an “expected”

number of upset cells when an incident ion strikes a sensitive transistor, the average number of upsets is calculated by considering the probability of each strike.

In the 1T1R architecture, it has been assumed that an SEU only occurs during the SET operation. Due to the selection transistor, only one bit may be flipped when a heavy ion strikes one of the sensitive locations. Then the average SEU rate (number of upsets) is

$$R_{SEU} = P_{SEU\_SET}\rho_{HRS}\rho_{SET}W_B\,, \tag{2.5}$$

where P_{SEU_SET} is the number of bits, either zero or one, flipped for a given ion LET (see

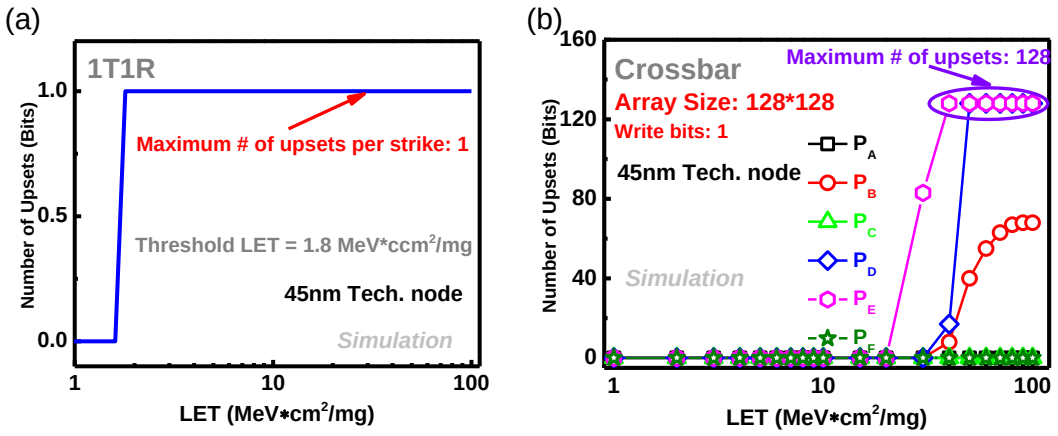


Fig. 2.11(a)), ρ_{HRS} is the probability that a RRAM cell not accessed and in the same column is in HRS, ρ_{SET} is the probability that a cell programming process is performing the SET operation, and W_B is the number of bits to be written in parallel into the array in each address loop. The value for W_B determines the number of BLs being activated in the SET operation. The average SEU rate for the 1T1R architecture is not dependent on the sub-array size because it only involves SBUs.

In the crossbar architecture, an SEU can occur during both SET and RESET operations. Due to the MBUs, the number of upsets is dependent on the size of sub-array (i.e. N × N) and the number of activated BLs (i.e. W_B) in both SET and RESET operations. During

SET operation, when an ion hits one of the "OFF" transistors, the average SEU rate in an N × N array is

$$R_{SEU\_SET} = [(N-1)P_A\rho_{LRS} + (N-W_B)P_B\rho_{HRS} + P_C\rho_{LRS} + W_BP_D\rho_{LRS}]/2N \quad (2.6)$$

where ρ_{LRS} is the probability of a RRAM cell in LRS. Similarly, the average SEU rate for the RESET operation is

$$R_{SEU\_RESET} = [(N-1)P_A\rho_{LRS} + (N-W_B)P_B\rho_{HRS} + P_E\rho_{HRS} + W_BP_F\rho_{HRS}]/2N \quad (2.7)$$

If the probability of a write process performing SET is ρ_SET and performing RESET is ρ_{RESET}, then the average SEU rate in a write process in the crossbar architecture is

$$R_{SEU} = R_{SEU\_SET}\rho_{SET} + R_{SEU\_RESET}\rho_{RESET} \quad (2.8)$$

Where,

$$\rho_{LRS} + \rho_{HRS} = 1 \quad (2.9)$$

$$\rho_{SET} + \rho_{RESET} = 1 \quad (2.10)$$

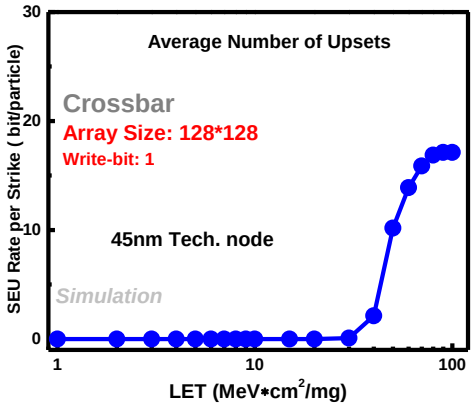


Fig. 2.13 Average SEU rate in a 128 × 128 array with one bit written in the crossbar architecture [44]. © 2016 IOP.

We assume that RRAM cells have the equal probability to be in HRS and LRS and the SET operation has the same frequency as the RESET operation. Thus, the average SEU rate in the 1T1R and in the crossbar can be determined using the above equations. Fig. 2.13 shows the average SEU rate of a crossbar array with a size of 128 × 128 when only one bit

is written into the array. In practice, the size of the driver varies with array size and the number of bits to be written into an array simultaneously in one address loop.

Table 2.6 Driver Size of Crossbar for Different Array Size

Array Size	Number of Activated Cells in an Array	W/L of NMOS
128 × 128	1	1
	4	1
	16	3
256 × 256	1	1
	4	2
	16	4
512 × 512	1	2
	4	3
	16	6

Table 2.6 shows the size of the NMOS of the driver for crossbar arrays with different array size and different numbers of bits to be written in parallel into an array. The PMOS of the driver is twice as large as the NMOS. In the photocurrent model, the photocurrent will scale with drain area of the device when the drain width and length are smaller than 200 nm, which is comparable with the ion radius (~100 nm) [37]. Fig. 2.14(a) shows that the SEU rate increases with more bits written into an array in parallel. This is mainly because a larger driver is used to provide enough drive current when more bits are written into an array simultaneously. For a given incident ion LET, the larger driver can generate larger photocurrent and incur larger transient voltage on the output of the driver, resulting in more upsets. The driver size may also be increased with larger array size. However, the parasitic resistances and capacitances of the wires and the loading effect of other cells increase with the array size as well. As a result, the generated spike attenuates faster as it

propagates along WL or BL in a larger array than that in a smaller array. Therefore, a trade-off exists between SEU rate and array size in the crossbar architecture. The 256 × 256 array presents the lowest SEU rate, as shown in Fig. 2.14(b).

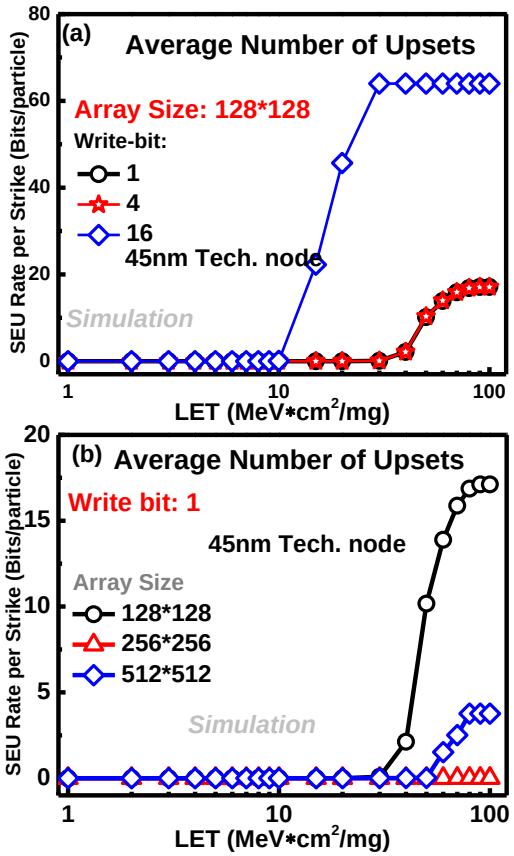


Fig. 2.14 (a) Average SEU rate increases with increasing the number of bits to be written in the 128 × 128 array. (b) Average SEU rate varies with array size for a given number of bits to be written. The 256 × 256 array presents the lowest SEU rate [44]. © 2016 IOP.

2.6.2 Bit Error Rate per Day Estimation

BER per day at the system level caused by SEU can be calculated by integrating the SEU rate, the ion flux, the sensitive area and the susceptible time window as

$$BER \cdot Day(LET) = \iint R_{SEU}(LET)\phi(LET)dAdt \tag{2.11}$$

where R_{SEU} (LET) is the SEU rate (from Eq. (4) for 1T1R or (6) for crossbar), ϕ(LET) is the integral ion flux. The susceptible time window is only when the programming operation

(either SET or RESET) is performed in the RRAM system. In this work, we employed an integral LET spectra for galactic cosmic rays during solar minimum to calculate the SEU-induced BER in the RRAM system (Fig. 6 in work [45]). For the 1T1R architecture, as mentioned above, the sensitive locations are the drain-body junctions of the "OFF" transistors tied to the selected BL in SET operation. Therefore, the sensitive area is the total area of the drains of the "OFF" transistors connected to the selected BL. Then for an N × N array with W_B bits to be written simultaneously, the sensitive area can be calculated as

$$Area_{1T1R} = A_{D\_N}(N-1)W_B n \tag{2.12}$$

where N-1 is number of "OFF" transistors tied to a selected BL, and W_B is number of selected BLs, and n is the number of activated arrays in the RRAM system. If I/O bit width of the system is N_{IO}, then

$$n = N_{IO}/W_B \tag{2.13}$$

Thus the sensitive area of the 1T1R becomes

$$Area_{1T1R} = A_{D\_N}(N-1)N_{IO} \tag{2.14}$$

which is only determined by the I/O width no matter how many sub-arrays are activated. For the crossbar architecture, in contrast, the sensitive area is the total area of the drains of the both N-type and P-type "OFF" transistors of the edge drivers. For an N × N crossbar array, the total number of drivers is 2N. In the SET operation, all the WL are biased to Vw or 1/2Vw and all the N-type transistors are "OFF"; W_B of BLs are grounded if W_B bits are written into the simultaneously, hence there are N-W_B N-type and W_B P-type "OFF" transistors on the BL. In total, 2N-W_B N-type and W_B P-type "OFF" transistors in SET operation. The analysis is similar in the RESET operation. Thus the sensitive area is calculated as

$$Area_{crossbar} = \{[(A_{D\_N}(2N - W_B) + A_{D\_P}W_B)\rho_{SET}] + [(A_{D\_N}(2N - 1) + A_{D\_P})\rho_{RESET}]\}N_{IO}/W_B \quad (2.15)$$

which is not only determined by the I/O width but also how many sub-arrays are activated. For example, for the 64 bits I/O used in the system, if we write 8 bits into one sub-array in parallel, we need to activate eight sub-arrays.

In this work, we assume the RRAM cell write latency is 20 ns. Then the write latency of a page (e.g. 4kB) is same as Ref. [46]. In addition, the most write-intensive workload in a solid-state-drive (SSD) in Ref. [47] is used to calculate the total write time window for the system, which is 308.9 s/day. Taking those assumptions into account, Fig. 2.16(a) and (b) show the maximum BER/day for the 1T1R array and the crossbar array, respectively. In the 1T1R architecture, it indicates that the maximum BER per day is exponentially increased with the sub-array size. However, it has no dependency on the bits to be written simultaneously in a sub-array (Fig. 2.16(a)). The reason is that the sensitive area is only as a function of I/O bit width of a system (from Eq. (2.13)). In the crossbar architecture, however, both the number of bits to be written in one sub-array and the array size can affect the maximum BER per day even for a given I/O bit width (i.e. 64 bit) and total write time window (i.e. 308.9 s/day). It suggests that array size of 256 × 256 has the lowest maximum BER per day if a single bit is written into one array as shown in Fig. 2.16(b).

If we compare the 1T1R and the crossbar architectures, the crossbar array does not have transistor in the array (only at the edge), thus it has smaller sensitive area than the 1T1R array. On the other hand, the 1T1R only shows SBU, the crossbar array may have MBUs, thus once the crossbar is struck, it would generate more bit errors. However, the threshold LET of the crossbar structure is much higher than that of the 1T1R structure. In addition, the integral LET spectra for galactic cosmic rays is decreased exponentially with ion LET

(see Fig. 2.15). As a result of all the aforementioned factors, our analysis shows that the crossbar architecture still has a smaller maximum BER/day as compared to the 1T1R architecture.

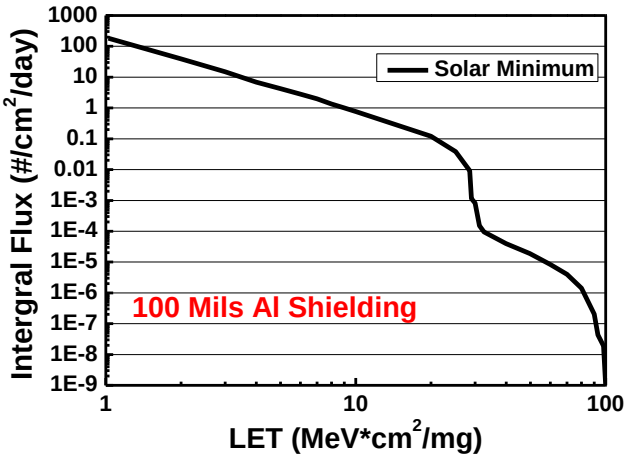


Fig. 2.15 An integral LET spectra for galactic cosmic rays during solar minimum [45]. © 2016 IOP.

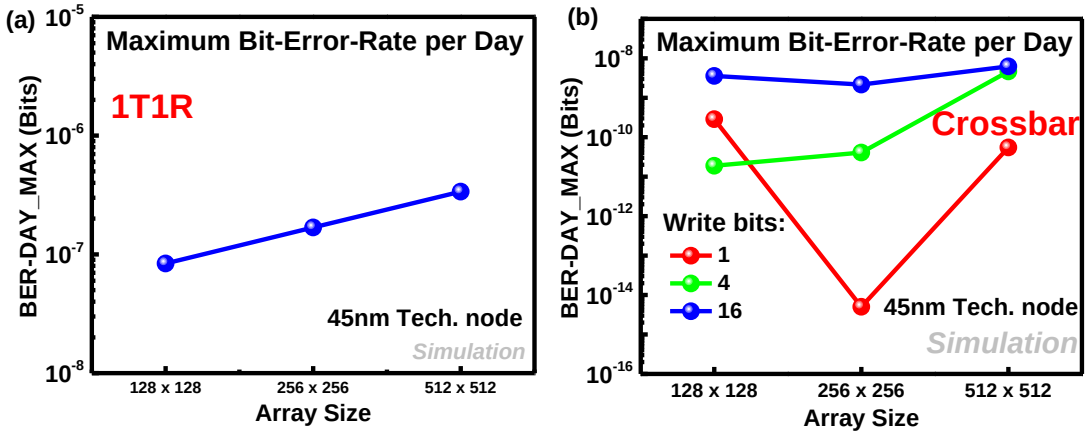


Fig. 2.16 Maximum Bit-Error-Rate per day for various array sizes and numbers of bits to be written in one array for (a) 1T1R array and (b) the crossbar array, respectively. © 2016 IOP.

2.7 Summary

When an incident ion strikes a silicon device, a current or voltage transient can occur at the circuit node, which may cause temporary state change of memory. For the DRAM

and SRAM, only the trend of soft errors with technology node scaling is summarized since they are mature memory technologies—DRAM system reliability has remained roughly constant over many generations, however, SRAM system reliability become severer with each generation.

For emerging NVM technology, we investigated the radiation induced soft error systematically. First, an RRAM SPICE model and a photocurrent model for incident particle strikes with various LETs were used to analyze and gain insights into underlying mechanisms behind the soft errors. In 1T1R, the half selected cell is only vulnerable in HRS, when a particle strikes the drain of the NMOS during the SET process. In contrast, the cell in the crossbar array has the potential to be flipped in both HRS and LRS. The radiation hardness of the crossbar array is found to be strongly dependent on the write voltage. The crossbar array exhibits extremely high robustness to the heavy ion strike if the half of the write voltage is larger than the PN junction turn on voltage (~0.7V). However, if the half of the write voltage is comparable to 0.7V, then one strike at the edge of the crossbar array may cause a sequence of bits flipping along the rows or columns, resulting in MBUs. By contrast, the susceptibility of the 1T1R array does not show such strong dependence on the write voltage. When comparing 1T1R with crossbar, the 1T1R array has a larger sensitive area since all the transistors in the array have the possibility of being struck, while the crossbar array is transistor-free inside the array with transistors only at the edge. However, the MBUs in a crossbar at low voltage operation effectively increases the sensitive area. Second, to compare the sensitivity for single-event upset, a methodology to calculate the BER/day at the system level was developed. The average SEU rate was obtained to simulate the expected the number of upset bits considering the possibilities of

ion strikes at the “OFF” transistors sharing same BLs as the selected RRAM cells in the 1T1R architecture and the “OFF” transistors of edge drivers in the crossbar architecture. SPICE simulation results indicate the number of upsets increases as the number of bits to be written in parallel increases in the crossbar architecture. There is a trade-off between number of upset bits and array size. At the system-level, BER/day is evaluated in a RRAM system with I/O of 64 bits used in the write-intensive applications. In general, the crossbar architecture presents a stronger tolerance to SEEs than the 1T1R architecture.

3 MEMORIES FOR HARDWARE SECURITY APPLICATIONS

3.1 Overview

Electronic information exchange between mobile devices and cloud on the data center is now pervasive in our everyday life, such as electronic-commerce and mobile-banking. Unfortunately, this increases the identity and secure information leaks since it provides more opportunities for the adversaries to access user's secure and private information. The security problem is likely to be exacerbated in the Internet-of-Things (IoT) era where millions of devices in our homes, offices and cars are digitally connected [48]. Every connected IoT device provides more attack possibilities and increases the potential risk. Therefore, it is necessary to equip each device with a unique and secure device signature (like the fingerprint) from the hardware itself during authentication through the cloud [49].

Physical Unclonable Function (PUF) has been widely investigated as a promising hardware security primitive, which employs the inherent randomness in a physical system (e.g. the intrinsic semiconductor manufacturing variability). By inquiring the PUF device with a challenge (input), a unique response (output) is produced correspondingly [50]. This challenge-response pairing behavior is device-specific and easy to be evaluated but prohibitively difficult to be predicted. In general, PUF is used for two applications: cryptographic key generation and device authentication, as shown in Fig. 3.1. Hence, the PUFs are classified into two types: 1) weak PUFs and 2) strong PUFs. Weak PUFs are typically used for cryptographic key generation and have limited number of challenge-response pairs (CRPs) [51]. In contrast, strong PUFs are used for device authentication and they require a very large number of possible CRPs to make it not feasible to measure or

predict all CRPs within a short time frame [51]. In addition, the used CRPs are deleted from the database and are never reused in the future.

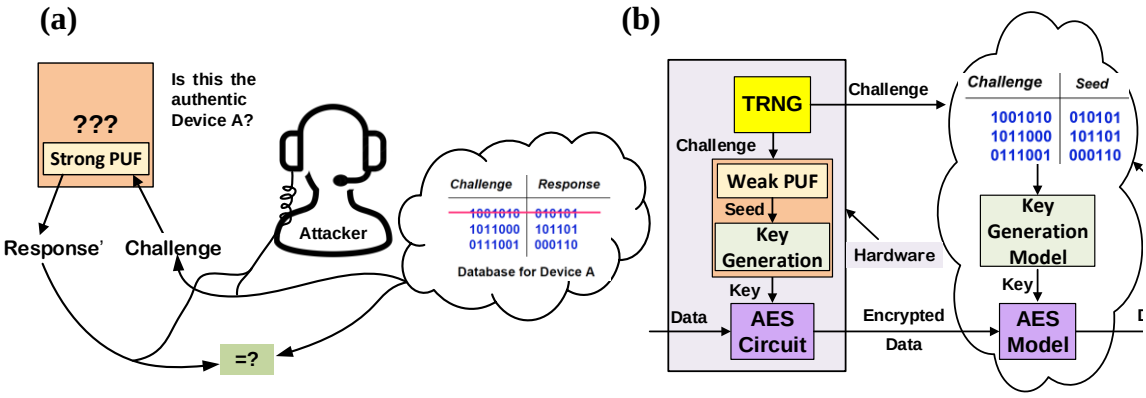


Fig. 3.1 (a) PUF-based authentication protocol; (b) PUF-based encryption protocol.

3.2 PUF Basics

There are a few important metrics to assess the fundamental PUF performance, consisting of uniformity, diffuseness, uniqueness, and reliability.

Uniformity is an indicator of the ratio of "1" and "0" in the response vector. An ideal PUF should have the equal probability of "1" and "0" in response. The uniformity is defined as:

$$Uniformity = \frac{1}{n}\sum_{j=1}^{n} r_{i,j} \times 100\% \tag{3.1}$$

where $r_{i,j}$ is the j^{th} binary bit of an n-bit response from a response vector i. The ideal value is 50%.

Diffuseness is the degree of variations among responses for different challenges applied to the same PUF. When the CRP space is too large, diffuseness can be measured by calculating the mean of hamming distance (HD) of a random sample of response vectors generated by the same PUF. The diffuseness is defined as:

$$\text{Diffuseness} = \frac{2}{m(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \frac{HD(r_i, r_j)}{n} \times 100\% \tag{3.2}$$

where m is the number of response vectors randomly selected from the CRP space. r_i and r_j are two different n-bit response vectors corresponding to 2 different challenges. The ideal value is 50%. A poor diffuseness results in collision in responses.

Uniqueness measures the difference between the response vectors which are evaluated from the same challenge on different PUF instances. The uniqueness is indicated by the inter-hamming distance (inter-HD) with an ideal value of 50%. The uniqueness of k PUFs is defined as:

$$\text{Uniqueness} = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \frac{HD(r_i, r_j)}{n} \times 100\% \tag{3.3}$$

where r_i and r_j are two different n-bit response vectors generated from 2 different PUF entities for the same challenge.

Reliability represents how well a PUF can reproduce the response bits in different operating conditions and measurement trials. The reliability is measured by intra-hamming distance (intra-HD) which should be close to 0% in the ideal case.

$$\text{intraHD} = \frac{1}{m} \sum_{t=1}^{m} \frac{HD(r_{i,ref}, r_{i,t})}{n} \times 100\% \tag{3.4}$$

where $r_{(i,ref)}$ is the reference response which is recorded at the normal operating condition, $r_{(i,t)}$ is the t^{th} measured response at a different operation condition, m is the total number of measurement trials.

3.3 SRAM PUF Implementations

There are several implementations of PUFs with CMOS technology. The most common ones are delay-based PUFs and memory-based PUFs. The arbiter PUF [52] is a delay-based PUF where the difference in the delay of the two paths is used to determine whether the

output is a 0 or a 1. While this PUF is easy to be implemented using standard CMOS logic circuits, it can be characterized by a linear delay model and the output response can be predicted by modeling attacks (e.g. the machine learning algorithms) [53] or side-channel attacks [54]. To make the delay model non-linear, XOR arbiter PUF [52] and lightweight arbiter PUF [55][20] have been introduced. Nevertheless, these variants of the delay-based PUFs (including the Ring-Oscillator PUF) are not inherently immune to modeling attacks or side-channel attacks.

The memory-based PUFs are typically SRAM or Flip-Flop based PUFs. In this design, then randomness or the entropy source comes from the randomness of startup values of the cross coupled inverters of a 6T SRAM cell due to the fabrication variation [56]. The SRAM PUFs and its variants such as Latch or Flip-flops [57] all suffer from semi-invasive or invasive tampering attacks. For example, the SRAM PUF has been characterized by photon emission analysis and cloned by Focused Ion Beam (FIB) circuit edit [58]. In addition, many of the aforementioned PUFs' response is not very stable under environmental variations such as supply voltage or temperature variations. Therefore, additional units such as error correction [51] or fuzzy extractors [59] are needed to stabilize the PUF's response. Unfortunately, the use of helper data in these units may leak sensitive information in the PUF's response as demonstrated in [60].

Therefore, it is important and necessary to develop new PUF primitives that can mitigate the threats from these attacks.

3.4 RRAM Weak PUF Design

Recently, emerging non-volatile memory based PUFs have been proposed, including PCM PUFs [61], STT-MRAM PUFs [62] RRAM, or memristor PUFs [63-65]. Next, we

will discuss an implementation of RRAM PUFs for both cryptographic key generation. The performance and reliability will be also investigated.

3.4.1 Entropy Source in RRAM

Experimentally, the RRAM device presents a relatively large variability in resistance distribution, which poses a significant design challenge for NVM applications. However, this intrinsic resistance randomness can be exploited to as the entropy source in hardware security applications. Here we leverage RRAM's resistance variability to design a weak PUF. The physical mechanism of oxide-based RRAM swiching is generally attributed to the formation and rupture of conductive filaments with oxygen vacancies between two metal electrodes. Due to the randomness of the oxygen vacancies' generation and annihilation, the dimension and composition of the conductive filament inevitably vary from cell to cell, and even from cycle to cycle for a given RRAM cell. Therefore, the resistance variability of RRAM is the combined outcome of inherent randomness in its physical mechanism and the manufacturing process variation. Since conduction in the HRS is dominated by the tunneling mechanism between the tip of the residual filament and the electrode, a small variation of tunneling gap distance results in a significant variation in HRS resistance, which provides a sufficient entropy for PUF application. Therefore the larger variability in HRS (rather than in LRS) is used as the entropy source to implement RRAM weak PUF in our design.

3.4.2 RRAM Weak PUF Architecture

Fig. 3.2 shows the circuit diagram of the proposed weak PUF design with 1T1R structure. Different from the conventional memory design, there are two set of sense amplifiers (SAs), split sense amplifier which is dedicated for the construction phase and

normal sense amplifiers which are used in the operation phase. During the construction phase, a forming process using voltage pulse is first performed on each RRAM cell to initiate the subsequent switching. And a current criteria is reinforced in this process to ensure all the cells are initiated to the almost uniform resistance. Sequentially, all the cells in the array are attempted to reset to HRS one by one with an exactly same voltage pulse,

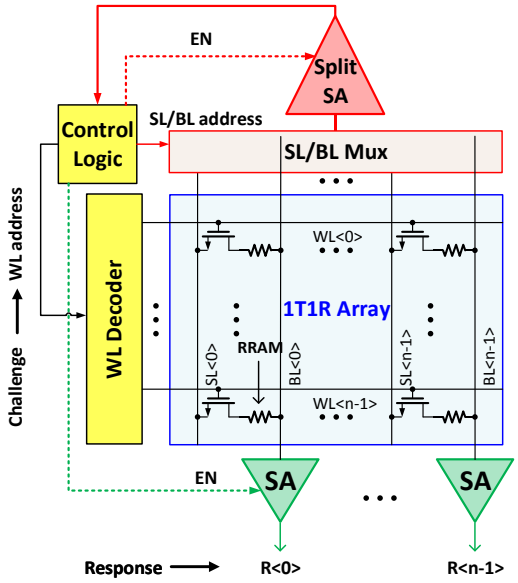


Fig. 3.2 Circuit Diagram of the RRAM weak PUF design [66]. © 2015 IEEE.

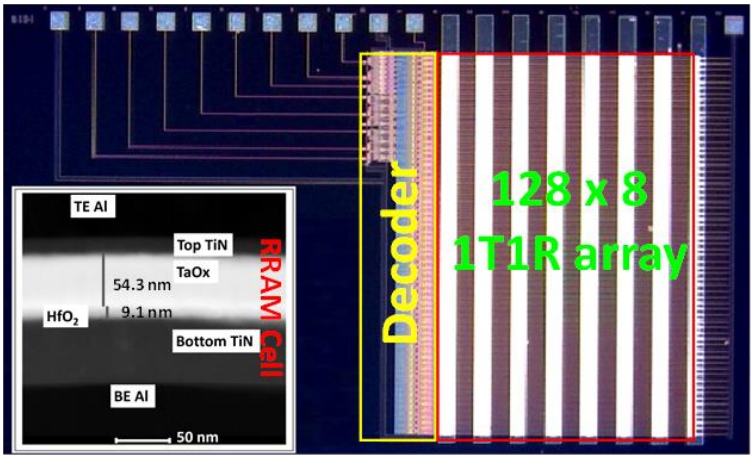


Fig. 3.3 Top view of the fabricated 128×8 1kb 1T1R RRAM array with a built-in decoder under the microscope. The inset is cross-sectional microscopic image of TiN/TaOx/HfO_2/TiN RRAM device [66]. © 2015 IEEE.

thus the variation that occurs in the first-time RESET process becomes the entropy source for the RRAM PUF. Then a read operation is conducted to measure the read currents for all RRAM cells in the array. We demonstrated this construction phase on an 1kb 1T1R array (see Fig. 3.3). Fig. 3.4(a) shows the measured read current distribution after forming process and first time RESET process, indicating significant randomness (i.e. wide resistance range). Next, a split reference current (Ref_Split) is selected within the distribution. The cells with currents larger than the reference are SET into LRS, as Fig. 3.4(b). This split process aims to digitize the randomness and improve the PUF's reliability against resistance noises [63]. Fig. 3.4(c) and (d) show the analog data pattern right after the first RESET operation and digital data pattern after split, respectively. The large window between the two split states are advisable such that the design of sense amplifier

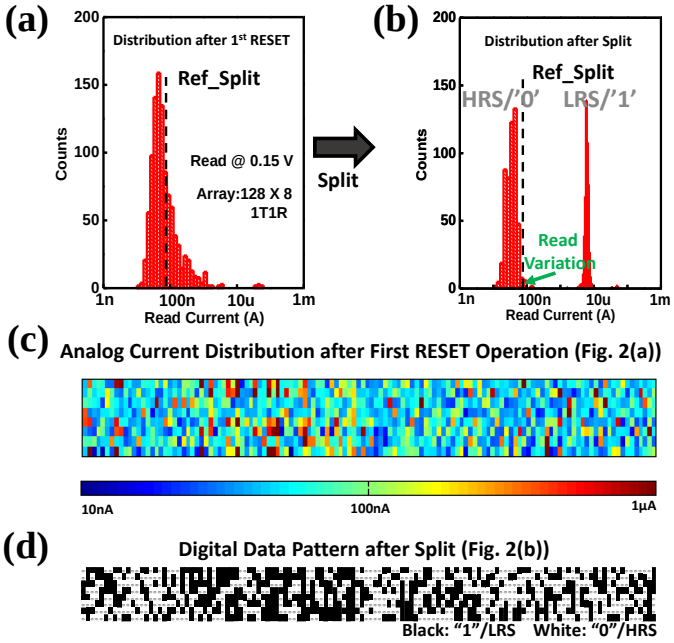


Fig. 3.4 (a) Initial distribution of read current of HRS in an RRAM array and (b) distribution after a part of cells are programmed into LRS according to the reference. (c) Analog data pattern after the first RESET operation. (d) Digital data pattern after split.[66] © 2015 IEEE.

could be relaxed to differentiate the read current. Ultimately, the challenge-response pairs (CRPs) are measured and enrolled in the database for future use and the RRAM PUF construction is completed. In this design, the challenge is the address applied to the WL decoder and response is the digital outputs of the read sense amplifiers. During the operation phase when deployed in the field, only read operations will be performed.

3.4.3 Performance Evaluation of RRAM Weak PUF

To evaluate the uniqueness experimentally, 40 PUF instances are prepared through five 1 kb 1T1R arrays with size of 128 × 8. By applying the same challenge inputs (activating all rows one by one), 128-bit responses are measured. Then the uniqueness is evaluated by inter-HD of the responses pair-wisely compared across the 40 PUF instances. In addition, we also investigated the impact of non-ideal factors of peripheral circuits on the performance of the proposed RRAM weak PUF.

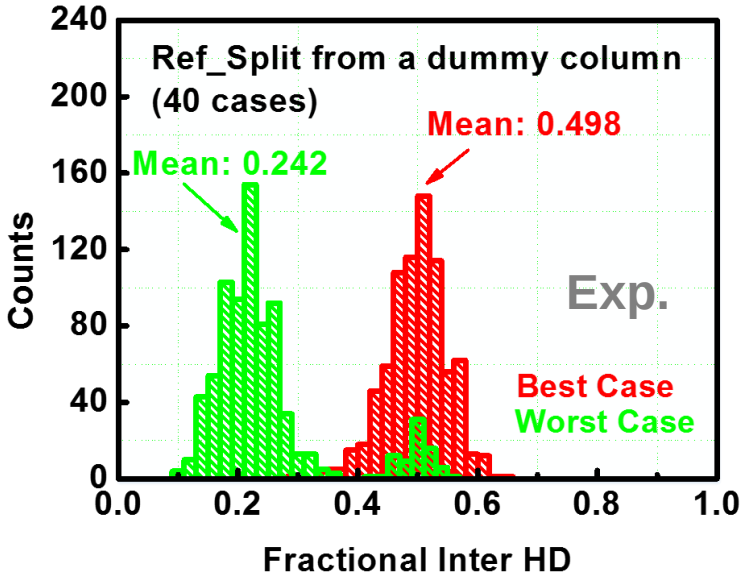


Fig. 3.5 Distribution of fractional inter-Hamming distance (HD) of 128-bit responses with split reference current obtained from a dummy column. The best and worst cases of 40 split references from 40 dummy columns are shown [67]. © 2015 IEEE.

3.4.4 Impact of Split Reference on RRAM Weak PUF's Uniqueness

One factor that affects the RRAM PUF's uniqueness is the split reference used in the split process. Ideally, the reference should a read current that can make an equal 50% probability of generating "0" or "1", although this restriction will reduce the possible configurations of the response bit stream. In experiments, we used a dummy column to generate the split reference. Then the split reference is set as the median current of the 128 dummy cells in one column. We prepared 40 dummy columns. Hence we have 40 possible split references. Due to variations between column and column, the generated 40 possible references distribute in a wide range from (74 nA to 238 nA). We choose one with smallest deviation from the ideal reference and the other one with largest deviation from ideal reference to conduct the split process. Fig. 3.5 shows the fractional inter-HD distributions when using these two split references. When the split reference is closer to the ideal, a good average inter-HD is can be achieved, e.g. ~49.8% with a tight distribution. However, when split reference is further away from the ideal value, the average inter-HD is bad, e.g. ~24.2%. This suggests the importance of generating a good split reference.

Impact of Split S/A on RRAM Weak PUF's Uniqueness

Split sense amplifier is used as a comparator in the split process. Under ideal conditions, an idea S/A should be able to amplify a very small input differential signals correctly. In reality, however, process variations in the transistors of an S/A introduce an input offset, which results in a skewed preference to generate "1" or "0". A voltage mode sense amplifier (Fig. 3.6(a)) is employed in the split process as a case of study to investigate the impact of offset on the uniqueness. The two differential inputs are V_DL and V_REF. At first, pre-charge transistor (Q10) is turned on and the BL is charged to V_read. Then Q10 is turned off and BL is being discharged through RRAM cell for a short period of time to

develop a voltage sense margin. Depending on the RRAM cell's current, V_DL's decay can be fast or slow. Finally, SAEN is turned on and the difference between V_DL and V_REF is amplified by the latch based load and the digital output is generated in SA1 and SA2. In a naive implementation, all the transistors in S/A can be minimum sized. To assess the input offset of this S/A, 1000 Monte Carlo simulation runs were performed in Cadence Spectre in TSMC 65 nm node using library "TSMC65-GP-1p9m_6X1Z1U_ALRDL_2.0". The simulation shows that the S/A with minimum sized transistors has an offset voltage σ of 25.9 mV. If S/A with 3σ input offset voltage is used in the split process, it might have a much skewed preference to generate more "0"s or "1"s. As a result, the distribution of the fractional inter-HD decreases to 30.6% as shown in Fig. 3.6(b). Therefore, minimizing S/A offset voltage is necessary in the split process.

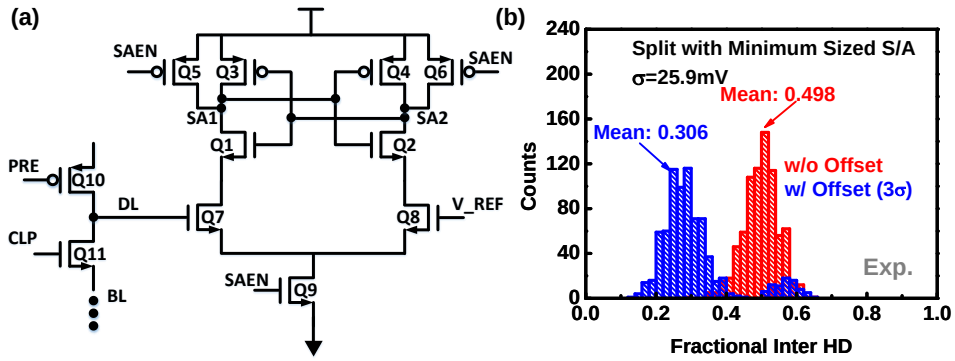


Fig. 3.6 (a) Schematic of a voltage mode sense amplifier (S/A) used in the split process as a comparator. (b) Distribution of fractional inter-HD of 128-bit responses without or with S/A offset (σ= 25.9 mV). Minimum sized transistors are used. [67] © 2015 IEEE.

Impact of RRAM Retention Failure on RRAM Weak PUF's Reliability

Reliability of RRAM PUFs requires an excellent data retention even at elevated temperature conditions. Once the retention failure occurs in the RRAM cell, it will introduce an error in the response bits, thus increasing the intra-HD. To evaluate the RRAM's data retention, a high temperature (150 °C) is used to accelerate the failure in our

measurement. Fig. 3.7 shows the 1 kb RRAM array's read current degradation without voltage bias at 150 °C. The experimental result shows that the tail bits in HRS and tail bits in LRS crossed-over in less than 2 hours (or equivalently less than a 25 days at 85 °C), which means errors occur in the PUF's response if a PUF response bit is represented by a single RRAM cell. This experimental result illustrates the necessity of improving RRAM PUF's reliability.

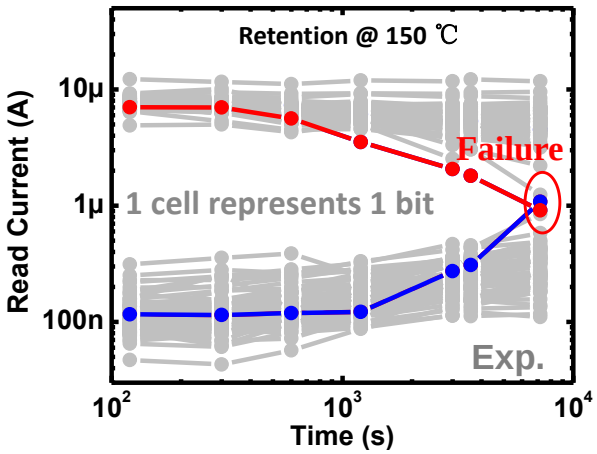


Fig. 3.7 Measured retention degradation of 1 kb RRAM array baking at 150 °C. Error occurs within 2 hours if a single cell represents a PUF response bit [66]. © 2015 IEEE.

3.4.5 Improving RRAM Weak PUF's Performance and Reliability

In this section, we will discuss some design strategies to improve RRAM PUF's uniqueness and reliability issues as pointed out in the previous section. In addition, we will employ a layout obfuscation technique to enhance its tamper resistance.

Accurate Split Reference Generation by Dummy Array

In order to generate a more accurate split reference, more dummy cells are needed to average out the cell to cell variations. For example, the dummy cells can be obtained from an array (with eight columns each array) instead of one column. We prepared five split references from five dummy arrays. Table 3.1 lists the mean values and standard deviations of the fractional inter-HD when using these five split references. The distribution of inter-

HD is centered at 47.78% with a small σ of 5.56% in the worst case. In the practical design, there are two ways to obtain a good split reference. First, it can be obtained by off-chip pre-calibration. A dummy array (or a few dummy arrays) can be manufactured in same batch. The same programming conditions are performed on the dummy array. It is easy to find the median of read current by a simple sorting algorithm off-chip. Second, a dummy array are designed adjacent to the real array on-chip, and a custom circuit is needed to do the sorting. Since the split process is only done once in the PUF construction phase, finding a good split reference from off-chip pre-calibration is more efficient in terms of area and energy.

Table 3.1 Uniqueness Evaluation with Ref_Split Generated from A Dummy Array

Uniqueness	Ref_Split generated from Array No.				
	1st	2nd	3rd	4th	5th
Mean (%)	49.48	48.97	49.79	47.77	49.80
Std (%)	4.90	5.06	4.87	5.56	4.86

Accurate Split Reference Generation by Dummy Array

As technology node scales down, the input offset of S/A increases due to the overall increase in local (i.e. within-die) process variation, e.g. random dopant fluctuation (RDF). It is known that the standard deviation of the transistor's threshold voltage (Vth) distribution is proportional to $1/(WL)^{1/2}$ [18]. Sizing the transistors is a flexible option and is employed in this work. The key contributor to the offset is from the input differential pair (Q7 and Q8 in Fig. 3.6(a)), thus their sizes should be increased most. Besides Q7 and Q8, Q1, Q2, Q3, Q4 in the latch based load and Q9 in the bottom current source are also critical transistors that should increase sizes. Table 3.2 and Table 3.3 list two sets of transistor's sizes that can reduce the offset σ to 7.868 mV and 6.511 mV respectively. In

addition, in order to reduce the input offset from layout point of view, symmetrical and common centroid layout design is employed. Fig. 3.8(a) is the distribution of input voltage offset obtained by running 1000 Monte Carlo simulations with transistor sizes listed in Table 3.3. When the standard deviation of input offset is reduced to 6.511 mV, the average inter-HD can be improved to 42% as shown in Fig. 3.8(b). Such a relaxed design of split S/A does not increase the total area of RRAM PUF macro too much because there is only one split S/A per PUF used in the construction phase, while other read S/A to generate response bits used in the operation phase can still be minimum sized.

Table 3.2 Split S/A Transistor Sizing to Reduce Offset to 7.858 mV

Transistor	Q1/Q2	Q3/Q4	Q5/Q6	Q7/Q8	Q9	Q10/Q11
Gate Length (nm)	60	60	60	180	60	60
Width (nm)	240	240	120	900	120	120

Table 3.3 Split S/A Transistor Sizing to Reduce Offset to 6.511 mV

Transistor	Q1/Q2	Q3/Q4	Q5/Q6	Q7/Q8	Q9	Q10/Q11
Gate Length (nm)	60	60	60	180	60	60
Width (nm)	240	240	120	1800	240	120

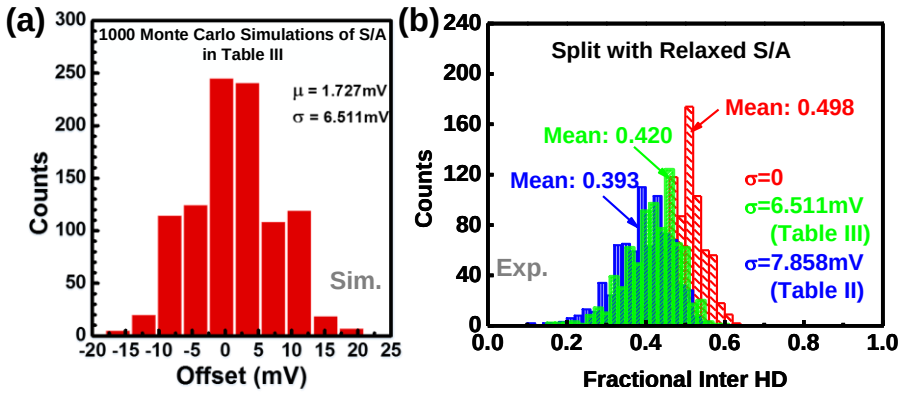


Fig. 3.8 (a) Distribution of split S/A voltage offset from 1000 Monte Carlo simulations with sizing the transistors in Table 3.3. (b) Distribution of fractional inter-HD with considering S/A different 3σ voltage offsets [67]. © 2015 IEEE.

Multi-Cell-Per-Bit to Improve Reliability

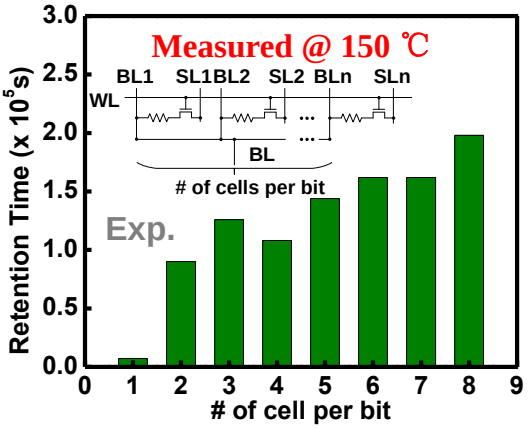


Fig. 3.9 Measured retention time when a PUF response bit is represented by different number of cells [67]. © 2015 IEEE.

To improve the retention properties, we propose to use multiple RRAM cells to produce a response bit. The concept behind is that if multiple RRAM cells in parallel are wired as one group, the read-out current will be added up. Due to inherent cell to cell variations, some cells may fail later than others, and the redundancy can minimize the probability of early lifetime failure for the whole group. In the practical design, multiple BLs can be wired together before sending the BL current to the read S/A. In the PUF construction phase, we can program each cell (not including the redundant cells) individually using separate source lines (SLs). Then both the cell and redundant cells should be programmed to the same state as a group according to the comparison result with the split reference. Therefore, we do not average out the variation by grouping the cells together. Fig. 3.9 shows the retention time for different number of RRAM cells representing one PUF response bit that is measured at 150 °C. In general, longer retention time can be achieved with more redundant cells as expected. When each response bit is represented by 8 parallel RRAM cells, it can be sustained for more than 50 hours at 150 °C for a given PUF instance with high reliability (Fig. 3.10(a)). The on/off ratio of readout the currents for the tail bits is larger than 2.5×,

which can be reliably sensed by the read S/A. Fig. 3.10(b) shows the equivalent retention time extrapolated to 85 °C and 27 °C using the 1/kT extrapolation with activation energy (Ea=1.15 eV, determined in another experiment). Eight RRAM cells in parallel can possibly generate a highly reliable response for 1.75 years at 85 °C, and 10 years at 69 °C. In addition, we have examined that the multiple-cell-per-bit approach has negligible impact on the PUF's uniqueness.

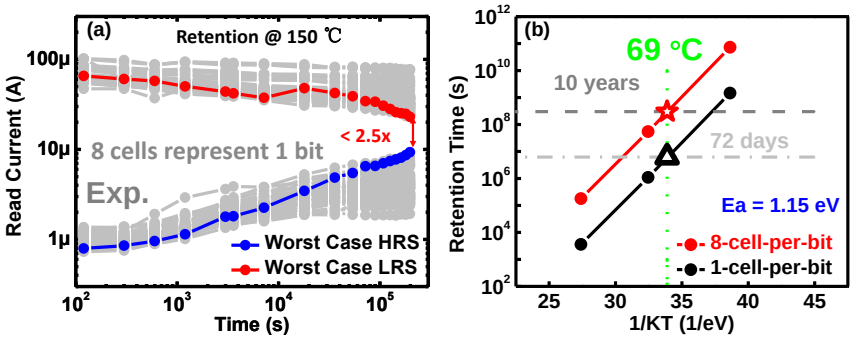


Fig. 3.10 (a) Measured retention degradation of 1 kb RRAM array baking at 150 °C if 8 cells represent one response bit. (b) Extrapolated retention time using Ea=1.15 eV. 8 cells per bit can possibly ensure 10-year lifetime at 69 °C [67]. © 2015 IEEE.

Layout Obfuscation for Tamper Resistance

A basic requirement for a weak PUF is that the adversary should not have access to the response bits, as the number of CRPs in a weak PUF is limited. However, the adversary can perform semi-invasive or invasive tampering attacks to obtain the response bits. For example, the SRAM's data pattern can be seen under near-infrared imaging because the hot carriers in the transistors emit photons. It is expected that RRAM's conduction in oxide does not emit photons under laser or X-ray scanning (at. least not reported yet). However, the digital responses of RRAM PUF are still read out through the S/A. Hence, the read S/A might be a potential weak spot that an adversary can micro-probe to access the output and read out the secret information. Fig. 3.11 shows RRAM PUF architecture with 1T1R

memory array. Eight cells are grouped together to generate one response bit. The conventional design places the read S/A at the edge of the array, thus they are easy to be identified under the microscope thus vulnerable to the probing attack.

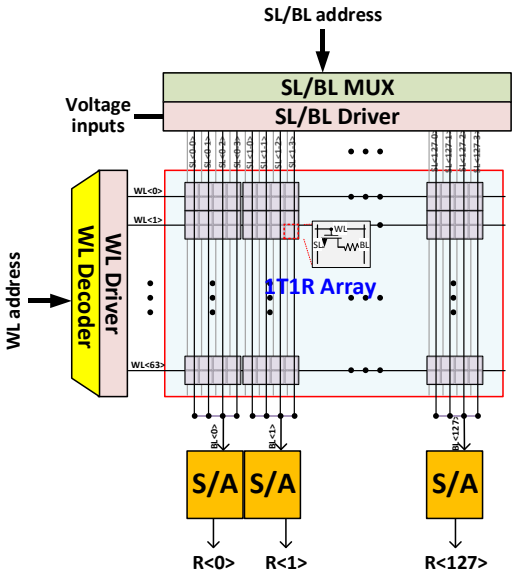


Fig. 3.11 RRAM PUF architecture with 1T1R memory array. Eight cells are grouped together to generate one response bit. The conventional design places read S/A at the edge of the array, thus vulnerable to the probing attack [67]. © 2015 IEEE.

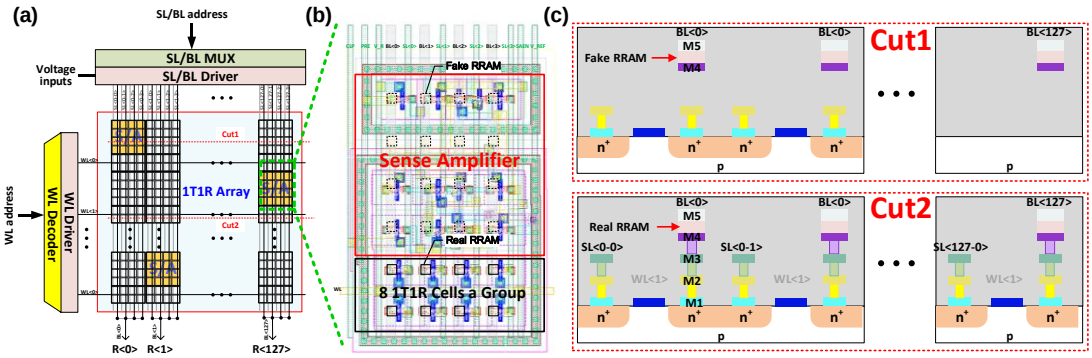


Fig. 3.12 (a) Tamper-resistant RRAM PUF architecture with read S/A randomly embedded into the array and hiding underneath a sea of real and fake RRAM cells. (b) Layout obfuscation of a block including read S/A, 8 real RRAM cells and 12 fake RRAM cells. (c) Cross-section view for cutting through S/A and cutting through the real RRAM cells respectively. [67] © 2015 IEEE.

In order to obfuscate the adversary, we propose to hide the S/A within the 1T1R array and randomize the locations of S/A, as shown in Fig. 3.12(a). Between M4 and M5, we uniformly place the RRAM contact vias across the array. Fig. 3.12(b) shows the Cadence layout of a block including S/A, 8 real RRAM cells and 16 fake RRAM cells on top of S/A. The RRAM contact vias on top of the 1T1R are the real RRAM cells, while the RRAM contact vias on top of the S/A are fake RRAM cells. Fig. 3.12(c) presents the cross-section of the die for a cut through S/A with fake RRAM cells and a cut through region with the real RRAM cells respectively. When an adversary attempts to probe the S/A's output underneath the uniformly distributed RRAM cells, it is difficult for him to differentiate between the real RRAM cells and the fake RRAM cells from the top-view. The real RRAM cells which implement a secure key storage might be permanently destroyed when the adversary tries to invasively probe, thus the proposed layout obfuscation enables a "self-destructive" feature for the RRAM PUF.

Area Cost and Performance Overhead Analysis

All the proposed design strategies such as relaxing split S/A's transistor sizes, multiple-cell-per-bit, and layout obfuscation with S/A hiding are associated with hardware overhead including more area, larger latency and energy consumption. In order to evaluate the overhead, we use Cadence and HSPICE to evaluate the area cost and performance of a 64 × 128 RRAM PUF macro. Three designs are evaluated. The first one is 1-cell-per-bit without S/A hiding as the baseline, which has the poorest reliability and the lowest security. The second one is 8-cell-per-bit without S/A hiding, which is highly reliable but not tamper resistant. The last one is 8-cell-per-bit with S/A hiding, which is of highest reliability and tamper resistance. All the designs are benchmarked at TSMC 65 nm node using "TSMC65-GP-1p9m_6X1Z1U_ALRDL_2.0" library. Table 3.4 shows the benchmark results.

Compared to the baseline, the highly reliable design introduce 1.52× latency, 1.55× energy, and 4.70× area, and the highly reliable plus tamper-resistant design introduce 3.88× latency, 1.84× energy, and 24.53× area. Depending on the application scenarios, the designers can choose the appropriate design strategies. For example, if the security is a not topmost requirement but still 10-year lifetime is necessary, the highly reliable design but without S/A hiding may be sufficient.

Table 3.4 Area and Performance of RRAM Weak PUF with Array Size of 64 × 128

Architecture	S/A hiding (w/ or w/o)	Latency (ns)	Energy (pJ)	Area (mm^2)
1-cell-per-bit	w/o	4.24	9.59	0.0083
8-cell-per-bit	w/o	6.46	14.87	0.0390
	w/	16.45	17.69	0.2036

3.5 RRAM Strong PUF Design

3.5.1 Entropy Source in RRAM

The typical PUF-based device authentication protocol is shown in Fig. 3.1(a). During the enrollment phase, a large number of responses are collected by applying a random set of challenges to an authentic PUF entity when a trusted party is in physical possession of the device. When the enrollment phase is finished, these collected CRPs are safely stored in a database for future authentication use. During the deployment phase in the field, to check the authenticity of a device to be authenticated, a recorded but unused challenge is selected from the database and is sent to the device by the trusted party. Then a PUF response is received from the device side and compared with the corresponding response previously recorded in the database. If they are close enough, the device is authenticated. Otherwise, it will be denied. In the authentication application, the used CRPs will be

deleted from the database to protect against the man-in-the-middle attacks in the communication channel as shown in Fig. 3.1 (a). Therefore, a large number of CRPs are needed for device authentication.

For emerging NVM based PUFs, one or a few memory cells are usually used to generate a single response bit. Therefore, it cannot provide a sufficient number of CRPs and be used as strong PUF for device authentication, due to the limited capacity of memory cells in the emerging NVM arrays. To achieve a large CRP space, we propose a strong PUF design by exploiting the sneak paths in the X-point array. An X-point array is essentially a resistor network if there is no isolation transistors or selectors in the RRAM array. When bias voltages are applied on selected rows/columns and unselected rows/columns are floated, sneak currents flow through the entire array. For a given set of bias voltages applied to the array, if the resistance pattern of the resistor network is random, the currents measured from the end of columns are also random. Such randomness could be amplified when there are more rows and/or columns left floating. For memory applications, such sneak paths are detrimental to the read-out sense margin [5], however, we take an advantage of sneak paths for PUF design in this design.

3.5.2 X-point PUF Architecture

Fig. 3.13 shows the proposed architecture of X-point PUF. For an N (rows) × N (columns) array, the resistance pattern is constructed before the enrollment phase and will be discussed in detail in the next section. During the enrollment phase and deployment phase, the binary challenge vector (i.e. an N-bit vector) decides the bias voltages of the rows or wordlines (WLs) in the N × N array. For a given challenge vector, if i^{th} element in the N-bit challenge is "1", a read voltage is applied to the i^{th} row; if j^{th} element is "0", then

the corresponding row is floating. For example, if the N-bit challenge vector is '10010…001'', a positive read voltage is applied to 1st row since the first element of the challenge vector is "1", and the 2nd row is floating because the second element of the challenge vector is "0". The portion of "1" in a challenge vector is defined as WL activity. Then the N-bit response vector can be generated from the columns or bitlines (BLs) by a sense amplifier (S/A). In practice, multiple columns may share one S/A, thus a multiplexer (MUX) may be used to do the time-multiplexing. Specifically, selected columns are connected to S/A, and the other unselected columns are floating. Then the current from each column (including the sneak path current) is measured and converted into binary fashion "0" or "1" by S/A according to a reference. The generated digital response element

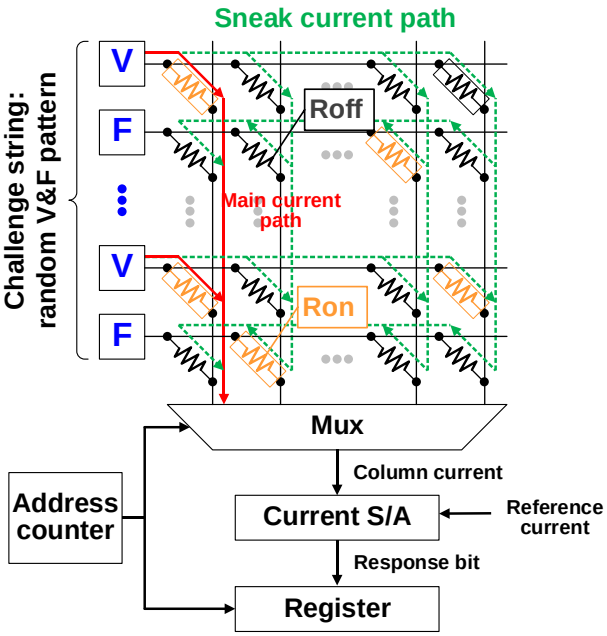


Fig. 3.13 Proposed resistive X-point strong PUF circuit leveraging the sneak paths (green) to create CRPs [68]. © 2018 IEEE.

is stored in the register before the element is to be read out. After a few read cycles, all the N response elements are collected from S/A and stored in the register. This means a challenge-response pair generation is finished and ready for use, e.g. authentication. In this

work, we assume the current mode S/A is used and the current reference (I_ref) is a current which is set to be the median value of the column output current distribution. In practice, the current reference could be obtained by off-chip pre-calibration. A bare array (or a few bare arrays) can be manufactured in the same batch. The same programming conditions are performed on these arrays. It is straightforward to find the median of read current by a simple sorting algorithm off-chip. Since the split process is only done once in the PUF construction phase, finding a good split reference from off-chip pre-calibration is more efficient in terms of area and energy.

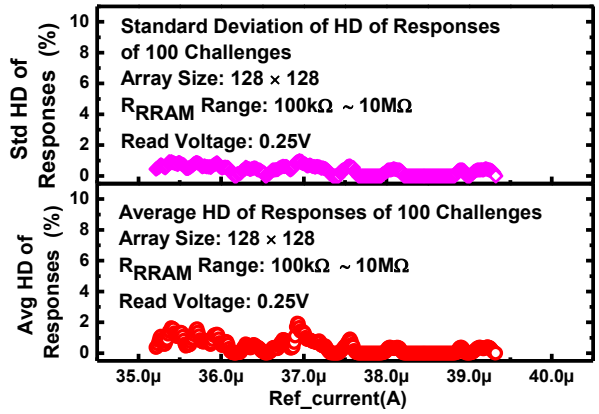


Fig. 3.14 Collison problem in "analog" X-point PUF: (a) standard deviation and (b) average hamming distance (HD) of responses over 100 challenges with different reference currents of SA for the same PUF. The resistance of the cells is randomly selected from a log-normal distribution within a range from 100kΩ to 10MΩ [68]. © 2018 IEEE.

3.5.3 Security Issue in "Analog" X-point PUF

In prior work [69], we used "analog" resistance distribution of RRAM cells in the array to implement the X-point PUF. The experimental results showed good uniqueness and reliability for a 12 × 12 array, however, this design has a potential security problem when the array size is larger. Fig. 3.14 shows the simulated diffuseness of responses over 100 different challenges collected from the same X-point PUF instance when the array size is

128 × 128. In the simulation, we assumed the cell resistances are log-normally distributed within a range from 100kΩ to 10MΩ. The diffuseness is always less than 1% irrespective of the magnitude of I_ref used for S/A. This means the responses for difference challenges are similar, namely collision problem. This is because the BL current is dominated by the sneak paths in a large array. As a result, the adversary can easily guess the possible response for any challenge based on the known CRPs of the same PUF. This is undesired for a secure PUF design.

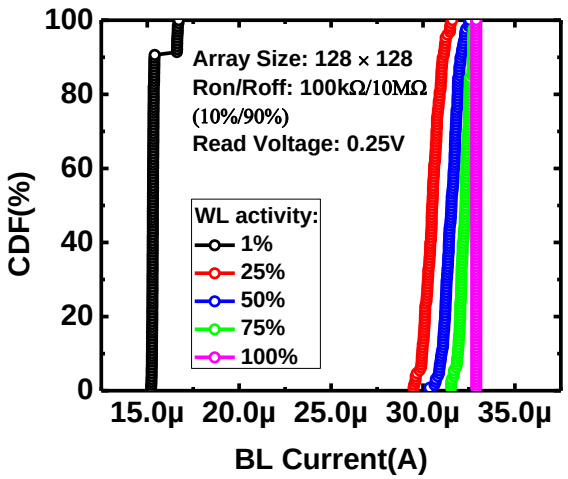


Fig. 3.15 BL current distribution over 100 different challenges with different WL activities in "digital" X-point PUF. © 2018 IEEE.

To solve this collision problem, the use of "digital" resistance distribution is proposed. Generally, the forming operation is performed for all the RRAM devices to initiate the switching behaviors. After the forming, an identical programming condition is applied to every cell in the array to RESET it to high resistance state (HRS). Due the inherent variability of switching dynamics in RRAM device, the RRAM devices in the array are programmed to different HRS values. To prepare the "digital" resistance pattern, a split resistance reference is chosen within the HRS distribution. Then the cells with resistance

lower than the split reference are further SET into LRS with resistance R_on and the cells with resistance above the reference are further RESET to off-state with resistance R_off.

3.5.4 CRP Space of X-point PUF

Theoretically, in this design, the maximum number of CRP of an N × N array is 2N since each WL has two options: either applying read voltage or floating. For different WL activities, however, the readout column currents present different current distribution ranges as shown in Fig. 3.15. Lower WL activity shows lower readout column currents and higher WL activity shows higher readout column currents. Consequently, the response bit is more likely to be "0" (or "1") with lower (or higher) WL activity for a given I_ref, which is not secure for a PUF design. Hence, the activity of WL should be specified to generate a comparable range of column currents for different challenges. Thus the preference of the response bit to be "1" or "0" is not decided by the WL activity. To achieve the largest CRP space, the WL 50% activity is employed in this work. Then the number of CRPs is $C_N^{N/2}$ for an N × N array. When the array size is 128 ×128, the CRP space is around 2×10^{37}, which is quite large against the brute-force guessing within a short time frame.

3.5.5 Ron Activity of X-point PUF

In this section, extensive simulations were performed to investigate the dependence of the proposed "digital" X-point PUF's performance on the R_on activity, which is defined as the percentage of the on-state cells in the array. The simulation study was carried out at circuit level using HSPICE. To simplify the simulation at this stage, interconnect resistance and resistance variation were not considered in the simulation, which will be considered later in Section V as non-ideal factors. As a case of study, we assumed the RRAM array was 128 × 128 for the X-point PUF design and R_on = 100 kΩ and R_off = 10 MΩ for

typical RRAM devices, where the R_on and R_off were randomly distributed in the array for a given R_on activity. To save the simulation time, we prepare 100 challenge-response pairs to evaluate each performance metric in most cases since the difference between the evaluation results with 100 CRPs and a larger set of CRPs (e.g. 1000) are negligible (i.e. less than 0.1%) as found in our simulations.

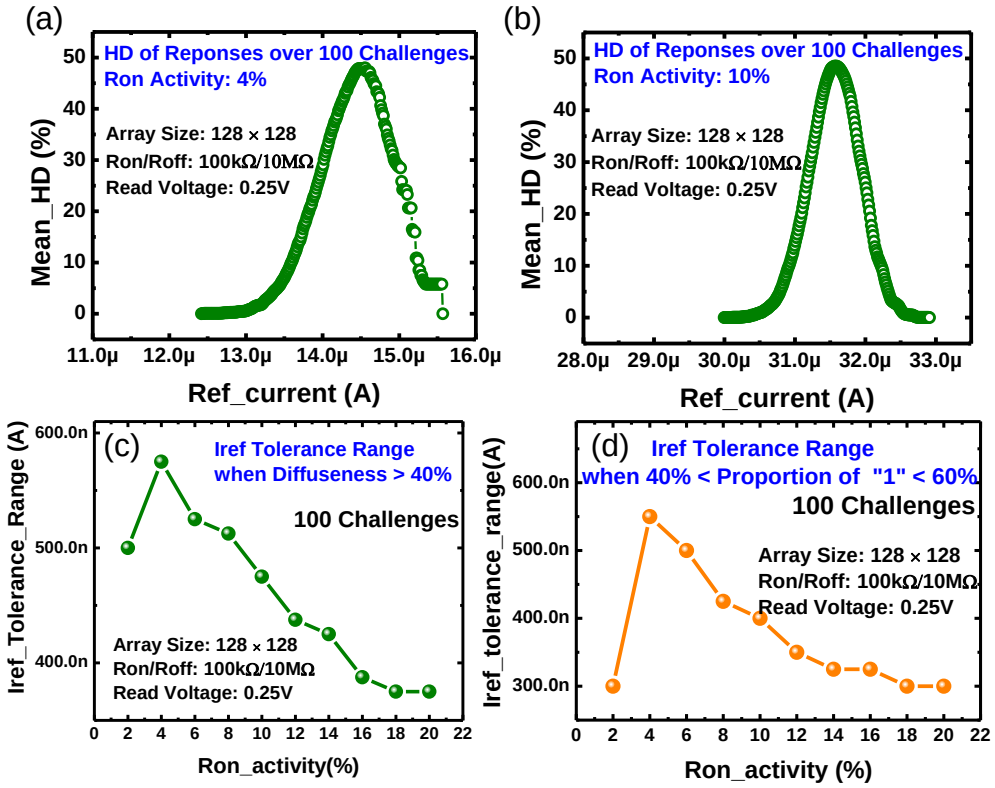


Fig. 3.16 Mean of HD over 100 random response vectors generated by the same X-point PUF with different I_ref when R_on activity is (a) 4% and (b) 10%. I_ref tolerance range for different R_on activity when (c) diffuseness is larger than 40% and (c) the portion of "1" is in range of (40%, 60%) [68]. © 2018 IEEE.

Diffuseness and Uniformity

To evaluate the diffuseness and uniformity of the "digital" resistance distributed X-point PUF, 100 response vectors are generated by applying 100 random challenge vectors to one PUF. Since the WL activity is 50%, there are 64 "1"s out of 128 bits in each

challenge vector. Subsequently, a 128-bit response is generated by sensing the column current via S/A with a given reference current (I_ref). Fig. 3.16(a) and (b) show the distribution of average HD over 100 random responses generated by the same PUF for different I_ref and R_on activities (4% and 10% respectively). The simulation results indicate that the diffuseness of X-point PUF is significantly impacted by the I_ref. By choosing a good I_ref, the diffuseness of the "digital" X-point PUF can be close to its ideal value (i.e. 50%), thus eliminating the severe collision problem as found in the "analog" design of X-point PUF. However, when the I_ref varies at different operating conditions the diffuseness of the "digital" X-point PUF might be degraded. Therefore, high resistance against I_ref variation is preferred for a secure and robust PUF design. If we set a target to the diffuseness, e.g. mean of HD should be larger than 40%, then an I_ref tolerance range can be extracted for different R_on activities, as shown in Fig. 3.16(c). For example, with R_on activity=10%, the I_ref tolerance range is from 31.4 µA to 31.8 µA, which is 0.4 µA, and with R_on activity = 4%, the I_ref tolerance range is from 14.2 µA to 14.8 µA, which is 0.6 µA. Similarly, the uniformity of the responses presents dependence on the I_ref and R_on activity. Fig. 3.16(d) shows the extracted I_ref tolerance range for different R_on activities when the probability of "1" in the response bits is in range of 0.4 and 0.6. From Fig. 3.16(c) and (d) we can see that when the R_on activity is 4%, an excellent diffuseness can be achieved for a good given I_ref: diffuseness is around 49.5% and the standard deviation is 1.8%. In addition, the uniformity is centered at 50.4% with a standard deviation of 3.5%.

Uniqueness

To study the dependence of uniqueness of the "digital" X-point PUF on R_on activity, we prepared 100 different X-point PUF instances for a given R_on activity. The uniqueness

is a function of I_ref and R_on activity. The extracted I_ref tolerance range for different R_on activities is shown in Fig. 3.17 when the target average inter-HD is set to >40%.

The simulation results again suggest that 4% is an optimal R_on activity to achieve the strongest robustness against I_ref variation's effect on the uniqueness. Therefore, in the rest of paper, 4% R_on activity (and corresponding I_ref=14.5 μA) will be used in the simulation by default if not specified.

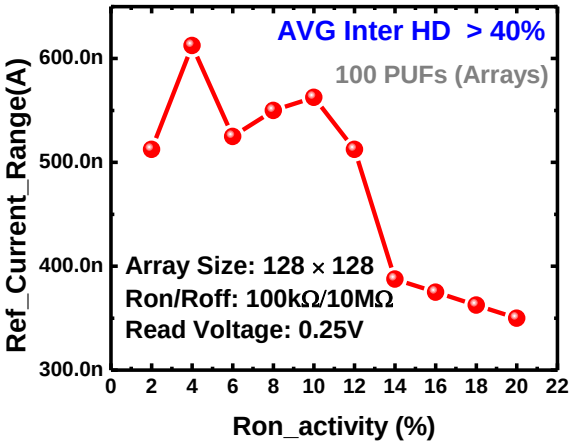


Fig. 3.17 I_ref tolerance range for different R_on activity average inter-HD is larger than 40% [68]. © 2018 IEEE.

3.5.6 The Effect of Non-ideal Factors on X-point PUF's Performance

Interconnect Resistance

In reality, the wire resistance of interconnect is non-zero, which will cause IR drop along the wires. In order to investigate the effect of interconnect resistance on the X-point PUF, we assume both WL and BL wire pitches are 2F. As a case of study, copper is assumed as the interconnect wire material and the circuit is fabricated at F=65 nm and 22 nm technology node, respectively. Then the interconnect resistance can be estimated using the corresponding resistivity of copper (i.e. 4.51 μΩ*cm at 65 nm and 5.85 μΩ*cm at 22 nm) and etching aspect ratio (i.e. 1.9 at 65 nm and 2.0 at 22 nm) predicted by the ITRS and

the geometry parameters [43]. The calculated wire resistance per segment of the array is 0.73 Ω for 65 nm and 2.42 Ω for 22 nm. Table 3.5 shows the performance evaluation results (i.e. diffuseness, uniformity and uniqueness) of the X-point PUF without interconnect resistance and with interconnect resistance at 65 nm and 22m, respectively. The same I_ref (i.e. 14.5 μA) is used in the evaluation for all the cases. This result suggests that the change of performance is negligible for 65 nm technology node. However, uniqueness and uniformity of X-point PUF present noticeable degradation due to the higher interconnect resistance when the technology node is scaled to 22 nm. The higher interconnect BL currents without interconnect resistance, which cannot be neglected anymore. It is more reasonable to scale down the I_ref with the technology node. The performance evaluation

Table 3.5 X-point PUF's Performance with Wire Resistance at 65 nm and 22 nm

Metrics	Tech. node (nm)	Mean (%)	Standard deviation (%)
Diffuseness (%)	w/o R_wire	48.36	5.27
	65	48.10	3.46
	22	47.67	5.17
Uniformity (%)	w/o R_wire	50.05	3.47
	65	48.10	3.46
	22	43.14	3.61
Uniqueness (%)	w/o R_wire	50.44	4.20
	65	50.44	4.20
	22	49.77	4.06

Table 3.6 X-point PUF's Performance with Scaled I_ref (i.e. 14.38 μA) at 22 nm

Metrics	Mean (%)	Standard deviation (%)
Diffuseness (%)	50.05	4.48
Uniformity (%)	50.07	3.36
Uniqueness (%)	50.36	4.16

results are shown in Table 3.6 for 22 nm technology node with a scaled I_ref (i.e. 14.38 μA). The performance is as good as the results without interconnect resistance. In summary, to achieve good performance, the I_ref should be scaled with different technology nodes since the interconnect resistance will cause BL current reduction.

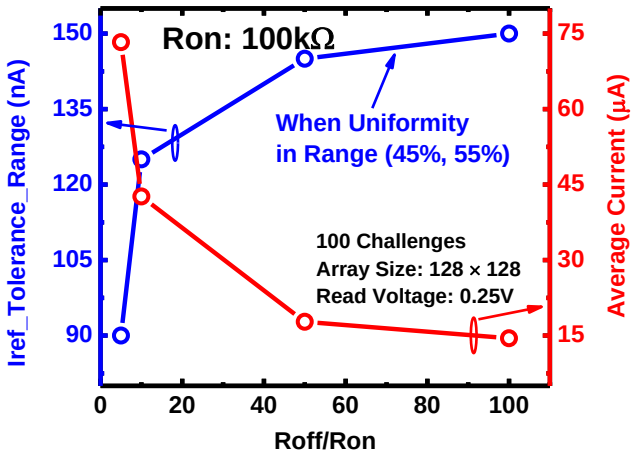


Fig. 3.18 I_ref tolerance range (in blue) when the uniformity is in range of (45%, 65%) and the average current (in red) for different RRAM on/off ratios [68]. © 2018 IEEE.

RRAM On/Off Ratio

In fact, the RRAM device may have different resistance ranges and different on/off ratios (i.e. ratio of R_off to R_on) with different material stacks. Therefore, it is necessary to investigate their impact on the performance and robustness of the X-point PUF. To save energy consumption in practical operation, the high resistance of R_on cell is preferred and 100 kΩ is assumed in our simulation. Depending on the on/off ratio, the resistance of R_off is changed correspondingly. In practice, a larger I_ref tolerance is preferred for a secure and robust PUF design as aforementioned in section IV. For a given performance target, e.g. all the HD values should be within the range of 45% to 55%, the simulation results show that the uniformity presents stricter requirement since the I_ref tolerance range is the lowest. Therefore, we employ the I_ref tolerance range of a given uniformity target as the

evaluation metric to study the effect of RRAM's on/off ratio on the performance and robustness of the X-point PUF, as shown in Fig. 3.18. The I_ref tolerance range decreases as the RRAM's on/off ratio decreases, which becomes more severe when the on/off ratio is less than 10. This means the lower switching ratio make the X-point PUF more sensitive to the operating conditions. In addition, lower on/off ratio for a given R_on resistance results in larger column current (Fig. 3.18), which is not desired for the low-power applications.

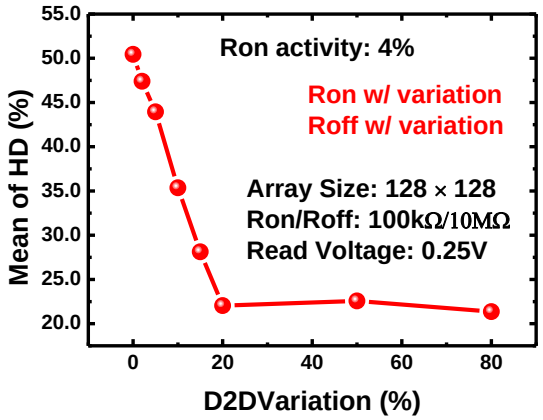


Fig. 3.19 Average HD of 100 random responses generated from one X-point PUF with different device-to-device variation on R_off and R_on. The I_ref is 14.5 μA [68]. © 2018 IEEE.

RRAM Resistance Variation

When we prepare the digital resistance pattern even with the write-verify programming, not all the cells could be programmed to the precise R_on and R_off resistance levels (i.e. 100kΩ and 10MΩ in our case). This could result in device-to-device variations in the array. The effect of device-to-device resistance variation on the diffuseness of X-point PUF is shown in Fig. 3.19. Device-to-device resistance variation will cause degradation of diffuseness and it becomes more severe as the device-to-device resistance variation is larger. Fig. 3.20 shows the diffuseness with 10% R_on and 10% R_off variations and with

10% R_off variation only. We can see that the uniqueness degradation is mainly caused by the variation of R_on cells (instead of R_off cells), because R_on cells dominate the BL current. The device-to-device resistance variation has the similar impact on the other performance metrics (e.g. uniformity and uniqueness) and the simulation results are shown in Table 3.7.

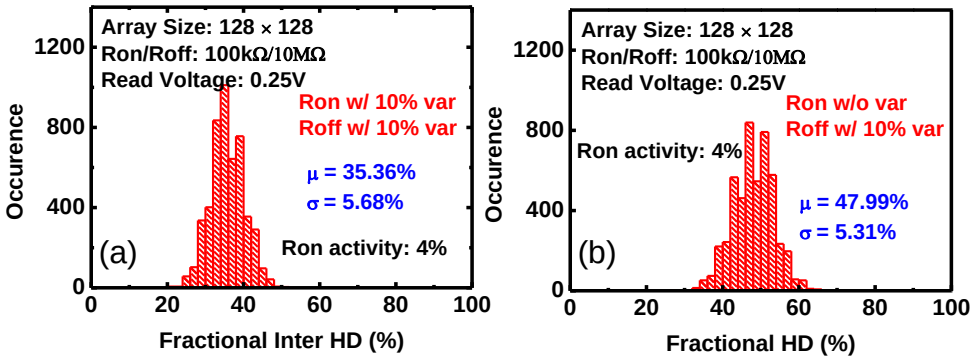


Fig. 3.20 Fractional HD distributions of 100 random responses generated from one X-point PUF with 10% device-to-device variation on (a) both on-state and off-state cells and (b) only the on-state cells [68]. The I_ref is 14.5 μA. © 2018 IEEE.

Table 3.7 Average Uniformity and Uniqueness with Different Device-to-Device Variation

Variation (%)	5	10	15	20
Uniformity (%)	52.45	56.07	61.08	65.64
Uniqueness (%)	49.95	48.74	46.97	44.18

To mitigate the performance degradation, the device-to-device resistance variation, at least for the on-state cells, should be tightly controlled during the resistance pattern preparation stage. We propose to employ the write-verify programming scheme to suppress the variation. For example, when we SET the cells into on-states in the split process and the target resistance of on-state is assumed to be 100 kΩ, a read operation is conducted to read out the current resistance and is compared with the target resistance. If the readout resistance is within the proper resistance range (e.g. 95~105 kΩ) for a given variation

tolerance criterion (e.g. 5%), the SET is successful. Otherwise, more SET-and-read processes are conducted until the readout resistance is within the desired range. A similar process can be done for the cells in off-states, but as discussed earlier, the variation in off-state may not be critical.

Reliability of RRAM Device

For PUF application, RRAM device reliability has two aspects: data retention and read disturb. Data retention refers to how long the RRAM resistance states can be maintained. Generally, it is expected that an RRAM device can maintain the resistance state for longer than 10 years (~3×10^8 s) for nonvolatile memory applications. To evaluate the retention time, temperature-accelerated failure method is typically employed: the device is baked at elevated temperatures over a span of time and read out the resistance by applying read voltage at specific times. With the recorded time-to-failure at each temperature, the lifetime of the expected operating temperature can be extrapolated by the Arrhenius (1/kT) plot. With material engineering and programming scheme optimization, the RRAM device is able to maintain its resistance state for more than 10 years at 85°C (i.e., the operating temperature on chip) and even higher temperatures [70, 71]. Read disturb refers to the resistance of RRAM cell changing over time caused by the continuous read operations. During the deployment phase, only read operation is performed to generate the response. Typically, a read voltage with a specific polarity is applied to RRAM array in the X-point PUF circuit. This will cause the RRAM resistance drifting over time. For example, if positive read voltage is employed to read out the column currents of the X-point PUF it can enable the filament growth gradually, thus resulting in resistance gradually decreasing over time. To avoid the read disturb in the RRAM array, a low voltage (e.g. 0.25 V) is preferred in the read operation to inhibit the read disturb process. Also the I-V nonlinearity

is not a significant issue at such low read voltages. Furthermore, we also propose to use an RRAM dummy column to generate the I_ref to mitigate the read disturb error occurrence since the I_ref will drift along with the RRAM cells following the same trend over continual reading.

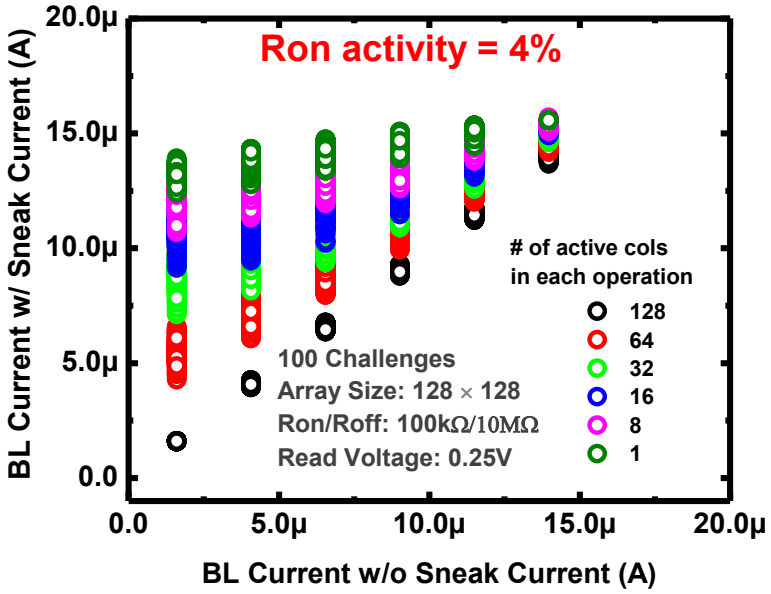


Fig. 3.21 Correlation between column currents with sneak path currents and without sneak path currents of 100 different challenges for different number of active columns per read operation [68]. © 2018 IEEE.

3.5.7 Security Analysis of X-point PUF

Correlation

We investigated the correlation between BL currents with and without sneak paths for different numbers of active columns per read operation, as shown in Fig. 3.21. Here BL currents with sneak paths are obtained by SPICE simulation while BL currents without sneak paths are obtained by hand calculating the sum of all the cells' current along the column. The simulation results show the BL currents with sneak paths tend to be larger when the BL currents without sneak paths are larger. This correlation between BL currents with and without sneak paths becomes stronger when the number of active columns in each

read operation increases, because there are fewer columns left floating. In the extreme case (i.e. active all the columns simultaneously in one operation), the BL currents with sneak path currents are almost the same as the currents without sneak path currents since BL currents are dominated by the main currents instead of sneak path current. This will pose a security issue for X-point PUF because the adversary may be able to microprobe the cell resistance of the array and then use superposition principle to estimate the BL currents for a given challenge. This means that the adversary can produce the response for any challenge by a simple hand calculation. To eliminate this vulnerability, a lower number of active columns per read operation could be employed since the sneak path currents are comparable with the main currents, presenting less correlation between BL currents with and without sneak paths. For example, the BL current without sneak paths = 9.1 μA, the corresponding BL currents with sneak paths may spread between 13.9 μA to 15.1 μA. If the I_ref of SA is 14.5 μA, the response bit may be either "1" or "0", thus unpredictable from a simple superposition.

The other potential security issue is the correlation between challenges and responses. Since the BL current is decided by the challenge vector, the difference between the BL currents generated by two different challenges is more likely to be small if the HD of the two challenge vectors is small. As a result, the response is more likely to be similar. We prepared 100 random CRPs from one PUF instance to investigate the correlation between challenges and responses and results are shown in Fig. 3.22. Although the HD of challenges spreads across a wide range, the HD of responses is always centered at around 50% with some spread. This means that the responses have no dependence on the challenges and it will not introduce undesired security problem for X-point PUF design.

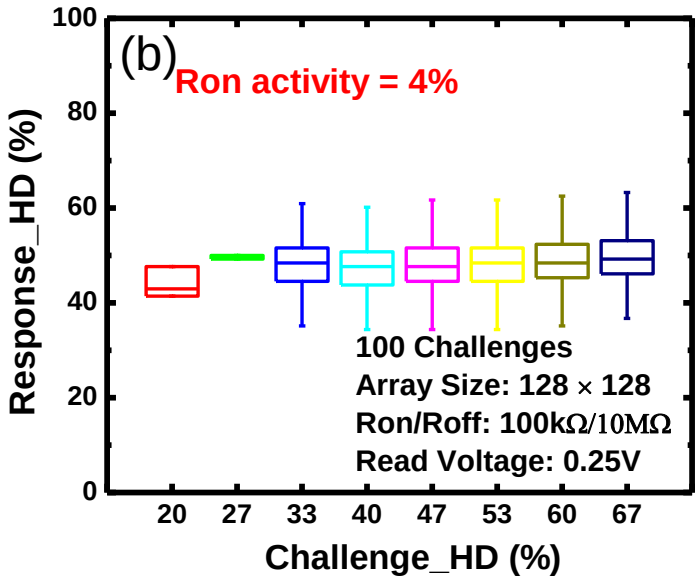


Fig. 3.22 Correlation between hamming distances of challenges and associated responses [68]. The I_ref is 14.5uA. © 2018 IEEE.

Numerical Modeling Attack

If the adversary is presumed to collect the resistance values of the RRAM cell, e.g. by microprobing of the X-point array (though may not be an easy job), he/she might build a SPICE model to simulate the authentic PUF. From our HSPICE simulation, which was carried out with Intel Quadcore i7-4790 CPU, it takes 1.37 s to generate 1 bit of response vector without considering interconnect resistance and RRAM device variation for 128×128 array. Even with the aggressive parallelization and running the 128 bits simultaneously, it still requires around 1.37 s to generate all the 128 bits. However, to evaluate the X-point PUF in the field, it only takes around 10's ns to generate 1 bit of the response by on-chip read circuitry. Then it may take less than 2 μs to readout 128 bits of the response. Therefore, the simulation time is significantly larger than the evaluation time in the field. The public authentication protocol [72] that utilizes the response time

difference between a real PUF device that can sense response within 2 μs on-chip and an adversary who has to run the time-consuming (>seconds) SPICE simulation, can be used.

Numerical Modeling Attack

Delay-based strong PUFs (i.e. Arbiter PUF) are prone to the machine learning attack and the predication rates could be even larger than 99% [53], therefore it is necessary to check the proposed X-point PUF's resistance against such modeling attacks. We used a well-established machine learning (ML) algorithm: multi-layer perceptron (MLP) with backpropagation for this purpose. In this work, a 3-layer MLP is employed to attack the RRAM PUF with a relatively large network topology (i.e. 128-256-256-128) to achieve a good learning capability. The challenge-response pairs (CRPs) used in our ML experiments were generated in the following procedure: first, a set of challenges (e.g. 1,000) was selected randomly from all possible challenges; finally, the corresponding responses were simulated by HSPICE. To train the MLP algorithm, 10, 000 CRPs were prepared as the training set. 1,000 CRPs which do not appear in the training set were used as the testing set. Fig. 3.23 shows the ML attack results with different sizes of training set. The prediction rate of correct bits in the response vectors fluctuates around 50% even with increasing the training set to 10,000. 50% correct rate of a single bit means a pure random guess of the response vectors. In addition, we also investigated the machine learning attack of X-point array with a 5% device-to-device variation under the assumption that the write-verify programming scheme is able to suppress the device-to-device variation within 5%. The simulation results suggest that the security of X-point PUF is not degraded although device-to-device variation might cause performance degradation. In summary, the X-point strong PUF has very high resistance against the machine learning attack.

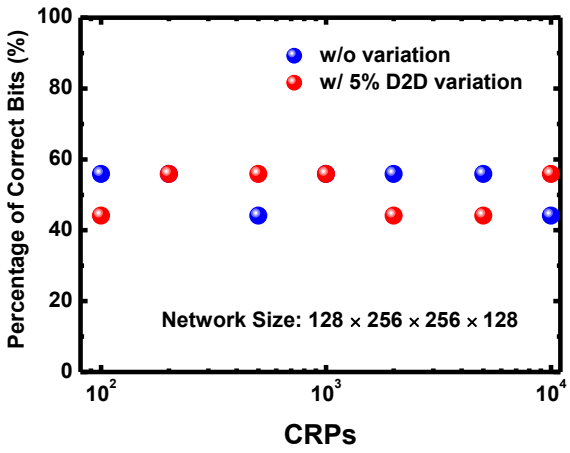


Fig. 3.23 Prediction rate of correct bits in response vectors with size of training set varied from 1,000 to 10,000. The CRPs are simulated without and with 5% device-to-device variation, respectively. The prediction rate does not increase as the number of training set increases for both cases. © 2018 IEEE.

3.5.8 Comparison with Other Strong PUFs

We developed a circuit-level macro model that can be used estimate the area, latency, energy consumption for both X-point PUF and Arbiter PUFs, following the principle of NeuroSim [73]. The hierarchy of the simulator consists of different levels of abstraction from the RRAM cell parameters and transistor technology parameters, to the gate-level sub-circuit modules and then to the array architecture including the peripheral circuits. In the evaluation, the circuit modules for the X-point PUFs include: RRAM array, WL switch matrix, BL MUX and the mux decoder and sense amplifiers; the circuit modules for the Arbiter PUFs include: multiplexor chain, latch, switch matrix and XOR circuit which is only considered in the 4-XOR PUF. Table 3.8 summarizes the estimation results of X-point PUF with one active column and eight active columns in each operation, Arbiter PUF and 4-XOR Arbiter PUF [50, 53] at 65 nm. Compared to Arbiter PUF and 4-XOR Arbiter PUF, the X-point PUF presents a significant advantage in area because the area of RRAM arrays in X-point PUFs is much smaller than the area of CMOS logic gates based multiplexer

chains in the Arbiter PUFs. However, the latency of X-point PUF is larger than that of the Arbiter PUFs. The latency of X-point PUF is dominated by the S/A, which is assumed to be one ns in the simulation. Compare to X-point PUFs with 1-column active, the X-point PUFs with 8-column active presents significant advantages in area, latency and energy. The RRAM array (i.e. 128×128) and WL switch matrix are same for both designs. However, BL mux, the mux decoder and the number of S/As cause the area difference. For PUF with 1-column active, only one S/A is required and shared among 128 columns. Since only one of 128 columns is selected to connect with S/A each time, a 7-bit (27=128) mux decoder is required. While for PUF with 8-column active, 8 out of 128 columns are active and eight S/As are required. This can be considered that 128 columns are split into eight groups and each group contains 16 columns, such that, the mux decoder is only 4-bit (24=16). In summary, the PUF with 1-column active requires one S/A and a 7-bit mux decoder; the PUF with 8-column active needs eight S/As and a 4-bit mux decoder. With the normalized the area, 1) each S/A contributes ~1 unit area; 2) 4-bit mux decoder contributes ~10 unit area and the area for the associated BL mux and routing is ~34 unit area; 3) 7-bit mux decoder requires 109 unit area and the area for the associated BL mux and routing is ~269 unit area. Overall, the area of S/As, mux decoder, BL mux and routing is 379 unit area for the PUF with 1-column active and 52 unit area for the PUF with 8-column active, respectively. Hence, the PUF with 1-column active occupies more area than the one with 8-columns active. Considering the tradeoff between latency and security (correlation between column currents with and without sneak path currents as discussed), eight active columns in each operation might be an optimized choice for X-point PUF, since it can improve the latency by ~8X in comparison to one active column, and it also shows

negligible correlation degradation between column currents with and without sneak path currents (Fig. 3.21). In addition, the energy consumption is also reduced by ~ 7X if eight columns instead of one column are activated in each operation. Compared with 4-XOR Arbiter PUF, the X-point PUF with 8 active columns is able to reduce the area by a factor ~215X, reduce energy by a factor ~ 18X, while increase the latency by a factor of less than 3X.

Table 3.8 Benchmark Results of X-point PUF and Arbiter PUFs at 65 nm

Performance	X-point		Arbiter	4-XNOR Arbiter
	1-column active	8-column active		
Area (μm^2)	7890.69	4503.67	242133	971294
Latency (ns)	128.387	16.25	5.43	5.47
Energy (pJ)	94.79	13.17	59.02	233.87

3.6 Summary

In this chapter, we presented two RRAM based PUF implementations: one for key generation and the other for device authentication.

For the RRAM weak PUF design, we experimentally evaluated RRAM PUF's characteristics such as uniqueness and reliability on 1 kb 1T1R arrays. Design strategies to improve uniqueness, reliability and security have also been proposed. The uniqueness of RRAM PUF can be improved by selecting a more accurate split reference from more dummy cells and minimizing the input offset of the split S/A with relaxed transistor's sizes. The reliability of RRAM PUF can be improved by using multiple RRAM cells to generate one response bit. The security in terms of tamper resistance can be improved by layout obfuscation of hiding S/A into the array and underneath fake RRAM cells. As these proposed strategies come with the expense of latency, energy consumption and area

efficiency, trade-offs should be considered given the application's priorities. The realistic data measured from the RRAM arrays in this work will be valuable for system designers to develop the practical protocols using the RRAM PUF at the system level.

For the digital X-point PUF, the sneak path currents in the X-point array were employed as the entropy source. To improve the poor diffuseness in the "analog" X-point PUF design, we proposed to digitize resistance distribution into on-states and off-states. In order to avoid the dependence of the preference of the response bit to be "1" or "0" on the WL activity, a fixed WL activity (i.e. 50%) was used with the largest CRP space. In each evaluating operation, only one column should be activated to greatly mitigate the correlation between BL current with and without sneak path current. The PUF's characteristics, such as diffuseness, uniformity, and uniqueness, were comprehensively evaluated on 128×128 X-point arrays by SPICE simulation. The simulation results showed that the performance of the proposed PUF design was strongly dependent on the R_on activity and I_ref of S/A. 4% R_on activity presented as an optimal design since it showed the strongest resistance against I_ref variation. On the other hand, the effect of non-ideal properties of the X-point array and RRAM devices on the performance of X-point strong PUF were investigated as well. The interconnect resistance reduces column currents, which become more severe for more advanced technology nodes. However, its impact could be mitigated if the I_ref is scaled by a certain factor for different technology nodes. Higher on/off ratio of the RRAM devices is preferred to maintain a good robustness against I_ref variation. The device-to-device variation might cause a significant degradation in the performance of X-point PUF design and we proposed to employ low read voltage and write-verify programming scheme during resistance preparation phase to mitigate the

effect. In the end, we also discussed the security of X-point PUF through numerical SPICE modeling and machine learning attacks. It showed that the X-point PUF possesses a very high resistance against the numerical SPICE modeling and the machine learning attack. We also compared X-point PUF with Arbiter PUF and 4-XOR PUF in terms of area, latency and energy. Compared with 4-XOR Arbiter PUF, the X-point PUF with 8 active columns could reduce the area by a factor ~215X, reduce energy by a factor ~ 18X, while increase the latency by a factor of less than 3X.

4 MEMORIES FOR MACHINE LEARNING APPLICATIONS

4.1 Overview

4.1.1 Machine Learning

Machine learning is a large sub-field within artificial intelligence (AI), which uses statistical techniques to give computers the ability to learn with data, without being explicitly programmed. Within the machine learning field, there is an area that is often referred to as brain-inspired computation. The brain-inspired computation is a program or algorithm that emulates some aspects of its basic form or functionality that the brain works. It is generally believe that the main computational element of the brain is the neuron. The neurons themselves are connected together with a number of elements entering them called dendrites and an element leaving them leaving them called an axon. These input and output signals are called activations. The connection from a neuron to a neighboring neuron is referred to as a synapse. There are approximately 86 billion neurons 1000 trillion synapses, respectively, in the average human brain. A key characteristic of the synapse is that it can modulate the input signal crossing it. That modulation factor can be referred to as a weight. The brain is believed to conduct learning in a way of adjusting the weights associated with the synapses. Thus different weights result in different responses to an input. This characteristic makes the brain an excellent inspiration for a machine-learning-style algorithm.

Within the brain-inspired computation, there is a sub-area called spiking computing. The network of this area is generally referred to as the spiking neural network (SNN). SNNs take their inspiration from the biological learning rules in brain, such as spike-timing-dependent plasticity (STDP), and the communication on the dendrites and axons

are spike-like pulses. Note that the information being conveyed is not only based on a spike's amplitude but also the time when the pulse arrives. A well-known example of a project that was inspired by the SNN is the IBM's TrueNorth [74]. In contrast to spiking computing, another sub-area of brain-inspired computing is called artificial neural network (ANN). ANN takes their inspiration from the notion that a neuron's computation involves a weighted sum of the input values. These weighted sums correspond to the value scaling performed by the synapses and the combining of those values in the neuron. Furthermore, the output of each artificial neuron is generated only if the weighted sum cross some threshold, which is implemented by a non-linear function. In ANN, typically, artificial neurons are aggregated into layers. Different layers may perform different kinds of transformations on their inputs. Fig. 4.1 shows a picture of a simple 2-layer neural network. The neurons in the input layer receive some values and propagate them to the neurons in the middle layer (also called hidden layer). The outputs of hidden layer are ultimately propagated to the output layer.

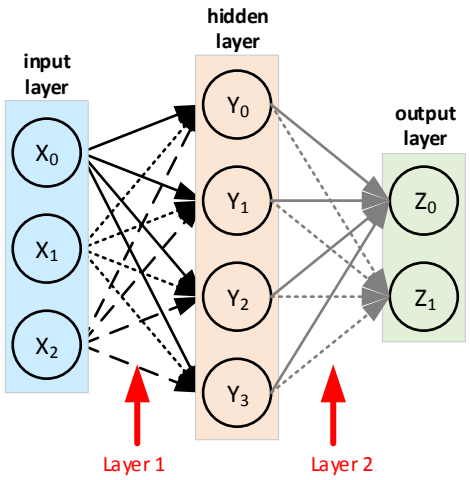


Fig. 4.1 2-layer neural network.

Within the artificial neural networks, there is an area called deep learning, in which the neural networks have more than three layers, i.e. more than hidden layers. The neural

networks used in deep learning are referred to as deep neural networks (DNNs). DNNs are capable of learning high-level features with more complexity and abstraction than shallower neural networks. For example, in an image recognition application, the raw input may be a matrix of pixels; the first representational layer may abstract the pixels and encode edges and lines; the subsequent layers may combine the edges and lines into a higher level features (e.g. shapes). Finally, the last layer may recognize that the image contains an object or scene. Deep learning has had waxed and waned history, which was first proposed in 1940s. The first practical application happened until 1989 with LeNet network for recognizing handwritten digits [75]. These systems are widely used by ATMs for digit recognition on checks. Thanks to advances in computer hardware and software infrastructure and availability of big training data, the deep learning resurges currently, for example, Microsoft's speech recognition system in 2011 [76] and AlexNet system for image recognition in 2012 [77]. The successes of these early DNN applications opened the floodgates of algorithm development. It has also inspired the development of several frameworks, such as Caffe, Tensorflow, Torch, Theano and etc. The existence of such frameworks are not only a convenient aid for DNN researchers and application designers, but also invaluable source of workloads for hardware researchers for exploring hardware-software trade-offs.

4.1.2 Training and Inference

The machine learning algorithms usually have two phases—training and inference. In the specific case of DNNs, the training involves learning and determining the value of the weights (and bias) in the network. Once trained, the program can perform its task by computing the output of the network using the weights determined during training, which

is referred to as inference. In this dissertation, we will focus on the efficient processing of DNN inference instead of training.

4.1.3 Popular Datasets for Image Classification

Image classification is the most common task, which involves being given an image, and selecting one of N classes to which the image most likely belongs. MNIST is widely used dataset for digit classification [78]. It consists of 28 × 28 pixel grayscale images of handwritten digits. There are 10 classes and 60,000 training images and 10,000 test images. CIFAR is a dataset that consists of 32 × 32 pixel colored images of various objects [79]. The CIFAR-10 dataset consists of 60,000 images, in 10 classes, with 6000 images per class. There are 50,000 training images and 10,000 test images. The CIFAR-100 dataset has 100 classes containing 600 images each. There are 500 training images and 100 testing images per class. ImageNet is a large scale dataset that consists of 256 × 256 pixel colored image with 1000 classes [77]. There are 1.3M training images and 100,000 testing images (100 per class) and 50,000 validation images (50 per class). MNIST is a fairly easy dataset, while ImageNet is a challenging one. Therefore, it is important to consider the dataset on which the accuracy is measured when we evaluate the accuracy of a given DNN model.

4.1.4 Overview of DNNs

DNNs have shown remarkable improvements in various intelligent applications such as image classification [77], speech classification [80] and object localization and detection [80]. DNNs can be composed solely of fully-connected layers (FCLs), which is also referred to as multilayer perceptron, or MLP, as shown in the rightmost layers of Fig. 4.2.

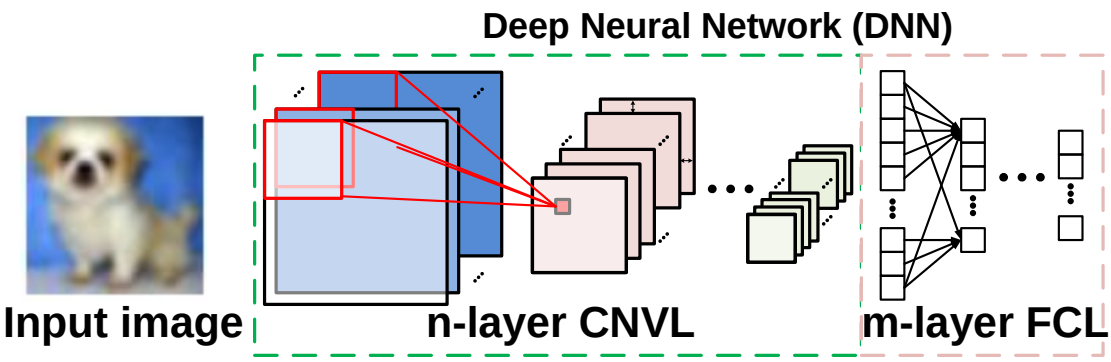


Fig. 4.2 A DNN with convolution layers (CNVL) and fully-connected layers (FCL).

In a FCL, all the output neurons from previous layer are fully connected to each neuron in the next layer as inputs. Hence, the output activations of the next layer are determined by a weighted sum of all input activations of the previous layer. This requires a significant amount of storage and computation. Recently, a common form of DNNs is Convolutional Neural Networks (CNNs), which are composed of multiple convolutional layers (CNVLs) to learn the important features, followed by a small number (e.g. 1 to 3) of FCL for classification, as shown in Fig. 4.2. In a CNVL, an output feature map (IFM) is the result of multiply-and-accumulate (MAC) operations on a collection of weights (or filters, donated by K) operating in a sliding window fashion over a set of the input feature maps (OFMs), each of which is called a channel, as shown in Fig. 4.3. On the other hand, the computation of a CNVL can be defined as

$$OFM(m,x,y) = B(m) + \sum_{c=0}^{C-1} \sum_{i=0}^{K_x-1} \sum_{j=0}^{K_y-1} IFM(c,x+i,y+j) \times K(m,c,i,j) \qquad (4.1)$$

where $IFM(c,x,y)$ is the activation at position (x,y) of $c^{th}(0 \le c \le C)$ input feature map; $K(m,c,i,j)$ is the weight at position (i,j) of c^{th} filter in the set of filters associated with m^{th}output feature map; $OFM(m,x,y)$ is the result of MAC of activation at position (x,y) of $m^{th}(0 \le m \le M)$ output feature map; $B(m)$ is the bias for m^{th} output feature map.

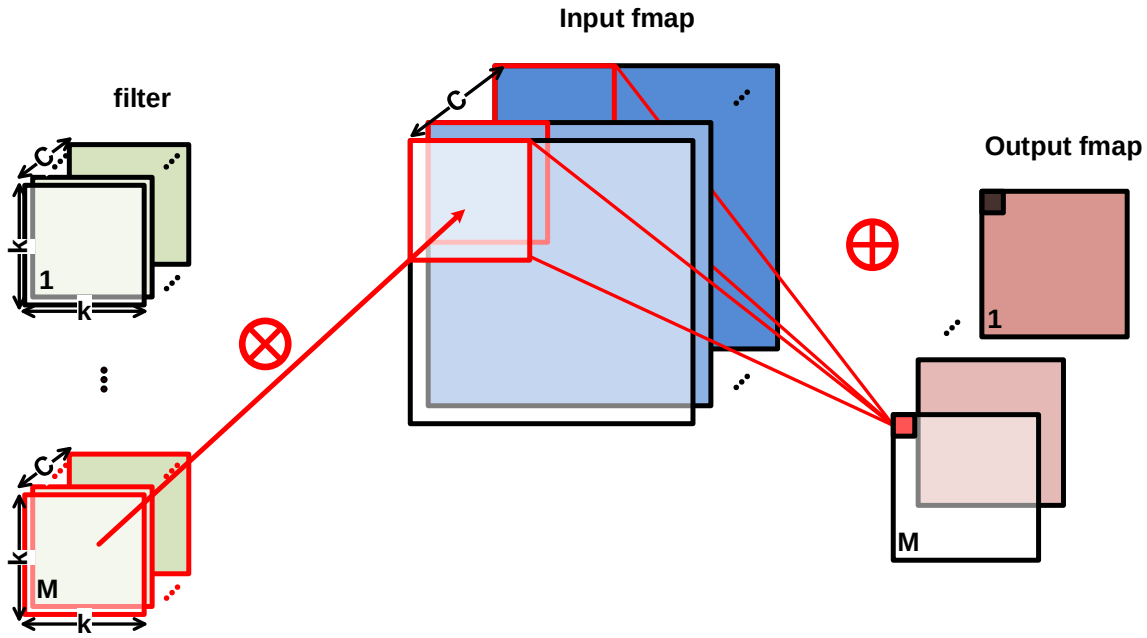


Fig. 4.3 Three-dimensional convolutions with multiple channels in CNNs. k, C and M refer as to the size of filter, number of channels of input featur maps and number of channels of output feature maps, respectively.

A non-linear activation function is typically applied after each CNVL and FCL are used to introduce non-linearity into the DNN, including sigmoid or hyperbolic tangent as well as rectified linear unit (ReLU) [81]. ReLU has become popular in recent years due to its simplicity and its ability to enable fast training. Beside, a pooling layer can be found in a DNN to reduce the dimensionality of a feature map. Pooling, which is applied to each channel separately, enables the network to be robust and invariant to small shifts and distortions. There are two typical pooling techniques, i.e. max pooling and average pooling. It can be configured based on the size of its receptive field (e.g., 2×2) and typically occurs on non-overlapping blocks (i.e., the stride is equal to the size of the pooling). Furthermore, a normalization layer can also be found in a DNN to control the input distribution across layers and can help to significantly speed up training and improve accuracy.

Many DNN models have been developed over the past decade. Each of these models has a different 'network architecture' in terms of number of layers, layer types, layer shapes (i.e., filter size, number of channels and filters), and connections between layers. Table 4.1 summaries the popular DNNs. Increasing the depth of the network tends to provide higher accuracy. Furthermore, most of the computation has been placed on CNVLs rather than FCLs. In addition, the number of weights in the FCLs is reduced and in most recent networks (since GoogLeNet) and the CNVLs dominate in terms of weights. Thus, the focus of hardware implementations should be on addressing the efficiency of the CNVLs, which in many domains are increasingly important.

Table 4.1 Summary of Popular DNNs, adopted from [82].

Metrics	LeNet 5	AlexNet	VGG 16	GoogLeNet (v1)	ResNet 50
Accuracy	n/a	16.4	7.4	6.7	5.3
CONV Layers	2	5	16	21	49
Weights	2.6k	2.3M	14.7M	6.0M	23.5M
MACs	283k	666M	15.3G	1.43G	3.86G
FC Layers	2	3	3	1	1
Weights	58k	58.6M	124M	1M	2M
MACs	58k	58.6M	124M	1M	2M
Total Weights	60k	61M	138M	7M	25.5M
Total MACs	341k	724M	15.5G	1.43G	3.9G

4.2 Hardware Platforms for DNN Processing

The most fundamental and intensive computation in DNN (CNVLs and FCLs) are the MAC operations. In order to achieve high performance, it is common to devote to parallelize the MAC operations. CPUs or GPUs employ a variety of techniques to improve parallelism such as vectors (SIMD) or parallel threads (SIMT). All the arithmetic-logic units (ALUs) share the same control and on-chip memory (register file). These ALUs can

only fetch data from the memory hierarchy (Fig. 1.2) and cannot communicate directly with each other. On the other hand, the popular DNN modes, as shown in Table 4.1, requires tens to hundreds of megabytes of parameter for the millions of MAC operations. For example, VGG-16 network [83] requires 138M parameters and requires 15.5G floating-point precision MAC operations to classify one 224 × 224 input image. This creates significant data movement from on-chip and off-chip memories to support the computation. In fact, the data movement between memories can be more energy-consuming than computation [84]. Therefore, the processing of DNNs has to not only provide high computation parallelism for high throughput but also optimize the data movement to achieve high energy efficiency. To optimize the data movement, it is important to understand the memory access energy in the memory hierarchy, i.e. the access energy is higher when the memory is farther from the ALUs. For example, the energy cost of DRAM access is as ~200× much as the energy cost of on-chip memory access. Some projects have taken significant strides in this direction. In contrast to the temporal architecture used in general-purpose processor, like CPUs and GPUs. The spatial architecture are commonly used nowadays for DNNs in ASIC and FPGA-based designs, where the on-chip memory is distribute to each ALU such that it is closer to the computation unites, such as Eyeriss architecture [85]. There are also been efforts to move the data and compute closer to reduce data movement, thus reducing the memory access cost. For example, advanced memory technology can reduce the access energy for high density memories, such as DRAM. DaDianNao architecture used an embedded DRAM (eDRAM) to bring the high density memory on-chip in their architecture design [86]. The eDRAM is 321× more energy efficient than DRAM (DDR3). eDRAM also offers higher

bandwidth and lower latency. In addition, the DRAM can also be stacked on the top of the chip using through silicon vias (TSV). This technology is often referred to as 3D memory. Tetris [87] demonstrated a DNN accelerator with 3D DRAMs. The same concept has also been explored with 3D SRAM to further reduce the memory access latency [88].

However, the architectures of aforementioned accelerators are still von Neumann architecture and the challenge in memory access degrades the overall performance and energy efficiency of the system. In the context, a new computing paradigm has emerged in recent years as an attractive alternative, which is referred to as compute-in-memory (CIM) or in-memory computing. In contrast to the separation of memory and computation in von Neumann architecture, the CIM architecture integrate the computation into memory, thus reducing the energy of memory access significantly. In the next section, we will discuss the CIM-based hardware accelerator designs.

4.3 Compute-in-Memory Based Hardware Accelerator Design

In earlier work, the DNN models were designed to maximize the accuracy without much consideration of the implementation complexity. However, the high demands on memory storage capacity and computational resources make it challenging to implement and deploy state-of-the-art DNNs on resource-limited platforms such as embedded and mobile devices. It becomes even more challenging as the DNN models trend to be deeper. Various techniques such as network pruning [89] and fixed-point precision [90] were proposed to reduce the energy and area cost of the storage. Recently, it is demonstrated that the precision can be aggressively reduced to 1-bit in Binary Neural Networks (BNNs) [91, 92], which are still able to achieve a reasonable classification accuracy on representative image datasets (e.g., MNIST, CIFAR-10, and ImageNet). Since both the

weights and neuron activations are binarized to +1/-1, thus 1) the memory storage requirement for these BNNs is dramatically reduced; 2) computational resources are significantly reduced as high-precision MAC operations are replaced by XNOR and bit-counting operations. Therefore, BNNs provide a promising solution for on-chip implementation of DNNs. In BNNs, both the weights and activations are constrained to +1/-1. Hence, multiplications between activations and weights can be simplified as bitwise XNOR operations and accumulation of their products are equivalent to bit-counting operation. We refer this type of BNN to as XNOR-BNN. We also constrain the weights and activations in a different binary fashion—weights are binarized to +1/-1 while neuron activations are binarized to 0/1. We refer this type of BNN as hybrid BNN (HBNN). Then the multiplications between activations and weights can be replaced by bitwise multiplications and accumulation of their products are equivalent to bit-counting operation as well. Both BNNs have advantages and disadvantages, targeting to different applications.

In this work, we trained both types of BNNs using similar algorithms proposed in [2,3] on the Theano platform. Note that the binarization function for binary activations in HBNN is:

$$x^b = Sign(x) = \begin{cases} 1 & \text{if } x \geq 0, \\ 0 & \text{otherwise,} \end{cases} \tag{4.2}$$

The binarization function for binary weights in XNOR-BNN is:

$$x^b = Sign'(x) = \begin{cases} +1 & \text{if } x \geq 0, \\ -1 & \text{otherwise,} \end{cases} \tag{4.3}$$

where x^b is the binarized weight or activation and x is the real-valued variable. A convolutional neural network (CNN) (refer to as inspired VGG-16) with 6 convolution layers and 3 fully-connected layers are trained and binarized for evaluations on the CIFAR-10 dataset. The corresponding classification accuracy with floating point (FL) precision is

89.99%. The classification accuracy slightly drops from 89.98% to 88.48% for HBNN and to 88.34% for XNOR-BNN.

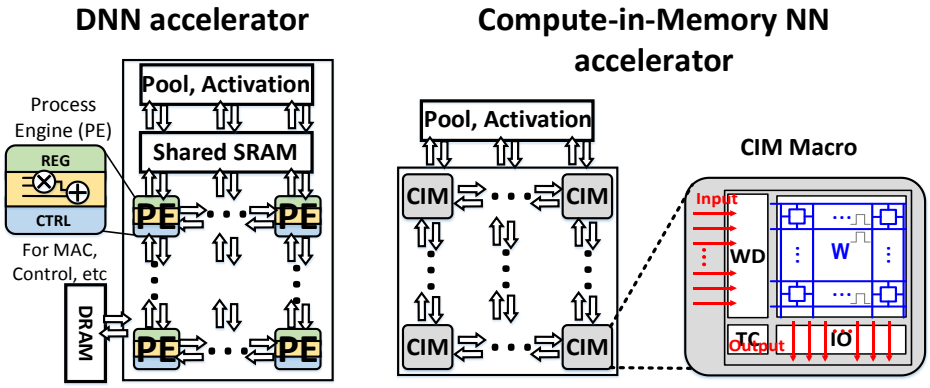


Fig. 4.4 (a) The conventional deep neural network accelerator where the processing element (PE) arrays exploit parallelized computation but with inefficient row-by-row access to the weights stored in shared buffers (i.e. SRAM). (b) The diagram of compute-in-memory (CIM) architecture where the input vectors activate multiple rows and the dot-product output is obtained as column voltage or current [93]. © 2018 ACM

In the DNN accelerator designs with von Neumann architecture, SRAM cache is commonly utilized to store the synaptic weights; however, the extensive computation such as MAC is performed using other logic circuits, e.g. processing element (PE) [85, 94], as shown in Fig. 4.4(a). To improve the data utilization efficiency, parallelized computation is exploited across multiple PE arrays but still with inefficient row-by-row access to the weights stored in the shared SRAM buffers. Therefore, it is more attractive to integrate the computation into the memory array itself (i.e. CIM) as shown in Fig. 4.4(b). The CIM technology can enabled parallel vector-matrix multiplication that the input vectors are send to memory row and activate multiple rows, thus the weighted sum is obtained as column voltage or current. To design the CIM based hardware accelerator, there are various memory technologies are available, e.g. SRAM, DRAM, FLASH and eNVM. Due to the disadvantage in requiring interval “refresh” and large voltage operations in DRAM and

FLASH, respectively, SRAM and eNVM are chosen as the memory technologies to design CIM based hardware accelerator for XNOR-BNN and HBNN, respectively, in this dissertation.

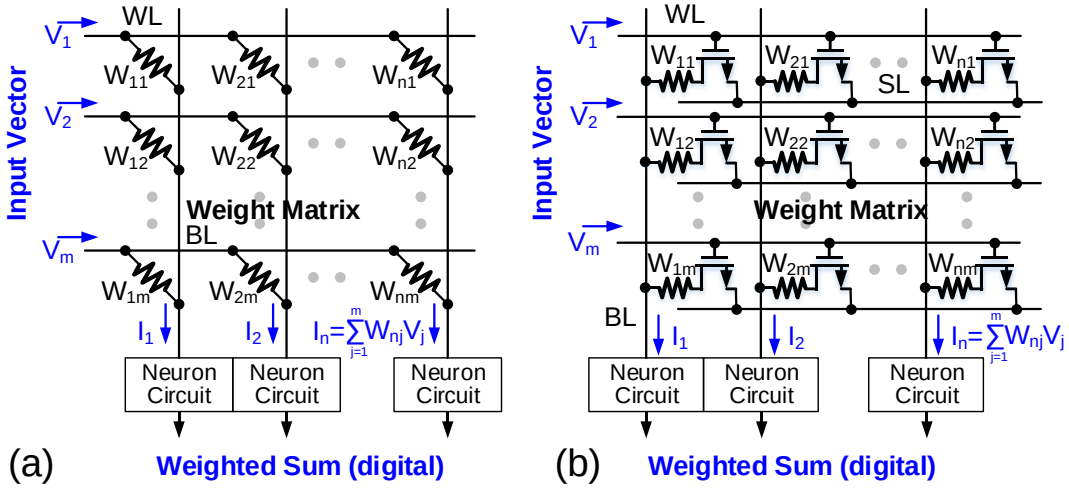


Fig. 4.5 (a) Crossbar eNVM array architecture. (b) Pseudo-crossbar array by 90^o rotation of SL to enable weighted sum operation.

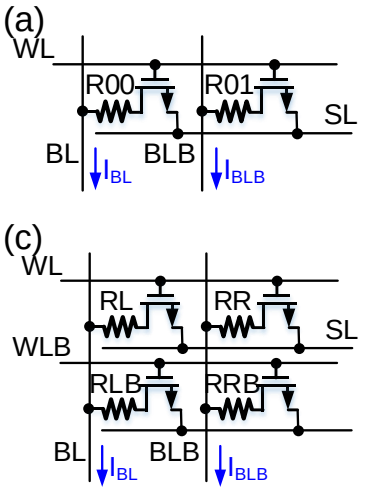
4.3.1 Synapse Cell Design

eNVM

eNVM based synaptic devices can represent the weight with their conductance states, HRS or LRS for binary use. The most compact eNVM based synaptic array structure is the crossbar array structure (Fig. 4.5(a)). Although the crossbar array architecture is simple, it suffers from the write disturbance issue, as there is no isolation between cells, thus leading to inaccurate weighted sum. To eliminate the write disturbance, "pseudo-crossbar" are proposed by rotating source lines (SLs) by 90^o of the existing conventional 1-transistor-1-resister (1T1R) array (Fig. 4.5(b)).

For HBNN, two 1T1R cells are used as the unit synapse cell as shown in Fig. 4.6(a) [95]. We can represent weight +1 as RL = HRS and RR = LRS and the reversed pattern is used for weight -1. The binary neuron activation of 0/1 can be represented by WL of 0/1,

correspondingly. In this way, the value of the discharge current or voltage along the BL during the read-out is dependent on the combination of WL input pattern and bit-cell pattern. Fig. 4.6(b) lists the coding schemes and truth table with possible combination patterns of binary neuron and weight. The corresponding bitwise multiplication results are also shown. The current or voltage difference BL and BLB is effectively a product of input neuron value and the weight value stored in the unit synapse cell. For XNOR-BNN, four 1T1R cells are used as the unit synapse cell as shown in Fig. 4.6(c) [96]. The coding schemes are shown in Fig. 2(d), which is slightly different from the ones for the HBNN.



(b)

Input		Weight			MUL (Product)	
Value	WL	Value	RL	RR	Value	BL vs BLB
1	1	+1	HRS	LRS	+1	$I_{BL} < I_{BLB}$
1	1	-1	LRS	HRS	-1	$I_{BL} > I_{BLB}$
0	0	+1	HRS	LRS	0	No current
0	0	-1	LRS	HRS	0	No current

(d)

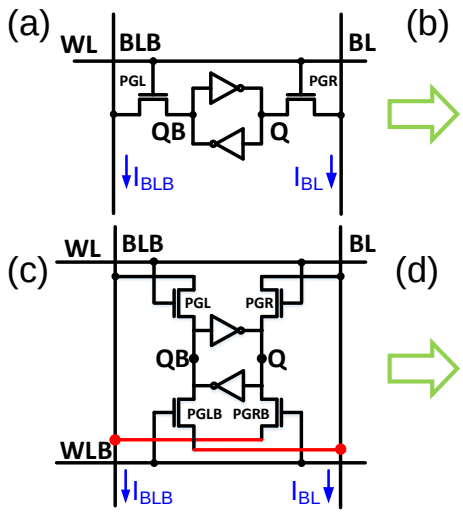
Input			Weight			MUL (Product)	
Value	WL	WLB	Value	RL	RR	Value	BL vs BLB
+1	1	0	+1	HRS	LRS	+1	$I_{BL} < I_{BLB}$
+1	1	0	-1	LRS	HRS	-1	$I_{BL} > I_{BLB}$
-1	0	1	+1	HRS	LRS	-1	$I_{BL} > I_{BLB}$
-1	0	1	-1	LRS	HRS	+1	$I_{BL} < I_{BLB}$

NOTE: RLB = RR and RRB = RL

Fig. 4.6 eNVM based unit synapse cell designs for HBNN and XNOR-BNN, respectively. (a) and (c) are the schematics, respectively. (c) and (d) are the coding schemes, respectively. Current difference of BL and BLB is taken as the output.

SRAM

Although eNVMs hold great advantages on area-efficiency and standby power reduction, those eNVM technologies are still premature for large-scale integration at this moment due to the manufacturing challenges such as the yield, variability, and reliability. In contrast, SRAM has reached the industrial maturity. Therefore, we also proposed two SRAM based CIM designs for DNNs [93, 97].



(b)

Input		Weight			MUL (Product)		
Value	WL	Value	Q	QB	Value	BL	BLB
1	1	+1	1	0	+1	iC	iD
1	1	-1	0	1	-1	iD	iC
0	0	+1	1	0	0	No current	
0	0	-1	0	1	0	No current	

(d)

Input			Weight			XNOR (Product)		
Value	WL	WLB	Value	Q	QB	Value	BL	BLB
+1	1	0	+1	1	0	+1	iC	iD
+1	1	0	-1	0	1	-1	iD	iC
-1	0	1	+1	1	0	-1	iD	iC
-1	0	1	-1	0	1	+1	iC	iD

Fig. 4.7 SRAM based unit synapse cell designs for HBNN and XNOR-BNN, respectively. (a) and (c) are the schematics, respectively. (c) and (d) are the coding schemes, respectively. Current difference of BL and BLB is taken as the output. [93] © 2018 ACM

SRAM stores bi-stable information in one cell. For the HBNN, a conventional 6T SRAM cell is used as a unit synapse cell, as shown in Fig. 4.7(a). Fig. 4.7(b) presents the coding schemes of binary input and weight. For XNOR-BNN, we proposed a customized 8T SRAM cell as shown in Fig. 4.7(c). There are two complementary WLs and two pairs of pass gates (PGs). The first pair of PGs controlled by WL connects Q and QB to BL and BLB, respectively. In contrast, the second pair of PGs controlled by WLB connects Q and QB to BLB and BL, respectively. This design is also different from the conventional 8T SRAM that aims to improve the static noise margin. In our 8T SRAM design, the synaptic weight is stored in Q and QB similarly as in 6T SRAM. However, the input binary neuron is represented with a pair of complimentary WLs (Fig. 4.7(d)). To evaluate the multiplication or XNOR function with both 6T and 8T SRAM synapse cell, BL and BLB will be charged (iC) or discharged (iD) depending on the input and weight pattern since both BLs could decay below VDD. For the evaluation of weighted sum along a column,

we essentially compare the number of '1's coupled to BL and BLB since '1' results in current discharge and iD is significantly larger than iC.

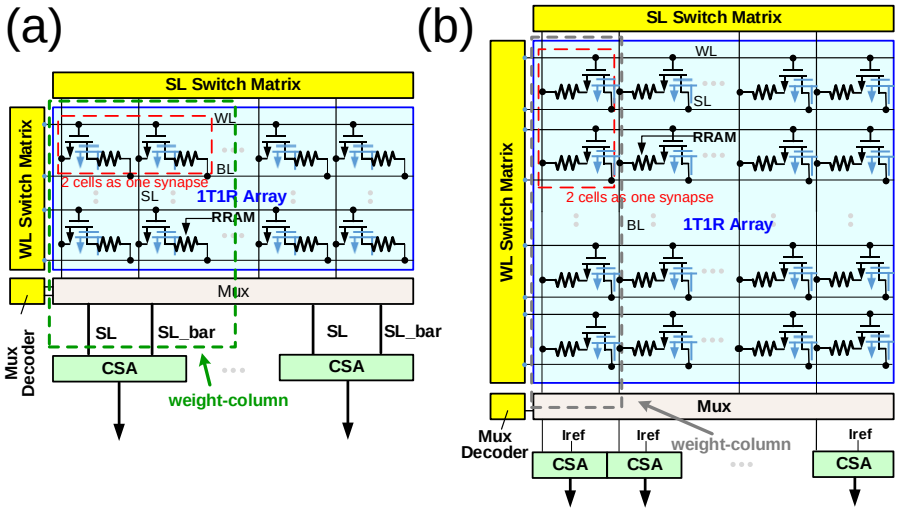


Fig. 4.8 Diagram of proposed eNVM CIM architectures with (a) 2-1T1R and (b) 4-1T1R unit synapse cells and peripheral circuits for activating multiple rows in parallel. © 2018 ACM.

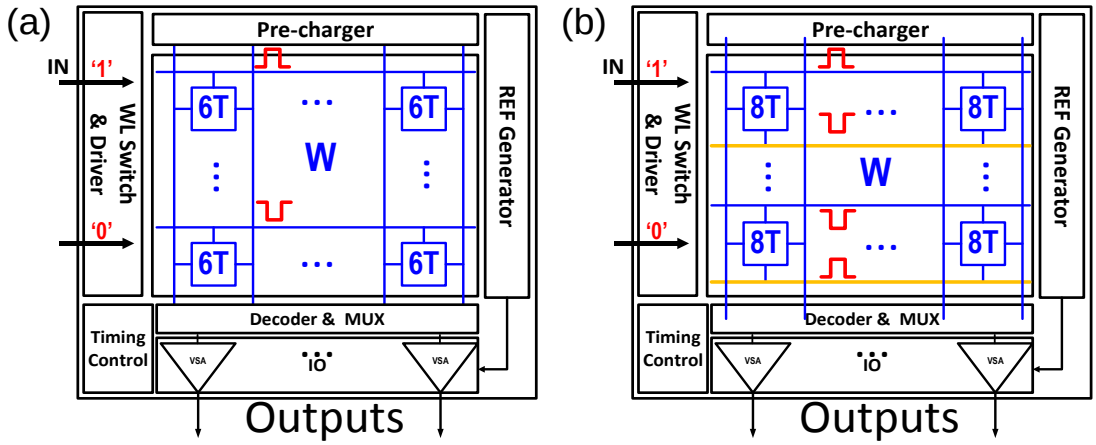


Fig. 4.9 Diagram of proposed SRAM CIM architectures with (a) 6T SRAM and (b) 8T XNOR-SRAM unit synapse cells and peripheral circuits for activating multiple rows in parallel [93]. © 2018 ACM.

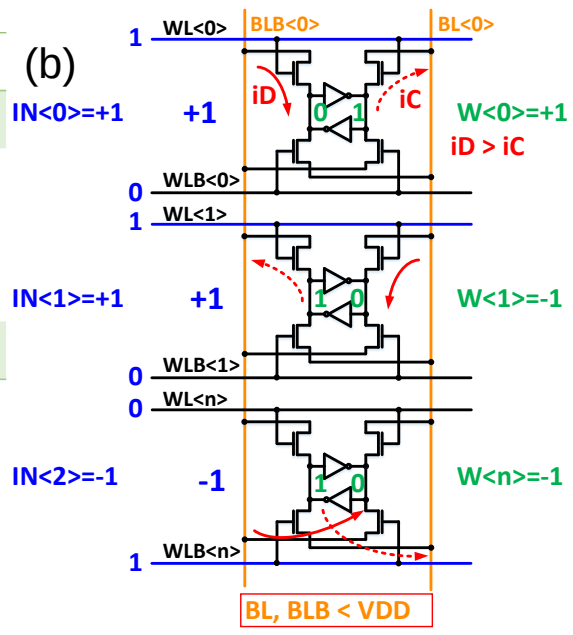
4.3.2 CIM Macro Design

With all the unit synapse cells discussed above, the MAC operation is replaced by bitwise multiplication for HBNN or XNOR for XNOR-BNN plus bit-counting operations. To parallelize the weighted sum operation, we activate multiple word lines in the both eNVM and SRAM arrays simultaneously and digitize the analog current or voltage developed along the BLs by a multi-level sense amplifiers. Fig. 4.8 and Fig. 4.9 show the circuit diagrams of proposed eNVM and SRAM CIM designs, respectively. Compared to conventional design for memory application, some peripheral circuits are different. We will take the SRAM based design as example to explain.

(a)

Input vector	Weight vector	XNOR (Product)		
		Value	BL	BLB
+1	+1	+1	iC	iD
+1	-1	-1	iD	iC
-1	-1	+1	iC	iD
Bit-counting		+1	iD+2*iC	2*iD+iC

(b)



(c)

$|Sum\_BL| + |sum\_BLB| = 3*iD + 3*iC$

$|Sum\_BLB| - |Sum\_BL| = iD - iC$

$Sign(|Sum\_BLB| - |Sum\_BL|) = +1$

$$Sign(x) = \begin{cases} +1, & if\ x \geq 0, \\ -1, & otherwise. \end{cases}$$

Fig. 4.10 An example of bit-wise XNOR and parallel bit-counting in an XNOR SRAM column. (a) Ideal value and corresponding currents on BL and BLB. (b) Current flowing directions along the column. Note iD>iC. (c) Comparison of total currents on BL and BLB.

First, a WL decoder is generally used to drive the WL in the conventional design and each time there is only one row is activated. To calculate the weighted sum, only one bit in the input neuron vector is fed into WL and the output from the BL (or BLB) with a fixed

current or voltage reference of a binary sense amplifier (SA) to determine the binary output. Then, the adders and registers are needed to perform row-by-row summation and store the partial sums, respectively, when the rows are activated consecutively. In our design, a WL switch matrix is employed to activate multiple WLs simultaneously according to the input neuron vector to enable the parallel read-out. Note that in the 8T SRAM design, each bit in the input vector is encoded to a pair of complementary signals to enable one WL and disable the other. In the parallel access design, the currents from multiple rows along the same column contribute together to discharge the bit line (BL or BLB). The MAC operation in the parallel access scheme is performed as follows. The total discharging current or residual voltage after discharge from BL or BLB depends on both input pattern but also weight pattern. Fig. 4.10 shows an example of bit-wise XNOR and parallel bit-counting in an 8T SRAM design. Assume we are doing the 3-bit vector and vector multiplication, and the weight vector is "+1-1-1" and the input neuron vector "+1+1-1". The ideal value of the weighted sum is +1. If we store the weight vector as the data stored in the SRAM cell in a column and we encode the input neuron vector as a voltage vector applied to WL, we will see different current or voltage from BL and BLB. Specifically, the BL will see current of 1 iD (iD: discharging current from one SRAM cell) and the BLB will see current of 2iD in this example. The current difference between BLB and BL is iD, which can be taken as the weighted in analog fashion. Fig. 4.10(b) shows the current flowing directions in the XNOR-SRAM column determined by the input and weight vectors. With binary activation function, the comparison result of total currents of BL and BLB determines the output polarity (Fig. 4.10(c)). Essentially, we compare the number of '1's coupled to BL and BLB since '1' means discharge and iD is significantly larger than iC.

Second, a binary sense amplifier (S/A) is generally used to read the binary data in the SRAM out. For large-scale matrices in FCLs or after unrolling hundreds of convolution kernels in multiple channels in CNVLs, the array partition is necessary to split a large matrix into multiple small sub-arrays. However, the accuracy may be substantially degraded if the binary activation is still used to accumulate the partial sums from the sub-arrays. Hence, the multi-level sense amplifier (MLSA) is employed as an analog to digital conversion (ADC) to maintain higher precision for partial sums, as shown in Fig. 4.11. We also design a reference generation circuit to generate the references to the MLSA. Theoretically, we need a 7-bit MLSA if the array size is 64 × 64 since the range of analog weighted sum is from -64iD to +64iD. This will pose a significant design challenge and area overhead to the overall design. In order to mitigate the design challenge of the MLSA, we proposed to quantize the partial sums for each sub-array.

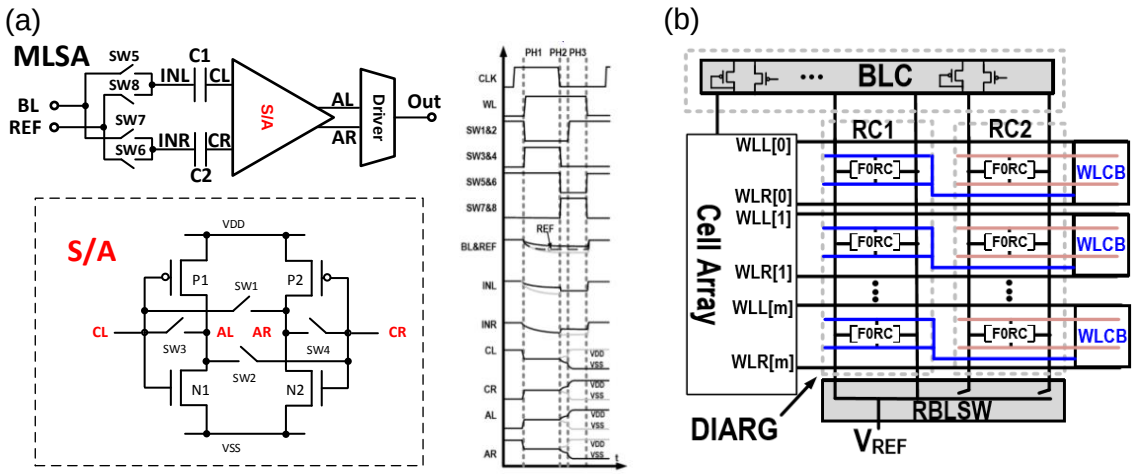


Fig. 4.11 (a) Schematic illustration and waveform of the multi-level sense amplifier (MLSA) and (b) Dynamic Input-Aware Reference Generation [97]. © 2018 IEEE.

4.3.3 Partial Sum Quantization

To determine a proper quantization method, it is necessary to analyze the distribution of partial sums. And the simulation results show that the partial sums present a Gaussian-

like distribution. To minimize the quantization error of the partial sums, we performed non-linear quantization where quantization edges (or references) are determined via Lloyd-Max algorithm [98]. The idea is to make the quantization edges denser in the center of the distribution thus each quantization level has roughly the same number of partial sums. Fig. 4.12 presents the distribution of partial sums for HBNN and XNOR-BNN with seven quantization edges (or references), and eight quantization levels acquired from the Lloyd-Max algorithm. Due to the reduced quantization error, nonlinear quantization achieves significantly better accuracy than linear quantization given the same number of quantization levels. For example, the XNOR-BNN with inspired VGG-16 on CIFAR-10 achieves an accuracy degradation of 0.88% with nonlinear quantization and of 74.07% with linear quantization for 8 quantization levels.

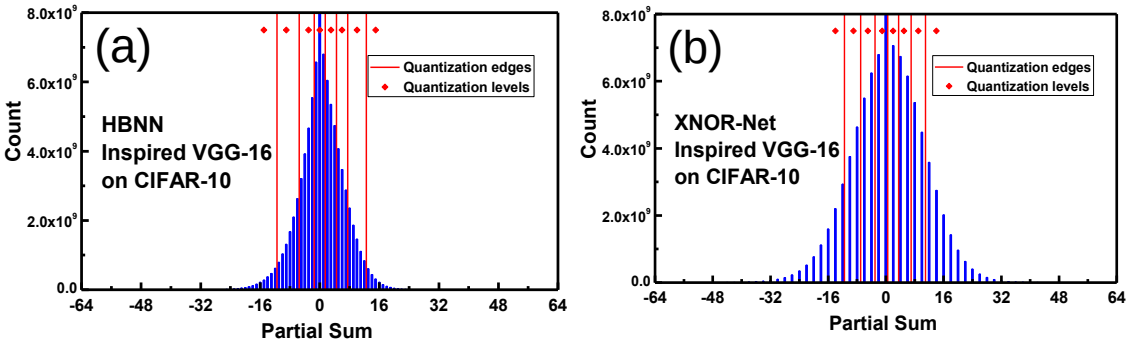


Fig. 4.12 Distribution of partial sums of (a) HBNN and (b) XNOR-BNN collected from inspired VGG-16 on CIFAR-10. Sub-arrays are assumed to be 64×64. Red lines are 7 nonlinear quantization edges (or references) and red diamonds indicate 8 quantization levels [93]. © 2018 ACM.

Fig. 4.13 shows the high-level architecture of the CIM system. MLSAs take the "non-linear" quantization edges as references and generate digital outputs, which then go through thermometer-to-binary (TM2B) encoders and look-up tables (LUTs) to be converted to the corresponding quantization values as partial sums. Those partial sums are added up with

adder trees to generate a final sum, which then goes through the binary activation to generate the neuron output.

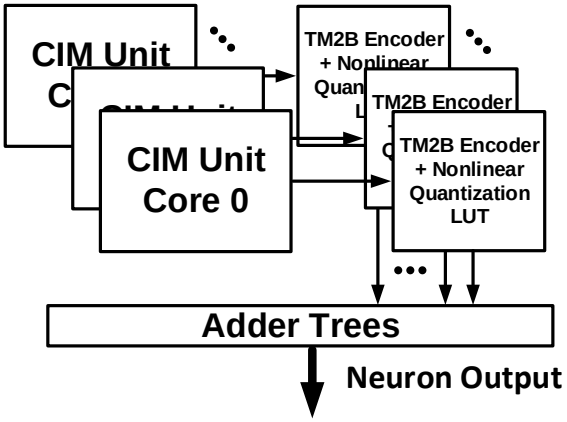


Fig. 4.13 Generic system diagram for implementing a large matrix with multiple small XNOR-SRAM unit cores [93]. © 2018 ACM.

We investigate the impact of quantization levels on the accuracy loss for both HBNN and XNOR-BNN as shown in Fig. 4.14. The simulation results indicate that the XNOR-BNN can achieve slightly better performance than HBNN on CIFAR-10. This is probably because that the XNOR-BNN will have sparser edges than HBNN. Note that the partial sum distribution for XNOR-BNN has an interval of two in between each other, while the interval for HBNN is one. However, the HBNN can achieve satisfying accuracy if the image dataset is relatively simple. For instance, for the same given system, HBNN can achieve an accuracy of 98.83% for LeNet-5 on MNIST, showing only 0.17% degradation compared to the accuracy of ideal BNN algorithm. We also investigate the dependence of classification accuracy degradation on the sub-array sizes (Fig. 4.14(c)). The simulation results suggest that the system with larger sub-array size will cause a slightly larger accuracy degradation than a system with a smaller sub-array size for a given quantization level.

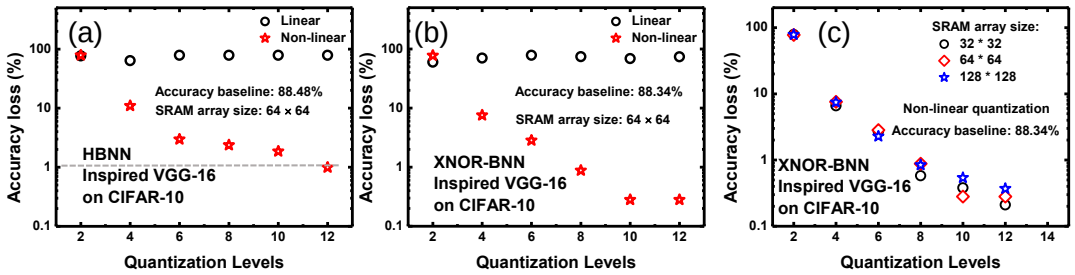


Fig. 4.14 (a) The classification accuracy degradation as a function of quantization levels when the partial sums are compressed by linear and non-linear quantization for (a) HBNN and (b) XNOR-BNN (c) different sub-array size, for inspired VGG-16 on CIFAR-10 [93]. © 2018 ACM.

4.3.4 Comparison between Row by Row and Parallel Computation

We customized a circuit-level macro model NeuroSim [15] that can be used to estimate the area, latency, and energy consumption of hardware accelerators implemented by SRAM synaptic arrays. The hierarchy of the simulator consists of different levels of abstraction from the memory cell parameters and transistor technology parameters, to the gate-level sub-circuit modules, and then to the array architecture. In this work, we estimated the area, latency, and energy-efficiency of conventional row-by-row 6T SRAM, parallel 6T SRAM for HBNN and 8T SRAM for XNOR. 3-bit MLSA is employed for partitioning a 512×512 weight matrix at 65 nm technology node. Standard 6T SRAM is assumed for row-by-row SRAM architecture. Regarding the MLSA design, two implementation strategies are explored here. First, the MLSA can be implemented by a 1-bit SA with different VREF in successive sensing cycles, which is referred to as MLSA_S. Second, the MLSA can be designed with a few stacked 1-bit SAs with different VREF in parallel, which is referred to as MLSA_P. The normalized simulation results in Fig. 7(a) suggest that—1) with MLSA_S, HBNN and XNOR-BNN can achieve both area and latency reduction and energy-efficiency improvement by a factor of >38%, >86%, and >7X;

2) with MLSA_P, latency and energy-efficiency can be improved further but with increased area overhead. Fig. 7(b) shows the comparison results of performance metrics for different sub-array sizes. The sub-array size of 32×32 is taken as the baseline for comparison in each case. With increasing the sub-array size, both area and latency can be further reduced while the energy-efficiency shows a slightly further improvement.

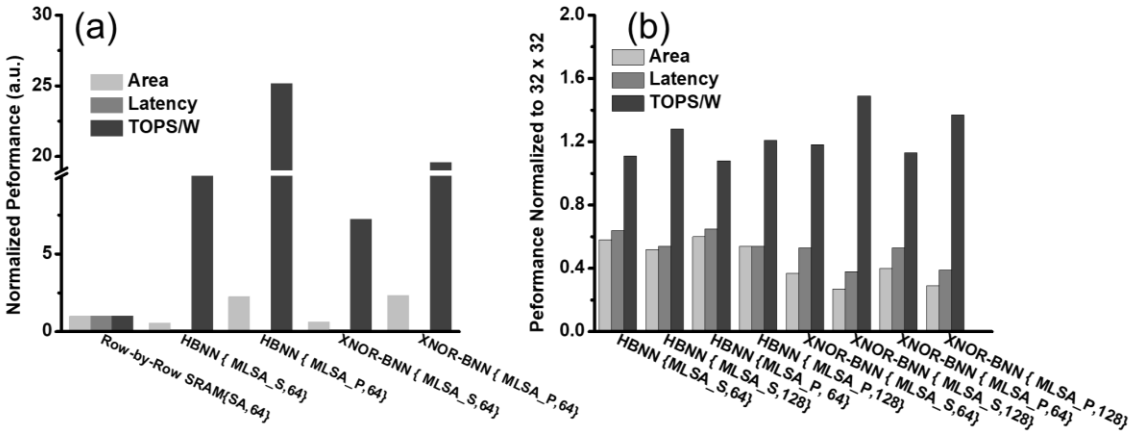
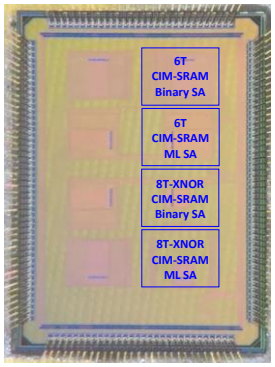


Fig. 4.15 (a) Comparison between different parallel access designs, and the results are normalized to the row-by-row access design. (b) Comparison between different sub-array sizes for parallel designs, and the results are normalized to 32×32 sub-array size [93]. © 2018 ACM.

We also taped-out and validated the proposed HBNN with 6T SRAM and XNOR-BNN with customized 8T SRAM in TSMC 65 nm process. Fig. 8(a) and (b) show the fabricated die photo and the summary of the design parameters and measured performance. The MLSA_S scheme was employed in our tape-out. Fig. 8(c) is the transient waveform of a 2-bit MLSA. In the tape-out, to further reduce energy consumption during read-out, we customized the 6T SRAM cell where we control the two pass gates with two different WLs and only one WL and associated BL are turned on during the read-out operation [16]. Therefore the energy consumption can be reduced by around 50% as compared to the

standard 6T SRAM. The measured silicon data shows that HBNN and XNOR-BNN can achieve energy-efficiency >100 TOPS/W and >50 TOPS/W, respectively, as WL is 0.8 V and VDD for other circuits is 1.0V. Here the one OP is defined as half a 1-bit MAC operation.

(a)



(b)

CHIP SUMMARY	
Technology	65nm CMOS
Unit-Macro Size	4Kb (64x64b)
6T SRAM bit-cell size	0.5μm X 1.05μm
8T XNOR-SRAM bit-cell size	0.51μm X 1.91μm
6T SRAM energy-efficiency	>100 TOPS/W
8T XNOR-SRAM energy-efficiency	>50 TOPS/W

(c)

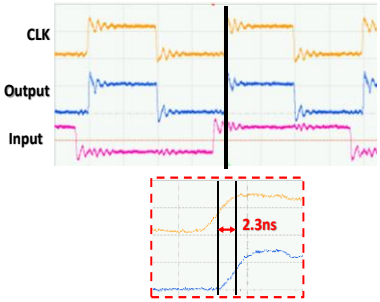


Fig. 4.16 (a) Die photo of the 6T SRAM macro for HBNN and 8T SRAM for XNOR, (b) Summary table of design parameters, and (c) the transient waveform diagram of 1-bit MLSA [93]. © 2018 ACM.

4.4 Summary

In this chapter, we have discussed the current hardware platforms and critical issues, i.e. throughput and energy efficiency, for the machine learning applications. To save the memory and computation cost, techniques, e.g. pruning and quantization, are proposed in the algorithm. From hardware, compute-in-memory is proposed to further reduce the data transfer, suppress intermediate data. We explored the design space of two BNNs, HBNN and XNOR-BNN with two memory technologies, i.e. SRAM and eNVM or RRAM. For HBNN, 6T SRAM is used as bit-cell. For XNOR-BNN, a customized 8T SRAM with complementary WLs is used as bit-cell. To parallelize the weighted sum operation, we activated multiple WLs in the SRAM array and digitize the accumulated analog voltage along BL by MLSA. Array partition was adopted for implementing large-matrices in DNNs. To reduce the quantization error, non-linear quantization was employed for the partial sums

collected from each SRAM array. The impact of the quantization levels on the classification accuracy was also analyzed. We also benchmarked of the area, latency, and energy for row-by-row access and parallel access, showing significant improvement in the parallel access design. Finally, we have validated the proposed HBNN [16] and XNOR-BNN designs with a tape-out in TSMC 65 nm process. A trade-off exists: 6T SRAM based HBNN design could achieve better energy-efficiency than 8T SRAM based XNOR-BNN design due to a simpler bit-cell. However, XNOR-BNN design shows a better accuracy for implementing deeper network on a larger dataset.

5 CONCLUSIONS

In this dissertation, we investigated the SEEs on oxide based RRAM with both 1T1R and crossbar array architectures from device-level, circuit-level and system. 1T1R array suffers from SBU cause by SEU during the set operation. Crossbar array, however, suffers MBU if other oxides are used that lower activation energies in pursuit of low operation voltage due to the propagation of radiation-induced transient spike on the driver at the edge of the array. To compare radiation resistance between 1T1R and crossbar, three factors are considered to evaluate system-level susceptibility: the upset rate, the sensitive area, and the vulnerable time window. Our analysis indicates that the crossbar architecture has a smaller maximum bit-error-rate (BER) per day as compared to the 1T1R architecture for a given sub-array size, I/O width and susceptible time window.

Second, a RRAM weak PUF and a RRAM strong PUF were proposed for cryptographic key generation and device authentication, respectively. The characteristics of RRAM weak PUF were experimentally evaluated with 1 kb HfO_2 based RRAM arrays. Design strategies to improve uniqueness, reliability and security have also been proposed. The uniqueness of RRAM PUF can be improved by selecting a more accurate split reference from more dummy cells and minimizing the input offset of the split S/A with relaxed transistor's sizes. The reliability of RRAM PUF can be improved by using multiple RRAM cells to generate one response bit. The security in terms of tamper resistance can be improved by layout obfuscation of hiding S/A into the array and underneath fake RRAM cells. As these proposed strategies come with the expense of latency, energy consumption and area efficiency, trade-offs should be considered given the application's priorities. The RRAM strong PUF's characteristics, such as diffuseness, uniformity, and uniqueness, were

comprehensively evaluated on 128×128 X-point arrays by SPICE simulation. The simulation results showed that the performance of the proposed PUF design was strongly dependent on the R_on activity and I_ref of S/A. 4% R_on activity presented as an optimal design since it showed the strongest resistance against I_ref variation. On the other hand, the effect of non-ideal properties of the X-point array and RRAM devices on the performance of X-point strong PUF were investigated as well. The interconnect resistance reduces column currents, which become more severe for more advanced technology nodes. However, its impact could be mitigated if the I_ref is scaled by a certain factor for different technology nodes. Higher on/off ratio of the RRAM devices is preferred to maintain a good robustness against I_ref variation. The device-to-device variation might cause a significant degradation in the performance of X-point PUF design and we proposed to employ low read voltage and write-verify programming scheme during resistance preparation phase to mitigate the effect. In the end, we also discussed the security of X-point PUF through numerical SPICE modeling and machine learning attacks. It showed that the X-point PUF possesses a very high resistance against the numerical SPICE modeling and the machine learning attack. We also compared X-point PUF with Arbiter PUF and 4-XOR PUF in terms of area, latency and energy. Compared with 4-XOR Arbiter PUF, the X-point PUF with 8 active columns could reduce the area by a factor ~215X, reduce energy by a factor ~ 18X, while increase the latency by a factor of less than 3X.

Third, we explored the design space of two BNNs, HBNN and XNOR-BNN and designed 4 CIM based hardware accelerator for those two types of BNNs with SRAM and RRAM technologies. With SRAM technology, 6T SRAM and custom 8T SRAM are proposed as bit cells for HBNN and XNOR-BNN implementations, respectively. With

RRAM technology, 2 1T1R cells and 4 1T1R cells are proposed as bit cells for HBNN and XNOR-BNN implementations, respectively. To parallelize the weighted sum operation, we activated multiple WLs in the SRAM array and digitize the accumulated analog voltage along BL by MLSA. Array partition was adopted for implementing large-matrices in DNNs. To reduce the quantization error, non-linear quantization was employed for the partial sums collected from each SRAM array. The impact of the quantization levels on the classification accuracy was also analyzed. We also benchmarked of the area, latency, and energy for row-by-row access and parallel access, showing significant improvement in the parallel access design. Finally, we have validated the proposed HBNN and XNOR-BNN designs with a tape-out in TSMC 65 nm process. A trade-off exists: 6T SRAM based HBNN design could achieve better energy-efficiency than 8T SRAM based XNOR-BNN design due to a simpler bit-cell. However, XNOR-BNN design shows a better accuracy for implementing deeper network on a larger dataset.

Overall, this dissertation explores the memory technologies' new applications beyond memory and data storage towards aerospace applications, secure and energy-efficient computing, and artificial intelligence hardware.

REFERENCES

[1] J. L. Hennessy, and D. A. Patterson, Computer architecture: a quantitative approach: Elsevier, 2011.

[2] J.-G. Zhu, "Magnetoresistive random access memory: The path to competitiveness and scalability," *Proceedings of the IEEE*, vol. 96, no. 11, pp. 1786-1798, 2008.

[3] H.-S. P. Wong, S. Raoux, S. Kim, J. Liang, J. P. Reifenberg, B. Rajendran, M. Asheghi, and K. E. Goodson, "Phase change memory," *Proceedings of the IEEE*, vol. 98, no. 12, pp. 2201-2227, 2010.

[4] H.-S. P. Wong, H.-Y. Lee, S. Yu, Y.-S. Chen, Y. Wu, P.-S. Chen, B. Lee, F. T. Chen, and M.-J. Tsai, "Metal–oxide RRAM," *Proceedings of the IEEE*, vol. 100, no. 6, pp. 1951-1970, 2012.

[5] S. Yu, and P.-Y. Chen, "Emerging memory technologies: Recent trends and prospects," *IEEE Solid-State Circuits Magazine*, vol. 8, no. 2, pp. 43-56, 2016.

[6] S. P. Park, S. Gupta, N. Mojumder, A. Raghunathan, and K. Roy, "Future cache design using STT MRAMs for improved energy efficiency: devices, circuits and architecture," *Proceedings of the 49th Annual Design Automation Conference (DAC)*, 2012, pp. 492-497.

[7] M. Jung, J. Shalf, and M. Kandemir, "Design of a large-scale storage-class RRAM system," *Proceedings of the 27th International ACM Conference on International Conference on Supercomputing*, 2013, pp. 103-114.

[8] R. Liu, D. Mahalanabis, H. J. Barnaby, and S. Yu, "Investigation of single-bit and multiple-bit upsets in oxide RRAM-based 1T1R and crossbar memory arrays," *IEEE Transactions on Nuclear Science*, vol. 62, no. 5, pp. 2294-2301, 2015.

[9] "What Are Cosmic Rays?," Sep 14, 2018; https://www.space.com/32644-cosmic-rays.html.

[10] "Van Allen radiation belt," Sep 14, 2018; https://en.wikipedia.org/wiki/Van_Allen_radiation_belt#Flux_values.

[11] T. C. May, and M. H. Woods, "Alpha-particle-induced soft errors in dynamic memories," *IEEE Transactions on Electron Devices*, vol. 26, no. 1, pp. 2-9, 1979.

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

[13] C. Hsieh, P. C. Murley, and R. O'brien, "A field-funneling effect on the collection of alpha-particle-generated carriers in silicon devices," *IEEE Electron Device Letters*, vol. 2, no. 4, pp. 103-105, 1981.

[14] D. Niu, C. Xu, N. Muralimanohar, N. P. Jouppi, and Y. Xie, "Design trade-offs for high density cross-point resistive memory," *Proceedings of the 2012 ACM/IEEE international symposium on Low power electronics and design*, 2012, pp. 209-214.

[15] W. Lee, J. Park, J. Shin, J. Woo, S. Kim, G. Choi, S. Jung, S. Park, D. Lee, and E. Cha, "Varistor-type bidirectional switch (J_{MAX}>10^7A/cm^2, selectivity~10^4) for 3D bipolar resistive memory arrays," *Symposium on VLSI Technology (VLSIT)*, 2012, pp. 37-38.

[16] J.-J. Huang, Y.-M. Tseng, C.-W. Hsu, and T.-H. Hou, "Bipolar Nonlinear Ni/TiO_2/Ni Selector for 1S1R Crossbar Array Applications," *IEEE Electron Device Letters*, vol. 32, no. 10, pp. 1427-1429, 2011.

[17] K. Virwani, G. Burr, R. Shenoy, C. Rettner, A. Padilla, T. Topuria, P. Rice, G. Ho, R. King, and K. Nguyen, "Sub-30nm scaling and high-speed operation of fully-confined access-devices for 3D crosspoint memory based on mixed-ionic-electronic-conduction (MIEC) materials," *IEEE International Electron Devices Meeting (IEDM)*, 2012, pp. 2.7. 1-2.7. 4.

[18] S. H. Jo, T. Kumar, S. Narayanan, W. D. Lu, and H. Nazarian, "3D-stackable crossbar resistive memory based on field assisted superlinear threshold (FAST) selector," *IEEE International Electron Devices Meeting (IEDM)*, 2014, pp. 6.7. 1-6.7. 4.

[19] S.-S. Sheu, M.-F. Chang, K.-F. Lin, C.-W. Wu, Y.-S. Chen, P.-F. Chiu, C.-C. Kuo, Y.-S. Yang, P.-C. Chiang, and W.-P. Lin, "A 4Mb embedded SLC resistive-RAM macro with 7.2 ns read-write random-access time and 160ns MLC-access capability," *IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, 2011, pp. 200-202.

[20] A. Kawahara, R. Azuma, Y. Ikeda, K. Kawai, Y. Katoh, Y. Hayakawa, K. Tsuji, S. Yoneda, A. Himeno, and K. Shimakawa, "An 8 Mb multi-layered cross-point ReRAM macro with 443 MB/s write throughput," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 1, pp. 178-185, 2013.

[21] T.-y. Liu, T. H. Yan, R. Scheuerlein, Y. Chen, J. K. Lee, G. Balakrishnan, G. Yee, H. Zhang, A. Yap, and J. Ouyang, "A 130.7- mm^2 2-Layer 32-Gb ReRAM Memory Device in 24-nm Technology," *IEEE Journal of Solid-State Circuits*, vol. 49, no. 1, pp. 140-153, 2014.

[22] S. Gerardin, M. Bagatin, A. Paccagnella, K. Grürmann, F. Gliem, T. Oldham, F. Irom, and D. N. Nguyen, "Radiation effects in flash memories," *IEEE Transactions on Nuclear Science*, vol. 60, no. 3, pp. 1953-1969, 2013.

[23] X. He, W. Wang, B. Butcher, S. Tanachutiwat, and R. E. Geer, "Superior TID hardness in TiN/HfO_2/TiN ReRAMs after proton radiation," *IEEE Transactions on Nuclear Science*, vol. 59, no. 5, pp. 2550-2555, 2012.

[24] M. J. Marinella, S. M. Dalton, P. R. Mickel, P. E. D. Dodd, M. R. Shaneyfelt, E. Bielejec, G. Vizkelethy, and P. G. Kotula, "Initial Assessment of the Effects of Radiationon the Electrical Characteristics of TaOx Memristive Memories," *IEEE Transactions on Nuclear Science*, vol. 59, no. 6, pp. 2987-2994, 2012.

[25] E. DeIonno, M. Looper, J. Osborn, and J. Palko, "Displacement Damage in TiO_2 Memristor Devices," *IEEE Transactions on Nuclear Science*, vol. 60, no. 2, pp. 1379-1383, 2013.

[26] W. M. Tong, J. J. Yang, P. J. Kuekes, D. R. Stewart, R. S. Williams, E. DeIonno, E. E. King, S. C. Witczak, M. D. Looper, and J. V. Osborn, "Radiation Hardness of TiO_2 Memristive Junctions," *IEEE Transactions on Nuclear Science*, vol. 57, no. 3, pp. 1640-1643, 2010.

[27] W. G. Bennett, N. C. Hooten, R. D. Schrimpf, R. A. Reed, M. L. Alles, E. X. Zhang, S. L. Weeden-Wright, D. Linten, M. Jurczak, and A. Fantini, "Dynamic Modeling of Radiation-Induced State Changes in HfO_2/Hf 1T1R RRAM," *IEEE Transactions on Nuclear Science*, vol. 61, no. 6, pp. 3497-3503, 2014.

[28] D. Chen, H. Kim, A. Phan, E. Wilcox, K. LaBel, S. Buchner, A. Khachatrian, and N. Roche, "Single-event effect performance of a commercial embedded ReRAM," *IEEE Transactions on Nuclear Science*, vol. 61, no. 6, pp. 3088-3094, 2014.

[29] D. Mahalanabis, H. J. Barnaby, M. N. Kozicki, V. Bharadwaj, and S. Rajabi, "Investigation of single event induced soft errors in programmable metallization cell memory," *IEEE Transactions on Nuclear Science*, vol. 61, no. 6, pp. 3557-3563, 2014.

[30] P.-Y. Chen, and S. Yu, "Compact modeling of RRAM devices and its applications in 1T1R and 1S1R array design," *IEEE Transactions on Electron Devices*, vol. 62, no. 12, pp. 4022-4028, 2015.

[31] Y. Y. Chen, R. Degraeve, S. Clima, B. Govoreanu, L. Goux, A. Fantini, G. S. Kar, G. Pourtois, G. Groeseneken, and D. J. Wouters, "Understanding of the endurance failure in scaled HfO_2-based 1T1R RRAM through vacancy mobility degradation," *IEEE International Electron Devices Meeting (IEDM)*, 2012, pp. 20.3. 1-20.3. 4.

[32] Y. Y. Chen, B. Govoreanu, L. Goux, R. Degraeve, A. Fantini, G. S. Kar, D. J. Wouters, G. Groeseneken, J. A. Kittl, and M. Jurczak, "Balancing SET/RESET Pulse for $>10^{10}$ Endurance in HfO_2/Hf 1T1R Bipolar RRAM," *IEEE Transactions on Electron devices*, vol. 59, no. 12, pp. 3243-3249, 2012.

[33] Y. Y. Chen, L. Goux, S. Clima, B. Govoreanu, R. Degraeve, G. S. Kar, A. Fantini, G. Groeseneken, D. J. Wouters, and M. Jurczak, "Endurance/retention trade-off on HfO_2/metal cap 1T1R bipolar RRAM," *IEEE Transactions on electron devices*, vol. 60, no. 3, pp. 1114-1121, 2013.

[34] "PTM," http://ptm.asu.edu/.

[35] O. Fageeha, J. Howard, and R. Block, "Distribution of radial energy deposition around the track of energetic charged particles in silicon," *Journal of Applied Physics,* vol. 75, no. 5, pp. 2317-2321, 1994.

[36] J. Wirth, and S. Rogers, "The transient response of transistors and diodes to ionizing radiation," *IEEE Transactions on Nuclear Science,* vol. 11, no. 5, pp. 24-38, 1964.

[37] P. E. Dodd, M. R. Shaneyfelt, J. A. Felix, and J. R. Schwank, "Production and propagation of single-event transients in high-speed digital logic ICs," *IEEE Transactions on Nuclear Science,* vol. 51, no. 6, pp. 3278-3284, 2004.

[38] J. D. Gleason, H. J. Barnaby, M. L. Alles, and G. J. Schlenvogt, "An examination of high-injection physics of silicon PN junctions with applications in photocurrent modeling," *IEEE Transactions on Nuclear Science,* vol. 60, no. 6, pp. 4570-4575, 2013.

[39] D. Mahalanabis, R. Liu, H. J. Barnaby, S. Yu, M. N. Kozicki, A. Mahmud, and E. Deionno, "Single event susceptibility analysis in CBRAM resistive memory arrays," *IEEE Transactions on Nuclear Science,* vol. 62, no. 6, pp. 2606-2612, 2015.

[40] W. G. Bennett, N. C. Hooten, R. D. Schrimpf, R. A. Reed, M. H. Mendenhall, M. L. Alles, J. Bi, E. X. Zhang, D. Linten, and M. Jurzak, "Single- and Multiple-Event Induced Upsets in HfO_2/Hf 1T1R RRAM," *IEEE Transactions on Nuclear Science,* vol. 61, no. 4, pp. 1717-1725, 2014.

[41] D. Ielmini, "Filamentary-switching model in RRAM for time, energy and scaling projections," *IEEE International Electron Devices Meeting (IEDM)* 2011, pp. 17.2. 1-17.2. 4.

[42] D. Ielmini, S. Larentis, and S. Balatti, "Physical modeling of voltage-driven resistive switching in oxide RRAM," *IEEE International Integrated Reliability Workshop Final Report (IRW)* 2012, pp. 9-15.

[43] "ITRS," http://www.itrs2.net/.

[44] R. Liu, H. J. Barnaby, and S. Yu, "System-level analysis of single event upset susceptibility in RRAM architectures," *Semiconductor Science and Technology,* vol. 31, no. 12, pp. 124005, 2016.

[45] S. Bourdarie, and M. Xapsos, "The near-earth space radiation environment," *IEEE transactions on nuclear science,* vol. 55, no. 4, pp. 1810-1832, 2008.

[46] R. Fackenthal, M. Kitagawa, W. Otsuka, K. Prall, D. Mills, K. Tsutsui, J. Javanifard, K. Tedrow, T. Tsushima, and Y. Shibahara, "A 16Gb ReRAM with 200MB/s write and 1GB/s read in 27nm technology," *IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC),* 2014, pp. 338-339.

[47] G. Soundararajan, V. Prabhakaran, M. Balakrishnan, and T. Wobber, “Extending SSD Lifetimes with Disk-Based Write Caches,” *FAST*, 2010, pp. 101-114.

[48] V. van der Leest, R. Maes, G.-J. Schrijen, and P. Tuyls, "Hardware intrinsic security to protect value in the mobile market," *ISSE 2014 Securing Electronic Business Processes*, pp. 188-198: Springer, 2014.

[49] M. Rostami, F. Koushanfar, and R. Karri, “A primer on hardware security: Models, methods, and metrics,” *Proceedings of the IEEE*, vol. 102, no. 8, pp. 1283-1295, 2014.

[50] C. Herder, M.-D. Yu, F. Koushanfar, and S. Devadas, “Physical unclonable functions and applications: A tutorial,” *Proceedings of the IEEE*, vol. 102, no. 8, pp. 1126-1141, 2014.

[51] G. E. Suh, and S. Devadas, “Physical unclonable functions for device authentication and secret key generation,” *Proceedings of the 44th Annual Design Automation Conference (DAC)*, 2007, pp. 9-14.

[52] B. Gassend, D. Clarke, M. Van Dijk, and S. Devadas, “Silicon physical random functions,” *Proceedings of the 9th ACM Conference on Computer and Communications Security*, 2002, pp. 148-160.

[53] U. Rührmair, J. Sölter, F. Sehnke, X. Xu, A. Mahmoud, V. Stoyanova, G. Dror, J. Schmidhuber, W. Burleson, and S. Devadas, “PUF modeling attacks on simulated and silicon data,” *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 11, pp. 1876-1891, 2013.

[54] J. Delvaux, and I. Verbauwhede, “Side channel modeling attacks on 65nm arbiter PUFs exploiting CMOS device noise,” *IEEE International Symposium on Hardware-Oriented Security and Trust (HOST)*, 2013, pp. 137-142.

[55] M. Majzoobi, F. Koushanfar, and M. Potkonjak, “Lightweight secure pufs,” *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 2008, pp. 670-673.

[56] J. Guajardo, S. S. Kumar, G.-J. Schrijen, and P. Tuyls, “FPGA intrinsic PUFs and their use for IP protection,” *International workshop on Cryptographic Hardware and Embedded Systems*, 2007, pp. 63-80.

[57] R. Maes, P. Tuyls, and I. Verbauwhede, “Intrinsic PUFs from flip-flops on reconfigurable devices,” *3rd Benelux workshop on information and system security (WISSec), 2008*, 2008, pp. 2008.

[58] C. Helfmeier, C. Boit, D. Nedospasov, and J.-P. Seifert, “Cloning physically unclonable functions,” *IEEE International Symposium on Hardware-Oriented Security and Trust (HOST)*, 2013, pp. 1-6.

[59] Y. Dodis, L. Reyzin, and A. Smith, "Fuzzy extractors: How to generate strong keys from biometrics and other noisy data," *International Conference on the Theory and Applications of Cryptographic Techniques*, 2004, pp. 523-540.

[60] J. Delvaux, and I. Verbauwhede, "Key-recovery attacks on various RO PUF constructions via helper data manipulation," *Proceedings of the Conference on Design, Automation & Test in Europe (DATE)*, 2014, pp. 72.

[61] L. Zhang, Z. H. Kong, C.-H. Chang, A. Cabrini, and G. Torelli, "Exploiting process variations and programming sensitivity of phase change memory for reconfigurable physical unclonable functions," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 6, pp. 921-932, 2014.

[62] J. Das, K. Scott, S. Rajaram, D. Burgett, and S. Bhanja, "MRAM PUF: A novel geometry based magnetic PUF with integrated CMOS," *IEEE Transactions on Nanotechnology*, vol. 14, no. 3, pp. 436-443, 2015.

[63] W. Che, J. Plusquellic, and S. Bhunia, "A non-volatile memory based physically unclonable function without helper data," *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 2014, pp. 148-153.

[64] A. Chen, "Utilizing the variability of resistive random access memory to implement reconfigurable physical unclonable functions," *IEEE Electron Device Letters*, vol. 36, no. 2, pp. 138-140, 2015.

[65] P.-Y. Chen, R. Fang, R. Liu, C. Chakrabarti, Y. Cao, and S. Yu, "Exploiting resistive cross-point array for compact design of physical unclonable function," *IEEE International Symposium on Hardware Oriented Security and Trust (HOST)*, 2015, pp. 26-31.

[66] R. Liu, H. Wu, Y. Pang, H. Qian, and S. Yu, "Experimental characterization of physical unclonable function based on 1 kb resistive random access memory arrays," *IEEE Electron Device Letters*, vol. 36, no. 12, pp. 1380-1383, 2015.

[67] R. Liu, H. Wu, Y. Pang, H. Qian, and S. Yu, "A highly reliable and tamper-resistant RRAM PUF: Design and experimental validation," *IEEE International Symposium on Hardware Oriented Security and Trust (HOST)*, 2016, pp. 13-18.

[68] R. Liu, P.-Y. Chen, X. Peng, and S. Yu, "X-Point PUF: Exploiting Sneak Paths for a Strong Physical Unclonable Function Design," *IEEE Transactions on Circuits and Systems I: Regular Papers*, no. 99, pp. 1-10, 2018.

[69] L. Gao, P.-Y. Chen, R. Liu, and S. Yu, "Physical unclonable function exploiting sneak paths in resistive cross-point array," *IEEE Transactions on Electron Devices*, vol. 63, no. 8, pp. 3109-3115, 2016.

[70] Y. Y. Chen, M. Komura, R. Degraeve, B. Govoreanu, L. Goux, A. Fantini, N. Raghavan, S. Clima, L. Zhang, and A. Belmonte, "Improvement of data retention

in HfO_2/Hf 1T1R RRAM cell under low operating current," *IEEE International Electron Devices Meeting (IEDM)*, 2013, pp. 10.1. 1-10.1. 4.

[71] B. Traoré, P. Blaise, E. Vianello, H. Grampeix, A. Bonnevialle, E. Jalaguier, G. Molas, S. Jeannot, L. Perniola, and B. DeSalvo, "Microscopic understanding of the low resistance state retention in HfO_2 and HfAlO based RRAM," *IEEE International Electron Devices Meeting (IEDM)* 2014, pp. 21.5. 1-21.5. 4.

[72] J. Rajendran, G. S. Rose, R. Karri, and M. Potkonjak, "Nano-PPUF: A memristor-based security primitive," *IEEE Computer Society Annual Symposium on VLSI (ISVLSI)* 2012, pp. 84-87.

[73] P.-Y. Chen, X. Peng, and S. Yu, "NeuroSim+: An integrated device-to-algorithm framework for benchmarking synaptic devices and array architectures," *IEEE International Electron Devices Meeting (IEDM)*, 2017, pp. 6.1. 1-6.1. 4.

[74] P. A. Merolla, J. V. Arthur, R. Alvarez-Icaza, A. S. Cassidy, J. Sawada, F. Akopyan, B. L. Jackson, N. Imam, C. Guo, and Y. Nakamura, "A million spiking-neuron integrated circuit with a scalable communication network and interface," *Science*, vol. 345, no. 6197, pp. 668-673, 2014.

[75] Y. LeCun, B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, and L. D. Jackel, "Handwritten digit recognition with a back-propagation network," *Advances in Neural Information Processing Systems*, 1990, pp. 396-404.

[76] L. Deng, J. Li, J.-T. Huang, K. Yao, D. Yu, F. Seide, M. L. Seltzer, G. Zweig, X. He, and J. D. Williams, "Recent advances in deep learning for speech research at Microsoft," *ICASSP*, 2013, pp. 64.

[77] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, 2012, pp. 1097-1105.

[78] Y. LeCun, C. Cortes, and C. J. C. Burges. "THE MNIST DATABASE of handwritten digits," http://yann.lecun.com/exdb/mnist/.

[79] A. Krizhevsky, V. Nair, and G. Hinton. "THE CIFAR dataset," https://www.cs.toronto.edu/~kriz/cifar.html.

[80] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, and T. N. Sainath, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82-97, 2012.

[81] V. Nair, and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 807-814.

[82] V. Sze, Y.-H. Chen, J. Emer, A. Suleiman, and Z. Zhang, "Hardware for machine learning: Challenges and opportunities," *IEEE Custom Integrated Circuits Conference (CICC)* 2018, pp. 1-8.

[83] K. Simonyan, and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[84] M. Horowitz, "1.1 Computing's energy problem (and what we can do about it)," *IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, 2014, pp. 10-14.

[85] Y.-H. Chen, T. Krishna, J. S. Emer, and V. Sze, "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 1, pp. 127-138, 2017.

[86] Y. Chen, T. Luo, S. Liu, S. Zhang, L. He, J. Wang, L. Li, T. Chen, Z. Xu, and N. Sun, "Dadiannao: A machine-learning supercomputer," *Proceedings of the 47th Annual IEEE/ACM International Symposium on Microarchitecture*, 2014, pp. 609-622.

[87] M. Gao, J. Pu, X. Yang, M. Horowitz, and C. Kozyrakis, "Tetris: Scalable and efficient neural network acceleration with 3d memory," *ACM SIGOPS Operating Systems Review*, vol. 51, no. 2, pp. 751-764, 2017.

[88] K. Ueyoshi, K. Ando, K. Hirose, S. Takamaeda-Yamazaki, J. Kadomoto, T. Miyata, M. Hamada, T. Kuroda, and M. Motomura, "QUEST: A 7.49 TOPS multi-purpose log-quantized DNN inference engine stacked on 96MB 3D SRAM using inductive-coupling technology in 40nm CMOS," *IEEE International Solid-State Circuits Conference (ISSCC)* 2018, pp. 216-218.

[89] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding," *arXiv preprint arXiv:1510.00149*, 2015.

[90] S. Gupta, A. Agrawal, K. Gopalakrishnan, and P. Narayanan, "Deep learning with limited numerical precision," *International Conference on Machine Learning*, 2015, pp. 1737-1746.

[91] M. Courbariaux, I. Hubara, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1," *arXiv preprint arXiv:1602.02830*, 2016.

[92] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, "Xnor-net: Imagenet classification using binary convolutional neural networks," *European Conference on Computer Vision*, 2016, pp. 525-542.

[93] R. Liu, X. Peng, X. Sun, W.-S. Khwa, X. Si, J.-J. Chen, J.-F. Li, M.-F. Chang, and S. Yu, "Parallelizing SRAM arrays with customized bit-cell for binary neural

networks," *Proceedings of the 55th Annual Design Automation Conference (DAC)*, 2018, pp. 21.

[94] N. P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, and A. Borchers, "In-datacenter performance analysis of a tensor processing unit," *Proceedings of the 44th Annual International Symposium on Computer Architecture*, 2017, pp. 1-12.

[95] X. Sun, X. Peng, P.-Y. Chen, R. Liu, J.-s. Seo, and S. Yu, "Fully parallel RRAM synaptic array for implementing binary neural network with (+ 1,− 1) weights and (+ 1, 0) neurons," *Proceedings of the 23rd Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2018, pp. 574-579.

[96] X. Sun, S. Yin, X. Peng, R. Liu, J. Seo, and S. Yu, "XNOR-RRAM: A scalable and parallel resistive synaptic architecture for binary neural networks," *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2018, pp. 1423-1428.

[97] W.-S. Khwa, J.-J. Chen, J.-F. Li, X. Si, E.-Y. Yang, X. Sun, R. Liu, P.-Y. Chen, Q. Li, and S. Yu, "A 65nm 4Kb algorithm-dependent computing-in-memory SRAM unit-macro with 2.3 ns and 55.8 TOPS/W fully parallel product-sum operation for binary DNN edge processors," *IEEE International Solid-State Circuits Conference (ISSCC)*, 2018, pp. 496-498.

[98] J. Max, "Quantizing for minimum distortion," *IRE Transactions on Information Theory* vol. 6, pp. 7-12, 1960.