FEFET BASED NONVOLATILE TCAM AND DRAM DEVELOPMENT

by

Ismail Bayram

Bachelor of Science in Computer Engineering

Istanbul University, 2009

Master of Science in Electrical Engineering

University of Pittsburgh, 2013

Submitted to the Graduate Faculty of

the Swanson School of Engineering in partial fullfilment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2017

UNIVERSITY OF PITTSBURGH

SWANSON SCHOOL OF ENGINEERING

This dissertation was presented

by

Ismail Bayram

It was defended on June 5, 2017

and approved by

Yiran Chen, Ph.D., Associate Professor

Department of Electrical and Computer Engineering

Mao Zhi-Hong, Ph.D., Associate Professor

Department of Electrical and Computer Engineering

Murat Akcakaya, Ph.D., Assistant Professor

Department of Electrical and Computer Engineering

Ervin Sejdic, Ph.D., Associate Professor

Department of Electrical and Computer Engineering

Mingui Sun, Ph.D., Professor

Department of Neurological Surgery

Dissertation Director: Yiran Chen, Ph.D., Associate Professor

Department of Electrical and Computer Engineering

FEFET BASED NONVOLATILE TCAM AND DRAM DEVELOPMENT

Ismail Bayram, PhD

University of Pittsburgh, 2017

Ferroelectric Field Effect Transistor (FeFET) is a promising nonvolatile device which provides high integration density, fast programming speed, and excellent CMOS compatibility. In general, the non-volatility of FeFET is impacted by its physical structure and there is a trade-off between data retention time and device endurance. To improve the cell endurance, for example, the ferroelectric layer of FeFET needs to be programmed to a low polarization level, leading to a short retention time. In ferroelectric DRAM (FeDRAM) design, degradation in FeFET retention time and write-read disturbance requires the FeDRAM cells to be periodically refreshed in order to prevent data loss. In this work, I propose a novel adaptive refreshing and read voltage control scheme to minimize the energy overheads associated with FeDRAM refreshing while still achieve high cell access reliability. In addition to the DRAM application FeFET based TCAM memory is also studied. TCAM (ternary content addressable memory) is a special memory type that can compare input search data with stored data, and return location (sometime, the associated content) of matched data. TCAM is widely used in microprocessor designs as well as communication chip, e.g., IP-routing. Following technology advances of emerging nonvolatile memories (eNVM), applying eNVM to TCAM designs becomes attractive to achieve high density and low standby power. In this work, I examined the applications of three promising eNVM tech-nologies, i.e., magnetic tunneling junction (MTJ), memristor, and ferroelectric memory field effect transistor (FeMFET), in the design of nonvolatile TCAM cells. All these technologies can achieve close-to-zero standby power though each of them has very different pros and cons.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

This work is completed and submitted by Ismail Bayram in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Electrical and Computer Engineering. It consists of the research work from August 2014 to May 2017. Prof. Yiran Chen from University of Pittsburgh is the Ph.D. advisor from 2013 to 2017.

This is an original work to the best of found knowledge, except where citations are made to the prior research. Any other thesis similar to this work which has been submitted for another degree at another university is not found by the author. Part of the work includes:

Published works:

1. **NVMSA 2014: I. Bayram** and Y. Chen, "NV-TCAM: Alternative interests and practices in NVM designs" Non-Volatile Memory Systems and Applications Symposium (NVMSA), 2014

2. **ICCAD2013: I. Bayram**, E. Eken, X. Wang, X. Sun, T.P. Ma and Y. Chen, "Adaptive refreshing and read voltage control scheme for FeDRAM" Circuits and Systems (ISCAS), 2016

Submitted journal publication:

1. **INTEGRATION, the VLSI journal 2017: I. Bayram** and Y. Chen, "NV-TCAM: Alternative Designs with NVM Devices" INTEGRATION, the VLSI journal, (Under Review)

ACKNOWLEDGEMENTS

Firstly I would like to state my sincere gratitude to my advisor Prof. Yiran Chen. His continuous and determined support for my Ph.D. research, his patience, his wast knowledge and the inspiration he gives showed me the way to complete this thesis. I feel thankful for having such an advisor and mentor for my study. He guided me through all obstacles about circuit design, emerging nonvolatile memory devices, architecture based simulations and evaluations and so on. I also would like acknowledge the financial support of the Republic of Turkey Ministry of National Education whose support made this work possible and helped me to complete my Ph.D. degree. My sincere thanks go to Professor Murat Akcakaya, Professor Ervin Sejdic, Professor Zhi-Hong Mao, and Professor Mingui Sun for being my committee members.

I want to show my appreciation to my ei-lab member friends for providing a great innovative environment and support during my research whenever I need. I am also thankful and happy to be a part of the Pitt community. Finally I would like to thank my Mother, Father, my wife Dila Torlak, and my sisters and brothers for their limitless support and prayers during my Ph.D. degree.

1.0 INTRODUCTION

As a result of the recent discovery on ferroelectric properties of HfO2, ferroelectric field effect transistor (FeFET) showed its importance as a strong future alternative for conventional memory technologies such as SRAM, DRAM and FLASH as well as a good competitor for other resistance based emerging NVM devices [19].

Conventional memories are used for different kind of specific applications due to their different advantages and disadvantages. SRAM achieves very fast read-write speeds thanks to its cell structure. Also it needs very low power to keep data [29]. This properties makes it convenient to use for speed oriented applications such a cash memory. However SRAM has a large cell size such as 4T, 6T, 8T, etc. which limits its use when large memory capacity is required. Also SRAM is not non-volatile and stored data will be lost if it cannot get necessary power [29]. Because of those disadvantages, SRAM is useful for performance or low power oriented applications such as caches or mobile devices.

A DRAM cell consists of one transistor and one capacitor so its cell size is very small compared to SRAM [29]. DRAM is useful for high memory density required applications because of its cell size. Disadvantage of DRAM is the data storage method. The data is stored in the capacitor as charge and due to nature of capacitor it leaks quickly. Thus, data in the DRAM cells must be refreshed frequently which causes high power consumption and data loss in case of power failure [29]. DRAM generally used as main memory in most systems.

FLASH memories, which are named as NAND or NOR Flash because of their cells connection scheme, are the recent replacement of mechanical hard drives which have been used as main storage in various systems since years [29]. Advantages such as non-volatility, high density and no leakage or refresh power consumption makes Flash memory the main

storage solution of mobile or low power devices today. Very slow write speeds and very low durability are main disadvantages of FLASH and limits its applications. Most FLASH memory has about 100.000 cycle wear free life time [29].

FeFET emerges as a promising candidate to replace the conventional main memory technology because it has all the advantages of SRAM, DRAM and FLASH [19]. The ability to use a non-volatile of memory in the system as main memory would reduce the power consumption as well as removing the risk of loosing valuable data in case of a power failure. FeFET is basically a regular transistor which has a ferroelectric layer in its gate stack [19]. As a result of its simple structure, it provides excellent CMOS compatibility and scalability.

Besides its advantages FeFET has also drawbacks. One of the issues of FeFET is gradual loss polarization on its ferroelectric layer due its physical features [19]. Also programming the device so many times or programming it to its highest possible polarization level reduces the endurance [19]. Thus, one has to choose the programming procedure carefully and employ different mechanisms in regard to FeFet's trade-offs.

In this work, I will delve into FeFET characteristics, disadvantages and investigate possible memory schemes to employ FeFET cells in TCAM structure. Also I propose new mechanisms to make FeFET convenient to use as memory in cell in DRAM structure.

2.0 NVTCAM: ALTERNATIVE INTERESTS AND PRACTICES IN NVM DESIGNS

TCAM (ternary content addressable memory) is a special memory type that can compare input search data with stored data, and return location (sometimes, the associated content) of matched data. TCAM is widely used in microprocessor designs as well as communication chip, e.g., IP-routing. Following technology advances of emerging nonvolatile memories (eNVM), applying eNVM to TCAM designs becomes attractive to achieve high density and low standby power. In this chapter, the applications of three promising eNVM technologies, i.e., magnetic tunneling junction (MTJ), memristor, and ferroelectric memory field effect transistor (FeMFET), in the design of nonvolatile TCAM cells are examined. All these technologies can achieve close-to-zero standby power though each of them has very different pros and cons.

2.1 INTRODUCTION

Ternary content addressable memory (TCAM) is a type of memory storage device which compares input search data with stored data, and returns location (sometime, the associated content) of matched values 0. With capability of performing parallel searching over all the cells in a single cycle, TCAM is widely used in applications that require high-speed searching, such as IP look-up in network routers and virus scanners in mobile devices [22]. However, to realize the high-parallelism of comparison and high-frequency of lookups, conventional CMOS TCAM design pays significant cost of integration density and standby power consumption. Many schemes have been proposed to overcome those drawbacks. As

one example, turning off or lowering power supply when the TCAM cells are in standby mode effectively reduces the leakage power of the design.

In recent years, many emerging memory technologies, such as magnetic memory (MRAM) , resistive memory (ReRAM), and ferroelectric memory (FeRAM), have been extensively studied. The relevant storage devices, including magnetic tunneling junction (MTJ), metal-oxide device (a.k.a memristor), and ferroelectric memory field effect transistor (FeMFET) also feature many advantages like zero leakage power, high integration density, fast access time, and excellent CMOS compatibility. Emerging memories have demonstrated great potentials in nonvolatile logic, memory design, on-chip and off-chip storage and computing applications [13, 29, 26, 5, 24, 8, 4, 27, 21] . The unique properties of these new devices, especially the non-volatility, make them perfect candidates for nonvolatile TCAM designs with ultra-low standby power.

In this chapter, I examined some possible TCAM designs based on the above three nonvolatile storage devices. Compared to traditional CMOS TCAM cell implementation with 12 or 16 transistors (12T or 16T), the topology of the TCAM cell is dramatically simplified, e.g., with only 2 to 4 transistors. As a result, the integration density is significantly improved. The utilization of nonvolatile storage device also greatly minimizes the standby power of TCAM cells [15, 14] . As shown in the simulation results, a sufficiently large voltage sensing margin is still maintained in such designs in order to realize a fully parallel search operation.

The rest of this chapter is organized as follows: Section 2.2 presents the basics of the introduced emerging memory technologies and TCAM design; Section 2.3 presents non-volatile TCAM designs with different emerging storage de-vices; Section 2.4 compares the introduced different designs and Section 2.5 concludes the chapter.

2.2 PRELIMINARY

2.2.1 TCAM

Different from memory cells, a ternary content addressable memory (TCAM) cell usually stores the memory locations of data but not the data itself. It takes an input as search word and returns the matched data which indicates a specific memory location [20] . A TCAM cell stores three different data values: logic 1, logic 0 and X. Here X represents a dont care value [20] . If the TCAM cell stores X value, the match-line of the corresponding cell stays high no matter what value of the search bit is. In conventional CMOS TCAM design, the three different data values are represented by two bits which are stored in two SRAM cells, as shown in Fig. 2.1.

A basic model for TCAM structure is shown in Fig. 2.2 [14]. A TCAM memory consists of TCAM cells which form word rows with sense amplifiers for each match-line (ML) and the necessary peripheral circuits such as row/column decoders and search-line/write-line drivers [14] . Each row represents a word and each TCAM cell represents one bit of the word. There are separate match-lines for each word and also separate search-line pairs for each bit of the word.

During search operation, all the match-lines will be first pre-charged to high assuming that all match-lines are in match state. After that, search-line (SL) pairs send the search

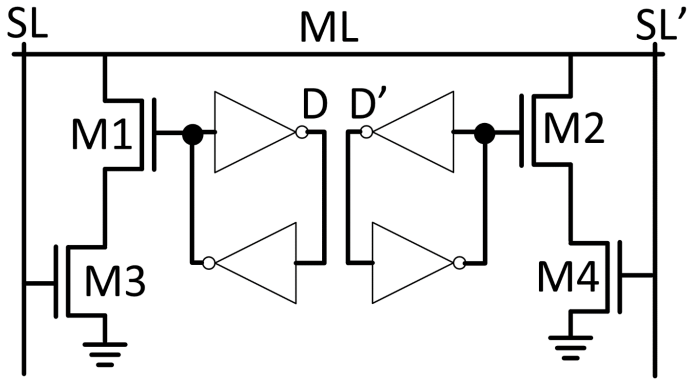


Figure 2.1: Conventional CMOS TCAM cell structure.

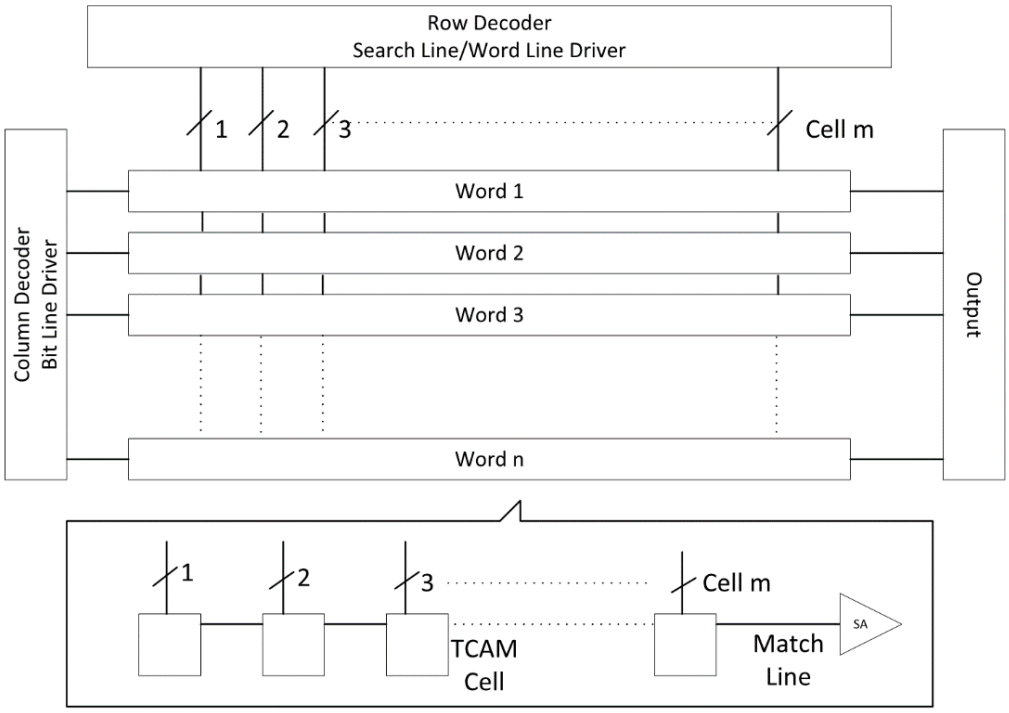


Figure 2.2: Conventional CMOS TCAM array structure.

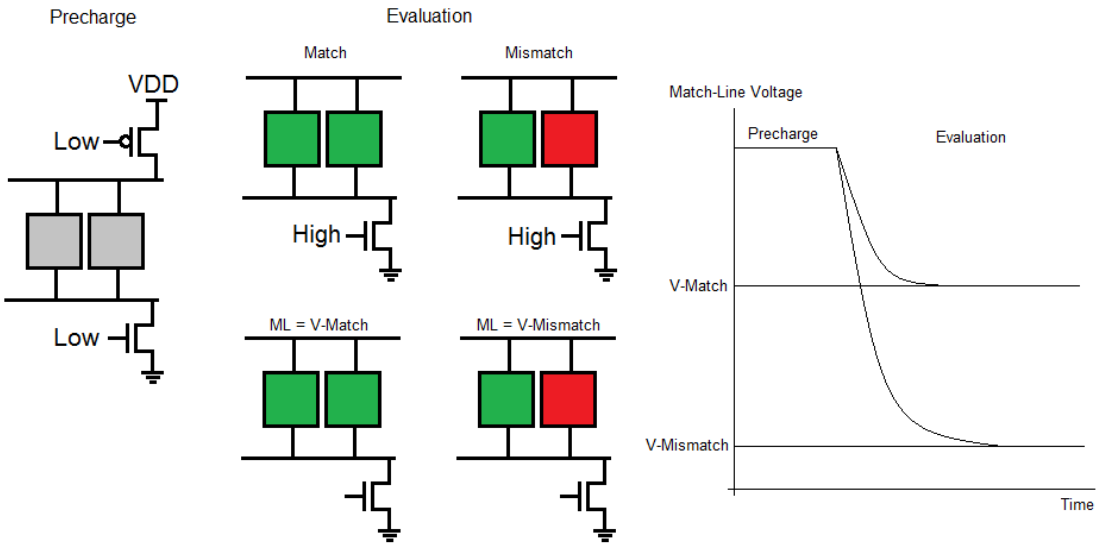


Figure 2.3: Conventional CMOS TCAM array structure.

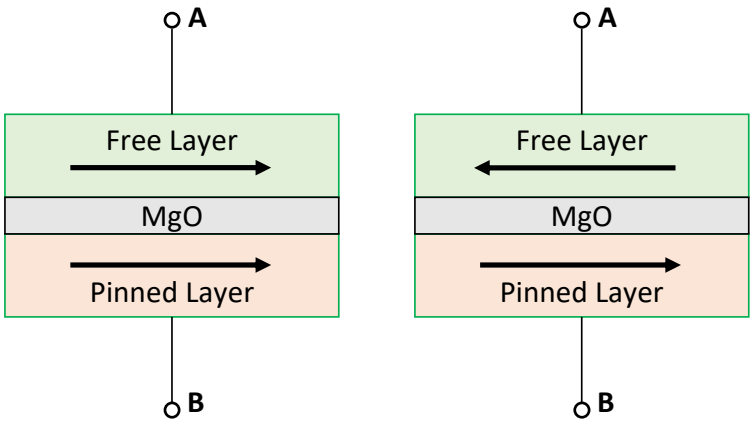


Figure 2.4: Magnetic tunneling junction device structure and its parallel and anti-parallel states.

word into the TCAM and each TCAM cell compares its corresponding bit of the search word with its stored data. If all the bits of the search word match the stored data in all the TCAM cells in a word, the corresponding match-line stays high, indicating a match. Contrarily, if there is any bit mis-matches the input data, the match-line of the corresponding word will be pulled down to ground, indicating a mismatch. Fig. 2.3 depicts the overall search operation.

2.2.2 Magnetic Tunneling Junction

Magnetic Tunneling Junction (MTJ) is a storage element which stores the data based on its resistance states [6] . As shown in Fig. 2.4, MTJ consists of three layers: an insulation barrier which is sandwiched between two ferromagnetic layers. One of the ferromagnetic layers, called reference layer, has a permanent magnetization direction while the other one, called free layer, has a switchable magnetization direction. The resistance state of the MTJ depends on the relative magnetization directions of these two ferromagnetic layers [6] : If the magnetization directions of the free layer and the reference layer are in parallel, the MTJ is in low resistance state, denoting logic 0; otherwise, the MTJ is in high resistance state, denoting logic 1.

The magnetization direction of the free layer can be switched by applying a polarized write current, which must be sufficiently large and/or last for sufficiently long time [6] .

When a write current is inserted from the free layer to the reference layer, for example, the magnetization direction of the free layer will switch to be in parallel to that of the reference layer and the MTJ resistance state will change to low; If the write current is inserted in the opposite direction, the MTJ resistance state will change to high. The small size and the simplicity of MTJ device make it a good candidate for future high density and high speed memory implementations, such as main memory, on-chip cache and CAM and TCAM [6] .

2.2.3 Memristor

In 1971, Prof. Leon Chua stated that since there are four fundamental circuit variables; current (i), voltage (v), charge (q) and magnetic flux (φ), there must be also six possible combinations of these variables. Two of them are the integral relationship between q and i (i.e. $q(t) = \int_{-\infty}^{t} i(\tau) * d(\tau)$) and the one between φ and v (i.e. $\varphi(t) = \int_{-\infty}^{t} v(\tau) * d(\tau)$) over time. In the other four relationships, three of them define the three basic circuit elements resistor (v and i), inductor (φ and i) and capacitor (q and v). Based on Chuas theory, the remaining relationship between φ and q defines the fourth basic circuit element, which is named by him as memristor. A memristor is a two terminal device which can be formulated as $d(\varphi)$ =Md(q) [2]. If the elements of the formula are considered as linear device, memristance (or the resistance state of a memristor) will become a constant and the memristor works as a regular resistor [25] . On the other hand, if we consider the M as a function of q, the memristor will become nonlinear. Different from regular resistor, a memristor has a hysteretic i/v curve. The memristance M is decided by the historical profile of the electrical excitation and changes based on the time, the state of the system and the current.

The first physical implementation of memristor was presented in 2008 [25] . The device has a metal-oxide-metal structure, as shown in Fig. 2.5 and two Pt terminals. The oxide, which is sandwiched between the two metal electrodes, is TiO2. The oxide itself has two different regions one region is doped with oxygen vacancies which make it conductive and the other region is un-doped and behaviors as an insulator. When a voltage difference is applied between the two metal electrodes, oxygen vacancies drift from one region to the other based

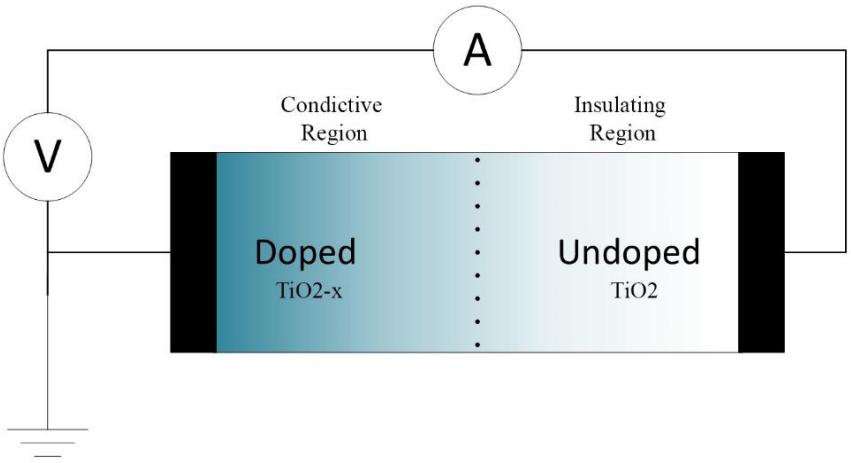


Figure 2.5: Memristor device structure.

on the applied voltages polarization. If the oxygen vacancies move to the un-doped region, the resistance of the device will decrease. On the contrary, if the oxygen vacancies move to the doped region, the resistance of the device will increase.

2.2.4 Ferroelectric Field Effect Transistor

Ferroelectric field effect transistor (FeFET) is a special transistor structure with a gate array that includes a ferroelectric film as memory element, as shown in Fig. 2.6 [1] . A common structure of the FeFET gate array is Metal-Ferroelectric-Insulator-Silicon (MFIS) [19] . Data is stored as the polarization direction of the ferroelectric film on the gate stack. Write operation is performed by applying different voltage pulses to the ferroelectric capacitor. The polarization of the ferroelectric film switches depending on the direction of the applied voltage [1] .

The polarized ferroelectric film shifts the threshold volt-age of the transistor [1]. For example, if the polarization direction is opposite to applied read voltage, which is usually from gate to channel, the threshold voltage of the transistor will increase, indicating logic 0; otherwise, the threshold voltage of the transistor will decrease, indicating logic 1. FeFET is a promising technology for high density, high-speed memory applications due to its simple structure.

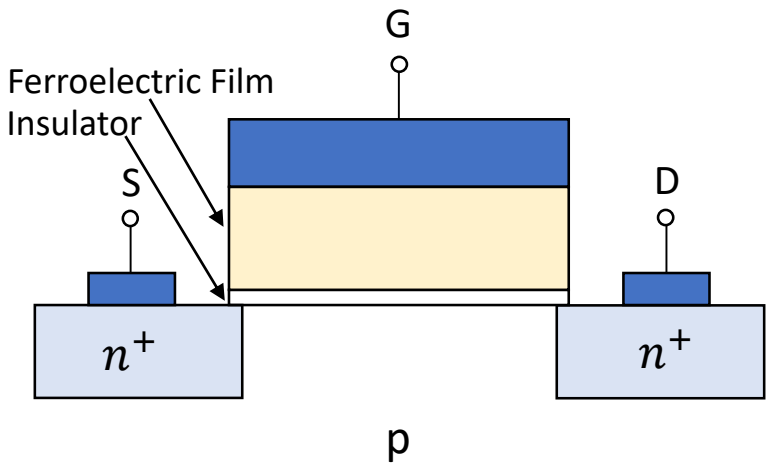


Figure 2.6: Ferroelectric memory field effect transistor device structure.

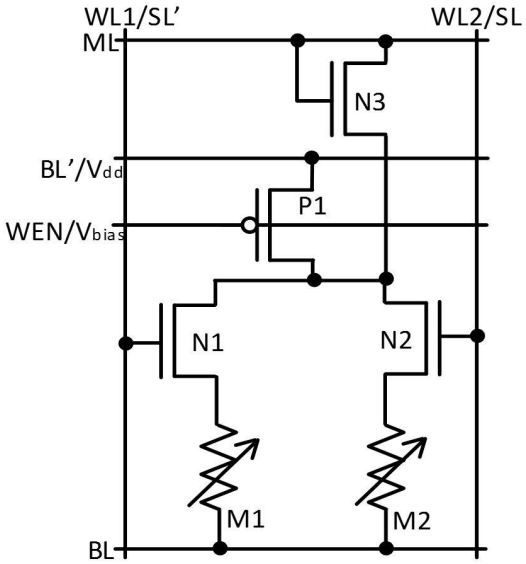
2.3 NON-VOLATILE TCAM DESIGNS

2.3.1 4T/2J Nonvolatile TCAM cell

A compact two-transistor-two-MTJ (2T/2J) nonvolatile TCAM cell was proposed in [14] . However, because of small match-line voltage difference between match and mismatch states, the proposed design can only operate in word-parallel bit-serial fashion [15] . As a result, it has a long search time and high energy consumption. In order to solve this problem, a 4T/2J nonvolatile TCAM cell is proposed in [15]. For comparison purpose, we reproduced this design using SMIC 65nm library.

Fig. 2.7 shows the simulated design where PMOS transistor P1 and NMOS transistor N3 are introduced to keep the match-line voltage difference between match and mis-match states sufficiently large so as to operate at fully parallel search mode even for a long word length. Here N1-N2 are select transistors and M1-M2 are MTJ devices.

Search operation of the MTJ-based TCAM shown in Fig. 2.7 is comprised of two phases, pre-charge and evaluation, as shown in Fig. 2.8 [15]. In pre-charge phase, the select transistors (N1 and N2) are at cut-off state and the match-line is charged to Vdd. In evaluation phase, TCAM cells are activated and the match-line voltage gradually pulls down through the TCAM cells. The logic state of the selected TCAM cell, which is stored as the resistance



Stored Data		SL	ML
M	M1,M2		
0	(0,1)	0	L
		1	H
1	(1,0)	0	H
		1	L
x	(1,1)	0	H
		1	H

Figure 2.7: 4T/2J TCAM cell structure and its storage states.

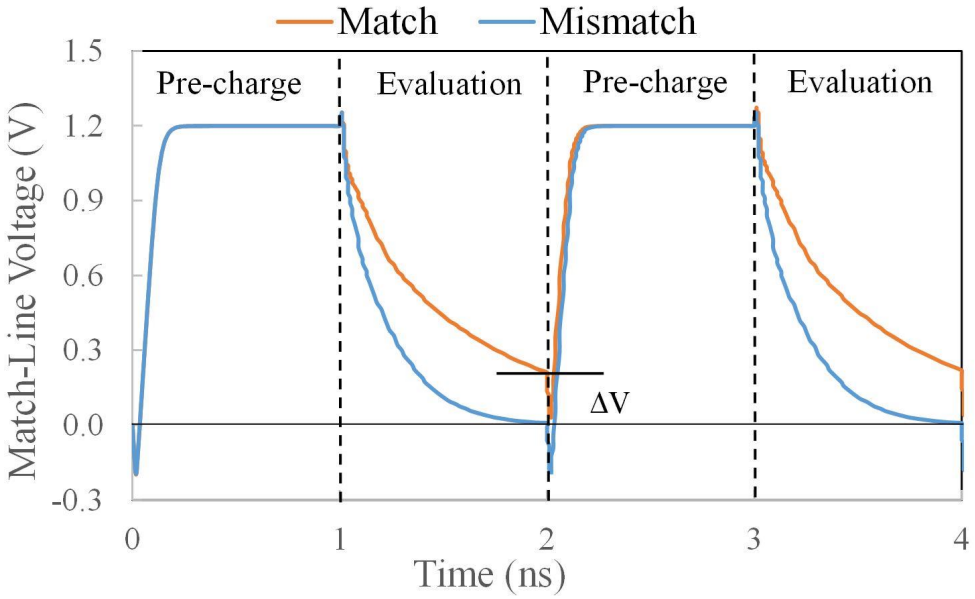


Figure 2.8: Operation phases and sample IV curve of NVM/MOS based non-volatile TCAM.

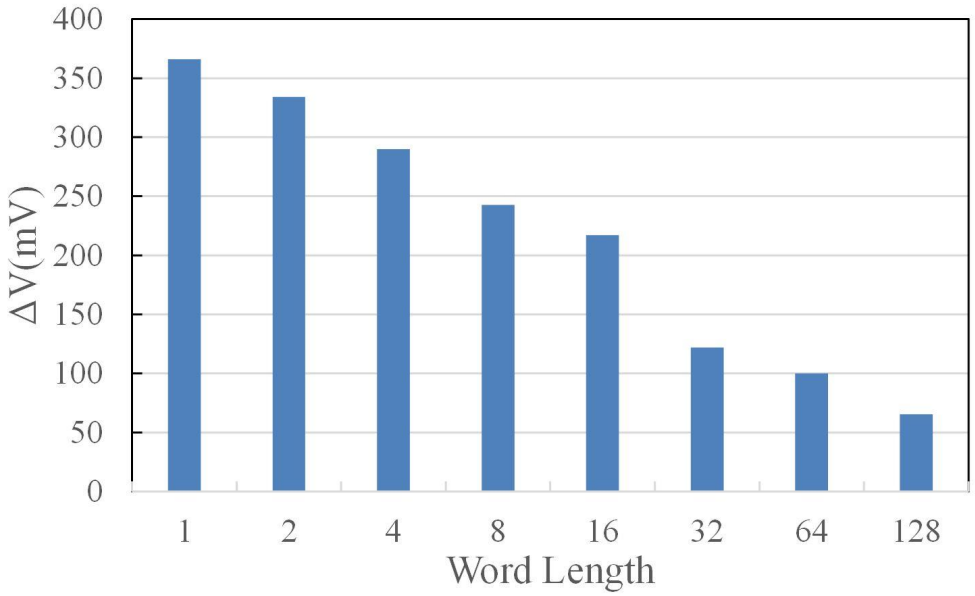


Figure 2.9: Match-line voltage difference between match and mismatch states for different word length.

states of the MTJs, determines the final value of the match-line voltage. In evaluation phase: If the search data and the stored data are equal, the discharge of the match-line is relatively slow so that the match-line keeps at a relatively high voltage level (Fig. 2.7), representing a match. On the contrary, if the search data and the stored data are not equal, the discharge of the match-line is relatively fast, pulling down to a low voltage level which represents a mismatch.

Fig. 2.9 shows that the 4T/2J cell structure can maintain a maximum match-line voltage difference of 0.1V or higher between match and mismatch states for a word length up to 128-bit. 2T/2J cell structure, however, cannot provide a ΔV of 0.1V even for a 4-bit word [15].

We note that the reliability of the TCAM design with MTJ is sensitive to the ratio between the high (R_{high}) and the low resistances (R_{low}) of the MTJ. In magnetic device research, we normally use tunneling magneto-resistance (TMR) ratio to measure the difference between the two resistance states of a MTJ, i.e., $TMR = (R_{high} - R_{low})/R_{low}$. As shown in Fig. 2.10, increasing the TMR from 200 to 400 (or a R_{high}/R_{low} ratio of 5) doubles

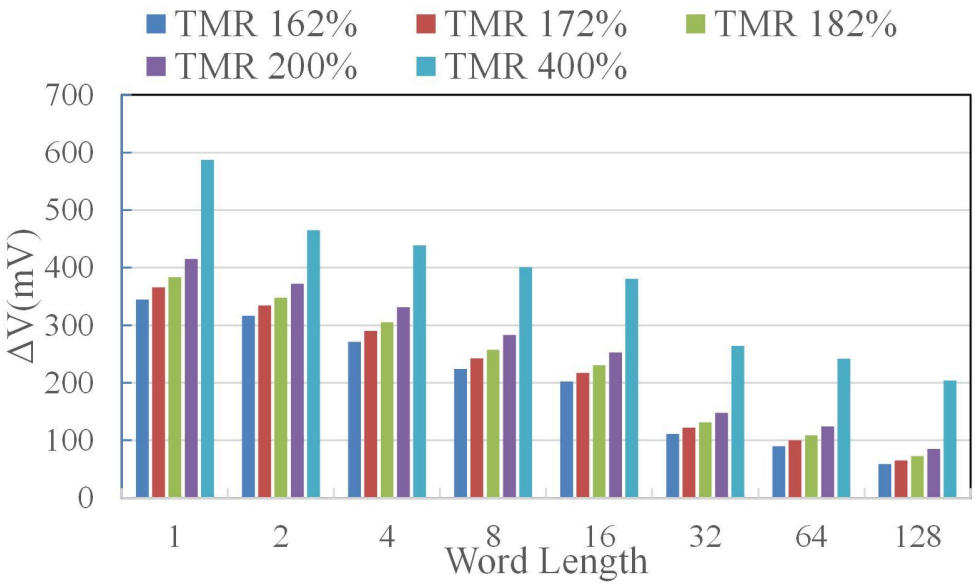


Figure 2.10: TMR dependency of match-line voltage difference between match and mismatch states.

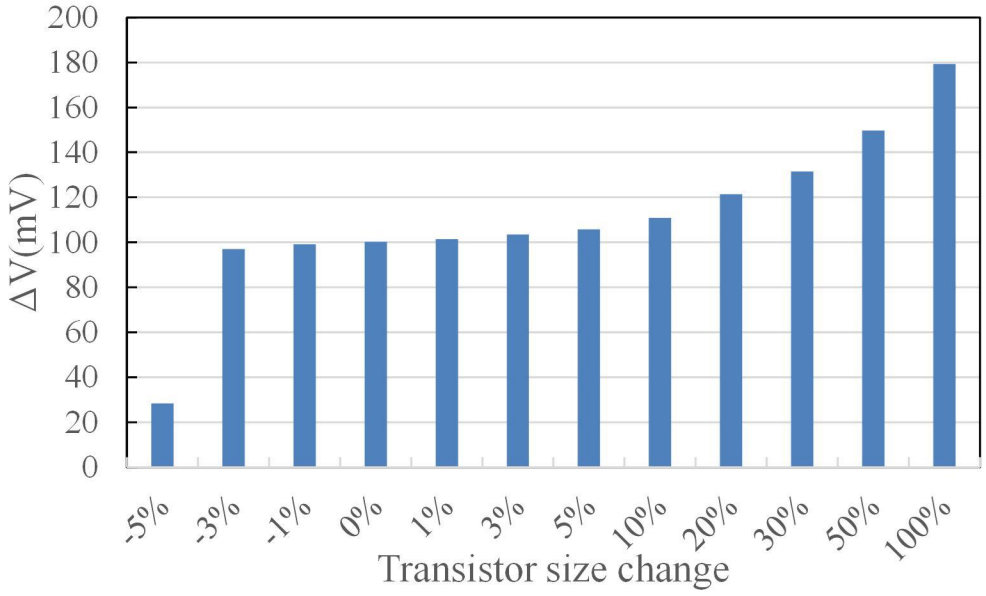


Figure 2.11: Effect of increasing transistor sizes on match-line voltage difference between match and mismatch states for 64 bit word. (Nominal transistor sizes: N1-N2=670nm, N3=350nm, P1=537nm.)

the match-line voltage difference between match and mismatch states, implying a possibly longer functional word length and/or a higher access performance in the design. Fig. 2.11 shows that increasing the sizes of the transistors (i.e., N1, N2, N3 and P1 in Fig. 2.7) in the TCAM cell can also enhance the match-line voltage difference between match and mismatch states. However, it will result in a larger TCAM cell area and degrades the design density.

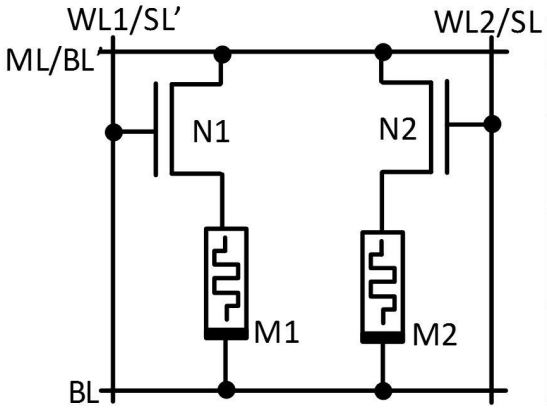
Even though the 4T/2J design in [15] can offer higher than 0.1V match-line voltage difference between match and mismatch states, some drawbacks exist in the design: the first issue is that the nominal transistor size must be in-creased as the word length increases to achieve the required match-line voltage difference. Hence, the cell size is considerably large for a word length of 128-bit (i.e., N1,N2=1.44 μm, N3=720nm, P1=1.1 μm). The second issue is that the design is very sensitive to some parameters such as PMOS transistor size and the applied bias voltage. For example, if the applied bias voltage increases by 5% then the match-line voltage difference reduces about 20%, say, from 97.74mV to 78.69mV, implying the vulnerability of the design against process variations.

2.3.2 2T/2M Nonvolatile TCAM cell

As a realistic TMR of MTJ devices is normally limited under 160%, we propose to design TCAM cell with memristors, which can have a R_{high}/R_{low} up to $1000x$ [25, 1, 19, 7] . Fig. 2.12 shows the designed 2T/2M nonvolatile TCAM cell and its logic states. The cell consists of only two select transistors (N1-N2) and two memristors (M1-M2), sharing the similar topology with the 2T/2J bit-serial TCAM proposed in [15] .

The memristors can be programmed by asserting their word-line signal and applying appropriate voltages to their bit-lines (BL and BL). The value being programmed is determined by the applied voltage direction (BL=0/BL'=1 or BL=1/BL=0). Similar to Section III.B, we implemented the 2T/2M TCAM cell with SMIC 65nm library.

As expected, the high R_{high}/R_{low} ratio of memristor helps to achieve a high match-line voltage difference between match and mismatch states for long word length, i.e., 0.1V, which is considered as the standard sensing margin that can be efficiently detected by a sense amplifier at high speed [14] . Fig. 2.13 shows the match-line voltage difference as the



Stored Data		SL	ML
M	M1,M2		
0	(0,1)	0	H
		1	L
1	(1,0)	0	L
		1	H
x	(0,0)	0	H
		1	H

Figure 2.12: 2T/2M cell structure and storage states.

word length varies. Here the select transistor size is set to 120nm. The resistance range of the memristors is between 1KΩ and 1MΩ, or a R_{high}/R_{low} ratio of 1000. The working voltage is set to 1.2V. Simulation results show that a 0.1V match-line voltage difference can be maintained in the 2T/2M design even the word length is beyond 256-bit. Note that the used transistor size is much smaller than that used in 4T/2J design (e.g., N1,N2=670nm, N3=350nm, P1=537nm for a 64-bit word length).

Fig. 2.14 shows that as R_{high}/R_{low} ratio decreases, the match-line voltage difference between match and mismatch states also decreases. The voltage swing drops below 0.1V when the ratio is under $600x$ for a word length of 64-bit. We note that the memristance (resistance) of the memristor is determined by the amplitude and pulse width of the programming voltage/current. Hence, we may trade the sense margin and consequently, operating speed of the 2T/2M TCAM cell for programming energy consumption saving. Similar to any other nano-scale devices, a memristor also suffer from process variations, which affect its resistance level as well as the R_{high}/R_{low} ratio. For instance, in Fig. 2.14, the TCAM cell can operate with a match-line voltage difference of 0.1V between match and mismatch states in a wide range of R_{high}/R_{low}, say, $600-1000x$. However, we could not draw the conclusion that the resilience of the 2T/2M TCAM cell to process variations is better than the TCAM design with MTJ because the R_{high}/R_{low} ratio is very sensitive to the variability of the lowest

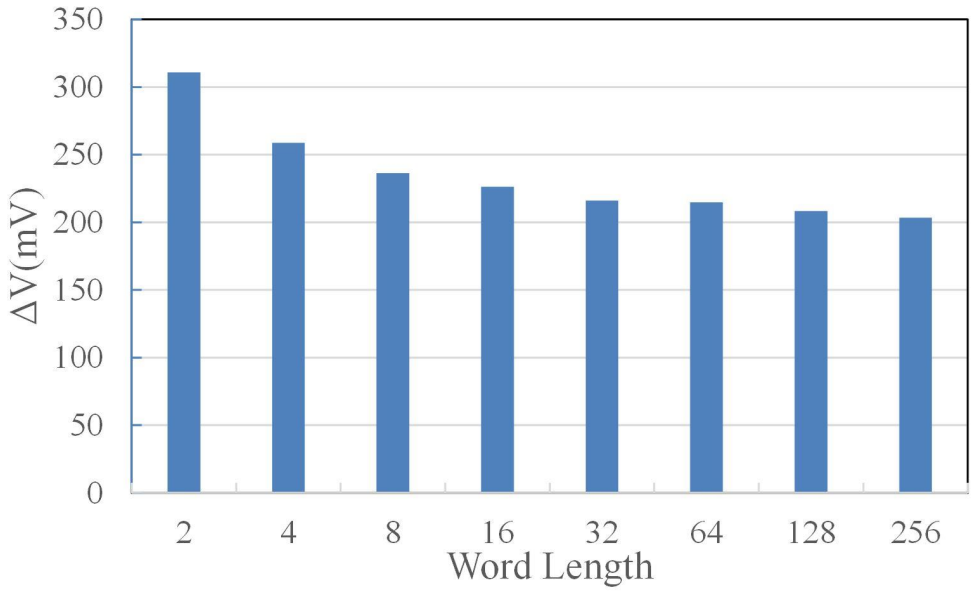


Figure 2.13: Match-line voltage difference between match and mismatch states for increasing word length. (N1,N2=120nm.)

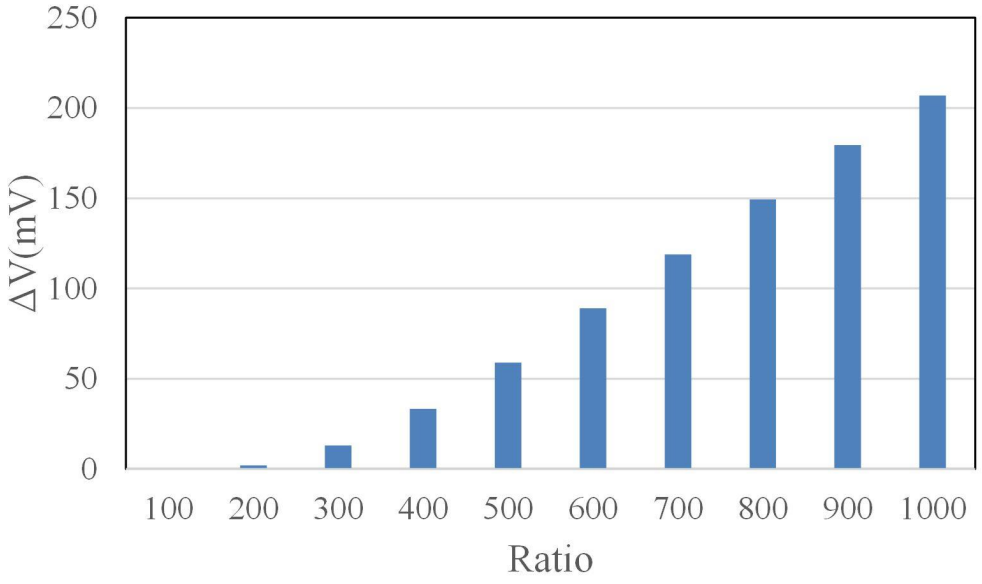


Figure 2.14: Match-line voltage difference between match and mismatch states at different R_{high}/R_{low} ratios for a word length of 64-bit. Select transistor (N1,N2) size = 120nm.

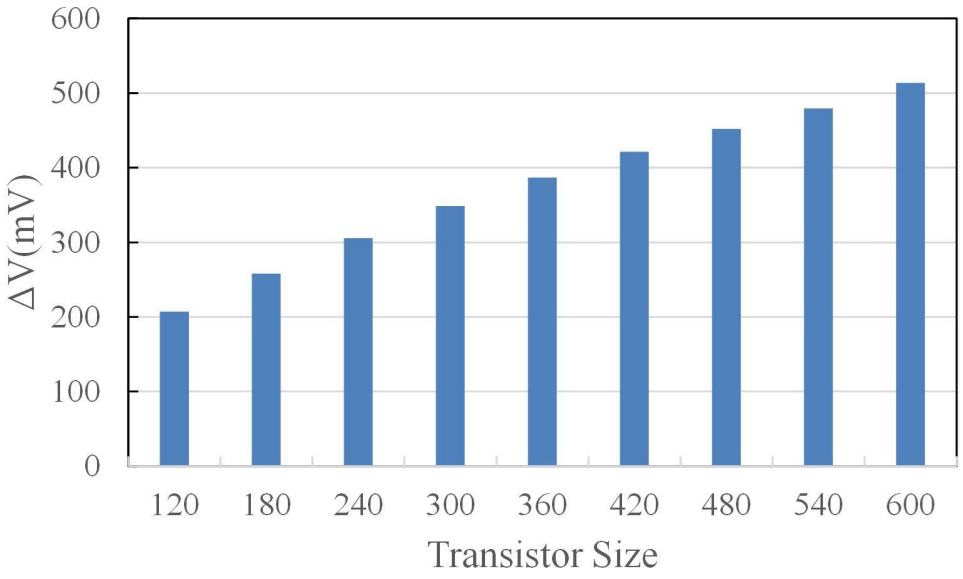
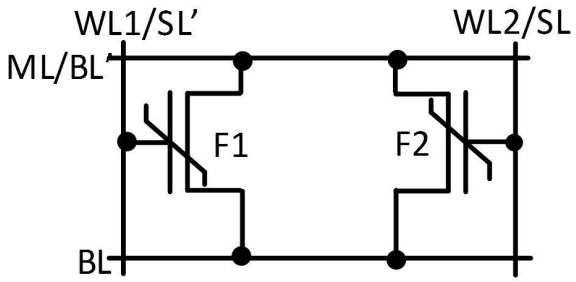


Figure 2.15: Effect of increasing select transistor size (N1,N2) on the match-line voltage difference between match and mismatch states for a 64-bit word length.

and the highest resistance levels of the memristor device. Memristor variation robustness research is not in scope of this work.

We note that increasing the size of select transistors (N1-N2) can also improve the match-line voltage difference between match and mismatch states. Fig. 2.15 shows that the match-line voltage difference between match and mismatch states for a 64-bit word length improves almost proportionally as the select transistor width increases, indicating a possible approach to enhance the robustness of the TCAM cell by scarifying the design density.

In design optimization of the 2T/2M TCAM cell, the impacts of each design parameters entangle with each other. For example, when the select transistor width changes from 120nm to 480nm, the minimum required R_{high}/R_{low} ratio ensuring a match-line voltage of 0.2V reduces from 1000 to 300 for a word length of 64-bit. It results in a higher tolerance to the memristor resistance variations. It also lowers programming energy consumption and improves programming speed of the memristors.



Stored Data		SL	ML
M	F1,F2		
0	(1,0)	1	H
		0	L
1	(0,1)	1	L
		0	H
x	(0,0)	0	H
		1	H

Figure 2.16: 2F TCAM cell structure and storage states.

2.3.3 2F Nonvolatile TCAM cell

Compared to 4T/2J design with MTJ, the enhancement on the match-line voltage difference between match and mismatch states in 2T/2M design with memristor is mainly from the improvement of R_{high}/R_{low} ratio of the nonvolatile devices. Nonetheless, these two designs all rely on the difference between the voltage decay rates corresponding to the different resistances of the nonvolatile devices. Such match-line discharging based mechanism limits the speed of the TCAM design. In this section, we propose a 2FeFET (2F) nonvolatile TCAM cell design which has not only a very simple structure but also a high operating speed.

Fig. 2.16 shows the schematic of the proposed 2F TCAM cell design. Write operations can be performed by applying the appropriate voltages to the bit line pair (BL and BL') and the word line (WL) of the targeted FeFET. The original data and its complement are stored in FeFETs F1 and F2, respectively. The search operation also includes two phases. Without loss of generality, here we assume the stored data is logic 1 where the F1 is programmed at a high threshold voltage while the F2 is programmed at a low threshold voltage. During pre-charge phase, the match-line is still charged to a high voltage level; during evaluation phase, a high search voltage level corresponding to logic 1 is applied on word-line WL1 while its complement a low voltage level (usually 0V) corresponding logic 0 is applied on word-line

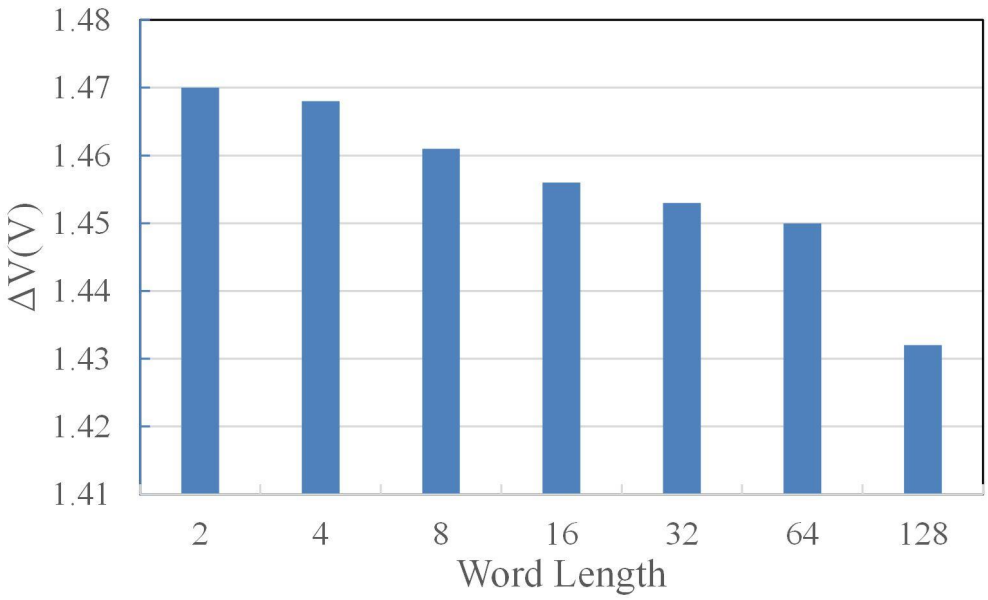


Figure 2.17: Match-line voltage difference between match and mismatch states for increasing word length. (F1,F2=120nm)

WL2. Since the threshold voltage of the F1 is high, the read voltage level could not turn on the F1 while the F2 is naturally turned off. As a result, the FeFETs in the TCAM cell are all in cut-off state and the pre-charged voltage on the match-line is retained. On the contrary, if the polarization of the search voltage level is opposite, say, a logic ‘0 is applied to WL1 and a logic 1 is applied to WL2, the F2 will be turned on and the match-line will be pulled down to ground.

The application of FeFETs in the TCAM cell offers not only a very high match-line voltage difference between match and mismatch states but also negligible performance degradation when the word length increases. As shown in Fig. 2.17, the match-line voltage difference between match and mismatch states is very close to the normal ope-rating voltage and degrades very slightly when the word length increases.

2.4 DISCUSSION ON EMPLOYING FEFET

As analyzed in Section III.B, the low R_{high}/R_{low} ratio of MTJ makes it very hard to design a TCAM cell with high sense margin and fast access performance. To compensate this drawback, more transistors (compared to the design with other types of nonvolatile devices) are introduced and large size transistors are often needed, resulting in a large cell area. Finally, the resistance switching of MTJ devices is indeed a stochastic process, which may cause some new reliability issues such as write failures and read disturbance [6] .

As an alternative to MTJ, memristor has a R_{high}/R_{low} ratio up to 1000 [7] . Such a high ratio greatly helps to improve the sense margin and access performance of the TCAM cell and minimize the required transistor sizes. However, memristors generally have a large R_{high}/R_{low} variation caused by manufacturing variability or even programming voltage fluctuations [25] . Hence, we may not conclude that the memristor-based design is better than the MTJ-based design before performing the detailed robustness evaluation using the realistic data of memristor and MTJ device variations.

A general design challenge in the TCAM design based on resistive device is the parallel connection between cells, which significantly reduces the effective resistance connected between match-line and ground. As a result, it reduces the distinction of the match-line voltage decay rates between match and mismatch states and degrades the corresponding match-line voltage difference. However, the application of FeFET solves this issue by introducing logic operation to perform the evaluation phase. The corresponding TCAM cell structure is also extremely simple only two transistors so as to allow a very high integration density [9] . Moreover, the amplitude of the match-line voltage difference between match and mismatch states is insensitive to the variation of the word length. However, as the retention time of FeFETs is often limited, the memory cell may require refreshing periodically [1]. The recent research shows that the MFIS structure of FeFET can retain data for 17 days in room temperature [19], which may satisfy the majority of normal computing applications. Also, FeFET requires relatively higher programming voltage than MTJ and memristor [19] . It may increase the write power consumption of the TCAM cell and require the costly on-chip charge pump.

2.5 EVALUATIONS OF EMPLOYING FEFET IN TCAM STRUCTURE

In this section 2F TCAM design's robustness against the variations, finding an optimal programing level based considering the memory window decay of FeFETs, the energy consumption and possible energy savings and finally search speed are examined and discussed. The proposed 2F nonvolatile TCAM design is implemented and evaluated using Cadence Virtuoso environment.

2.5.1 Finding Optimal Programming level

As discussed in the previous sections FeFET device can provide very high match-line voltage swing, when it is employed in simple 2F TCAM structure, thanks to its device structure. Employing ferroelectric layer in the gate stack changes the channel conductivity of the FeFET and provides large sensing margins between FeFET's memory states. Also as it is stated in the previous sections, a FeFET device can also be programmed to intermediate polarization levels. Hence a FeFET cell can be programmed either to full polarization level to provide the largest sensing margin for best performance or it can be programmed to a lower polarization level to provide possible energy savings and FeFET device endurance improvements while still providing sufficient sensing margin for proper TCAM function.

Since 2F design can provide very large sensing margin even for longer word lengths, one can find an optimal programming level to reduce programming energy consumption and improve the endurance of FeFET memory cell. To find an appropriate intermediate memory window, 2F TCAM with 128bit word length is evaluated.

Fig. 2.18 shows the simulation results for decreasing memory window on a 2F design 128bit word TCAM. Y axis indicates the Match-line voltage difference between match and mismatch states while X axis shows the memory window decay. As seen in the Fig. 2.18 2F design can sustain a large sensing margin until the the memory window reduces to %25. However that low memory window may suffer from errors due to device variations. To provide sufficient sensing margin range we choose the %41.6 which can provide 0.5mV sensing margin. This value is large enough for wide range of device variations and provides significant write energy savings.

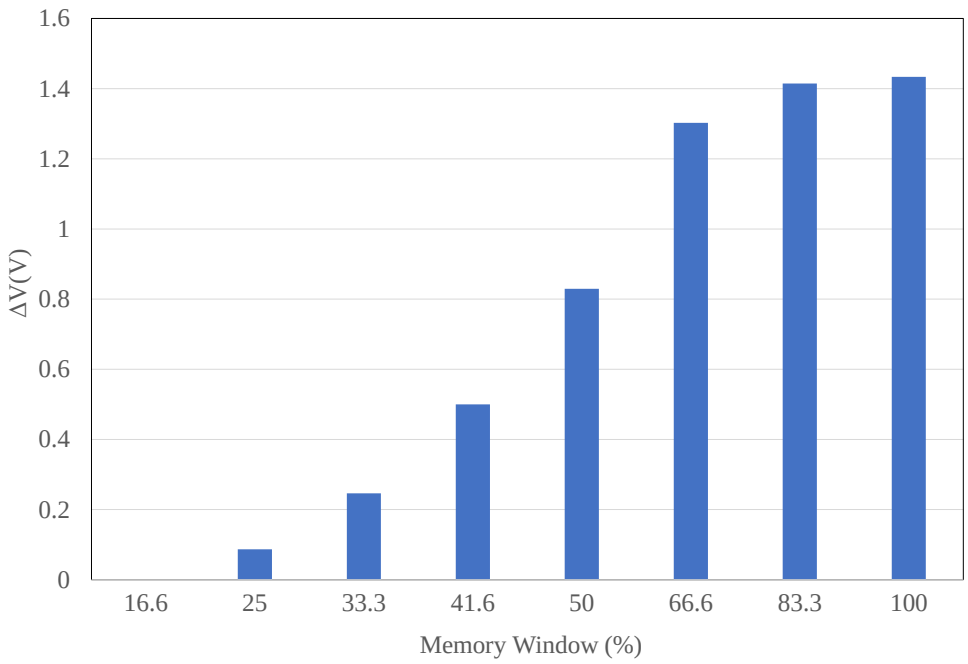


Figure 2.18: Match-line voltage difference between match and mismatch states for changing memory window levels for 128bit word lengtht.

2.5.2 Impact of Variations on Sensing Margin

Device variations can impact the performance of the TCAM significantly even it can make a memory row to provide wrong output. Previous sections discussed up to some level of variations due to transistor sizing during CMOS process. This sections focuses on 2F TCAM cell device variations.

Fig. 2.19 shows the memory window distribution under %10 device variation. Model created based on the [18] which states that the polarization variation follows a Gaussian like distribution. Such distribution can also be seen in the Fig. 2.19. As it is determined in the former section, a %41.6 of memory window is used during this evaluation. %33.3 memory window is also evaluated and it has been observed that the match-line voltage difference between match and mismatch states can drop well below 100mV which is stated as a lower limit for fast speed search operation. That indicated %33.3 memory window is not convenient to operate during a high speed parallel search.

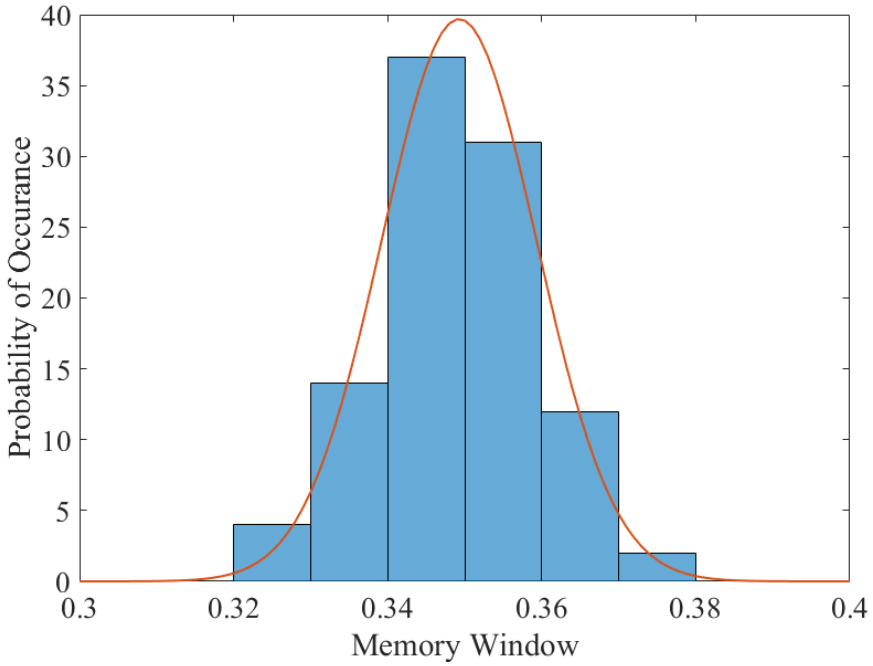


Figure 2.19: Memory window (mV) variation distribution.

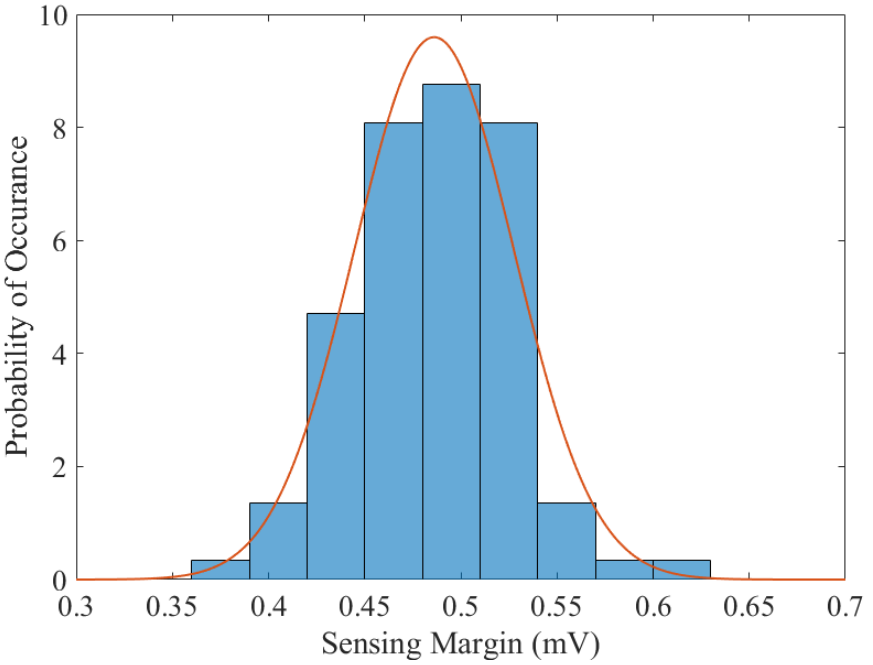


Figure 2.20: Match-line voltage difference between match and mismatch states distribution under %10 variation.

As expected the device variation has a significant effect on the match-line voltage swing as shown in the Fig. 2.20. Fig. 2.20 also proves that that as the memory window stays larger 2F design is more robust against the device variations. However to save energy %41.6 memory window is determined as a safe below limit since %33.3 memory window level can't provide large sensing margin in some situations. Fig. 2.20 indicates that the 2F design can maintain a large enough margin even for the worst case with %41.6 memory window (300mV at lowest level) which follows the Gaussian distribution.

2.5.3 Search Speed

The most significant performance metric is the search speed for the TCAM's operation performance. Search speed is determined by the time to reach appropriate sensing margin during read operation. Sensing speed's of the proposed schemes are compared with the baseline of MTJ based cell design. Fig. 2.21 shows the simulation results. A world length of 128 bit under the worst case condition ,which is the single data bit miss leads to minimal match-line discharging through a single transistor, used to obtain the sensing margin converge period.

MTJ based design is the slowest due to low R_{high}/R_{low} and very sensitive transistor size dependency. It can reach its full match-line voltage swing, which is lower than a 100mV, in close to 3ns. Memristor thanks to its high R_{high}/R_{low} ratio and simpler cell design, reduces the sensing speed below 1ns. Memristor based design can provide 200mV sensing margin within less than 1ns time period. FeFET based design greatly improves the performance by eliminating resistive elements and reduces the sensing speed under 0.1ns. 2F design can reach 200mV sensing margin within 0.1ns time period and it can reach its full match-line voltage swing in 1ns.

To further examine the FeFET based design, the effect of increasing word length and best-worst case operation conditions on sensing speed is also examined. Fig. 2.22 shows the simulation results for 128bit 2F TCAM. The worst case indicates the condition of only one bit mismatch between the stored and search data which leads to a slow discharge through a single transistor of the match-line. Best case indicates a full mismatch between search and

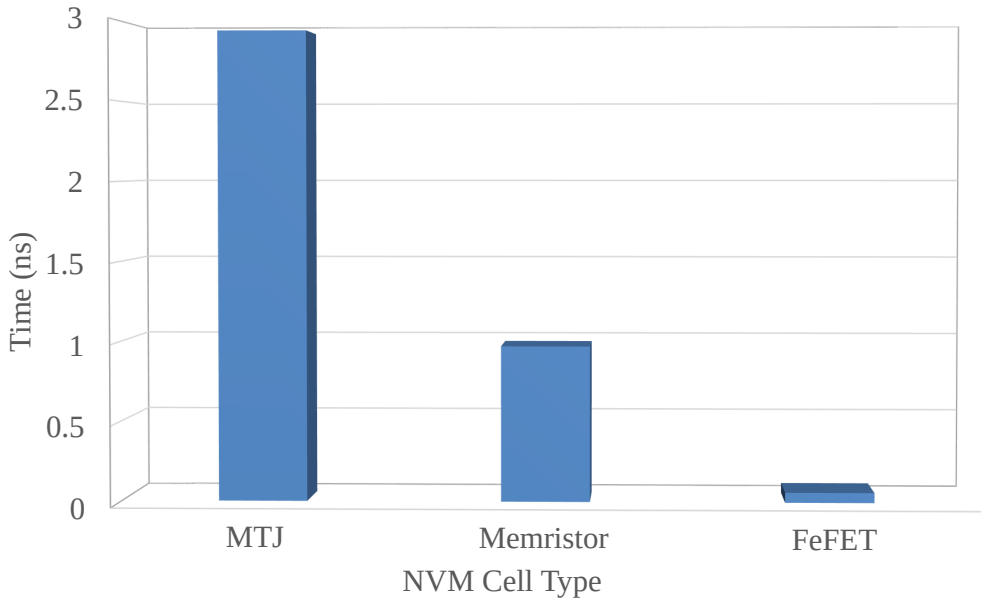


Figure 2.21: Sensing speed of 128 bit word during a single search operation.

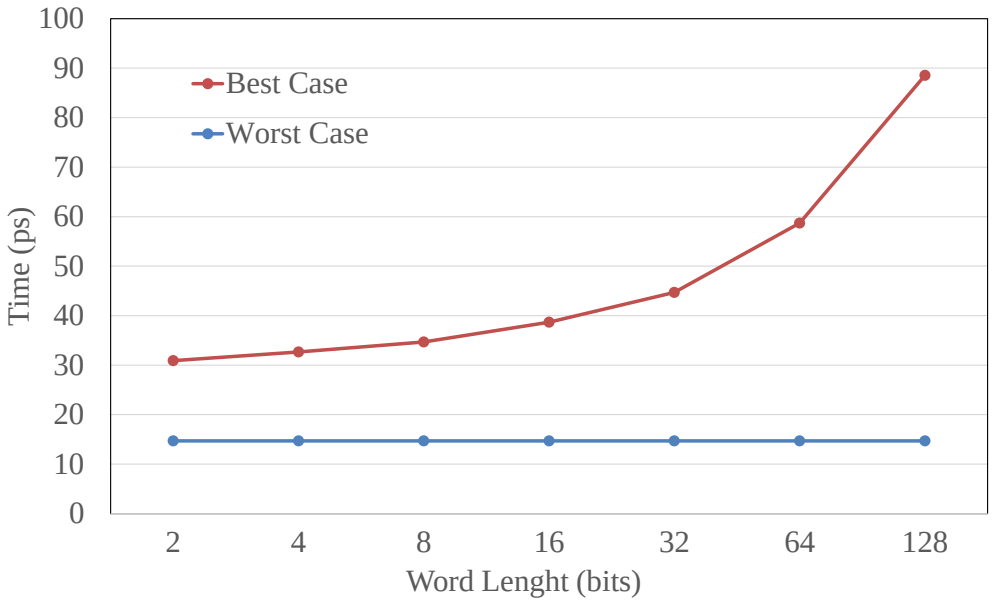


Figure 2.22: Effect of increasing word length on sensing speed for best and worst case match-mismatch conditions.

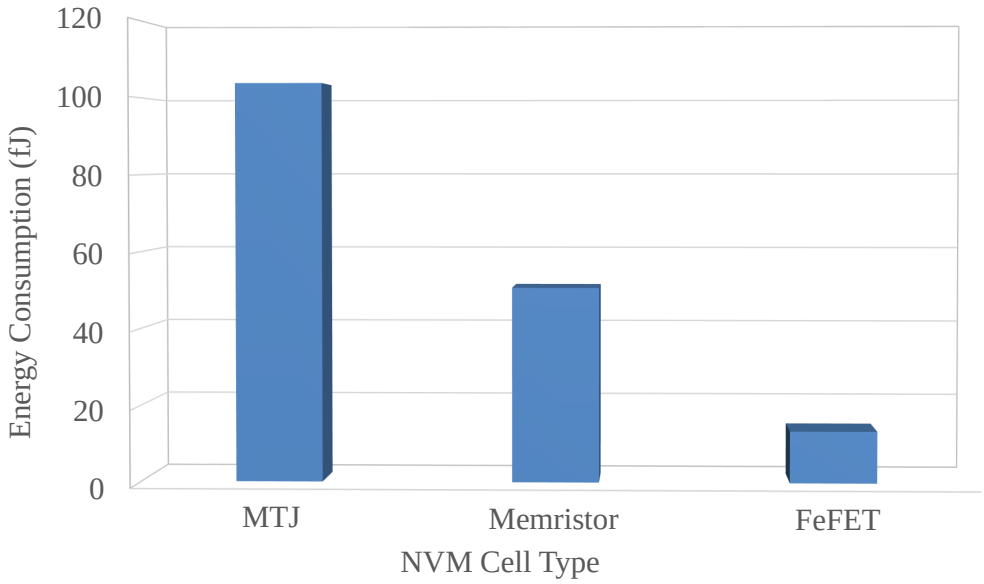


Figure 2.23: Per bit energy consumption of 128 bit word during a single search operation.

stored data which leads to a fast discharge through all transistors of the match-line voltage. Simulation results shows that sensing speed is not effected by the increased word length in the best case condition. Sensing speed reduces with the increasing word length under the worst case condition but it still says significantly faster than compared cell designs.

2.5.4 Energy Consumption

Reducing the energy consumption is the main objective of employing NVM based devices in TCAM designs. Non-volatile emerging devices can be turned-off when not in use and provide zero stand-by energy consumption compared to the conventional CMOS based TCAM design which leads to significant and obvious energy savings. The simplified cell structures of NVM based TCAMs also leads to a reduced number of transistors which translates into further energy savings during search operation. Fig. 2.23 shows per-bit energy consumption during a single row read operation of a 128 bit word TCAM. As shown in the results the proposed memristor based TCAM design provides significant power savings over MTJ based TCAM. FeFET based 2F TCAM design further improves the energy savings thanks to its simplified cell structure.

As an addition to the search energy consumption reduction, further energy saving can be provided by reducing the programing level of FeFET cells in a 2F TCAM design. As discussed in the former section we determined that it safe to program the FeFETs in 2F cell to %41.6 of full memory window. Such programming scheme indicates an average (writing logic '1' and logic '0' to FeFET requires different write voltage levels) %56.25 power saving during a write operation on 2F TCAM.

2.6 CONCLUSION

In this chapter I demonstrated three different hybrid eNVM/CMOS based nonvolatile TCAM cell designs built on MTJ, memristor and FeFET devices, respectively. The low R_{high}/R_{low} that the MTJ can provide makes it the least favorable candidate for nonvolatile TCAM design. Memristor, which can provide very high resistance ratio, can be used as a replacement to MTJ. Compared to resistive NVMs, the FeFET virtually removes the parallel connection problem of resistive memories and provides a very simple but promising solution to NVM based TCAM design.

3.0 ADAPTIVE REFRESHING AND READ VOLTAGE CONTROL SCHEME FOR FEDRAM

3.1 INTRODUCTION

When CMOS technology scaling approaches its physical limit, emerging nonvolatile memories, such as ferroelectric dynamic random access memory (FeDRAM), magnetic memory (MRAM) [4], resistive memory (RRAM), phase change memory (PCM), etc., start to draw attentions from both academia and industry. The attractive features of these technologies include nanosecond access time, small footprint, non-volatility (i.e., zero standby power), etc. In particular, the key device of FeDRAM – Ferroelectric field effect transistor (FeFET) has the structure similar to that of conventional CMOS transistor (except for a modified gate stack). Hence, FeFET promises excellent CMOS integration compatibility and manufacturing scalability [23]. Very recently, the discovery of HfO2 dielectric enables a ultra thin (e.g., 8nm [23]) layer, paving the long-term scaling path of FeFET device.

In a FeDRAM cell, the data is stored as the polarization state of the ferroelectric layer in the FeFET and reflected as the changeable channel conductivity [23]. In write operations, the polarization state can be programmed to different levels by applying appropriate voltages to the three terminals of the FeFET [23]. The data retention time of the FeDRAM cell, which denotes the time period for which the data can be retained without power supply, is mainly determined by two mechanisms: the depolarization field and the leakage [21]. Programming the polarization state to the maximum level can provide the longest retention time. The required high programming voltage, however, will cause severe endurance degradation of the memory cell [23, 16]. Here the endurance denotes the cycles for which the memory cell can be programmed.

Note that many applications of FeDRAM requires a high endurance cycle, e.g., main memory with frequent write accesses. In such a case, we can trade the non-volatility of the FeDRAM cell for its endurance, e.g., the polarization of the ferroelectric layer will not be programmed to its maximum level [10] so that the FeFET can be programmed for more times. However, such partial programming scheme reduces the memory window, shortens the data retention time, and harms the reliability of read operations. During system executions, the memory window of the FeDRAM cell can be also disturbed by the write and read accesses to its associated memory array. A refreshing scheme with often high energy cost must be implemented to prevent data loss and assure read robustness.

In this chapter, I proposed an adaptive refreshing scheme which tracks the degradation of memory window over time and refreshes the FeDRAM cell only when necessary. Proposed scheme reduces the number of refresh operations as well as the incurred energy consumption, and prolong the lifetime of the FeDRAM cells. The reduction of the refresh operations in the new scheme leads to significant energy saving and lifetime prolonging of the FeDRAM. The adaptive reference circuit can also be utilized to track the impact of read and write disturbances on the FeDRAM cells. The read reliability of the FeDRAM cells is then improved by adaptively adjusting the read voltage to overcome the degradation of the memory window. Finally, several programming schemes of the FeDRAM array with different write disturbance impacts are also discussed.

3.2 PRELIMINARY

3.2.1 FeFET Basics

The only difference between a FeFET and a conventional MOSFET is the additional ferroelectric layer placed between the metal and the insulator layers at the gate stack of the FeFET, as shown in Fig. 3.1(a) [10]. Data is stored by programming the ferroelectric layer to different polarization states. During write operations, an electrical field higher than a certain threshold (i.e., coercive field [10]), is applied. The direction of the electrical field determines the polarization of the ferroelectric layer, say, the final position of the dipoles. Fig. 3.1(a) and

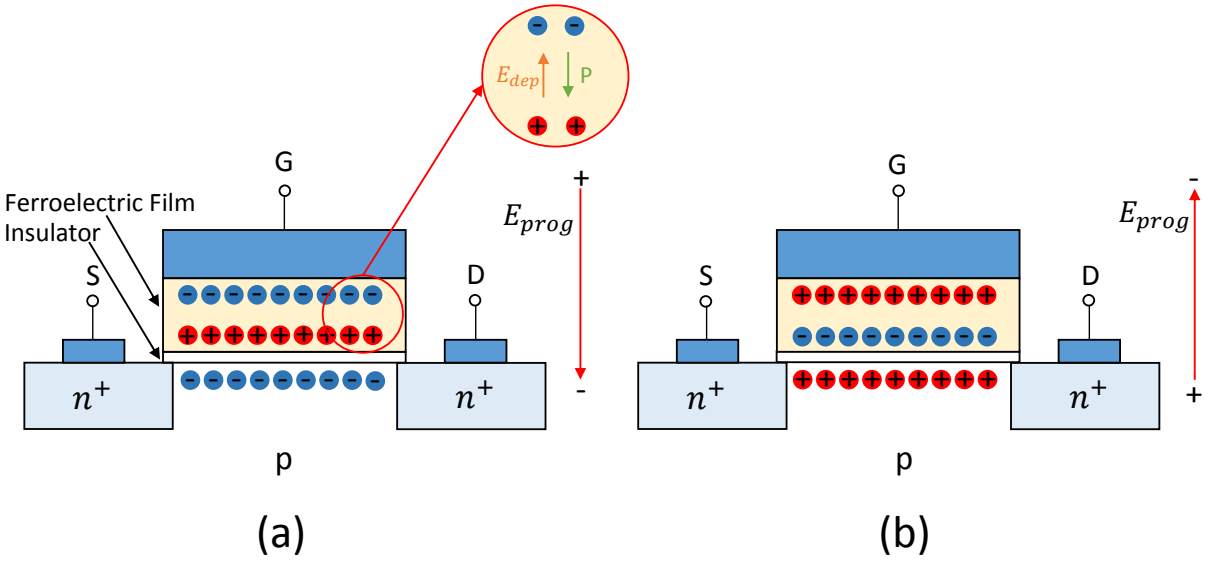


Figure 3.1: FeFET cell view when programming (a) logic '1' and (b) logic '0'.

(b) shows the biasing conditions when programming logic '1' and '0', respectively. Remnant polarization is used to denote the remaining polarization after the programming voltages are removed from all the terminals of the FeFET [10]. The two opposite remnant polarization states creates two different V_{th} states corresponding to logic '1' and '0'. Due to the V_{th} difference, the FeFET requires specific gate voltages based on its logic state to turn on, and the difference between those gate voltages is called *memory window* [23].

The polarization of the ferroelectric layer determines the conductivity of the FeFET, which represents the data stored on the FeFET. Since the body of the FeFET is p-type, the more negative ions are attracted to the FeFET's channel, the higher conductivity will be demonstrated when the same gate voltage is applied. Although there exist the maximum levels that the two polarization states of the ferroelectric layer can reach, the ferroelectric layer can be flexibly programmed to any intermediate polarization level between these two maximum levels through adjusting the amplitude and/or the pulse width of the programming voltage [10]. However, partial programming of the FeFET leads to the degradation of data retention time.

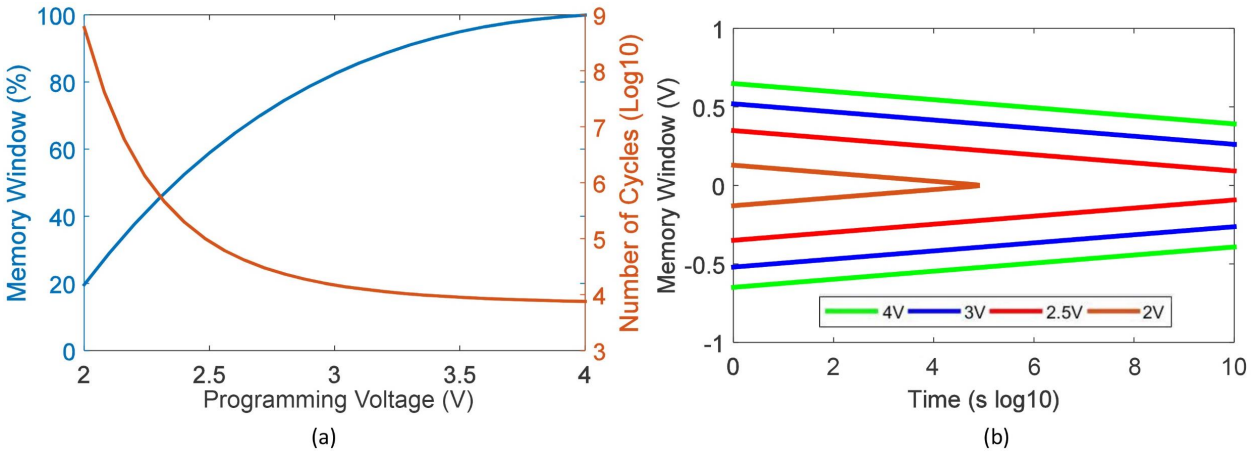


Figure 3.2: (a) Memory window, endurance relation based on the programming voltage (this graph obtained from [23, 28]) (b) Changing memory window in time for different programming voltages (linear MW loss model from [21])

3.2.2 Non-volatility vs. Endurance

The retention time of a FeFET is mainly affected by two factors:

1) The depolarization field E_{dep}, which is generated by the electric dipoles in the ferroelectric layer and negates the ferroelectric polarization [21].

2) The leakage current [21], which degrades the insulating ability of the insulation layer and causes the charge trapping in the ferroelectric layer [21, 28].

After the programming, if there are no compensation charges supplied to the ferroelectric surfaces, the polarization level will continuously degrade. The effect of depolarization field takes place in the time range much shorter than that of the leakage current, and is responsible for short-term memory loss [21].

The endurance of a FeFET is mainly determined by the number of polarization switching cycles and the programming voltage amplitude [21, 16]. As the programming cycle number increases, the memory window of the FeFET continues shrinking, causing the degradations of read reliability [23]. Also, a high programming voltage accelerates the insulation layer breakdown and charge trapping process in the dielectric stack of the FeFET [28]. Lowering

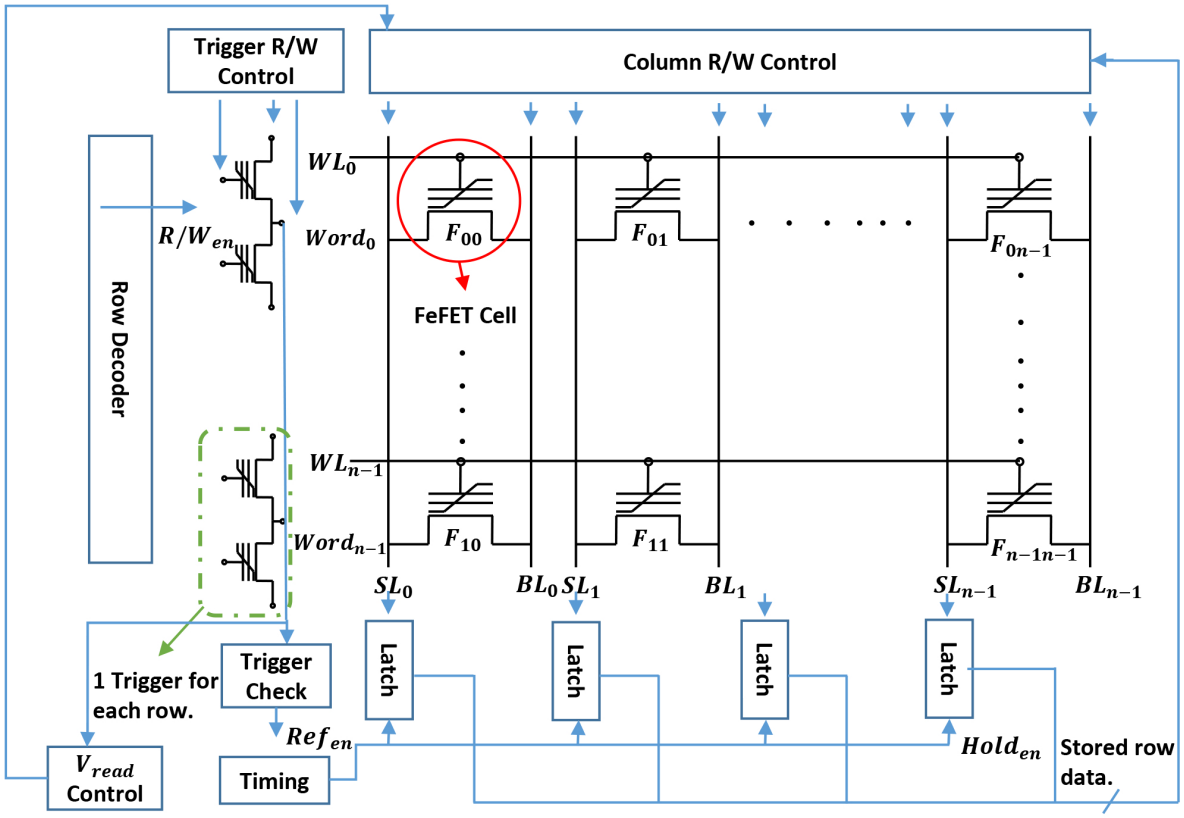


Figure 3.3: FeDRAM with "AND" type array structure.

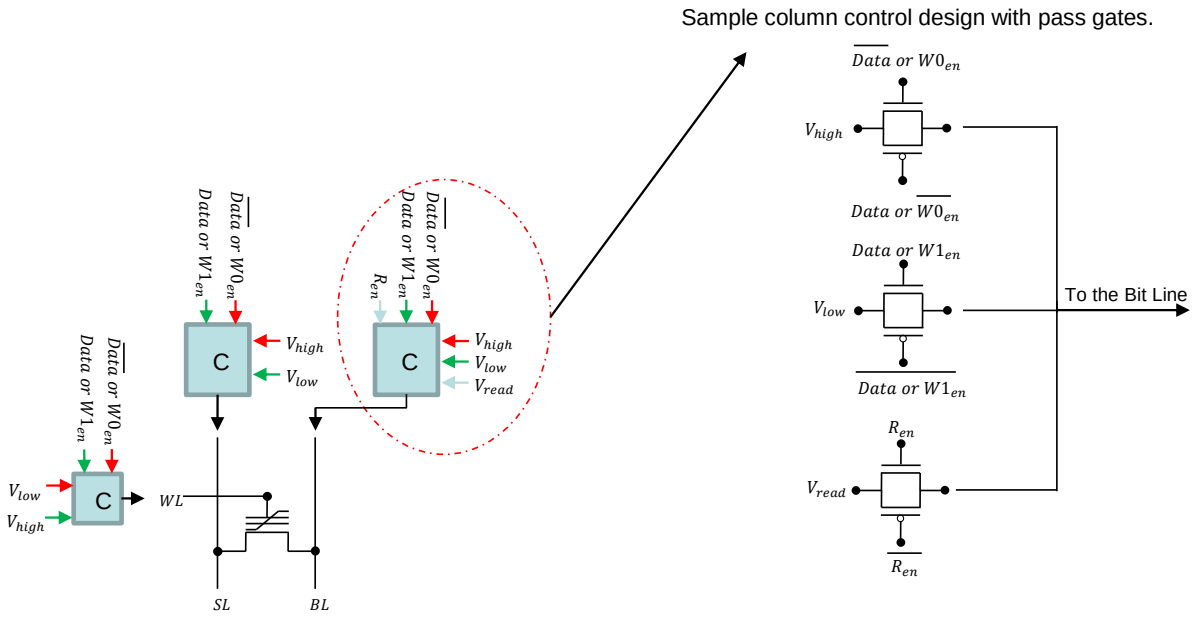


Figure 3.4: Column read-write control circuit design for FeDRAM array.

the programming voltage can improve the FeFET endurance, however, followed by a reduced memory window.

As shown in Fig. 3.2(a), when the programming voltage reduces from 4V to 2V, the ferroelectric layer's endurance improves from 10^4 cycles to 10^9 cycles while the memory window decreases from 100% to 20% of the maximum level. At the same time, we completely lose the memory window as short as $10^5 seconds$, as illustrated in Fig. 3.2(b). The memory window must be sufficiently large to sense the difference between the two logic states of the FeFET.

3.2.3 Array Structure and Read-write Disturbance

A popular FeDRAM memory array design is AND structure [12], which is depicted in Fig. 3.3. The AND structure chosen in this work due to area efficiency.

During a write operation, e.g., writing '1' to cell F_0, program voltage V_{prog} is usually applied to WL_0 while BL_0 and SL_0 are connected to Gnd; On the contrary, in writing '0',

V_{prog} is applied to BL_0 and SL_0 while WL_0 is connected to Gnd. Sample column control circuitry for FeDRAM array is shown in Fig. 3.4. In such a write operation, the FeDRAM cells that are not being written will still be effected by the voltages applied on the SL, BL and WL lines shared with the cell that is being programmed.

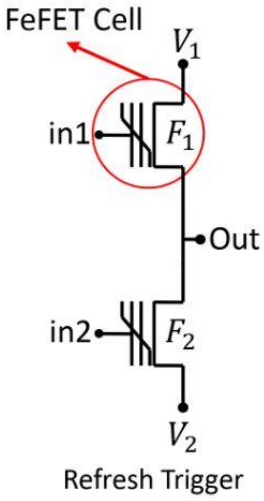
For example, as shown in Fig. 3.8, when writing '0' to cell F_0 in the first row, the unselected cells at the same column but different will observe the V_{prog} applied on their BL and SL lines too. To minimize the impact of such disturbances, the WL of cell F_0 of must be raised to a predefined voltage level, e.g., $V_{prog}/2$, to prevent the cell from being completely switched. In [17], the author also proposed a scheme that can reduce the disturbance to the unselected cells down to $V_{prog}/3$.

However, such biases asserted on the unselected cells cannot be completely eliminated during FeDRAM write operations, causing inevitable disturbance on the stored data: For a fully programmed FeFET, data loss occurs at about 10^3 cycles for a disturbance of 1/2 V_{prog}. When the amplitude of the disturbance reduces, the lifetime of the data will improve exponentially, e.g., 10^7 cycles if the disturbance reduced to 1/3 V_{prog}.

During a read operation, a predefined read voltage is applied on the selected WL. The BLs of the cells in the selected row are connected to a sense voltage and their SLs are connected to a SenAmp for data reading. Although the read voltage is generally much smaller than the write voltage, reading a FeDRAM cell may also disturb the unselected cells in the array and slowly change the polarization states of them [3]. A refresh operation is needed to overcome the memory window shrinking and the impact of write-read disturbance.

3.3 PROPOSED DESIGNS

To overcome the high energy cost of the refreshing operations, an adaptive scheme that refreshes the FeDRAM cells only when necessary is proposed. In addition, the read voltage is adjusted on-the-fly to accommodate the polarization state change of the FeFET for read reliability enhancement.



	In1	In2	V_1	V_2
Read	V_{read}	V_{read}	V_{rref}	G_{nd}
Write '0'	V_{high}	V_{low}	V_{high}	V_{high}
Write '1'	V_{high}	V_{low}	V_{low}	V_{low}

R/W Input Conditions

Figure 3.5: Refresh trigger and input values for different operations.

3.3.1 Refreshing Trigger Logic

In proposed scheme, the refresh operation is triggered by the refreshing trigger logic (RTL), which mainly consists of two FeFETs, as shown in Fig. 3.5. The FeFETs are connected in series and linked to two predetermined biases V_1 (e.g., V_{dd}) and V_2 (e.g., Gnd), respectively.

FeFETs F_1 and F_2 are initially programmed to '1' and '0', respectively; When the same read voltages are applied on their gates (i.e., In1 and In2), F_1 and F_2 demonstrate a low and a high resistances, respectively. After being programmed, the polarization level of F_1 and F_2 start decaying. Sample read-write control circuitry for refresh trigger is shown in Fig. 3.6. The resistances of both FeFETs change accordingly, resulting the reduction of the voltage at port "Out". Once the output voltage of the RTL meets a threshold, the refreshing operation will be triggered by considering the memory window has shrunken below a safety level.

Using FeFETs to implement the RTL ensures the degradation rate of the memory window close to that of the memory cells being monitored. Hence, the RTL must be initialized when the monitored memory cells is programmed. During the initialization, appropriate biases must be applied to the terminals of F_1 and F_2, similar to the write operations of FeDRAM cells.

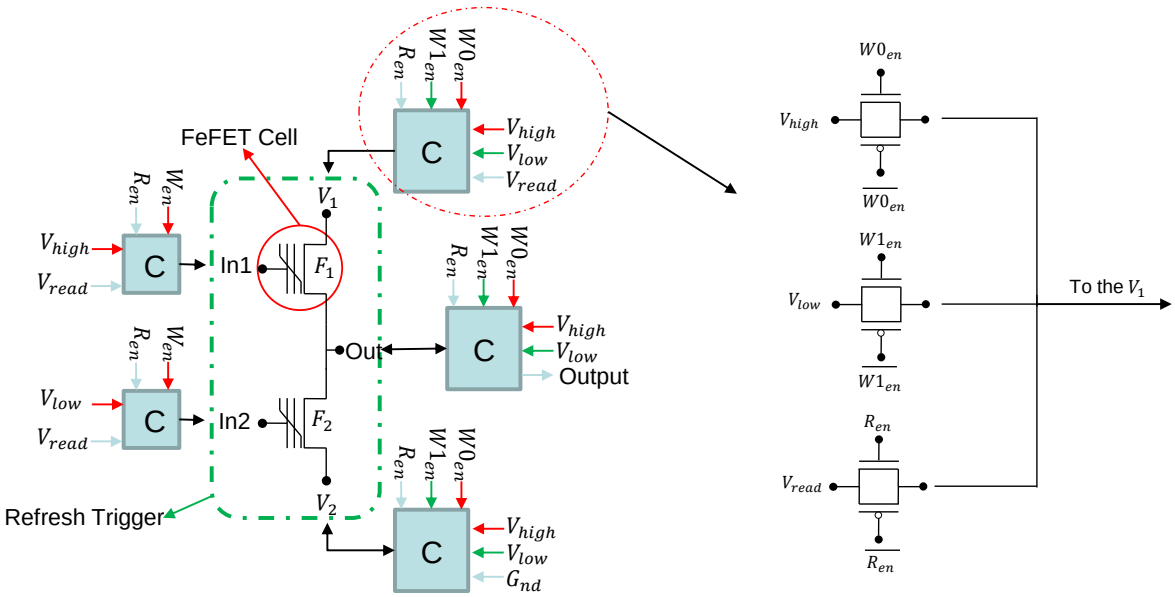


Figure 3.6: Read-write control circuit design for refresh trigger.

3.3.2 Adaptive Read Voltage Control

During the read operations of FeDRAM cells, a read voltage V_{read} is applied on the gate of the FeFET and the resistance of the FeFET is readout by sensing the current generated on the BL. Since the polarization level of the FeFET keeps reducing after being programmed, the equivalent V_{th} shift on the gate of the FeFET and the generated memory window also decrease over time.

I propose to use the output of the RTL as the guidance to adaptively adjust the read voltage for the corresponding memory row to compensate the reduction of V_{th} shift for read reliability enhancement. Note that the polarizations of the compensation voltage for logic '1' and '0' are different and during read operations, all FeRAM cells are under the same bias condition. Since it is not possible to know what is stored in each cell, we can only select one of the two logic values to compensate while the V_{th} shift of the memory cells with the other logic values is actually negatively impacted. Luckily, as I will show in Section 3.4.3, the resistance of the cells with the non-compensated logic value shifts negligibly if the compensation voltage

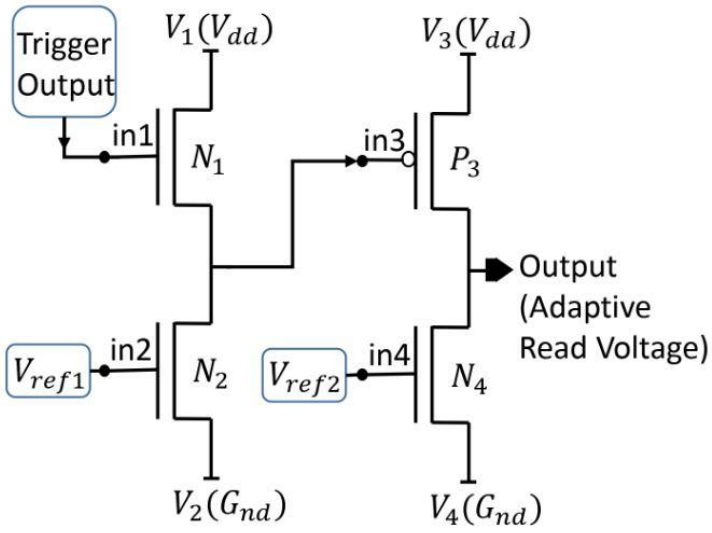


Figure 3.7: Adaptive read voltage circuit.

is carefully selected, thanks to the very nonlinear relationship between the resistance of the FeFET and the read voltage.

As shown in Fig. 3.7, our read voltage control (RVC) circuit for logic '1' consists of two cascaded voltage dividers: In the 1st stage, the gates of NMOS transistor N_1 and N_2 are controlled by the output from the refreshing trigger logic and an external bias V_1. Following the voltage reduction at the trigger output, the output of the 1st stage, which is connected to the gate of the PMOS transistor (P_3) in second stage, also reduces. It will gradually increases the output of the 2nd stage, say, the read voltage to compensate the V_{th} loss on the FeFET.

3.3.3 Array Design and Disturbance

In the modified FeDRAM design, each row in the array is augmented with one dedicated RTL, as shown in Fig. 3.3. In the read access of a memory row, its corresponding RTL will be checked first to decide if refresh is required and generate appropriate read voltage V_{read} for data sensing. For illustration purpose, Fig. 3.8 shows the conventional write biasing schemes with the maximum disturbance of $V_{prog}/2$ and $V_{prog}/3$ [17].

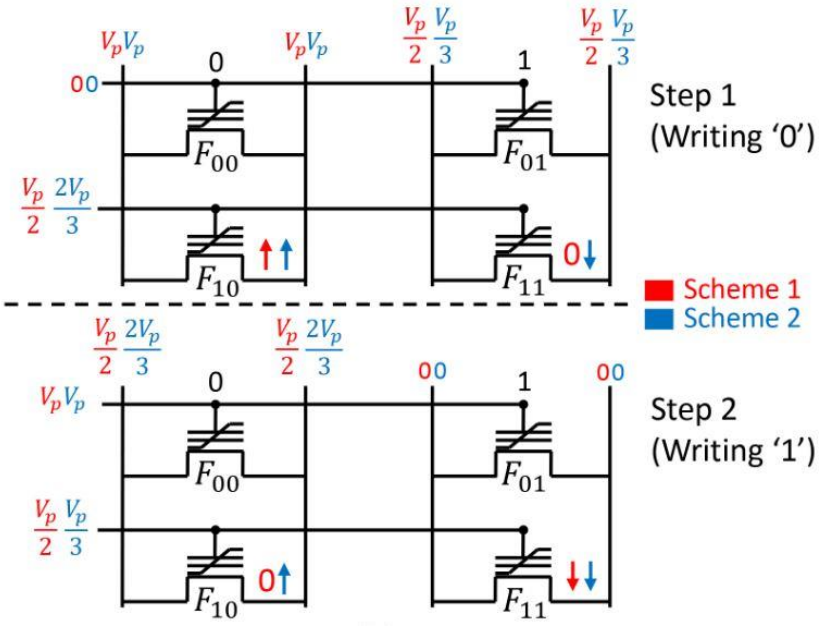


Figure 3.8: 1/3 Biasing scheme to reduce write disturbance.

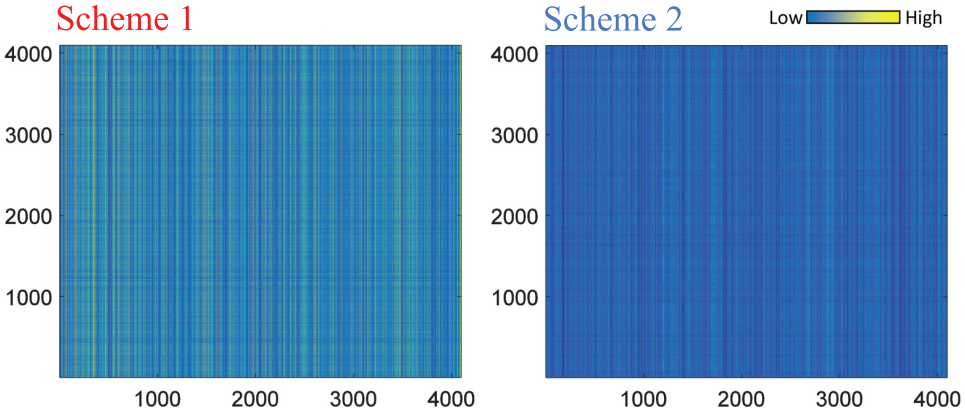


Figure 3.9: Sample write disturbance accumulation on a 4096 row array.

Note that besides tracking retention time limit of the FeDRAM cells, RTL is also able to track the impact of the write and read disturbances. For example, in the read access of a memory row, since the same read voltage is applied to the FeFETs in both the refreshing trigger logic and the memory cells, the impact of the read disturbance is naturally recorded. To accurately track the influence of the write disturbance, however, requires the same disturbing biases to be applied to the RTL as the one applied on the memory cells being disturbed. Fig. 3.9 shows a sample write disturbance accumulation a memory window array. Even though V/3 scheme reduces the disturbance accumulation there is still significant amount of accumulation which reduces the memory retention.

In the application of FeDRAM, there exists the possibility that some (though very rare) data may become erroneous even before its next read access. In such a case, periodic checking needs to be performed on these data to ensure that they can be refreshed on time. How to identify such type of the data will be extensively studied and included as future work.

3.4 EXPERIMENT RESULTS

3.4.1 Trigger Operation

As is is discussed in the previous sections, output of the refresh trigger will reduce based on the memory window decay. Since we plan to program the FeFET devices to increase the endurance of the memory cells, it is necessary to examine the if the trigger output can provide necessary outputs under memory windows decay. Fig. 3.10 shows the trigger can provide enough margin between different memory window levels which indicate it will perform properly for smaller memory window levels.

Also a variation analysis is done to show trigger can provide enough margin to sense if the memory window dropped the predetermined threshold level. Fig. 3.11 shows the trigger can provide enough margin between %100 and %75 memory window levels under a device variation which follows Gaussian like distribution as stated in previous work [18].

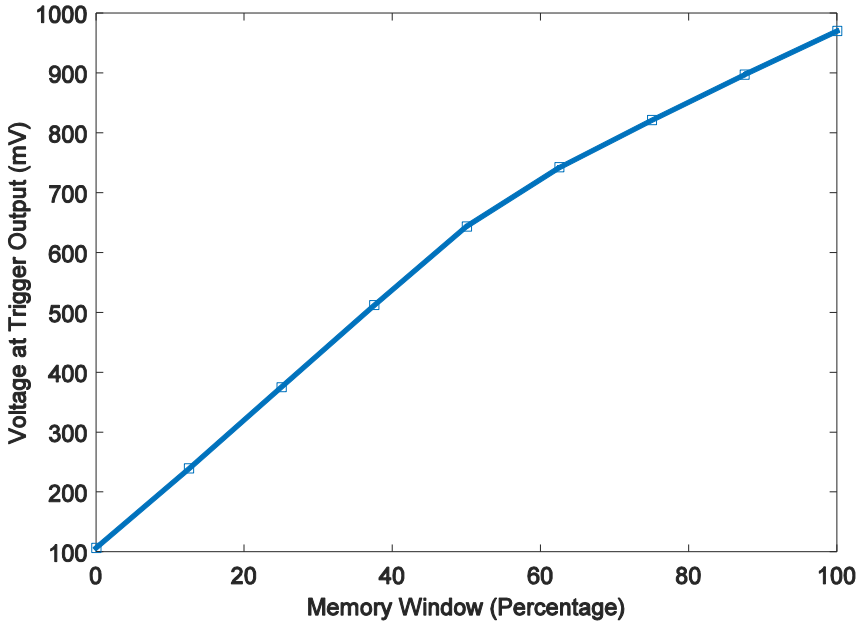


Figure 3.10: Trigger output steps for different memory window levels.

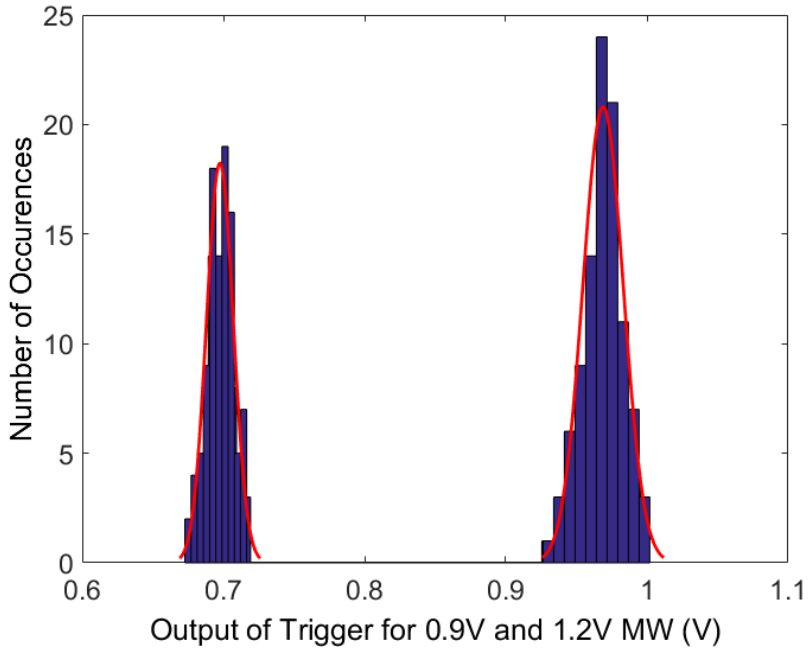


Figure 3.11: Trigger output steps for different memory window levels.

3.4.2 Refresh Reduction and Endurance Enhancement

I perform architectural experiments to evaluate the possible reduction of fresh operations when our adaptive refreshing scheme is applied. Table 3.1 depicts the architectural setup of the experiments. Here we assume the memory window is programmed to only 20% of the maximum level due to endurance concern for main memory application. Periodical refreshing must be performed to retain the FeDRAM data. As shown in Fig. 3.12(a), compared to the periodical refreshing baseline with the reduced memory window, adaptive refreshing averagely reduces 42% refreshes by removing the unnecessary refresh operations on the main memory rows, e.g., the data is overwritten before it wears out.

I also simulate the lifetime of the FeRAM when running different benchmarks. By reducing the programmed memory window by 80%, the lifetime of the FeDRAM is extended from minutes to months, as shown in Fig. 3.12(b). This result implies that FeDRAM may merely satisfy the endurance requirement of some applications but still need further improvement, even adaptive refreshing is enabled.

3.4.3 Adaptive Read Voltage

When memory window degrades, the BL voltage generated by the RVC circuit keeps increasing to compensate the V_{th} loss on the gate of the FeFET. As shown in Fig. 3.13(a), it slows down the effective gate voltage loss on the FeFET providing a more stable sensing margin for read operation.

Table 3.1: Architectural experimental setup

Core	4GHz, OOO 8-width, 64-energy LSQ, 64-entry instruction queue, 192-enrty ROB
Caches	32KB L1I, 2-way, 2-cycle, 64B line
	32 KB L1D, 4-way, 2-cycle, 64B line, write back
	2MB L2 cache, single-core, 32-way, 16 banks, 1 port, 64B line, write back, 20 MSHRs Read energy 0.241nJ, Read delay 12 Cycles, Write energy 0.882nJ, Write delay 30 Cycles
Main memory	2GB, DDR3-1600, 2-channel, open-page, FR-FCFS

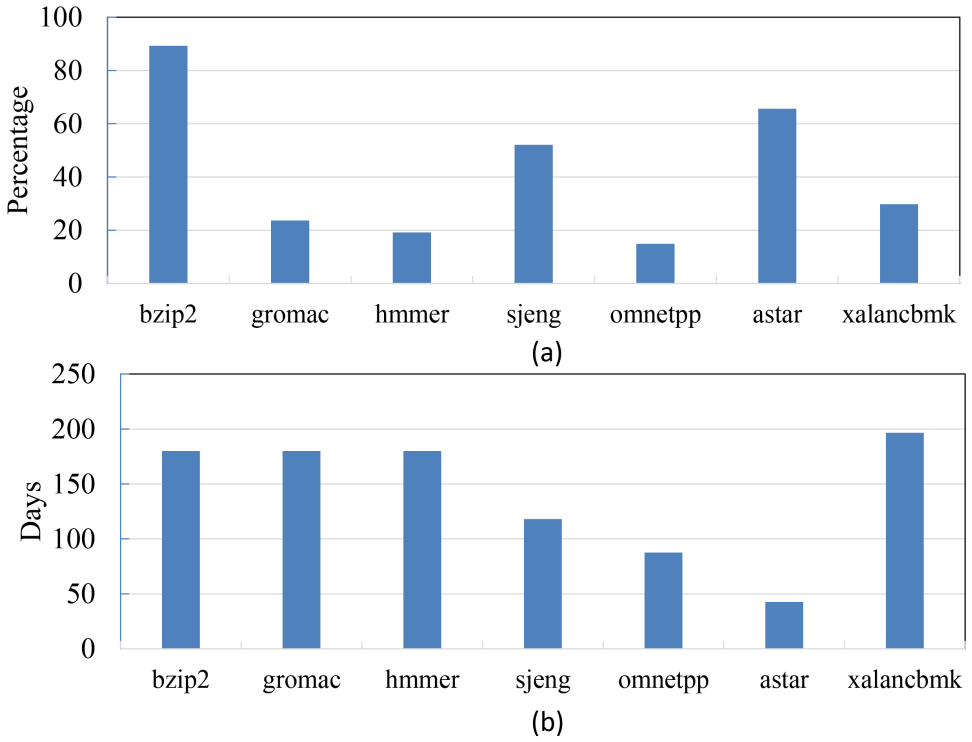


Figure 3.12: Simulation results of (a) Refresh reduction (b) Lifetime of FeDRAM.

The graph includes the margins for MW=1.2V and the MW=0.8V. 0.8V selected as the reduced MW value because the library used for simulation (SMIC 65nm) has V_{th} =430mV so the FeFET device will have slightly positive V_{th}, which means it is normally turned off, even if its programmed as Logic '1'. Hence it can respond to logic inputs (e.g. V_{read}).

Fig. 3.13(b) depicts that the change of the sensed current margin between two logic states of a FeDRAM cell is improved when RVC scheme is utilized, but still shrink with the degradation of memory window. The adaptive read voltage scheme is set to perform better on the earlier stage, which provides much larger current margin, of the MW reduction in order not to increase the logic "0" state SL current too much. As a result increment amount of the V_{read} levels and does not keep increasing with the MW loss.

The detailed analysis in Fig. 3.13(c) shows that, although a voltage compensating one logic state of the FeFET, it will harm the sensing of the other state. As can be seen in Fig. 3.13(a) there is improvement on the current margin between two states through out the MW range which means the increase of the logic "0" state's SL current stays negligible.

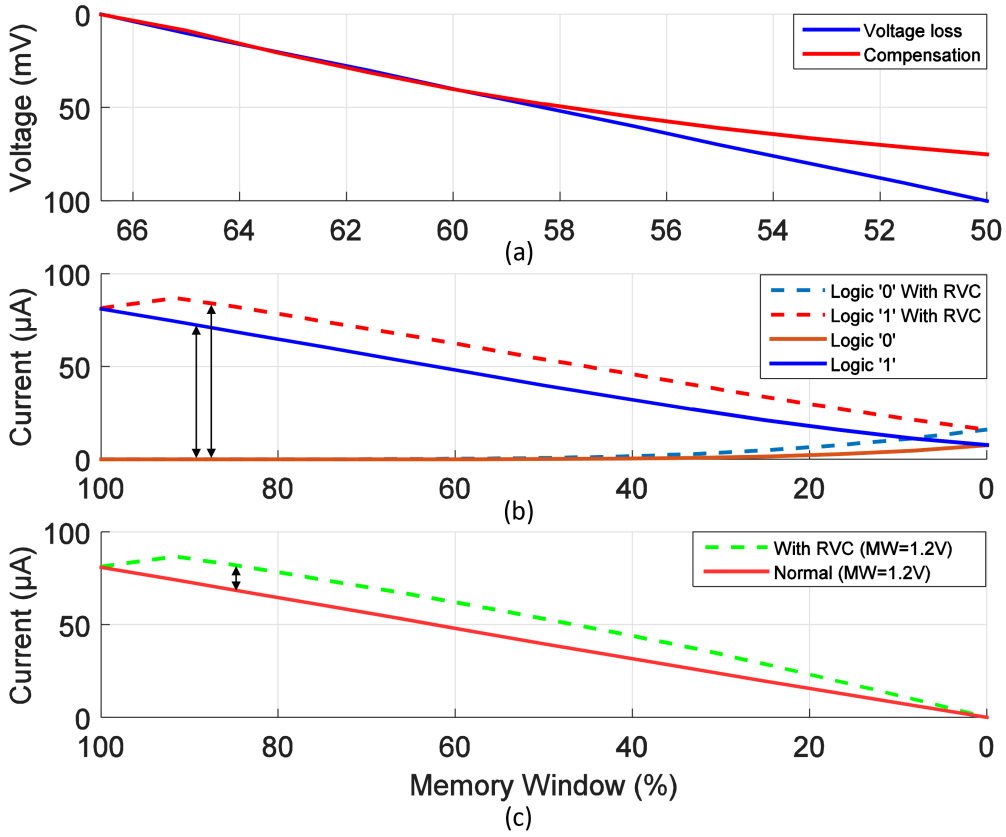


Figure 3.13: (a) V_{th} shift due to memory window loss and the increment amount of the V_{read}. (Graph is for the period where MW reduces from 0.8V down to 0.6V.) (b) Improved SL current difference between two logic states when RVC applied. (MW = 1.2V, V_{BL} = 0.6V) (c) Overall margin shift as a result of RVC. (MW = 1.2V, V_{BL} = 0.6V)

During this evaluation it is assumed that the FeFET is programmed to its maximum memory window, say, 1.2V.

3.5 CONCLUSION

The recent progress in FeFET technology development motivated the exploration of using FeFET to implement DRAM module. To achieve the endurance requirement in DRAM applications, we need to sacrifice the retention time of the FeDRAM cells by only partially programming the FeFET. As a result, the FeDRAM cells need to be refreshed periodically and become more sensitive to the write and read disturbance. In this work, I propose an adaptive refreshing and read voltage control scheme to minimize the large overhead of the refresh operations and enhance the read robustness of the FeDRAM, respectively, by taking into account the intrinsic degradation of the FeFET memory window and the write and read disturbances. Simulation results show that proposed scheme can effectively reduce the number of refresh operations on average by 42%, and achieve substantial sensing margin improvement of the FeDRAM cell.

4.0 IMPROVED REDUNDANT WRITE REDUCTION FOR FEDRAM

4.1 IMPROVED REDUNDANT WRITE REDUCTION

To further improve the endurance of the FeFET memory, conventional methods like redundant write reduction and wear leveling (row shifting, segment swapping) methods can be applied similar to the PCM which is formerly proposed in [30]. As proved by the previous work this methods can also greatly improve the life time of FeFET which is similar to PCM in terms of endurance. Especially redundant bit write reduction has a significant improvement and it is shown that it can reduce the number of writes on a memory cell an average of %75 [30].

In this section we proposed a novel method to further improve conventional redundant bit write reduction method. Conventional reduction methods is simply implemented as introducing a read operation before write and by comparing the old and new data. Then it cancels the write operation to the unnecessary memory cells based on the comparison decision. It is discussed in [30] that the performance loss introduced by extra read operation and hardware is negligible compared to the significant lifetime improvement. Also former work states that the write operations are not critical in terms of performance as the reads hence the delay introduced by the extra hardware can be stated as also not important compared to the enhancement. The FeDRAM scheme proposed in the former chapter already has a read operation to check if a refresh is required, so there will be no extra performance loss with the introduction of the new reduction scheme. Main idea of the new scheme is to store data as a more unified data stream say making all rows close to all zeros which means the new write data will be more similar to the already stored data. Thus there will be more redundant writes to reduce. A sample Hamming weight distribution of a block data which

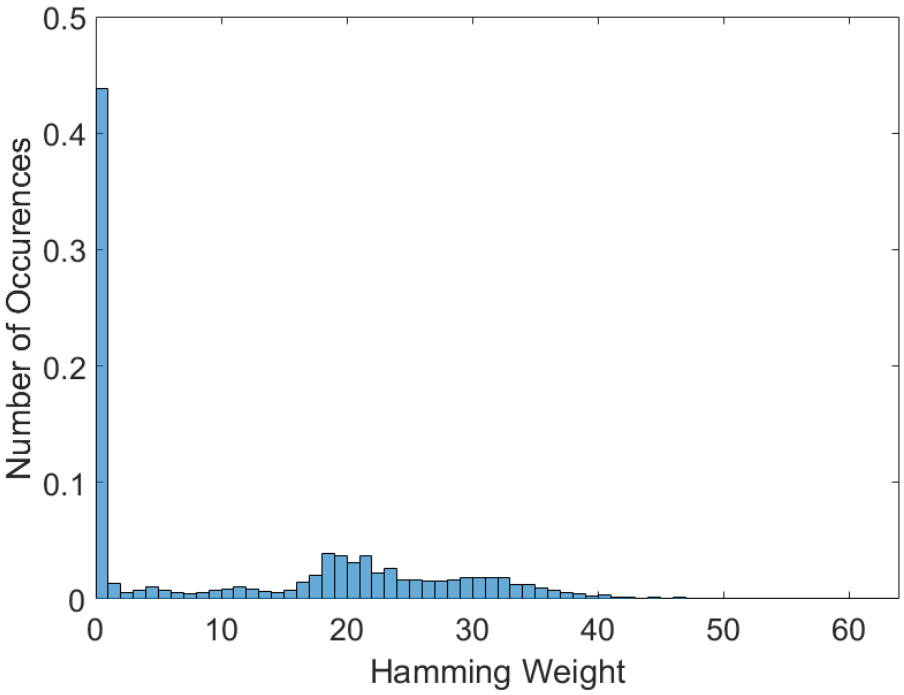


Figure 4.1: Hamming weight distribution of un-edited block data.

is obtained from a DRAM trace is shown in Fig. 4.1. It can be seen that a large range of(up to 64 bits of logic ones) weights visible in a dram data stream.

4.2 IMPLEMENTATION AND RESULTS

4.2.1 Decision Circuit

To make the data unified we simply invert data if there are more ones in it. The reason to store data as mostly zeros is switching from logic one to logic zero requires less programming voltage. We first need to check if the number of ones or zeros is higher in a write data. We proposed a decision circuit to determine if the hamming weight of the data block is larger than half of the number of bits. Fig. 4.2 shows the detection circuit. It consists of 2 voltage dividers. Voltage divider's top parts consist of multiple number of transistors which is equal to the block data size. First voltage divider's adjustable top part consists of PMOS transistors and Out1 stays higher than the Out2 while block data includes more

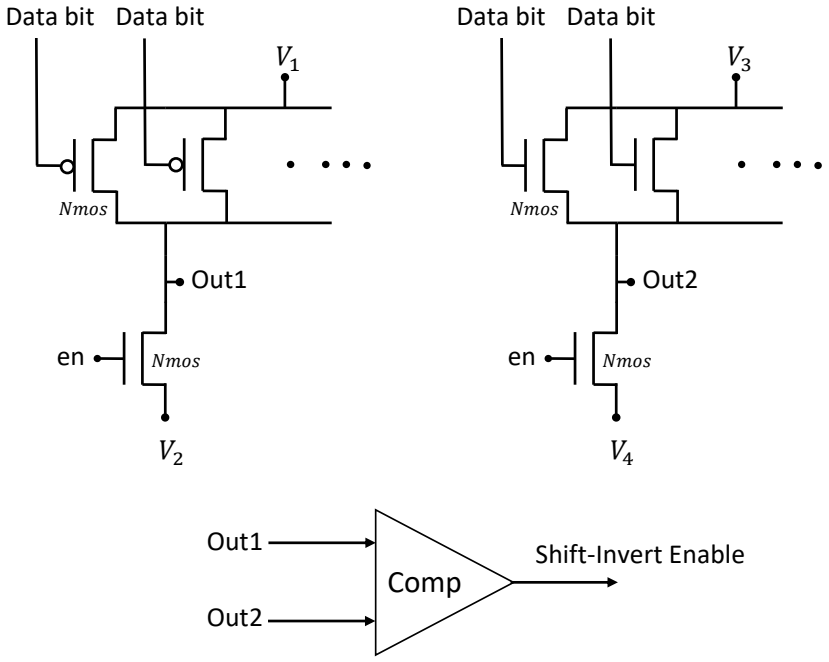


Figure 4.2: Detection circuit structure.

zeros than one. Second voltage divider's adjustable top part consists of NMOS transistors and its output stays higher than the Out1 while block data includes more logic one values than logic zero values. By comparing the Out1 and Out2, it can be determined that if the number of zeros is more than ones in a data block.

Since we do not need to know the exact Hamming Weight, a simple decision circuit structure chosen to reduce the hardware-delay overhead. Decision circuit can either be built by regular transistors or FeFET devices. However using FeFETs can provide more flexibility since the channel conductivity can be altered and a middle meeting point between two voltage dividers can be achieved more precisely for larger block sizes. Fig. 4.3 shows the 8bit and 16bit transistor stack simulation results. Reason to choose data size as 16bits will be explained in the following section. A distinct separation between more than half logic zero and more than half logic one states can be achieved as shown in the simulation results.

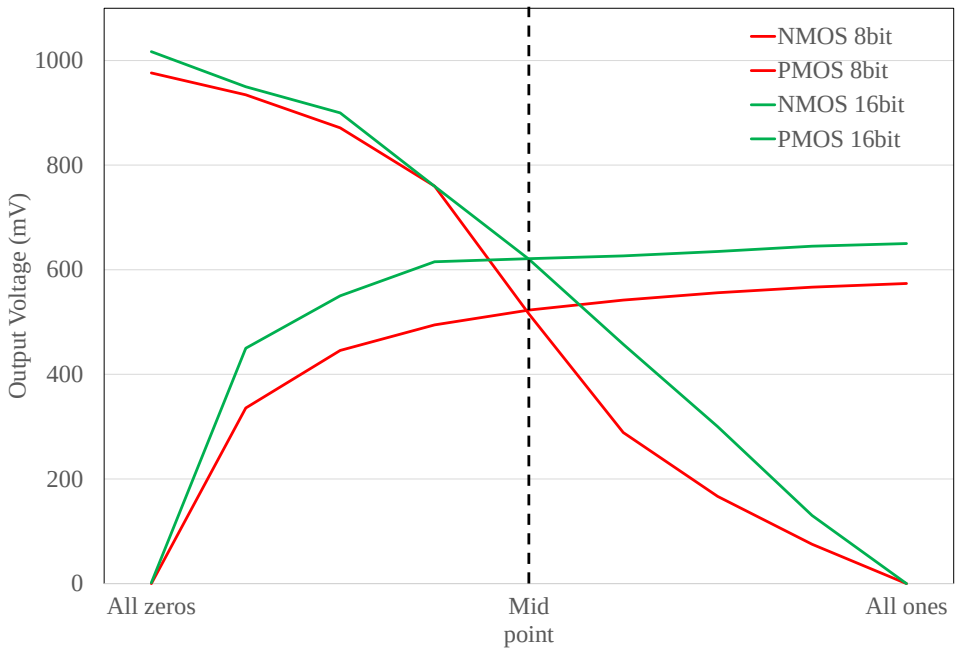


Figure 4.3: Detection circuit operating conditions. It can provide a mid meeting point for both 8bit and 16bit streams.

4.2.2 Inverting Operation and Simulation Results

Overall inverting only operation sequence is shown in the Fig. 4.4. Original data goes through the decision circuit and based on the number of logic ones and zeros, decision circuit will provide a select signal for the Inverter circuit. Inverter circuit consists of inverters and muxes. Based on the decision circuit output, Inverter circuit flips the '1' to '0' and '0' to '1'. After that the new data will be compared with the already stored data in the Comparator and corresponding write enable signals for the specific rows will be provided based on the match-mismatch between the stored and new data. Finally the Flip decision will be stored as extra bit for the corresponding row.

For the read operation, data will be read from the array including the previously stored Invert decision. Then the data will go through the output Inverter and will be inverted or not based on the stored Invert decision data. Whole sequence is performed on circuit and original data will be retrieved at the output. No other intermediate arrangement such as architectural or software required.

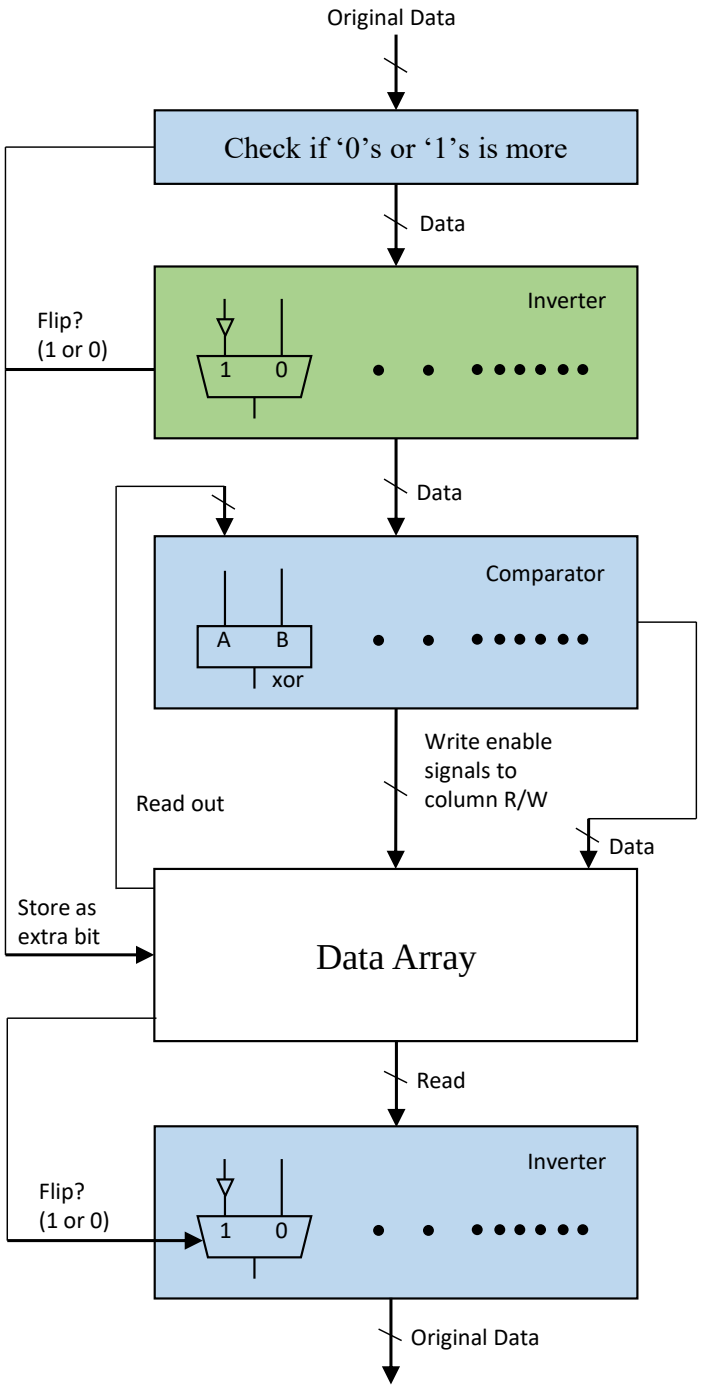


Figure 4.4: Operation sequence of the proposed bit write reduction scheme.

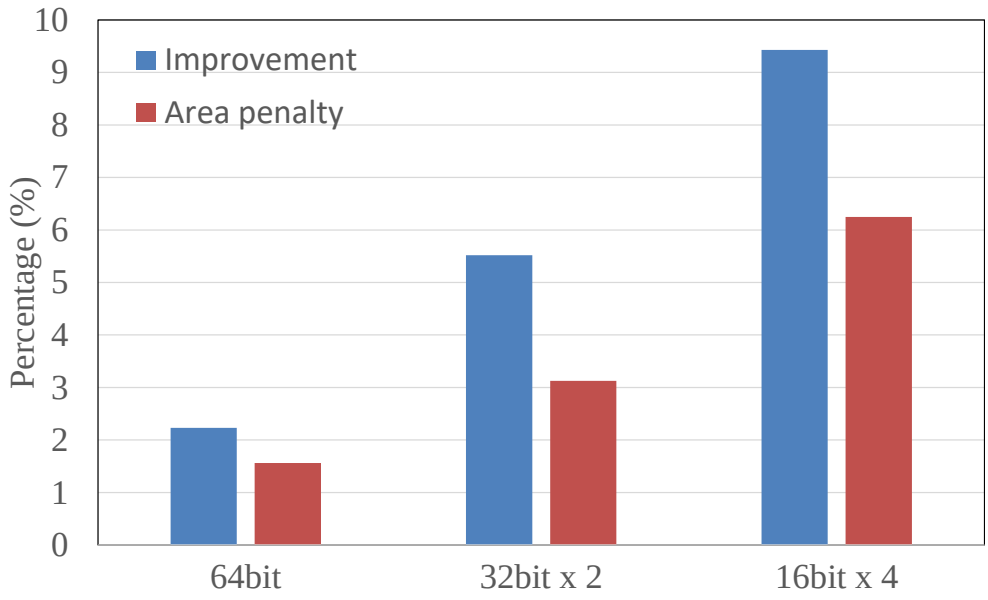


Figure 4.5: Improvement versus area penalty for 64bit block as a whole, divided by 2 and divided by 4.

Fig. 4.5 shows the evaluation of the proposed scheme based on data blocks of a DRAM trace data which consist of 64 bit blocks. Here the baseline is the state of the art redundant bit write reduction method. Percentages shows the improvement for different conditions over the regular reduction method. For the whole block only 1 extra bit required to store Invert decision, for half and quart divisions 2 and 4 extra bits required respectively. Since the simulation based on 64 bit blocks extra bits are shown as area penalty in Fig. 4.5. As shown in the simulation results, as the data blocks divided in to the smaller portions, improvement percentage increases significantly while the area penalty stays smaller.

4.2.3 Inverting-Switching Operation and Simulation Results

To further improve the elimination of redundant writes, a simple shift operation (called as switching) can also be introduced. New operation sequence is shown in the Fig. 4.6. Original data goes through the detection circuit based on the number of logic ones and zeros, detection circuit will provide a select signal for the Inverter circuit which flips the '1' to '0' and '0' to

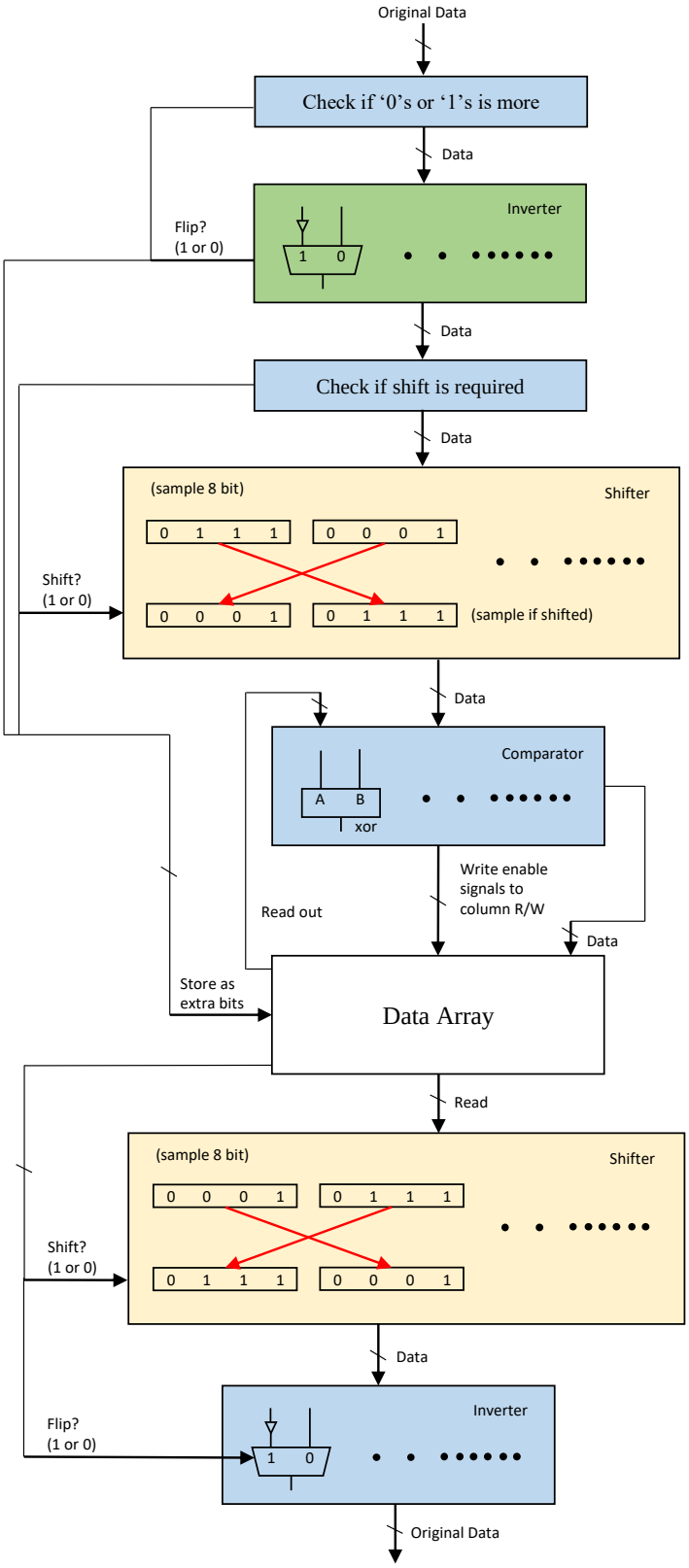


Figure 4.6: Operation sequence of the proposed bit write reduction scheme with shift operation.

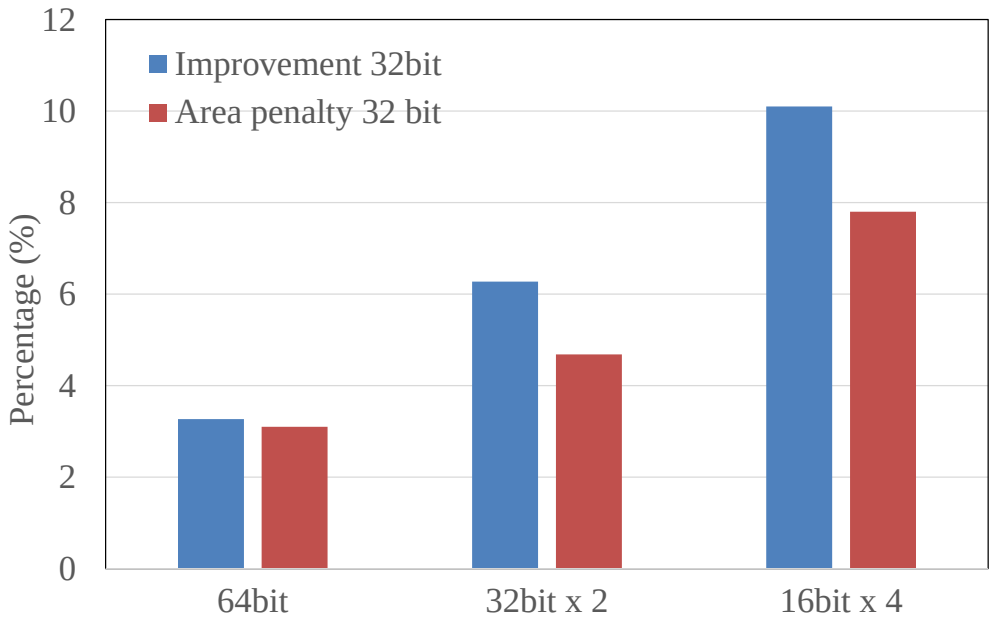


Figure 4.7: Improvement versus area penalty for 64bit block as a whole, divided by 2 and divided by 4 with 32 bit shift operation added.

'1'. After that the new data will be checked again based on the previously selected portions and detection circuit will provide a Shift decision signal. The shifter circuit basically switches first half potion of the data with the last half portion. Such operation is chosen to reduce hardware-delay overhead. Also during simulation it is observed that the too much editing does not make as much improvement. After that the new data will be compared with the already stored data in the Comparator and write enable signals for the specific rows will be provided based on the match mismatch between the stored and new data. The Invert and Shift decisions will be stored as extra bits.

For the read operation, data is read from the array at the same time with the previously stored Invert and Shift decisions. The data will go through the output Shifter and will be shifted or not based on the previously stored Shift decision. After that the data will go through the Inverter and will be flipped or not based on the stored Invert decision data. Whole sequence will be done on circuit and original data will be retrieved at the output.

Fig. 4.7 shows the evaluation of the proposed scheme with 32 bit shift operation based on DRAM trace which consist of 64 bit blocks. Here the baseline is the state of the art

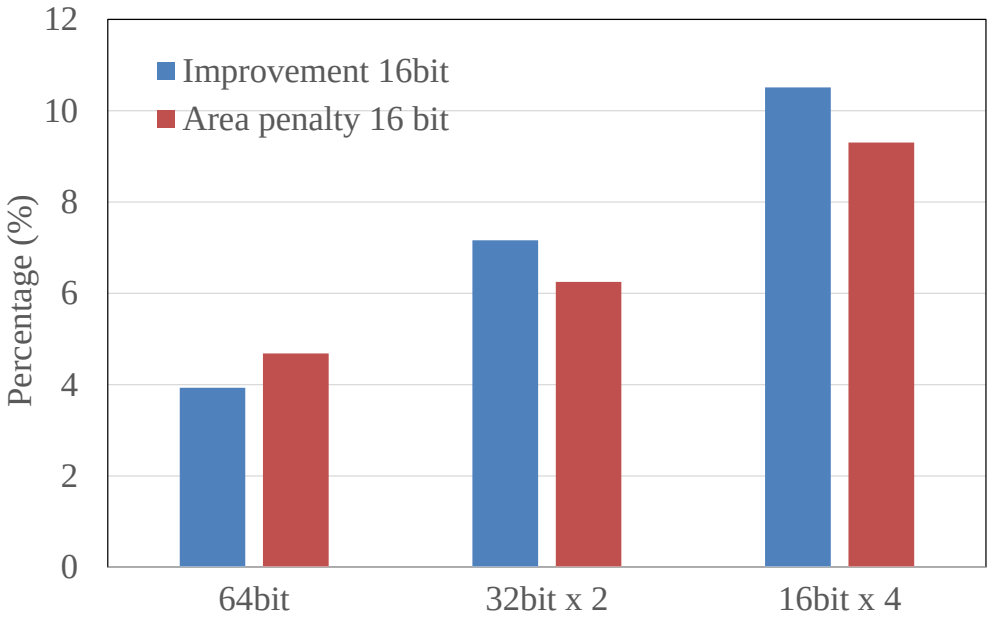


Figure 4.8: Improvement versus area penalty for 64bit block as a whole, divided by 2 and divided by 4 with 16 bit shift operation added.

redundant bit write reduction method similar to the former simulations. As shown in the simulation results, shifting the data block based on 32 bit divisions improves the performance of the redundant write reduction for all conditions while introducing the necessity of storing 1 extra bit which introduces extra area penalty.

To examine the effect of further partitioning the data block, 16bit shifts which introduces 2 extra bits to store also examined. As Fig. 4.8 shows 16bit shifts can further improve the performance for 32bit and 16bit Flip operations while area penalty percentage stays lover than the enhancement. For the 16bit shift total stored extra bits reach to 6 bits and exceeds the enhancement for a 64 bit non divided block data.

The proposed scheme with shift and flip operations can provide up to %10.5 improvement over conventional redundant bit write elimination method. Fig. 4.9 shows the Hamming Weight distribution after 16bit Flip and Shift applied on data trace. If compared with the Fig. 4.1 it can be observed that the improvement comes from the more unified data stream.It should be noted that introduced number of bits related the partition ratio which means a larger data block will also require only 4 bits if it is divided in to 4 partitions.

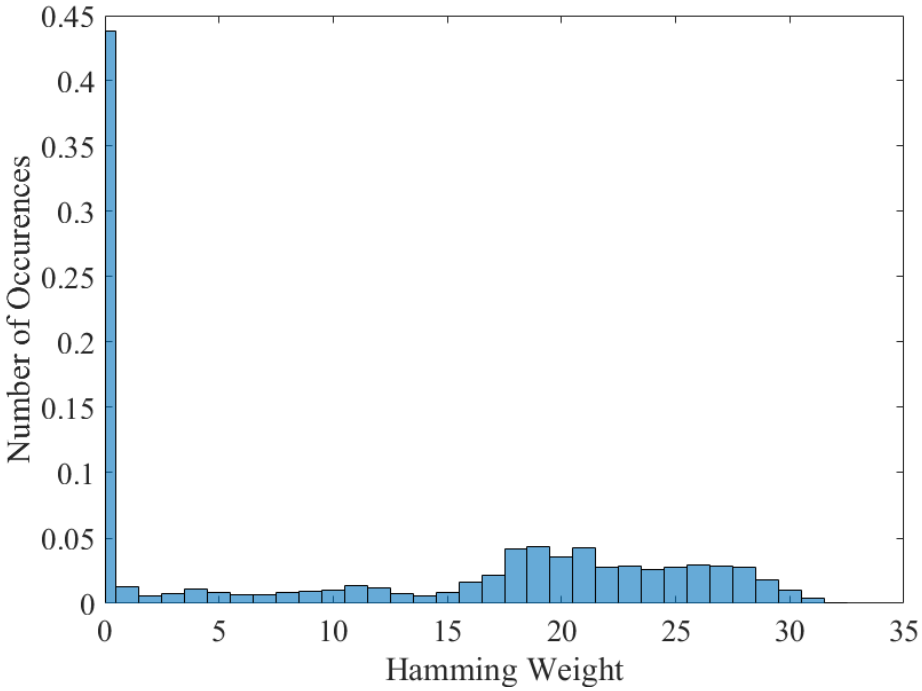


Figure 4.9: Hamming weight distribution after shift and flip operations.

4.3 CONCLUSION

FeFET technology recently emerged again as promising candidate to replace existing DRAM technology with the current improvements. To achieve the required endurance, additional methods should be applied as an addition to the Adaptive Refreshing and Read Voltage Scheme. In this work we proposed a novel scheme to improve conventional Redundant Bit Write reduction scheme. Our scheme can provide up to %10.5 improvement over the state of the art method while not introducing too much hardware-delay overhead. Also the method is not specific to FeFET device and can be applied to any kind of memory technology.

5.0 ON TIME REFRESH SCHEME

Adaptive refreshing scheme can greatly reduce the number of refreshes on FeDRAM. However since the time of refresh is dependent on the read-write access to a row, there is always a possibility that a row can lost the stored data before it receives an access. To overcome this issue we proposed On Time Refresh Scheme. In this section we explained and examined the the proposed scheme.

5.1 BASIC CONCEPT OF REFRESH FAILURE ELIMINATION

Overall scheme is based on having two physically separated data arrays [11]. Hence two arrays can be prevented from receiving same amount of write disturbance. The data rows will be separated in to two groups based on their read/write access frequency. In this way we have two separate data arrays such as frequently and rarely accessed as shown in Fig. 5.1. So we can arrange different refresh methods for the separated groups of rows based on their access schedule.

5.1.1 Data Separation

Basic data separation depends on when a row does not receive read or write access more than a predetermined period, then it will be moved to rarely accessed array. If a row has more frequent accesses than a predetermined period it will be placed in the frequently accessed array as shown in Fig. 5.2. The threshold for moving data between arrays can be determined based on both write disturbance accumulation and retention characteristics of FeFET device based on chosen programming level.

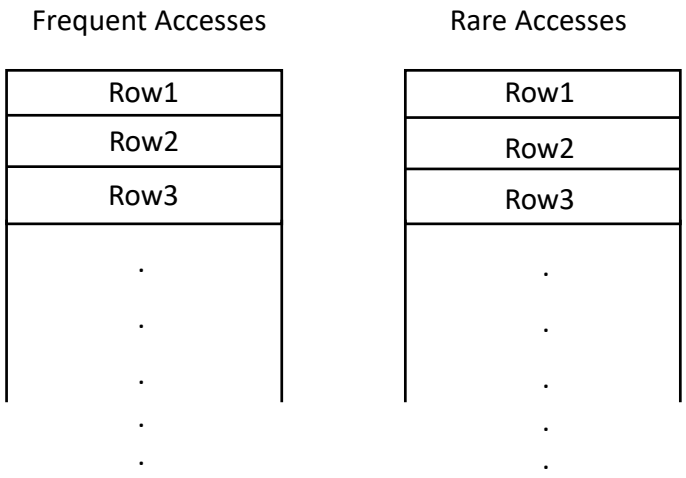


Figure 5.1: Seperated arrays based on access frequency.

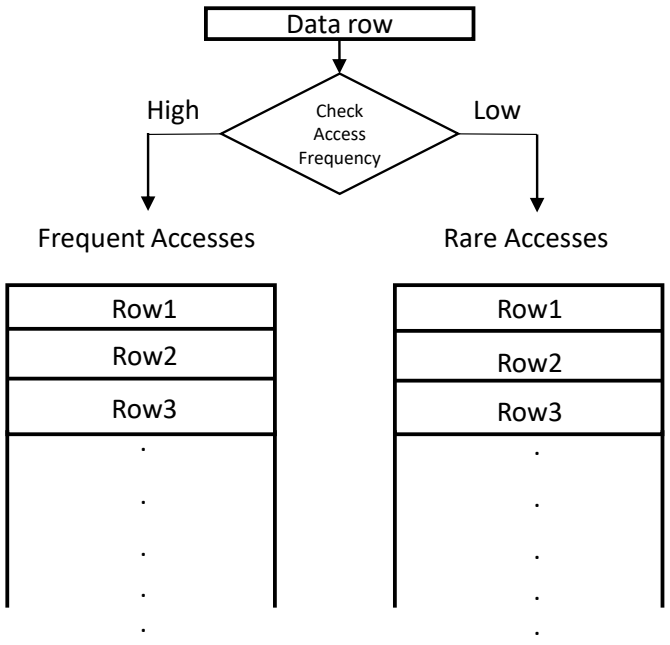


Figure 5.2: Data row placement based on access frequency.

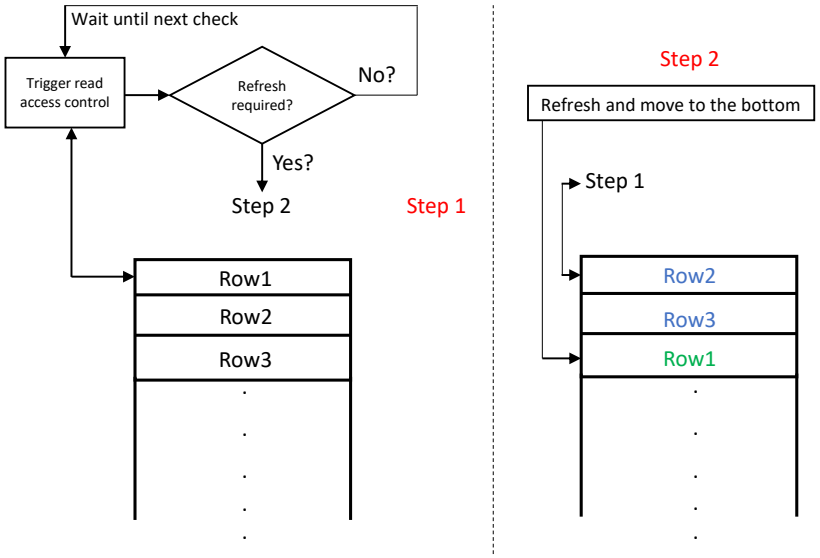


Figure 5.3: Refresh procedure for rarely accessed array.

Frequently accessed array does not need a complex assistive method since it receives the read or write access in required periods. Hence it can be directly refreshed based on refresh trigger check with the each access.

5.1.2 Refreshing the Rarely Accessed Array

Rarely accessed array's rows will not receive as much write disturbance as the frequently accessed array. Thus they can wait much longer than a frequently accessed data array. However they still need to be checked to make sure data will not be lost until a data access arrives.

To eliminate a DRAM like conventional cyclic refresh, we utilize a special storing and trigger check method for the rarely accessed data array. The data will be stored in to the rows of the rarely accessed array from top to bottom indicating that the first row is always the longest awaiting row with the smallest memory window. Based on the design requirements a trigger check frequency will be determined. The first row of the array's refresh trigger will be checked based on this frequency to determine if a refresh is required or not for the first row. If no refresh is required for the first row the array will wait until next trigger check.

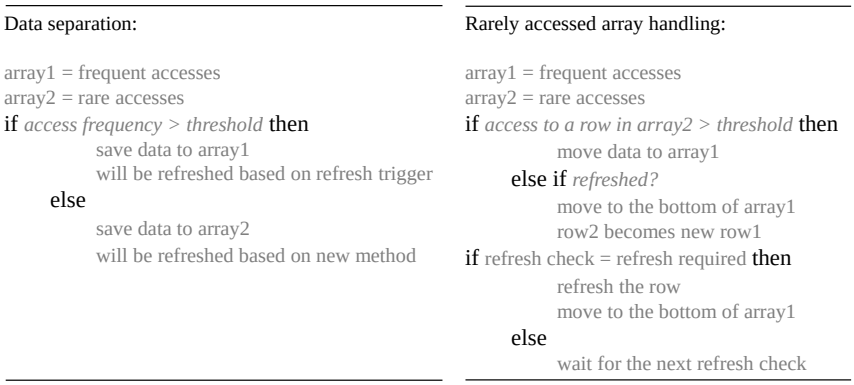


Figure 5.4: Refresh procedure flow for Failure Elimination scheme.

If a refresh is required, the first row will be refreshed and moved to bottom of the array. Then the following rows will be checked and if a refresh is required, they will be refreshed and moved to bottom. When the refresh operation reaches to a no refresh required row, the refresh operation will stop and last checked row will be the new first row of the array. Then the array will wait until next refresh check again. Fig. 5.3 explains the overall procedure.

When a row from the rarely accessed data array receives a write access or a read access its trigger will be checked and it will be refreshed if required. If a refresh occurs, the row will be moved to the bottom of the array since it has a fresh memory window.

If a row of the array receives more frequent accesses than a predetermined threshold, it will be migrated to the frequently accessed array [11]. Similarly frequently accessed array's rows will be moved to rarely accessed array in case they receive less than a threshold read or write accesses. Overall flow is shown in Fig. 5.4.

5.2 EVALUATION OF REFRESH FAILURE ELIMINATION SCHEME

To evaluate the efficiency of failure reduction method we used same benchmark suit based dram traces which obtained through a setup which is discussed in chapter 3. The methods are developed and evaluated based on DRAM trace data obtained from sample workloads of benchmark suite.

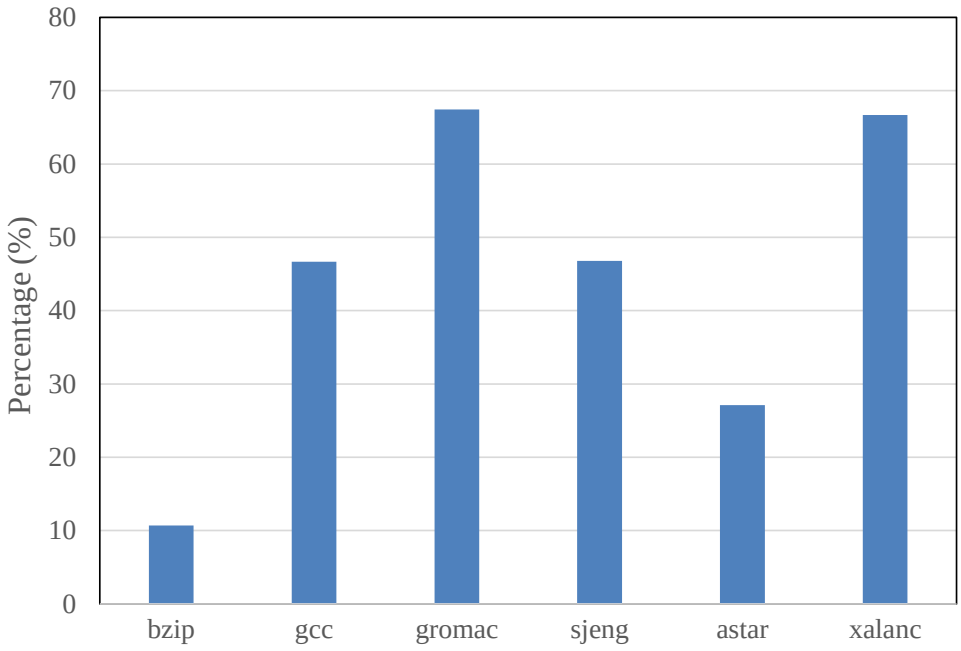


Figure 5.5: Data without any access can be saved in the rarely accessed array and wait for longer refresh period.

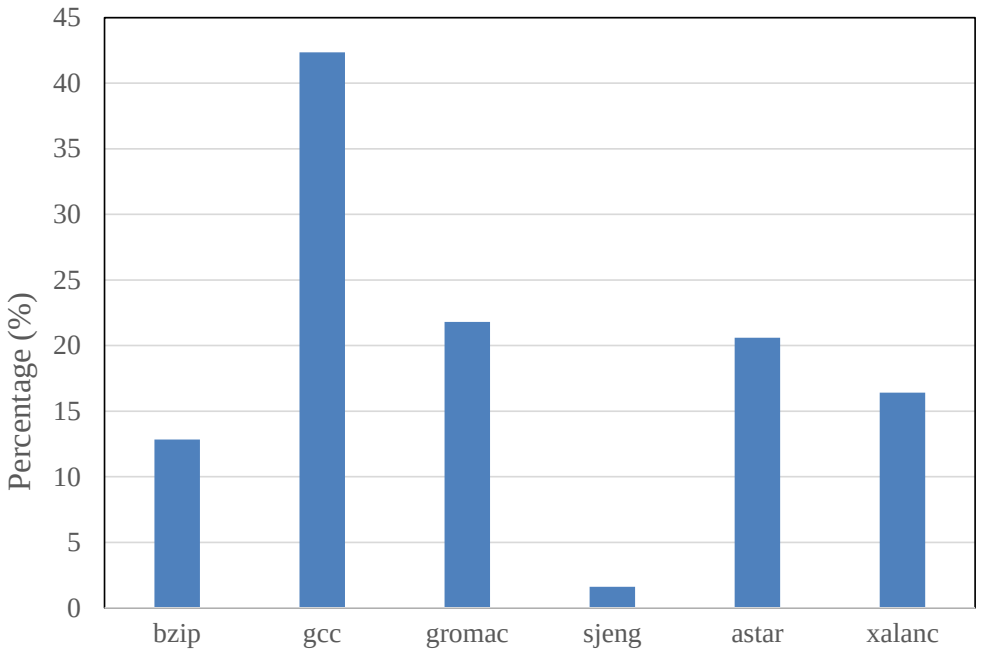


Figure 5.6: Scarcely accessed rows to be saved from failure of refreshing before the data loss.

Since the benchmarks are artificial loads for a certain amount of time, we obtained relatively large amount of un-accessed rows as shown in Fig. 5.5. The issues of benchmark traces discussed in [30]. The un-accessed rows can be assumed directly saved in the rarely accessed array and refreshed based on a much longer period than frequently accessed array. Simulation results shows that average %44.24 of the data can be directly saved in to the rarely accessed array and can be survived from unnecessary refreshes.

As and addition to the un-accessed rows, there are still row that have sparse accesses which means they can possibly loose the data before next read or write access arrives. These data row can also provide a more realistic evaluation than the large amount of un-accessed rows obtained due to trace data. Fig. 5.6 shows the percentage of saved data which suffers from sparse access. With conservatively determined thresholds, on time refresh scheme can provide an average of %19.26 failure reduction which basically means preventing all possible failures.

5.3 CONCLUSION

Adaptive Refreshing scheme can cause a possible failure of refreshing on time due to long waiting periods for a read or write access to a row. The proposed Refresh On Time scheme properly separates the data rows and ensures the appropriate refresh rates to be provided for the each specific row. Simulation results show that with a carefully determined operating threshold, failure free refresh operation can be achieved for FeDRAM with significant unnecessary refresh reduction compared to a conventional cyclic refresh scheme similar to DRAM.

BIBLIOGRAPHY

[1] Y. Arimoto and H. Ishiwara. Current status of ferroelectric random-access memory. *Mrs Bulletin*, 29(11):823–828, 2004.

[2] L. Chua. Memristor-the missing circuit element. *IEEE Transactions on circuit theory*, 18(5):507–519, 1971.

[3] D. Damjanovic. Hysteresis in piezoelectric and ferroelectric materials. *The science of hysteresis*, 3:337–465, 2006.

[4] E. Eken, Y. Zhang, W. Wen, R. Joshi, H. Li, and Y. Chen. A new field-assisted access scheme of stt-ram with self-reference capability. In *Proceedings of the 51st Annual Design Automation Conference*, pages 1–6. ACM, 2014.

[5] J. Hu, Q. Zhuge, C. J. Xue, W.-C. Tseng, and E. H.-M. Sha. Software enabled wear-leveling for hybrid pcm main memory on embedded systems. In *Design, Automation & Test in Europe Conference & Exhibition (DATE), 2013*, pages 599–602. IEEE, 2013.

[6] Y. Huai. Spin-transfer torque mram (stt-mram): Challenges and prospects. *AAPPS Bulletin*, 18(6):33–40, 2008.

[7] P. Huang, B. Chen, Y. Wang, F. Zhang, L. Shen, R. Liu, L. Zeng, G. Du, X. Zhang, B. Gao, et al. Analytic model of endurance degradation and its practical applications for operation scheme optimization in metal oxide based rram. In *2013 IEEE International Electron Devices Meeting*, pages 22–5. IEEE, 2013.

[8] Q. Li, J. Li, L. Shi, M. Zhao, C. J. Xue, and Y. He. Compiler-assisted stt-ram-based hybrid cache for energy efficient embedded systems. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 22(8):1829–1840, 2014.

[9] C. Lichtensteiger, J.-M. Triscone, J. Junquera, and P. Ghosez. Ferroelectricity and tetragonality in ultrathin p b t i o 3 films. *Physical review letters*, 94(4):047603, 2005.

[10] H.-T. Lue, C.-J. Wu, and T.-Y. Tseng. Device modeling of ferroelectric memory field-effect transistor for the application of ferroelectric random access memory. *Ultrasonics, Ferroelectrics, and Frequency Control, IEEE Transactions on*, 50(1):5–14, 2003.

[11] Y. Luo, Y. Cai, S. Ghose, J. Choi, and O. Mutlu. Warm: Improving nand flash memory lifetime with write-hotness aware retention management. In *Mass Storage Systems and Technologies (MSST), 2015 31st Symposium on*, pages 1–14. IEEE, 2015.

[12] T. Ma. Fedram: A capacitor-less dram based on ferroelectric-gated field-effect transistor. In *Memory Workshop (IMW), 2014 IEEE 6th International*, pages 1–4. IEEE, 2014.

[13] A. Makarov, V. Sverdlov, and S. Selberherr. Emerging memory technologies: Trends, challenges, and modeling methods. *Microelectronics Reliability*, 52(4):628–634, 2012.

[14] S. Matsunaga, A. Katsumata, M. Natsui, and T. Hanyu. Design of a low-energy nonvolatile fully-parallel ternary cam using a two-level segmented match-line scheme. In *2011 41st IEEE International Symposium on Multiple-Valued Logic*, pages 99–104. IEEE, 2011.

[15] S. Matsunaga, M. Natsui, K. Hiyama, T. Endoh, H. Ohno, and T. Hanyu. Fine-grained power-gating scheme of a metal–oxide–semiconductor and magnetic-tunnel-junction-hybrid bit-serial ternary content-addressable memory. *Japanese Journal of Applied Physics*, 49(4S):04DM05, 2010.

[16] T. Mihara, H. Watanabe, and C. A. P. de Araujo. Polarization fatigue characteristics of sol-gel ferroelectric pb (zr0. 4ti0. 6) o3 thin-film capacitors. *Japanese journal of applied physics*, 33(7R):3996, 1994.

[17] S. Mueller, J. Muller, R. Hoffmann, E. Yurchuk, T. Schlosser, R. Boschke, J. Paul, M. Goldbach, T. Herrmann, A. Zaka, et al. From mfm capacitors toward ferroelectric transistors: Endurance and disturb characteristics of-based fefet devices. *Electron Devices, IEEE Transactions on*, 60(12):4199–4205, 2013.

[18] S. Mueller, S. Slesazeck, T. Mikolajick, J. Müller, P. Polakowski, and S. Flachowsky. Next-generation ferroelectric memories based on fe-hfo 2. In *Applications of Ferroelectric, International Symposium on Integrated Functionalities and Piezoelectric Force Microscopy Workshop (ISAF/ISIF/PFM), 2015 Joint IEEE International Symposium on the*, pages 233–236. IEEE, 2015.

[19] J. Muller, T. S. Boscke, U. Schroder, R. Hoffmann, T. Mikolajick, and L. Frey. Nanosecond polarization switching and long retention in a novel mfis-fet based on ferroelectric. *IEEE Electron Device Letters*, 33(2):185–187, 2012.

[20] K. Pagiamtzis and A. Sheikholeslami. Content-addressable memory (cam) circuits and architectures: A tutorial and survey. *IEEE Journal of Solid-State Circuits*, 41(3):712–727, 2006.

[21] X. Pan and T. Ma. Retention mechanism study of the ferroelectric field effect transistor. *Applied Physics Letters*, 99(1):013505, 2011.

[22] G. Qin, S. Ata, I. Oka, and C. Fujiwara. Effective bit selection methods for improving performance of packet classifications on ip routers. In *Global Telecommunications Conference, 2002. GLOBECOM'02. IEEE*, volume 3, pages 2350–2354. IEEE, 2002.

[23] U. Schroeder, E. Yurchuk, S. Mueller, J. Mueller, S. Slesazeck, T. Schloesser, M. Trentzsch, and T. Mikolajick. Non-volatile data storage in hfo 2-based ferroelectric fets. In *Non-Volatile Memory Technology Symposium (NVMTS), 2012 12th Annual*, pages 60–63. IEEE, 2012.

[24] Z. Shao, Y. Liu, Y. Chen, and T. Li. Utilizing pcm for energy optimization in embedded systems. In *2012 IEEE computer society annual symposium on VLSI*, pages 398–403. IEEE, 2012.

[25] D. B. Strukov, G. S. Snider, D. R. Stewart, and R. S. Williams. The missing memristor found. *nature*, 453(7191):80–83, 2008.

[26] Y. Wang, Y. Liu, S. Li, D. Zhang, B. Zhao, M.-F. Chiang, Y. Yan, B. Sai, and H. Yang. A 3us wake-up time nonvolatile processor based on ferroelectric flip-flops. In *ESSCIRC (ESSCIRC), 2012 Proceedings of the*, pages 149–152. IEEE, 2012.

[27] C. J. Xue, Y. Zhang, Y. Chen, G. Sun, J. J. Yang, and H. Li. Emerging non-volatile memories: opportunities and challenges. In *Proceedings of the seventh IEEE/ACM/IFIP international conference on Hardware/software codesign and system synthesis*, pages 325–334, 2011.

[28] E. Yurchuk, J. Muller, J. Paul, T. Schlosser, D. Martin, R. Hoffmann, S. Mueller, S. Slesazeck, U. Schroeder, R. Boschke, et al. Impact of scaling on the performance of hfo 2-based ferroelectric field effect transistors. *Electron Devices, IEEE Transactions on*, 61(11):3699–3706, 2014.

[29] Y. Zhang, I. Bayram, Y. Wang, H. Li, and Y. Chen. Adams: asymmetric differential stt-ram cell structure for reliable and high-performance applications. In *Proceedings of the International Conference on Computer-Aided Design*, pages 9–16. IEEE Press, 2013.

[30] P. Zhou, B. Zhao, J. Yang, and Y. Zhang. A durable and energy efficient main memory using phase change memory technology. In *ACM SIGARCH computer architecture news*, volume 37, pages 14–23. ACM, 2009.