Memristive Circuits for On-Chip Memories

by

Ravi Patel

Submitted in Partial Fulfillment
of the
Requirements for the Degree
Doctor of Philosophy

Supervised by
Professor Eby G. Friedman

Department of Electrical and Computer Engineering
Arts, Sciences and Engineering
Edmund J. Hajim School of Engineering and Applied Sciences

University of Rochester
Rochester, New York
2013

Dedication

To my mother, father and little sister, without which my life would be a long series of blank pages.

Biographical Sketch

Ravi Patel was born in Schenectady, New York in 1986. He received his B.Sc, and M.Sc. degrees in Electrical and Computer Engineering in, respectively, 2008 and 2010 from the University of Rochester in Rochester, New York. Working towards a Ph.D degree in Electrical and Computer Engineering, he has studied high performance integrated circuits and resistive memory technologies under the tutelage of Professor Eby G. Friedman.

During the 2011 and 2013 summers, he interned at Freescale Semiconductor, where he developed methodologies for the extraction of parasitic bipolar transistors, and investigated transient induced faults in silicon-on-insulator technology. During the summer of 2014, he interned at IMEC in Leuven, Belgium, investigating the impact of metallization on power networks for 14 nm, 10 nm, and 7 nm CMOS FinFET technologies. His research interests include resistive RAM, magnetoresistive RAM, and emerging device technologies.

The following publications are the result of work conducted during his doctoral study:

Journal papers

- R. Patel, S. Kvatinsky, E. G. Friedman, and A. Kolodny, “STT-MRAM Based Multistate Register," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* (in submission).

- R. Patel, S. Kvatinsky, E. G. Friedman, and A. Kolodny, “Reducing Switching Latency and Energy in STT-MRAM Caches with Field-Assisted Writing," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* (in press).

- R. Patel, S. Kvatinsky, E. G. Friedman, and A. Kolodny, “Multistate Register Based on Resistive RAM," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, Vol.23, No.9, pp.1750-1759, September 2015.

- R. Patel, E. Ipek, and E. G. Friedman,“2T - 1R STT-MRAM Memory Cells for Enhanced Sense Margin and On/Off Current Ratio," *Microelectronics Journal*, Volume 45, Issue 2, pp. 133 - 143, February 2014.

Conference papers

- I. Richter, K. Pas, X. Guo, R. Patel, J. Liu, E. Ipek, and E. G. Friedman, “Memristive Accelerator for Extreme Scale Linear Solvers," *Proceedings of the Government Microcircuit Applications & Critical Technology Conference (GOMACTech)*, March 2015.

- R. Patel and E. G. Friedman, “Sub-Crosspoint RRAM Decoding for Improved Area Efficiency," *Proceedings of the ACM/IEEE International Symposium on Nanoscale Architectures*, pp. 98 - 103, July 2014.

- R. Patel, E. Ipek, and E. G. Friedman, “Field Driven STT-MRAM Cell for Reduced Switching Latency and Energy," *Proceedings of the IEEE Symposium on Circuits and Systems*, pp. 2173 - 2176, June 2014.

- Q. Guo, X. Guo, R. Patel, E. Ipek, and E. G. Friedman, “AC-DIMM: Associative Computing with STT-MRAM," *Proceedings of the International Symposium on Computer Architecture*, pp. 189 - 200, June 2013.

- R. Patel and E. G. Friedman, “Arithmetic Encoding for Memristive Multi-Bit Storage," *Proceedings of the FIP/IEEE International Conference on Very Large Scale Integration (VLSI-SoC)*, pp. 99 - 104, October 2012.

- R. Patel, E. Ipek, and E. G. Friedman, “STT-MRAM Memory Cells with Enhanced On/Off Ratio," *Proceedings of the IEEE International SoC Conference*, pp. 148 - 152, September 2012.

Acknowledgements

A graduating undergraduate is a lump of clay that is tossed into a chasm of unknown desires, undefined sensibilities, and unintended outcomes, with gravity pulling towards mediocrity and failure. I had a vague inclination of the direction I wanted to go, but I had no idea how to get there. Looking back, I am overwhelmed by the encouragement, patience, support, and effort given to me by my family, teachers, professors and close friends. All of you have helped me progress to where I am now and I am truly grateful.

I will always lack the words necessary to properly thank my advisor Professor Eby G. Friedman. When I began, I was unsure of almost everything. His uninterrupted support and patience gave me an environment in which to inquire, learn, and grow. His insight, experience, and interminable enthusiasm informed me not only in academic pursuits, but also of life—a model of the world that I will heed in career and non-career endeavors. Professor Friedman, thank you for everything.

I must also thank Professor Engin Ipek. When I began, I had little understanding of the research process. In my first project, Prof. Ipek demonstrated the consistency and persistent effort required to do high quality research. I learned a great deal through the many effective collaborations and will apply the same level of effort and creativity in all my future endeavors.

Thank you to Prof. Shahar Kvatinsky, Prof. Avinoam Kolodny, and Prof. Uri Weiser, for constructive feedback on research, and mountains of help during our

numerous successful projects. The friendly and supportive tone allowed our research to progress efficiently. I hope all my future collaborations will mirror the level of teamwork we developed.

I would like to thank Professor Paul Ampadu and Professor Erin Smith for the constructive feedback while serving on my committee. I would also like to thank Professor Michael Scott for serving as committee chair. Prof. Zeljko Ignjatovic, Prof. Marc Bocko, Prof. Ji Liu, Prof. Hanan Dery for thoughtful conversations throughout my stay at the University of Rochester.

The High Performance Integrated Circuit Laboratory consists of a diverse group of highly professional individuals from a variety of cultural backgrounds that made my Ph.D experience both creatively enriching and enjoyable. I would like to thank the previous and current members of the laboratory: Prof. Emre Salman, Dr. Renatas Jakushokas, Prof. Selcuk Kose, Prof. Ioannis Savidis, Dr. Inna Vaisband, Boris Vaisband, Alex Shapiro, Mohammed Kazemi, Kan Xi, Shen Ge, Jinhui Wang, and Albert Ciprut. Days in the lab were typified by scribbling ideas on the whiteboard and arguing, *i.e.*, converting random thoughts into concrete research efforts. I am fortunate to be connected with so many individuals that I respect, admire, and consider as my extended family.

I also want to thank Dr. Olin Hartin, Dr. Radu M. Secareanu, Dr. Dan Blomberg, Dr. Gerald Nivson, Dr. Vance Adams, and Dr. Qiang Li from Freescale Semiconductor, Inc. for helping me throughout two summer internships in Tempe, Az. I would also like to thank Dr. Praveen Ragavan, Odesseas Zografos, and Dimitrios Velenis for enabling a smooth and life changing internship experience at imec, Inc. in Leuven, Belgium. I have learned a great deal from you all and will maintain the bonds of friendship extended to me.

I would like to thank my closest friends. An incomplete list: Aaron Forisha, Alex Lee, Alice Nelson, Andrea Gordon. Ben Bodner, Brian Chia, Brittany McFee,

Caitlin Tennyson Peterson, Chris Hergott, Dan Keeley, Dan Snyder, Danner Hickman, David Leeds, Deven Patel, Diana Lee, Gerald Abt, John Henderson, Katie Karasek, Kayla Molnar, Liz Marilyn, Luck Shay, Marc Karasek, Matthew Storey, Mehdi Naz Bojnordi, Michael Brundige, Michael Willett, Mindy Hoftender, Mohan Ahluwalia, Nate Housel, Nate McBean, Olivia DeDad, (Prince) Ali Valimahomed, Qing Guo, Rani 'stealthMaster' Ghosh, Sarilyn Ivancic, Sheema Shayesteh, Shikha Rawat, Shivani Kumar, Steven Ivancic, Xiaochen Guo, Yang Chen, Yanwei Song, Whether it was late nights grinding at the lab, long phone calls catching up, or just a welcome distraction after a long week, all of you have helped me get through tough moments and enjoy some amazing ones. I've been fortunate to interact with so many capable, quality, and good-hearted individuals.

To conclude, I must thank my family for the unending support and joy throughout my tenure at the University of Rochester. I would like to thank my sister Sally Patel for demonstrating to me the true definition of tenacity while always finding ways to extract a laugh and a smile. I would like to thank my uncles and aunts (Rajesh, Minesh, Bhupendra, Vijay, Bharat) and Aunts (Daksha, Rina, Hina, Bhavana, Jayshree) for always supporting me. I would like to thank my grandparents (Nanu-bhai, Susila-bhen, Karsan-bhai, and Lalita-bhen), for always watching over me.

I would like to thank my parents Dilip and Jayana Patel for demonstrating humility, forgiveness, and unrelenting tenacity in the face of unreasonable circumstances. You are quite simply the reason I exist and my models of what ideal people are.

I would also like to thank my fiancé Henal Patel. While you only recently came into my life, you are the whirlwind storm that washes away the unimportant and makes me feel alive.

Abstract

In less then a decade, memristors have evolved from an emerging device technology, to a promising circuit concept, and now a commercial product. This accelerated development is due to the importance of memristor devices, which provide greater capacity while reducing power and latency in computer memories. The research described in this dissertation explores circuits composed of memristor devices and design methods to enhance the performance of memristor devices in memory systems.

The dissertation begins with an introduction to memristor device technologies. Physical descriptions of metal-oxide resistive RAM (RRAM) and spin torque transfer magnetoresistive RAM (STT-MRAM) are presented. Classic CMOS memory organization and technology are also reviewed.

Several memory cells, memory array topologies, and memristor based circuits are presented. A novel RRAM based flip flop is described. This circuit, coupled with a highly threaded architecture, demonstrates up to 40% improvement in performance with a modest area overhead of 2.5% as compared to conventional microprocessors. An STT-MRAM based cache is described with a magnetic field assistance mechanism that reduces the switching latency of an individual device by four. This cache reduces energy by 55% as compared to an SRAM subsystem, and a 20% improvement over STT-MRAM based caches discussed in the literature. Additional circuits, design methods, and physical topologies are presented to improve

the sense margin, reduce area, and increase bit density of memristor based memories.

Memristor devices are an emerging technology capable of reshaping the computational process. Insight into the design of memristor memories is provided in this dissertation with solutions to improve the performance of memristor based memories.

Contributors and funding sources

The research presented in this dissertation is supervised by a committee consisting of Professors Eby G. Friedman, Engin Ipek, and Paul Ampadu of the Department of Electrical and Computer Engineering, as well as Erin Smith of the Department of Finance of the Simon Business School. The committee is chaired by Professor Michael Scott of the Department of Computer Science.

The author, Ravi Patel, developed novel memory circuits based on memristor devices, design methods to physically structure memory cells and arrays, and simulations to evaluate the performance, power, and energy of each approach. Chapters one to three comprise introductory material based on the literature published by other researchers. The contributions of the co-authors are described below for each chapter.

Chapter 4: Ravi Patel is the principal author of the chapter contributing the circuit design, circuit performance evaluation, and layout. The development of this research was performed in collaboration with co-authors, S. Kvatinsky, A. Kolodny, and E. G. Friedman. The results are published in the *Proceedings of the International Cellular Nanoscale Networks and their Applications*, and *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*.

Chapter 5: Ravi Patel is the principal author of this chapter, contributing the cache design, device switching simulations and energy evaluation, as well as cell

layouts and circuit simulation. X. Guo and Q. Guo provided the architectural evaluation. The research and evaluation are supported by E. G. Friedman and E. Ipek. The results from this study are published in the *Proceedings of the IEEE International Symposium on Circuits and Systems*, as well as the *IEEE Transactions on Very Large Scale Intergration (VLSI) Systems*.

Chapter 6: Ravi Patel is the principal author of this chapter, providing the memory cell circuit design, circuit simulations, and cell layouts. The research and evaluation are supported by E. Ipek and E. G. Friedman. The results are published in the *Proceedings of the IEEE System-on-Chip Conference* and the *Microelectronics Journal*.

Chapter 7: Ravi Patel is the principal author of this chapter, contributing the compression approach, dynamic read and write circuitry, and circuit simulations. The research and evaluation are supported by E. G. Friedman. The results are published in the *Proceedings of the IFIP/IEEE International Conference on Very Large Scale Integration*.

Chapter 8: Ravi Patel is the primary author of this chapter contributing the design and circuit simulation. The research and evaluation are supported by E. G. Friedman. The results are published in the *Proceedings of the IEEE/ACM International Symposium on Nanoscale Architectures*.

Chapter 9: Ravi Patel is the primary author of this chapter contributing the circuit simulations, device simulations, and register layouts. The research and evaluation are supported by S. Kvatinsky, A. Kolodny, and E. G. Friedman. The results have been submitted to the *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*.

Chapters 10 and 11: The concluding and future work chapters are written by Ravi Patel with support from E. G. Friedman.

This graduate work was supported by a Dean's Fellowship from the University of Rochester and by grants from the National Science Foundation under under Grant Nos. CCF-0541206, CCF-0811317, CCF-0829915, CCF-1329374 and CCF-1054179, the Binational Science Foundation under Grant No. 2012139, New York State Office of Science, Technology, and Academic Research through the Center for Emerging and Innovative Sciences, Intel Corporation, Samsung Electronics, Cisco Corporation, and Qualcomm Corporation.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In 1971, an article appeared in *IEEE Transactions on Circuit Theory* in which *memristors* were proposed as the fourth fundamental electrical circuit element [1] in addition to *resistors*, *capacitors*, and *inductors*. Resistors, capacitors, and inductors are classical elements that exhibit a specific relationship among the physical variables of *voltage (v), current (i), charge (q)*, and *flux (ϕ)*. Resistance (R) is $dv = Rdi$, capacitance (C) is $dq = Cdv$, and inductance (L) is $d\phi = Ldi$. A missing relationship exists between $d\phi$ and dq ($d\phi = Mdq$) that was postulated to be the missing memristor. While memristors were understood as a conceptual model for many phenomena, such as neurons, joule heating in resistors, and as circuit elements in various chaotic oscillators, a practical device remained elusive for 40 years. In 2008, a team led by R. Stanley Williams investigated the conductance of titanium oxide thin films and determined that these devices exhibited properties that matched the proposed behavior of memristive devices [2]. The link between a concept fundamental to

electronics and a practical device prompted a surge of interest in resistive memory technologies, some already existing and some entirely novel. This new focus spawned an on-going period of rapid innovation in materials, fabrication, circuits, and architecture.

1.1 What is a memristor?

The discovery of the physical phenomenon behind memristance and the development of devices with a variable resistance dates back two centuries [3, 4] to the thermistor [5] and the electric arc [6]. While a multitude of devices based on thin films, magnetic films, and phase change devices have been developed, as mentioned previously, the concept of memristance was not postulated until the 1970s [1].

A memristor was initially defined as a linear relationship between charge ($d\phi$) and flux (dq), implying that the memristance M is a constant [1]. Later, in 1976, the theory was generalized to incorporate nonlinear relationships $M(q, v, t)$ [7], called memristive systems. The key characteristic of these devices is a hysteric I-V curve that crosses the origin, *i.e,* a pinched hysteresis, as illustrated in Fig. 1.1.

It was postulated that a practical memristive device would arise out of an electromagnetic structure due to the relationship between magnetic flux and charge in

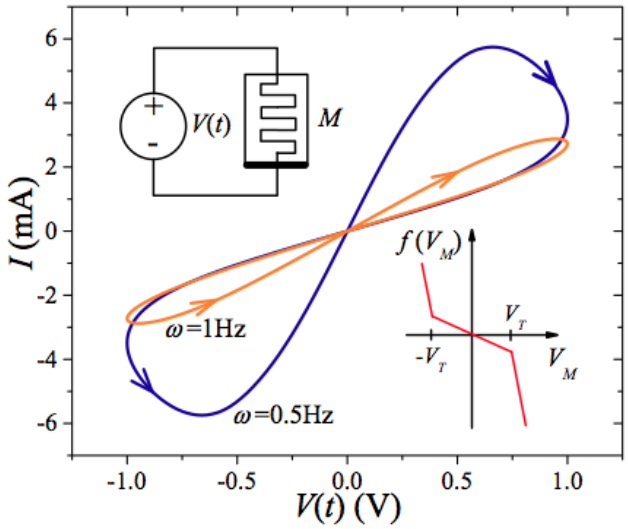


Figure 1.1: Numerical simulation of voltage controlled memristor subjected to a 1 Hz and 0.5 Hz sine wave [8]. The memristor I-V curve exhibits state retention while crossing the origin, a property known as a pinched hysteresis curve.

the mathematical definition of a memristor. In 2008, Williams' group at HP disrupted this notion with the discovery of a pinched hysteresis curve in TiO devices based on ionic conduction [2]. After this "re-discovery" of a practical memristor device, several other resistive memories that had been previously proposed and developed were labeled as memristors [9–13].

Significant controversy exists as to whether a resistive memory technology can be classified as a memristive device. Since the definition of a memristor is fairly broad, Williams and Chua posited that all resistive devices with memory are memristors, arguing that any device with a "pinched hysteresis" curve is a memristor (see Figure 1.1) [14]. This definition allows a wide assortment of disparate material and

physical systems to be classified as memristors such as ferromagnetic metals, insulators, and plastics. Notably, the Hodgkins-Huxley model of a biological neuron cell can be classified as a memristor based on this definition [7]. Other researchers have challenged this notion, arguing that several resistive memory technologies do not satisfy the original mathematical formulation of a memristor [15]. The definition of a memristor is of primary concern for intellectual property ownership. This debate is on-going. For the purposes of this discussion, the terms, memristor and resistive memory, are used interchangeably.

1.2 Memristors as a replacement for CMOS memory

Driving the technological development of computation has been the exponential growth in processing efficiency. This trend has already begun to slow. Mainstream CMOS memory exhibits diminishing returns on performance and cost reduction with each technology generation.

DRAM, the work horse of modern main memory, is increasingly less reliable and more difficult to shrink [16–18]. DRAM storage capacitors require complex fabrication. In 2001, the width-to-height ratio of a DRAM capacitor was approximately 7.6 [19]. The modern equivalent is greater than 100, as illustrated in Figure 1.2, and is predicted to double by 2017 [20]. Achieving sufficient capacitance requires a shift from standard silicon oxides to high-k dielectrics. Greater charge

leakage from DRAM capacitors has prompted more frequent refresh times, leading to increased power consumption and greater susceptibility to error.

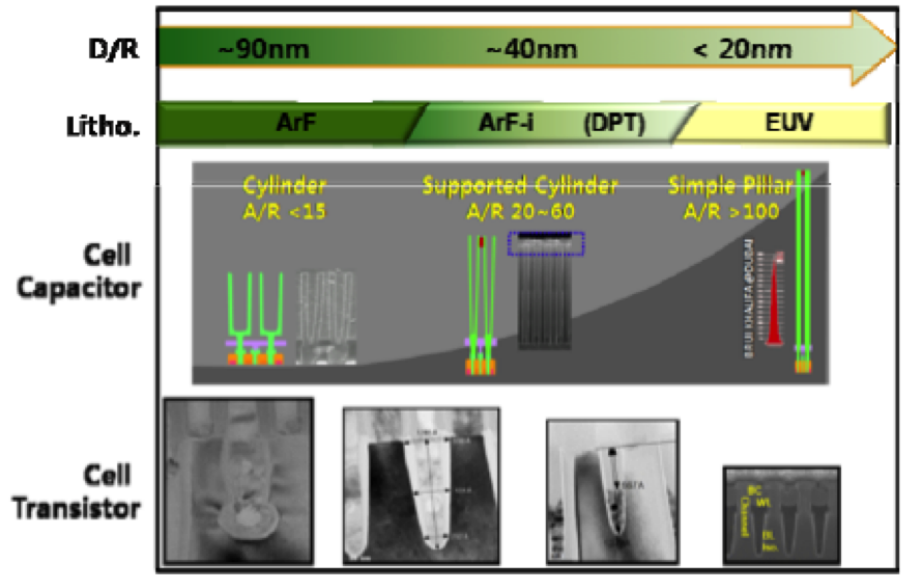
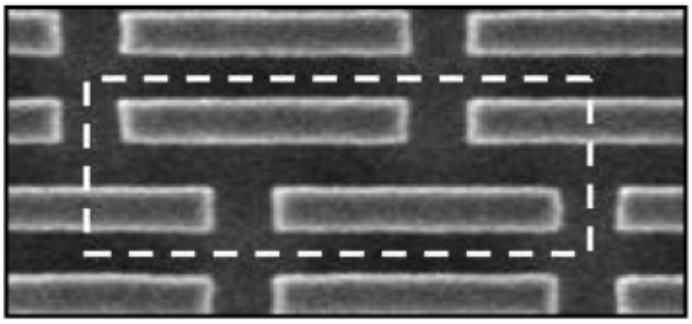
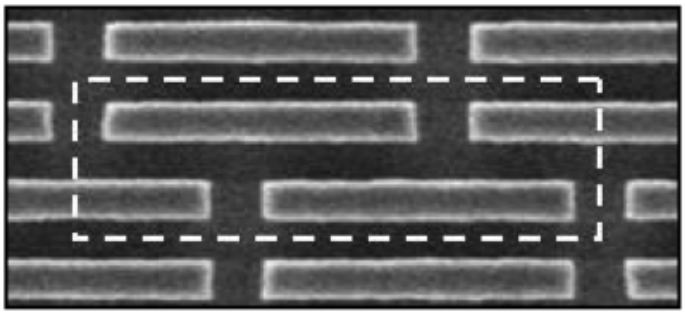


Figure 1.2: DRAM fabrication and scaling [21]

SRAM, extensively used as microprocessor cache memory, is already limited by a slower rate of technology scaling. In 2005, a high density six transistor SRAM cell was used for cache memory on an Intel microprocessor, exhibiting a density of 135 F^2 [22], where F is the lithographic feature size of the technology. By 2012, Intel processors used two different cell layouts for on-chip cache memory. A low voltage variant with a 223 F^2 was introduced to combat leakage current [23,24]. A high density variant exhibited 190 F^2 [23,24] (see Fig 1.3). These cells represent, respectively, a 65% and 41% reduction in density. Moreover, due to cell stability issues, "core memory [was] converted from 6-T traditional SRAM to 8-T SRAM [25]."



(a) 0.092 um^2 SRAM for high density applications

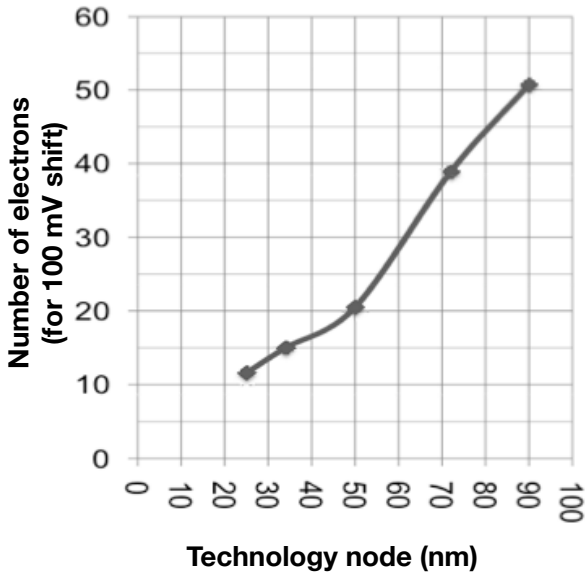


(b) 0.108 um^2 SRAM for low voltage applications

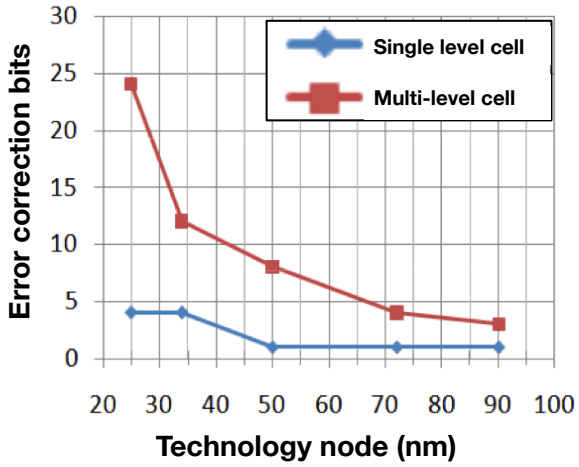
Figure 1.3: Variants of 6-T SRAM cells specialized for a) high density, and b) low voltage operation in 65 nm CMOS [24].

Flash memory, increasingly favored for high density solid state storage, has reached fundamental operational limits. With smaller devices, fewer electrons are available to store a state. As illustrated in Fig 1.4a, the number of electrons required to store a state has shrunk from 50 electrons to nearly single digits [26]. Moreover, system overhead for error correction has exploded, as depicted in Fig. 1.4b.

Memristors provide a possible solution to these issues with semiconductor memories. Fabrication of memristor devices is analogous to thin film deposition of an interlayer via as in standard CMOS processes [27,28]. Cell density is primarily limited by the lithography of the process, and requires only a few additional deposition



(a)



(b)

Figure 1.4: Scaling trends of flash memory [26]: a) the number of electrons required for a 100 mV shift of the threshold voltage within a flash transistor, and b) the number of additional error correction bits required for reliable state storage.

steps. Prototype memristive devices have been scaled to feature sizes beyond what is available with CMOS memories and exhibit greater cell density. A wide variety of CMOS compliant materials are available with memristive characteristics. Many of these materials are already in use in standard CMOS process technologies. As a result, memristors are compatible with existing semiconductor processes. Unlike CMOS memories that operate on charge storage, memristors are non-volatile and therefore do not leak current. The stored resistance of a memristor can also be tuned, supporting multi-bit memories [29].

Leveraging these features, however, requires mitigation of certain limitations of memristive devices. Memristors typically exhibit a long switching time. Due to the large current drawn for the entire duration of a write, switching a memristor typically consumes a large amount of energy. Memristors can also degrade after many write cycles. As a result of these issues and other factors, many material systems are being explored to determine which memristive technologies exhibit more desirable characteristics.

1.3 Outline of dissertation

To address the limitations of memristive technologies while exploring the strengths of each technology, novel memristor-based circuits and design methodologies

are required. Physical and circuit approaches to advance the performance of memories based on memristor technology are considered in this dissertation, leveraging the features of each technology to enhance existing memory systems.

The history, fabrication, and device physics of candidate memristor technologies are reviewed in Chapter 2. A common set of memristor properties are described to categorize each memristive technology. Spin torque transfer magnetic tunnel junctions as well as metal oxide RRAM technologies are considered.

Classic CMOS memory technologies are reviewed in Chapter 3. The circuit operation of SRAM and DRAM is discussed. The memory hierarchy and organization are summarized. Memristors are compared to CMOS to determine which technology is best suited for each level of the memory hierarchy.

In Chapter 4, a novel digital circuit, the multistate register, is proposed. The multistate register is different than conventional types of memory, and is used to store multiple data bits, where only a single bit is active and the remaining data bits are idle. The active bit is stored within a CMOS flip flop, while the idle bits are stored within an RRAM crossbar array co-located with the flip flop. Additional states require an area overhead of 1.4% per state for a 64 state register. The use of multistate registers as pipeline registers is demonstrated for a novel multithreading architecture—continuous flow multithreading (CFMT), where the total area overhead in a CPU pipeline is only 2.5% for 16 threads as compared to a single thread CMOS pipeline. The use of multistate registers in the CFMT microarchitecture

enables higher performance processors (40% average performance improvement) with relatively low energy (6.5% average energy reduction) and area overhead.

A field-assisted STT-MRAM cache is presented in Chapter 5 for use in high performance energy efficient microprocessors. Adding the field assistance reduces the switching latency by a factor of four. A model of an STT-MRAM array is used to evaluate the switching energy for different field currents and array sizes. Several STT-MRAM cells demonstrate a 55% energy reduction as compared to an SRAM cache subsystem. As compared to STT-MRAM caches with sub-bank buffering and differential writes, a field-assisted STT-MRAM cache memory improves system performance by more than 20%.

Novel spin torque transfer magnetic tunnel junction (STT-MTJ) based memory cell topologies are introduced in Chapter 6 to improve both the sense margin and the current ratio observed by the sense circuitry. These circuits utilize an additional transistor per cell in either a diode connected or gate connected manner and do not leak current. An order of magnitude increase in the current ratio as compared to a traditional 1T - 1R structure is observed. This improvement comes with a 61% and 117% increase in area, respectively, for the diode and gate connected cells.

A multi-bit memristive memory circuit architecture based on arithmetic coding is presented in Chapter 7. Both read and write circuits are presented which encode information into the memristive data cells. The proposed circuits provide fine control of the resistance within the memristor. The continuous resistance characteristic

of memristive devices is exploited by utilizing compression techniques to provide additional storage. This approach yields an increase in overall bit density of up to 20 bits per cell for a memristor-based data array as compared to a standard multi-bit cell array.

A methodology for the physical design of RRAM based crosspoint arrays is described in Chapter 8. Two sub-crosspoint physical topologies are proposed that places the RRAM decode circuitry beneath the RRAM crosspoint array. The first topology only integrates the row decode circuitry, while the second topology integrates both the row and column decoders. The topology for sub-crosspoint row decoding reduces area by up to 38.6% over the standard peripheral approach with an improvement in area efficiency of 21.6% for small arrays. Sub-crosspoint row and column decoding reduces the RRAM crosspoint area by 27.1% and improves area efficiency to nearly 100%.

In Chapter 9, the field assisted concept is applied to STT-MRAM to enable fast switching in multistate registers. An STT-MRAM based multistate register avoids endurance issues associated with metal-oxide RRAM devices. The proposed circuit compensates for the stochastic nature of MTJs by introducing additional logic to confirm successful writes and re-attempt on erroneous writes. This evaluation process introduces a small penalty to the overall latency of the system. Introducing field assisted STT-MRAM to the multistate register can significantly reduce the latency of the MTJs, but requires additional area and power as compared to an RRAM

based register circuit.

A brief summary and concluding remarks are offered in Chapter 10. The primary goal of this research effort has been to integrate memristive devices as memory into computing systems while improving system performance. The research has demonstrated improvements in bit density, array area, and read and write latency. Novel circuits have also been proposed that enhance microprocessor performance and employ memristors in heretofore new applications. Ultimately, this research provides insight into the use of memristors in computing systems, and offers methods to enhance system performance with existing and evolving memristor technologies.

Future research avenues are presented in Chapter 11. The suggested topics involve the application of memristor devices for analog circuits and systems. These future research paths provide a mechanism for achieving low variation, high performance analog circuits. Field programmable analog circuits based on memristors are also proposed for next generation computing and signal processing systems.

Chapter 2

Physical Behavior of Memristive Devices

A wide gamut of resistive memory technologies has been described in the literature and developed commercially [30]. In addition to different material systems and structures, these devices exhibit a wide variety of electrical and physical properties. The characteristics of each technology require a novel set of design approaches, both to address any performance limitations and to exploit available features, such as non-volatility and device density. Insight into memristor performance is achieved by assessing each technology in terms of a common set of properties which characterize and contrast the differences among memristors while providing a vocabulary to describe these technologies. General properties highlight the appropriateness of each technology for different applications and provide a roadmap to enable memristor-based memories in modern circuits. Addressing these challenges, however, requires an understanding of the physical mechanisms that govern each technology, as well as models of electrical operation.

Resistive random access memory (RRAM) and spin torque transfer magnetic random access memory (STT-MRAM) represent two modern and rapidly maturing memristive device technologies, both capable of replacing existing CMOS memories in modern integrated circuits (ICs). RRAM is a memory technology based on chemical restructuring in metal oxides. STT-MRAM devices are based on tunneling and physical momentum transfer of electrons within magnetic thin films. Both technologies are a radical departure from traditional semiconductor memories with properties that produce different system constraints, and require novel circuits.

In this chapter, a common set of properties to characterize the operation of memristive devices is outlined in Section 2.1. The physics and operation of spin torque transfer magnetic tunnel junctions are described in Section 2.2. The physics and operation of RRAM are outlined in Section 2.3. Some conclusions are offered in Section 2.4.

2.1 General properties of memristors

Memristive devices exhibit a common set of properties that can be used to compare and contrast the features, strengths, and operation of each technology for different memory applications. In the following section, common electrical properties are outlined in Section 2.1.1. Typical fabrication approaches are described in Section 2.1.2

2.1.1 Memristive device properties

A memristor is a two terminal resistive device represented by the symbol shown in Figure 2.1. The instantaneous resistance of the device behaves as a linear resistor. A large voltage or current bias, however, can change the resistance of the device. This change persists after the bias is removed, allowing the state information to be stored. These devices are therefore non-volatile. The devices can be categorized according to a basic set of properties that describe the limits of non-volatility and performance of memories for each technology.

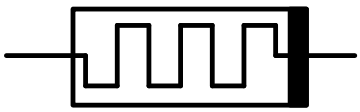


Figure 2.1: Circuit symbol of a memristor

2.1.1.1 Polarity

Memristors are often described by the device polarity, being either *bipolar* or *unipolar*. A bipolar device changes resistive state based on the direction of the applied bias, as illustrated in Fig. 2.2. A positive bias increases the resistance, while a negative bias causes the resistance to decrease. Larger biases typically increase the speed of the change in resistance. The resistance of a unipolar device is modulated by the magnitude of the applied bias. The direction of the applied bias does not

affect the device resistance. A large reset voltage is typically applied to return the device to the initial resistance state. A bias smaller than the reset voltage gradually changes the resistance of the device.

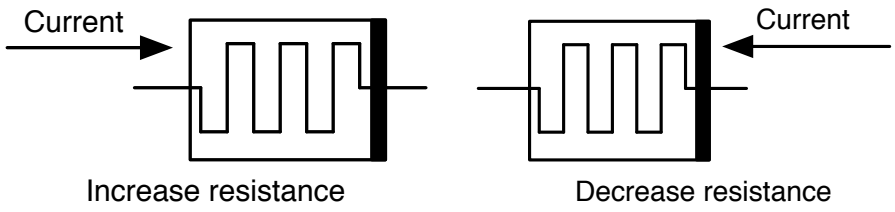


Figure 2.2: Bipolar memristive switching

Unipolar devices are advantageous since a single bias direction can set either state, enabling diodes as selector devices rather than transistors. Avoiding transistors within the memory array requires fewer bit lines, improving cell density. The additional degree of freedom in bipolar devices, however, enables fine grained tuning of the device resistance through a feedback mechanism. If a specific value of resistance is written into a memristor, and the write process "overshoots" this value, a negative bias can be applied to a bipolar resistor to adjust the memristance to the target value. With a unipolar device, overshooting a target resistance requires the device to be reset to the initial resistance before the write can proceed. Bipolar devices can also use tunnel barriers as selection devices, facilitating transistor-less memories [28,31–34]. Tunnel barriers, however, require one to two additional mask steps during fabrication.

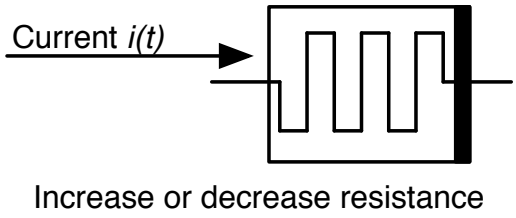


Figure 2.3: Unipolar memristor switching. A single bias direction writes the resistance.

2.1.1.2 Resistance Range and Ratio

Memristive devices and technologies also vary significantly in the range and high/low ratio of the resistance. The resistance ratio is

$$Ratio = \frac{R_{High}}{R_{Low}}. \tag{2.1}$$

Memristors exhibit a maximum (R_{High}) and minimum (R_{Low}) resistance that bounds the range of resistance exhibited by a memristor.

Devices with a large R_{High} are preferable for high density memories to enhance the ability to select a binary state. A larger resistance leads to lower current consumption during reads and therefore lower read energy. Moreover, a high off state resistance enables dynamic sensing of the memristor state. When the resistance is small, current comparators are required to detect the difference in the resistance state. Larger resistances, however, require a longer delay to charge and discharge the bit lines and therefore require a longer time to sense the memristor state.

While a large resistance ratio is generally beneficial, increasing the ratio beyond

a certain level yields diminishing returns. Intuitively, a larger change in resistance provides a larger difference in the on and off currents. For larger off resistances, however, the off current is comparable to leakage current. For smaller on resistances, the on current is limited by the supply voltage of the circuit and the resistance of the peripheral access circuitry. Additionally, a small on resistance requires more time and power to change the resistance. The low on resistance produces a small voltage drop across the memristive device, causing the device to switch more slowly.

2.1.1.3 Continuous vs Discrete Resistance Range

A memristive device is either *continuous* or *discrete*. A discrete memristor has a finite number of stable states, whereas a continuous memristor can occupy any resistance state between R_{High} and R_{Low}.

A continuous memristive device such as RRAM allows multiple levels to be stored within a single device, enabling multi-bit cells and improved memory density. This behavior, however, comes at the cost of greater complexity during reads and writes. Moreover, discrete devices exhibit the ability to relax to a stable state, which can improve reliability.

2.1.1.4 State retention

The stored state of a memristive device is susceptible to *retention* errors, where the stored state becomes unreliable over time. The mechanism of state retention is different from traditional CMOS soft errors and is highly dependent on the memristive technology.

In practical applications, the retention time is dependent on the system requirements. Shortening the retention time can reduce the write latency of a device [35]. Performance improvements resulting from reduced write latency generally improve the write energy of the device, but sacrifice some reliability. Retention errors are typically caused by temperature related processes [36–39]. The retention time is strongly dependent on the operating temperature of the device.

2.1.1.5 Write endurance

A key technology challenge in the development of memristors has been improving the *write endurance*. The write endurance of a memristor specifies the number of write cycles for which the device can be reliably written. Higher endurance devices like STT-MRAM are highly desirable to avoid reliability problems. Many memristive technologies exhibit a degradation in endurance which limits the use of these technologies to storage class applications. RRAM relies on the chemical breakdown and formation of molecular bonds, and exhibits degradation on subsequent writes. State of the art TaO RRAM exhibits 10^{12} endurance cycles before failure [27], which

is not sufficient for write intensive applications.

2.1.2 General fabrication approaches

Memristors are manufactured using patterned deposition of thin films. From a planar perspective, a device is constructed by patterning an oval, circular, or square thin film on a surface. Fabrication is typically limited by the minimum planar dimension of the lithographic technology. As compared to fabricating CMOS transistors, memristor technologies require fewer lithographic steps. Depending upon the material, structure, and fabrication approach, memristive devices require between one and five additional lithographic mask steps [40, 41]. STT-MRAM requires a minimum of three mask steps, while RRAM technologies require just a single additional mask. Proposed memristive prototypes can be stacked on top of traditional CMOS circuits or directly integrated with CMOS processes [42, 43].

Stacked technologies typically allow separate fabrication methods. This capability enables techniques, such as nanoimprint lithography [40, 44, 45], which can produce denser memristive devices. Stacked technologies, however, require precise alignment with the underlying CMOS technology. Denser lithographic technologies are typically not capable of producing circuits in high volume due to the serial nature of these patterning techniques and require substantial investment in manufacturing processes as compared to classic lithographic patterning techniques commonly used in CMOS fabrication processes.

Direct integration with CMOS requires the memristive materials to be compatible with silicon processes. Moreover, memristors are integrated in back-end-of-the-line (BEOL) processes that are primarily devoted to interconnect fabrication in CMOS circuits. Memristors can therefore be placed above the CMOS circuitry between the higher level metal layers, as illustrated in Figure 2.4 . BEOL processes are typically low temperature to avoid affecting the transistors fabricated within the CMOS substrate. The fabrication temperatures of memristor technologies integrated with CMOS must therefore be sufficiently low to not affect the underlying CMOS transistors.

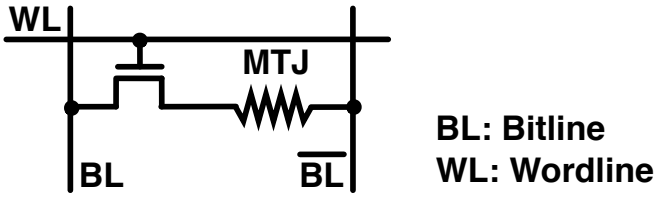
2.1.3 Circuit topology

Many cell topologies have been proposed for use in memristive memories. A cell typically consists of a selector device and a memristor. The selector device isolates the multiple devices connected to the same bit line. CMOS transistors are used as selector devices in a one transistor, one resistor (1T-1R) cell, as shown in Fig. 2.4

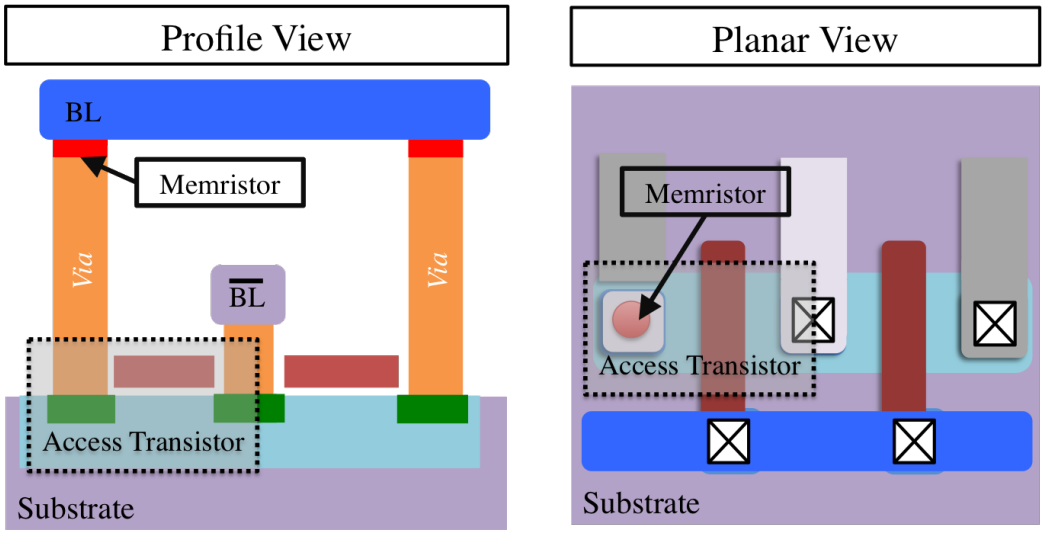
Crosspoint arrays efficiently organize memristive devices into a memory. Crosspoint arrays provide a high cell density by integrating a memristive device at the intersection of perpendicular metal lines on adjacent metal layers, as illustrated in Fig 2.5.

An individual bit is selected by biasing a row and grounding a column within an array through a sense amplifier. Selecting a single device, however, produces a

voltage drop across the unselected rows and columns. In addition to the selected cell, adjacent cells are also biased, causing an additional parasitic current to flow to the ground terminal (see Figure 2.5b). The resultant parasitic sneak currents can propagate through the unselected cells, causing a degradation in sense margin and an increase in power consumption [46]. These currents prohibit the use of memristor-only crosspoint arrays in all but the smallest arrays [46, 47]. Larger arrays utilize a selector device (*e.g.*, a tunneling barrier or diode) in series with the



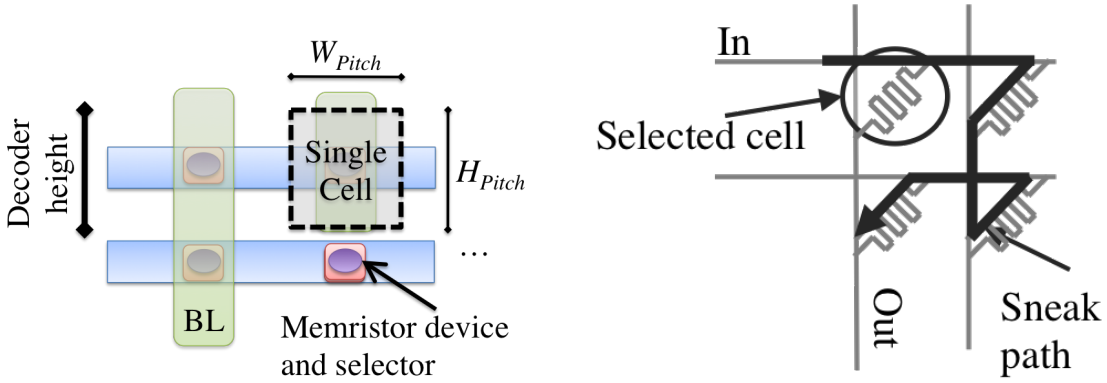
(a) Electrical topology



(b) Physical topology

Figure 2.4: 1T-1R memristor cell

memristor to ensure that only a small (leakage) current passes through the unselected rows [48]. Unlike traditional CMOS memories, crosspoint memories need to be bit addressable. Only a single bit can be written into a crosspoint array during a write operation due to the resistive load of the bit lines in large arrays. This characteristic requires additional area for the peripheral circuitry.



(a) Physical topology

(b) Electrical topology

Figure 2.5: Structure of a 2x2 crosspoint array, a) Physical topology, and b) circuit diagram of 2 x 2 crosspoint array with indicated sneak path.

2.2 Spin torque transfer magnetic tunnel junctions

Spin torque transfer magnetic tunnel junctions (STT-MTJ), the storage elements in STT-MRAM, are two terminal devices that operate on the principle of spin dependent conduction through magnetic domains [41,49–51]. The device is a bipolar

memristor with a small resistance range and a low resistive ratio [41]. MTJs are discrete devices that exhibit either a binary resistance, either high or low, and are particularly applicable to high utilization memories due to the near infinite endurance of STT-MRAM. An MTJ is structured as a stack of thin films where a thin oxide layer separates two ferromagnetic layers, as illustrated in Fig. 2.6 [52]. One of these ferromagnetic layers has a fixed spin polarity (the *fixed or hard* layer) that passes electrons of the same spin direction and reflects electrons with the opposite spin. The other layer (the *free or soft* layer) has a bistable magnetic polarity that is affected by the spin of the incoming electrons. By controlling the direction of the current through the device, either the passing electrons or the reflected electrons influence the free layer. Applying a large bias current to the STT-MTJ (approximately 35 μA to 300 μA) switches the polarity of the device [41, 53, 54]. In the following section, the physical structure, fabrication, switching process, and electrical characteristics of STT-MTJ are described.

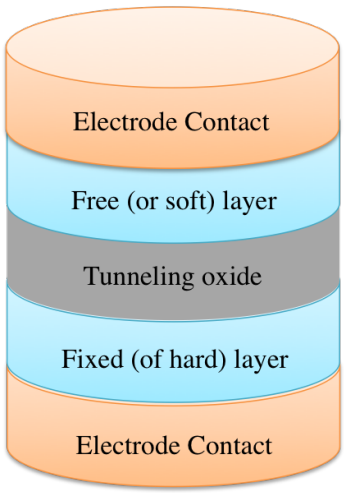


Figure 2.6: MTJ thin film stack

2.2.1 Physical structure and fabrication

Like most memristive devices, MTJ fabrication is based on thin film deposition [55–57]. During a BEOL process, individual MTJs are patterned on to metal contacts, as shown in Fig. 2.7. At a minimum, three layers are needed to construct a device, a ferromagnetic soft layer, an oxide layer, and a ferromagnetic hard layer. Both ferromagnetic layers are structured to act as single domain magnets. The relative hardness and softness of a layer describe the magnetic stability. The softer layer is susceptible to a reversal of polarity while the hard layer has a permanent or fixed polarity. The oxide layer acts as a tunnel barrier that enables resistive conduction through the device. Additional layers are deposited above and below these three layers to control the magnetic polarity during fabrication.

Individual domains are patterned as ovals to control the direction of polarity, as illustrated in Fig. 2.8. The magnetic polarity of both films are physically aligned

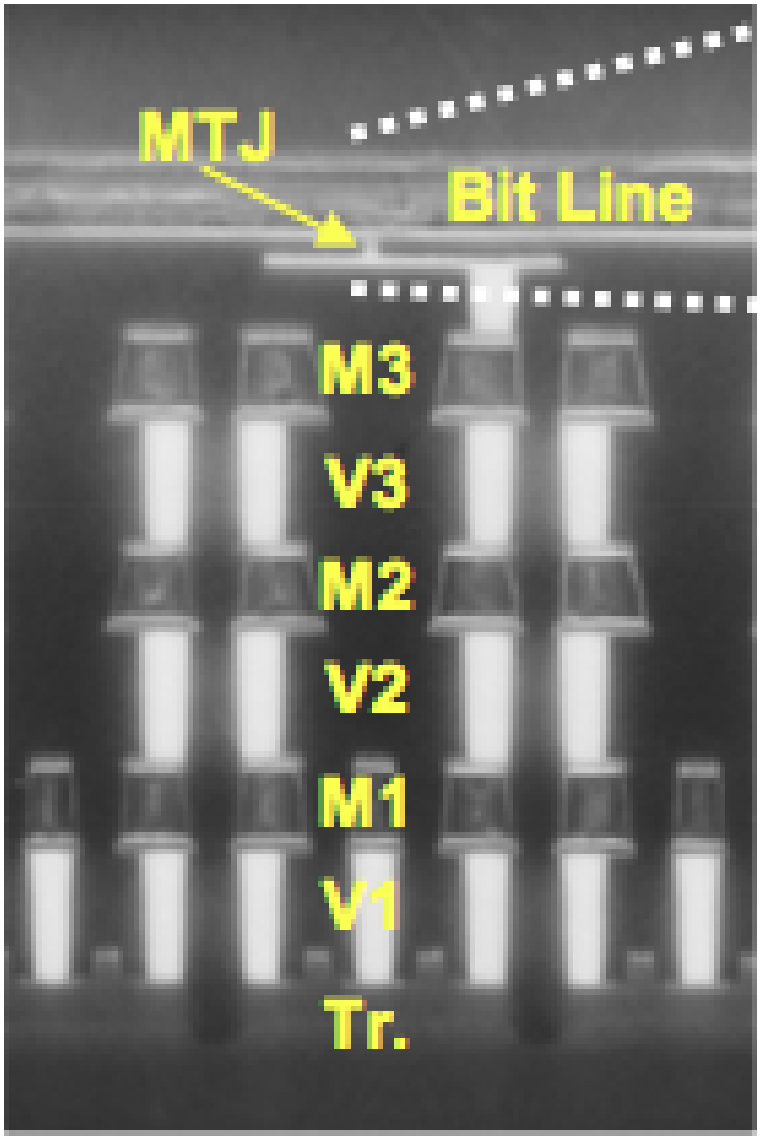


Figure 2.7: First demonstration of STT-MRAM device by Hosomi *et al.* at Sony Electronics [41]. The crosssectional SEM image depicts the MTJ patterned between a bit line and a metal via.

along the long axis of the oval as compared to the short axis, as this direction is the most energetically stable state. The long axis of both domains is aligned to ensure that the polarity of each domain is either parallel or anti-parallel. Parallel alignment refers to the case where both the free layer and fixed layer are oriented in the same

direction, whereas anti-parallel alignment occurs when the free layer is oriented opposite to the fixed layer. MTJs are also categorized as in-plane or perpendicular-to-plane, a distinction discussed in 2.2.2.6.

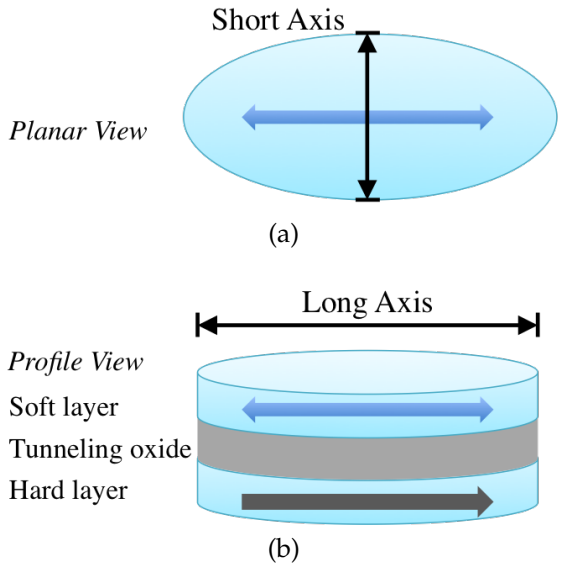


Figure 2.8: STT-MTJ device structure with a) planar, and b) profile views.

2.2.2 Behavior of Spin Torque Transfer Magnetic Tunnel Junctions

The behavior of an MTJ is based on the combination of two phenomenon, spin dependent tunneling and spin torque transfer. Spin dependent tunneling is the physical mechanism that produces a change in resistance in an MTJ under bias. The second mechanism, spin torque transfer, describes the physical interaction of electrons with a ferromagnetic material to facilitate switching.

2.2.2.1 Historical Perspective

The magnetoresistance phenomenon, the change in resistance under an applied magnetic field, traces the discovery to Lord Kelvin (William Thomson) in a letter published in the *Proceedings of the Royal Society of London* in 1856 [4]. The key discovery was that the resistance of a ferromagnetic material is affected by the magnitude and direction of an applied magnetic field. 114 years later, Tedrow and Meservey in 1970 [58] observed spin dependent tunneling at the interface of a ferromagnetic tunnel junction. Julliere later proposed the classic model for the change in resistance (δR) of spin dependent tunneling [59],

$$TMR = \frac{\Delta R}{R} = \frac{2PP'}{1 + PP'} = \frac{R_{ap} - R_p}{R_p}, \tag{2.2}$$

where P and P' represent the spin polarization of electrons in the two ferromagnetic metals of a tunnel junction and the low and high resistance states of an MTJ are described, respectively, as R_{ap} and R_p. The tunneling magnetoresistance ratio (TMR) describes the change in resistance for an individual MTJ stack.

These experiments were conducted at ultra-low temperatures (4 K) and thus had limited practical utility. In 1988, independent teams led by A. Fert [60] and P. Grünberg [61] made the discovery of room temperature Giantmagnetoresistance (GMR) in stacked magnetic monolayers. GMR devices are almost identical to MTJs except that a metallic spacer is used in GMR devices as compared to the oxide spacer

used in MTJs. This discovery forms the basis for all high density magnetic drives, a development for which both teams would win the Nobel Prize [62, 63].

Magnetic tunnel junctions continued to be developed in parallel with GMR based devices. Subsequently, many enhancements have been developed to improve the TMR of individual devices [55, 64, 65]. GMR devices became prevalent in high density hard drives. The relatively small on-conductance of theses devices (on the order of $\mu\Omega$) prohibits the use of GMR devices as a mainstream replacement for high performance DRAM or semiconductor memories. Tunnel junctions, however, exhibit a resistance on the order of $\mathrm{k}\Omega$, which is comparable with the resistance of CMOS transistors. This characteristic has allowed MTJ devices to become an industrial research focus as a potential replacement for DRAM [66–68].

During this era, field switched MRAM was the first MRAM technology to be developed [69]. Similar to modern STT-MRAM, a bit is stored by flipping the magnetic state of the free domain within the MTJ. The switching mechanism, however, is due to the application of current generated magnetic fields.

Crosspoint MRAM arrays were initially explored to provide high density memory arrays. The lack of a transistor or diode to act as a gating element, however, limits the practical use of MRAM crosspoint arrays. Without a selection device, sneak currents, illustrated in Fig. 2.5b, consume the bulk of the energy of a read access, and reduce the observable resistance ratio to noise [67, 70, 71]. As a result, MRAM-based crosspoint arrays were abandoned in favor of memories with one

transistor and one resistor (1T-1R) cells.

Field switched MRAM continued to develop until a key issue limited scalability. To switch the free layer of an MTJ using current induced fields, two large currents are applied perpendicular to the array. At the intersection of these two currents, a single MTJ is switched. Large write currents, typically on the order of milliampheres, require high power. This approach introduced the half-select problem [72–75]. As each current is applied to the array, any unselected cells along the access paths are partially exposed to stray magnetic fields. These cells may inadvertently switch during a write. As a result, large arrays are susceptible to reliability and retention problems. Note that as the technology is further scaled, the magnitude of the magnetic fields did not commensurately change, causing additional MTJs to be exposed to these stray magnetic fields. While field mode MRAM has become a commercially successful non-volatile technology [76], this technology has been relegated to embedded niche applications [66]. STT-MRAM, a second generation MRAM technology entered development during the mid 2000s and is now being considered as a potential replacement for both DRAM and SRAM [77,78].

2.2.2.2 Spin dependent tunneling

This change in the MTJ resistance arises due to the quantum mechanical phenomenon of spin dependent tunneling [79]. An individual magnetic domain acts as a reservoir of spin polarized charge. At the edge of the domain, *i.e.*, the interface

between a ferromagnet and an adjacent material, electrons encounter a potential barrier, as in classical tunnelling. The key difference, however, is that the potential barrier encountered by an incident electron is dependent on the spin orientation.

An individual ferromagnetic domain subjected to a current will act as a spin filter, as illustrated in Fig. 2.9. Those electrons that encounter the surface of the domain scatter more if the electron spin polarity is opposite to the polarity of the domain (anti-parallel electron spin). Electrons with the same polarity scatter less and are more likely to transfer unimpeded through the domain (parallel electron spin). The selective scattering of electrons generates a spin polarized current on the far side of the domain. This behavior can be modeled as two conduction paths, one for electrons with parallel alignment and one for electrons with anti-parallel alignment [69].

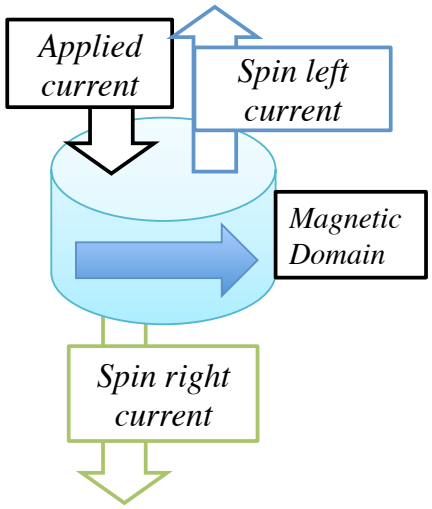
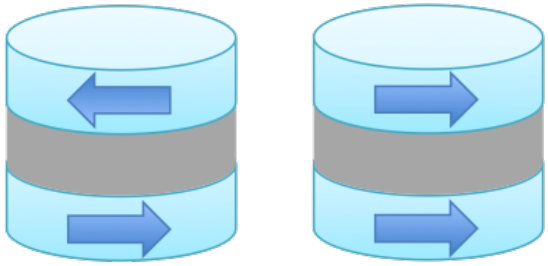


Figure 2.9: Spin polarization of a magnetic domain

An MTJ is a stack of two magnetic domains separated by an insulator, as illustrated in Fig. 2.10. If both domains are the same polarity, the device exhibits a low resistance (R_{low}). With anti-parallel alignment of the domain, the MTJ exhibits a high resistance (R_{high}).



(a) $\mathrm{R_{high}}$ (b) $\mathrm{R_{low}}$

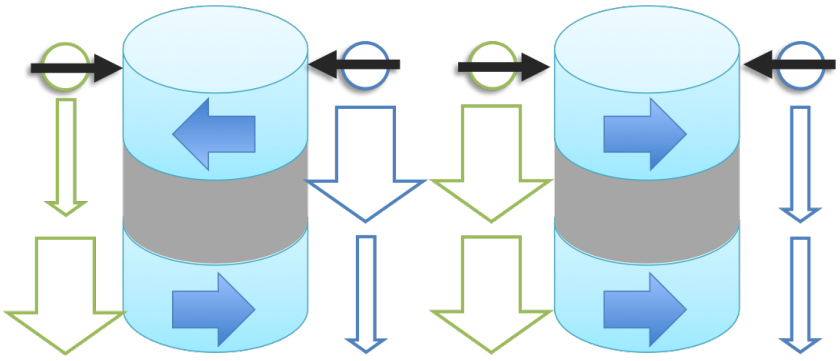
Figure 2.10: STT-MTJ device polarity in the a) high and b) low resistance states.

This characteristic can be understood by considering a single electron, as illustrated in Fig. 2.11. For R_{low}, a parallel electron passes through a domain without scattering, and encounters a second parallel domain. A low likelihood of scattering is exhibited along the conduction path. Note that an anti-parallel electron scatters at the interface of both domains. For an MTJ set to R_{high}, a parallel electron is scattered preferentially by the first domain. Anti-parallel electrons pass through the first domain but scatter at the interface of the second domain. For an MTJ set to R_{low}, only those electrons with anti-parallel orientation scatter, whereas R_{low} exhibits high scattering along both conduction paths.

Slonczewski proposed a model for a tunnel MTJ based on Schrodinger's equation [80]. In this model, the conductance of an MTJ is

$$G = G_0(1 + P_f P'_f cos(\theta)), \tag{2.3}$$

where P_f and P'_f are the current polarization due to, respectively, the free layer and fixed layer, G_0 is the mean conductance of an MTJ at zero bias, and θ is the angle of separation between the magnetization of the free layer and the fixed layer. For example, if the free layer and fixed later are parallel, $\theta = 0°$. If the free layer and fixed layer are anti-parallel, $\theta = 180°$. By solving for the dependence of the conductance on the angular separation between the polarity of the free layer domain, the output conductance of the transient behavior of an MTJ is determined.



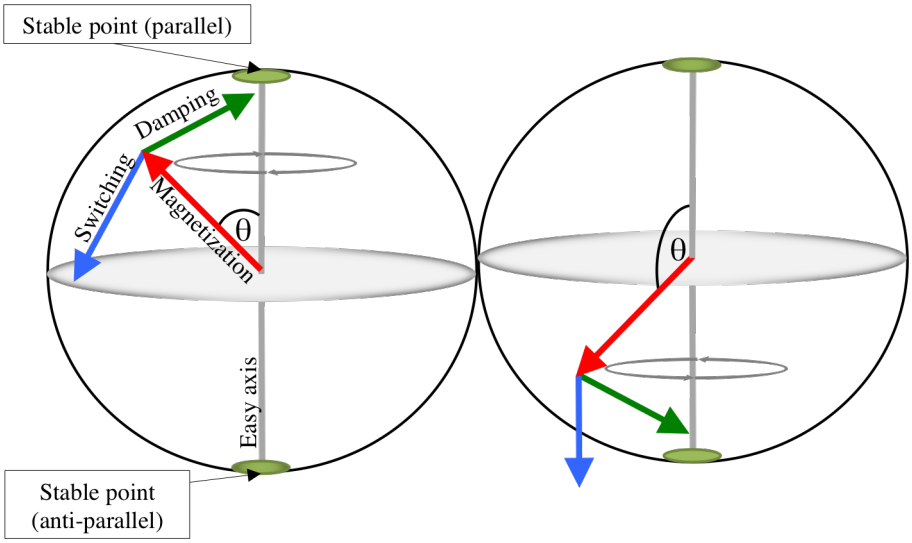
(a) Anti-parallel (b) Parallel

Figure 2.11: Spin dependent electron transmission and reflection in an MTJ in the a) anti-parallel, and b) parallel states.

2.2.2.3 Free Layer Switching Dynamics

Switching an MTJ requires changing the polarity of the free layer. The dynamic behavior of the magnetization polarity is classically modeled as a sphere, as illustrated in Fig. 2.12. The arrow is the magnetic polarity or magnetization of the MTJ. The vertical line spanning the sphere represents the easy axis of the free layer. The magnetization of the free layer is most stable along this easy axis. Intuitively, the two stable points on the sphere correspond to either the parallel or anti-parallel polarity in an MTJ.

If the magnetization deviates from the easy axis, a damping torque acts against the perturbation to restore the torque to the nearest stable point. To switch the MTJ state, a switching torque must overcome the damping torque to ensure that the



(a) Initial state (b) Final state

Figure 2.12: Magnetic model of the free layer of an MTJ during switching in the a) initial, and b) final states.

magnetization crosses the equator of the sphere. If this condition is satisfied, the damping torque switches direction and stabilizes the magnetization at the opposite side, as illustrated in Fig 2.12b.

As the magnetization deviates from a stable point, the magnetization begins to oscillate parallel to the hard axis plane, as illustrated by the circular arrows and the vector $\gamma\left[\mathbf{M} \times \mathbf{H}_{\mathbf{eff}}\right]$ in Figure 2.14. This oscillation is called *precession*. This effect can be ignored in magnetic memories as the oscillation has little impact on the resistance of the MTJ in the final or initial states. The effect is, however, important for other proposed applications of MTJs such as high frequency oscillators [81–83]. The angle θ corresponds to the angle of the magnetization with respect to the easy axis. In MTJs, this angle also corresponds to the angle of separation between the free layer axis and the fixed layer axis. Based on the Slonczewicki's conductance model [84], the electrical switching dynamics of an MTJ are illustrated in Fig. 2.13.

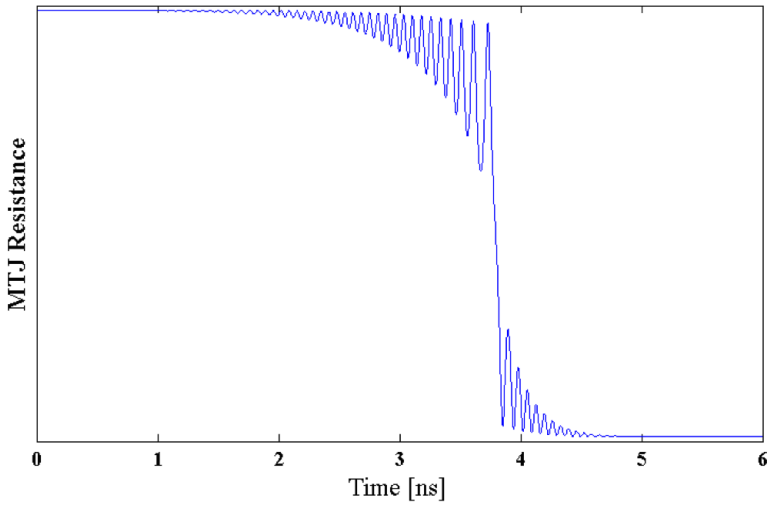


Figure 2.13: Electrical switching behavior of an MTJ

A tug-of-war occurs between the switching torque and the damping torque, as illustrated in the figure. This conflicting behavior causes the MTJ to oscillate until the magnetization finally switches to the opposite polarity.

2.2.2.3.1 Landau-Lifshitz-Gilbert equation of motion A model that describes this dynamic oscillatory behavior of an MTJ is the Landau-Lifshitz-Gilbert (LLG) equation [85, 86],

$$\frac{d\mathbf{M}}{dt} = \gamma \left[\mathbf{M} \times \mathbf{H_{eff}}\right] + \frac{\alpha}{M_s} \left[\mathbf{M} \times \mathbf{H_{eff}}\right] + \tau_{ext}, \tag{2.4}$$

where M is the magnetization vector, M_s is the scalar saturation magnetization, H_{eff} is the effective magnetic field, γ is the gyromagnetic constant, and α is the damping parameter.

The first term in the expression describes the precessional motion of the magnetization around the vertical axis of H_{eff}. The second term describes the damping torque that pulls the magnetization towards H_{eff}, as illustrated in Fig. 2.14. The term τ_{ext} represents an external torque induced on the MTJ. This external torque can be generated either by a magnetic field or other phenomena. The relevant case of current induced torques is described in Section 2.2.2.4.

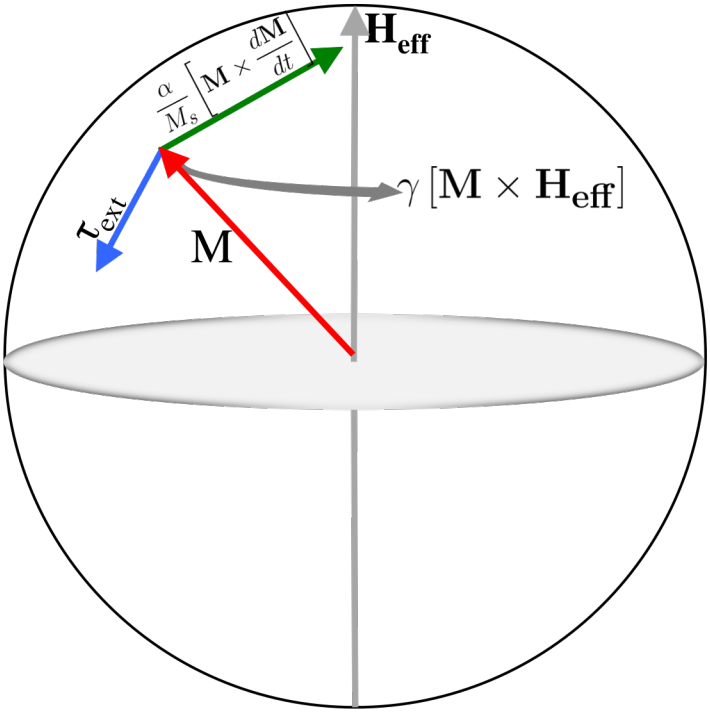


Figure 2.14: Torque components within the LLG equation

2.2.2.4 Spin torque transfer

An external stimulus is required to switch the state of a magnetic domain. Magnetic fields are classically used to switch state. *Spin torque transfer* (STT), a phenomenon first independently proposed by Slonczeweki and Berger in 1996, facilitates current induced switching of a magnetic domain [87,88]. In this phenomenon, electrons incident on a magnetic domain transfer angular momentum to the domain. The transfer of angular momentum exerts a force on the magnetization of the domain, as illustrated in Fig. 2.15. A sufficiently large force overcomes the damping torque of the domain and switches the polarity. A classical current exhibits a

random distribution of electron polarities. A spin polarized current, however, contains a majority of electrons with a net spin. Application of a spin polarized current to a magnetic domain exerts a torque on the magnetic domain.

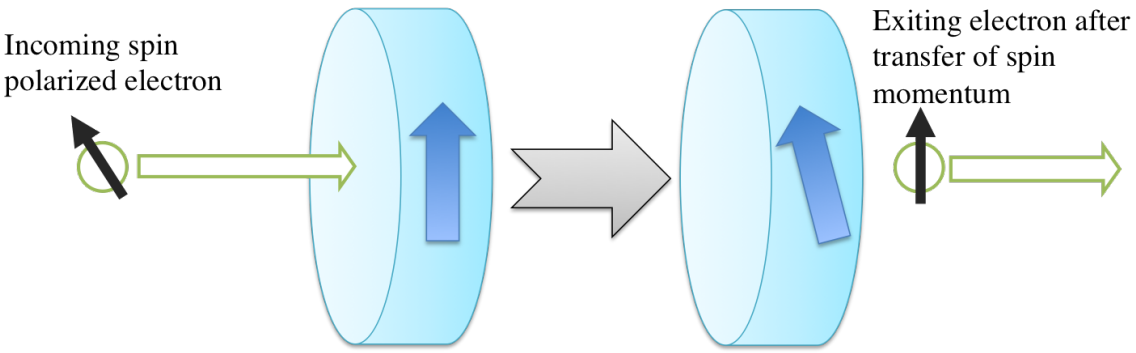
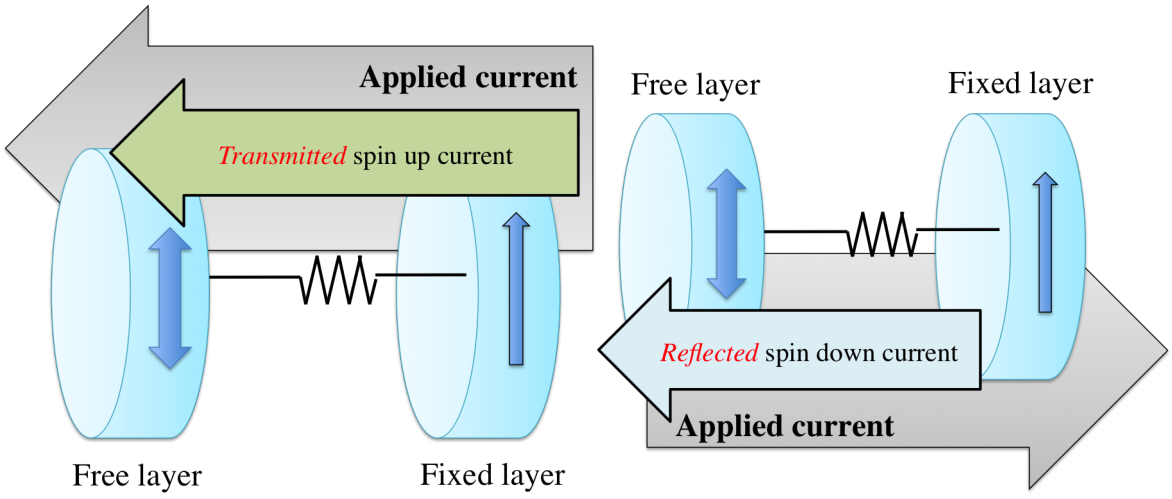


Figure 2.15: Transfer of angular momentum from the electron to magnetic domains.

As described in Section 2.2.2.2, a magnetic domain behaves as a spin filter for an applied current. The current passing through a domain attains a spin polarity parallel to the domain, and electrons spinning with the opposite polarity are reflected off the domain. In this manner, the pinned layer of an MTJ produces a spin polarized current, as illustrated in Fig. 2.16. A current that passes through the pinned layer first attains the spin direction of the pinned layer. If the current is sufficiently large, the free layer switches to the same state as the pinned layer. The process in the reverse direction is similar, except that the reflected spin current interacts with the free layer. Intuitively, the free layer attains a polarity opposite to the pinned layer.



(a) Transmitted (b) Reflected

Figure 2.16: Current induced torques in an MTJ during a) transmission, and b) reflection.

2.2.2.4.1 Landau-Lifshitz-Gilbert-Slonczewski equation Slonczewski modified the LLG expression to account for the contribution of the current induced torque as

$$\tau_s = \frac{\gamma\hbar}{2eM_sV}\mathbf{M}\times(\mathbf{M}\times\mathbf{I_s})\,, \tag{2.5}$$

where $\mathbf{I_s}$ is the current injected into the free layer, and V is the volume of the free layer. $\mathbf{I_s}$ is a vector quantity that describes the magnitude and direction of the net spin associated with the inbound current. Note that the torque is dependent on the crossproduct of the current and the magnetization. While the direction of $\mathbf{I_s}$ does not change, the crossproduct of these terms changes as the magnetization moves farther from the easy axis. This term is added to the classic LLG equation to describe the behavior of a single domain with an applied spin current.

2.2.2.5 Switching statistics and randomness of MTJs

Thus far the magnetization dynamics have been described as a deterministic process, where a current above some critical current induces switching. Practically, the magnetization of a ferromagnetic domain undergoes temperature induced perturbation [89]. This situation is manifested as a random torque applied to the magnetization. This effect is typically modeled using a Langvin random field [90].

The critical current density of an STT-MTJ characterizes the switching statistics of a device. The zero Kelvin critical current density is [91]

$$J_{c0} = \frac{2e\alpha M_s t_F (H_{eff} + 2\pi M_s)}{\hbar \eta}, \tag{2.6}$$

where t_F is the thickness of the free layer, and η is the efficiency at which angular momentum is transfered from a spin current to a magnetic domain. At a finite temperature, the critical current density is [91]

$$J_c = J_{c0}\left[1 - \frac{k_B T}{K_F V} ln\left(\frac{t_p}{\tau_0}\right)\right], \tag{2.7}$$

where τ_0 is the relaxation time, and t_p is the duration of the applied current pulse. Note that $\frac{K_F V}{k_B T}$ is often mentioned as the thermal activation factor (Δ). All of these parameters are material and geometry based. The switching statistics of an STT-MTJ are dependent on this critical current density as well as the amplitude and pulse width of the current applied to the device. In current induced switching, an

MTJ switches into one of two regimes, precessional mode and thermally activated switching [91].

Precessional mode switching is the high speed switching regime (below 5 ns) where the current passing through the MTJ (J) is much greater than the critical current density ($J > 4J_{c0}$). In this case, the current torque is significantly larger than the damping torque as well as the thermally induced random torque. The probability of MTJ switching for a current pulse of duration t_p is [91]

$$P(t_p) \propto exp\left[\frac{H_K M_s V}{2k_B T}(1 - cos^2\phi)\right](J - J_{c0})sin^2(\phi)\phi = \frac{\pi}{2}exp\left[-\frac{\eta\mu_B}{eM_s t_F}(J - J_{c0}t_p)\right]. \tag{2.8}$$

Thermally activated switching occurs when the applied current J is smaller than the critical current density ($J < 0.75J_{c0}$). A sufficiently large random field torque is required to assist the switching process. The switching probability in this regime is [92]

$$P(t_p) = 1 - exp\left\{-\frac{t_p}{t_0}exp\left[-\frac{K_F V}{k_B T}(1 - \frac{J}{J_c 0})\right]\right\}. \tag{2.9}$$

In the intermediate region ($0.75J_{c0} < J < 5J_{c0}$), switching is a combination of precessional mode and thermally activated switching. A closed-form equation for a device operating within this regime has been difficult to produce [91].

2.2.2.6 In-plane vs perpendicular to plane MTJs

STT-MTJs have been developed assuming two basic geometries, a planar device and an perpendicular device. The key difference between these devices is the direction and physical origin of the magnetic easy axis. An in-plane MTJ is patterned to ensure the oval geometry of the free layer produces a magnetic easy axis along the long axis of the oval within the plane of the device. Rather than geometric patterning, a perpendicular device relies on the crystalline orientation of the magnetic thin film to produce the MTJ easy axis.

This geometry reduces the stored magnetic energy of the device, lowering the required switching current. Intuitively, the magnetization of an in-plane MTJ is constrained along the plane of the device. The magnetization of a perpendicular MTJ is free to move in any direction, and the magnitude of the damping torque is therefore smaller.

2.2.2.6.1 Macrospin approximation vs micromagnetic modeling The free layer is assumed to behave as one contiguous unit. Practically, however, every atom in a ferromagnetic material acts as a small bar magnet and can be described using the LLG equation, as illustrated in Figure 2.17a. The assumption of a single contiguous domain, known as the *macrospin approximation*, neglects several physical effects that arise in the ferromagnetic domain.

Micromagnetic modeling of ferromagnetic domains considers the interactions

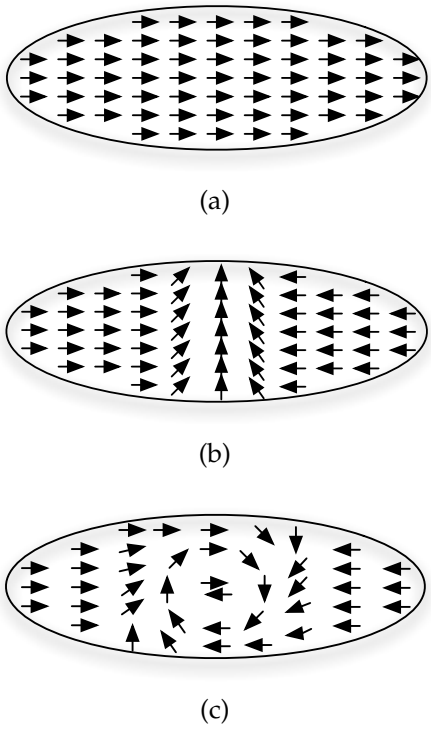


Figure 2.17: MTJ free layer magnetization and macrospin approximation failure states: a) mono-domain, b) domain wall pinning and c) magnetic vortex state

within the magnetic material as well as edge effects caused by the geometry of the domain [93]. Micromagnetic modeling is typically used to gauge the stability of an individual magnetic domain to avoid failure states caused by magnetic forces internal to the layer. One such state may cause the magnetic layer to fracture into multiple domains where the sections of the free layer exhibit different magnetization directions and a domain wall is pinned within the MTJ free layer, as illustrated in Figure 2.17b [94–96]. Another failure condition, known as a vortex state, occurs when the internal magnetic fields cancel, causing the domain to exhibit zero effective magnetization [94]. Each state is illustrated in Figure 2.17.

While numerical analysis of these effects produces a more accurate transient

simulation, the high computational complexity impedes the use of this technique in circuit simulation. A macroscopic simulation of a single magnetic domain typically requires many seconds or minutes, whereas a micromagnetic simulation may require multiple days. Some effort, however, has been dedicated to exploiting these alternative magnetization states for novel devices [97–99]. In MTJs, however, these additional physical faults are modeled by micromagnetic simulation and are considered to be error modes. Furthermore, as the free layer shrinks with device scaling, the macrospin approximation becomes more accurate. The smaller free layer volume reduces the likelihood of domain wall migration [100, 101].

2.2.2.7 Simplified DC model of an STT-MTJ

An STT-MTJ typically exhibits a peak TMR between 80% to 150%, corresponding to roughly a 100% (or 2x) change in resistance. The peak TMR is determined with a near zero voltage bias across the MTJ, which decreases with increasing voltage across the device [102].

An STT-MTJ, however, cannot be treated as an ideal resistor. These devices maintain a voltage dependent resistance that significantly lowers R_{OFF} with increasing bias. This effect can be modeled as an effective TMR,

$$TMR(V_{MTJ}) = \frac{TMR_0}{1 + \frac{V_{MTJ}^2}{V_h^2}}, \tag{2.10}$$

where V_{MTJ} is the voltage across the device, and V_h is the voltage bias across the MTJ where the TMR is reduced by 50% [103]. The bias degradation in the TMR is primarily observed when the device is in the anti-parallel state (R_{off}); therefore, R_{on} is typically assumed to be constant [104]. This basic model captures the DC operation of an MTJ and is valid in all cases where V_{MTJ} is less than the minimum write voltage of a device. Due to this bias dependence, the sense margin is degraded as compared to the ideal case. Notably, the transient characteristics of an MTJ have little effect on the observed sense margin. The switching process of an MTJ is a discrete, random event where the resistance settles to either R_{on} or R_{off}.

2.3 Metal Oxide RRAM

Resistive random access memory (RRAM) is a memristor technology based on modulated conduction through a metal oxide. The device is either a bipolar or unipolar memristor with a large resistance range and a large resistive ratio. RRAM devices are continuous devices, which are of particular interest for high density memories due to the interest in multi-bit memories and endurance lifetimes that exceed NAND flash devices [105, 106].

These devices operate on the principle of dopant migration through the crystal lattice of a metal oxide. Applying a voltage bias in the positive (or negative) direction increases (or decreases) the resistance of the device. In the following section,

the physical structure, fabrication, switching process, and electrical characteristics of RRAM devices are described.

2.3.1 Historical perspective

Resistive switching has been observed in oxide insulator films as early as the 1960s [107, 108], including TiO based devices in 1968 [109]. Much research was conducted on these devices including the "filamentary model" [110], which presented the first description of resistive switching based on filaments. This research focus, however, stalled due to the immature fabrication methods of the era and the commercial success of DRAM and SRAM for semiconductor memories.

Since this period, confusion has arisen in nomenclature as many different physical materials and mechanisms have been labeled as RRAM devices, including magnetic materials [111], reduction-oxidation (redox) memories [112], as well as metal oxide based devices. Alternative names such as OxRAM and ReRAM have muddled the literature as well [27]. Consensus, however, has emerged that the monicker, RRAM, refers to metal oxide memristors.

Binary oxides, such as HfO_2, TiO_2, and TaO, have been the focus of recent research activity. This interest is due in large part because of materials compatibility and low cost in integrating these metal oxide materials with existing CMOS fabrication processes. Significant research and development exist for these materials as

alternative dielectric materials for high-k CMOS gates [113–115] and DRAM capacitors [116, 117]. Much of this research has been repurposed to develop fabrication processes for RRAM devices.

In 2004, Samusung presented a 1T-1R RRAM "[confirming that RRAM] is highly compatible with the conventional CMOS process such that no other dedicated facility or process is necessary" [118]. Recent efforts have focused on evaluating different material systems to determine the metal oxide technology most suitable for commercial production.

2.3.2 Physical structure and fabrication

RRAM devices are patterned as simple thin film devices stacks, as described in Section 2.1.2. Early devices used a simple three layer structure, with an oxide layer sandwiched between a metal capping layer and an electrode [27], as illustrated in Figure 2.18. The memristor behavior occurs within the defect rich oxide layer, which is typically amorphous or polycrystalline. The metal capping layer provides a reservoir from which oxygen vacancies are extracted and stored. Each stack is patterned on a metal via between two metal layers.

As compared to other memristive devices, many RRAM devices require an initial *electroforming* step to initialize the device [27, 119]. Electroforming is physically equivalent to a soft dielectric breakdown process in the context of gate and DRAM dielectric reliability [27, 119]. In this process, a high electric field is applied across

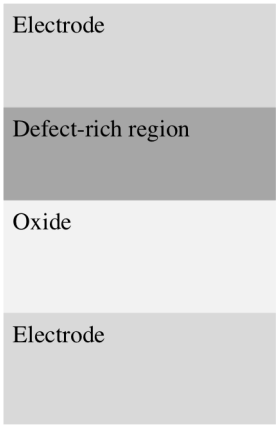


Figure 2.18: Material stack of RRAM devices

the oxide, placing stress on the material. Under high electric fields, oxygen atoms migrate to the edge of the oxide, forming chains of oxygen vacancies within the dielectric. These chains, referred to as *conductive filaments*, are electrical conduction paths [120–122]. Conductive filaments typically form along grain boundaries. To create these conductive filaments, a formative step is required, where a high voltage is applied across the memristor. Writing to the memristor is a process that either breaks or reconnects these filaments. This initial high voltage formative step reduces the voltage for subsequent filament formations during following writes.

Devices have been fabricated that either reduce the required voltage or avoid the formative step [123–125]. A procedure has recently been developed for forming free HfO_x devices [126]. It is anticipated that this additional step will be avoided with further manufacturing developments [123–125].

2.3.3 Behavior of RRAM switching

RRAM switching can be generally understood as controlled and reversible oxide breakdown. Switching an RRAM is based on one of two processes: 1) filament formation (set to R_{low}), and 2) filament breaking (reset to R_{high}). These processes change the potential barrier of the material and alter conduction through the device. The mechanisms behind each of these effects are described in the following section.

2.3.3.1 Filament formation (set to R_{low})

The formation step is physically similar to the electroforming step, as described in Section 2.3.2. High electric fields ($> 10\ MV$) [27] are applied to the oxide, producing a conductive filament across the oxide. Note that the voltage applied to an RRAM during a write is much smaller than the applied voltage during electroforming.

This formation process is physically similar to soft dielectric breakdown [127]. Under a voltage stress, oxygen vacancies drift from the enriched electrode into the oxide, as illustrated in Fig. 2.19. Individual oxygen vacancies migrate in the oxide during the drift and diffusion mechanisms, analogous to carrier migration in CMOS semiconductors [128], although the carrier mobilities are many orders of magnitude smaller than silicon. The applied electrostatic potential and the electrochemical potential form gradients in the material that give rise to ion migration (see

[128,129] for a complete treatment). Chains of oxygen vacancies form and produce a low resistance path through the oxide.

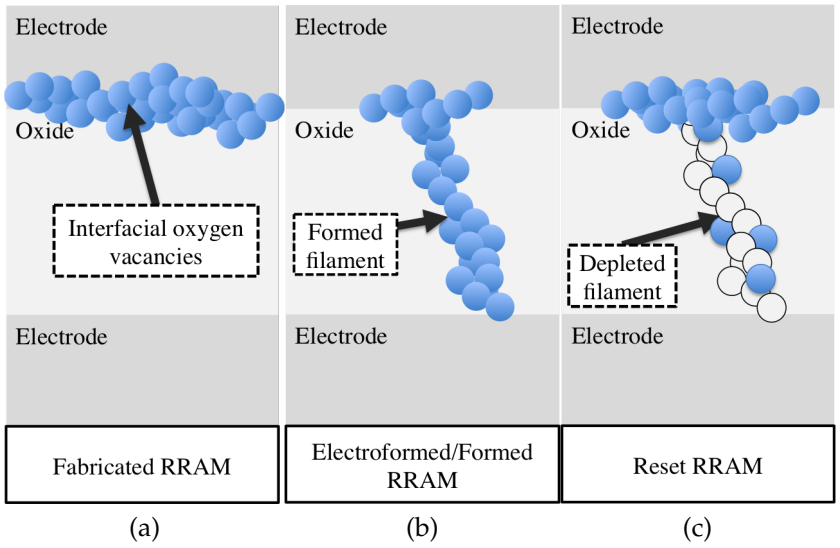


Figure 2.19: RRAM state after a) initial fabrication, b) filimant formation, and c) reset.

2.3.3.2 Filament destruction (reset to R_{High})

The physical mechanism behind filament destruction is a controversial topic [27]. A complete physical model of switching for both unipolar and bipolar devices remains elusive. It is known, however, that oxygen migration to break the filaments is the mechanism behind switching, as illustrated in Fig 2.19c. Controversy exists as to whether electrical or thermal mechanisms are the core mechanism for oxygen vacancy migration [27]. High temperatures are thought to enhance oxygen migration. Local heating in a junction is predicted to increase by hundreds of degrees due to the large current flowing through a relatively narrow filament

[37,130]. The electric fields applied during switching exceed tens of megavolts due to the small size of the device. The switching mode of the device (unipolar versus bipolar) is dependent on the electrode type. Devices with non-oxidizing electrodes exhibit unipolar behavior. If an oxidizing electrode is introduced, these devices exhibit bipolar behavior; Ti or TiN are typical electrode materials [27]. Many physical models have been proposed [131,132], however, consensus has yet to be reached.

2.3.3.3 Conduction through oxide films

While the memristive switching mechanism remains controversial, the conduction mechanism is well known. An oxide RRAM acts as a metal-insulator-metal tunneling barrier. Consider a biased MIM potential barrier. Several paths exist for an electron to tunnel into the cathode, each of which is illustrated in Figure 2.20 [27]. Schottky emission processes elevate electrons to the conduction band through thermal activation [133,134]. At high electric fields, Fowler-Nordheim (F-N) tunneling allows electrons to tunnel through the reduced triangular portion of the barrier [135,136]. Direct tunneling is classical tunneling through a potential barrier which occurs when a potential barrier is sufficiently thin [135]. Oxygen vacancies also behave as electron traps to enable conduction through the oxide. The dominant conduction mechanism is highly dependent on the properties of the material (*e.g*, bandgap, crystallinity), fabrication conditions (defect density, annealing temperature), and the region of operation.

2.3.4 Simplified model of RRAM

Metal oxide RRAM (*e.g.* Ta_2O_5, HfO, TiO_2,) generally exhibit two physical mechanisms to conduct current. Series conduction TiO_2 material systems [137] exhibit resistive switching based on metal filaments that protrude across an insulator. Memristive behavior is caused by migration of oxygen vacancies that either form an electrical filament to "short out" the oxide and create a low resistance path, or cause an existing filament to "break," creating a potential barrier. I-V characteristics follow the description of a classical metal-insulator-metal (MIM) tunnel diode,

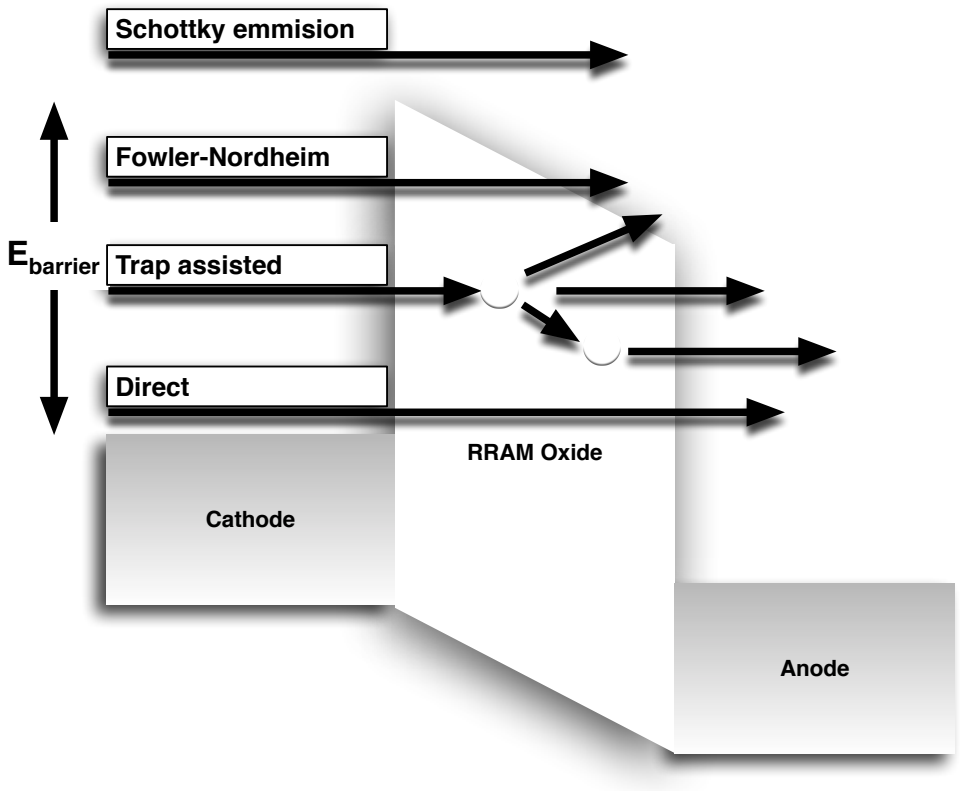


Figure 2.20: Tunneling mechanisms in metal-insulator-metal junctions. [27]

where the barrier thickness is modulated, as illustrated in Fig. 2.21. Intuitively, a gap forms between the end of the filament and the electrode. This gap serves as an insulator within the diode. These technologies utilize an initial electroforming step to produce a complete filament before the diode-like memristive behavior is realized [138].

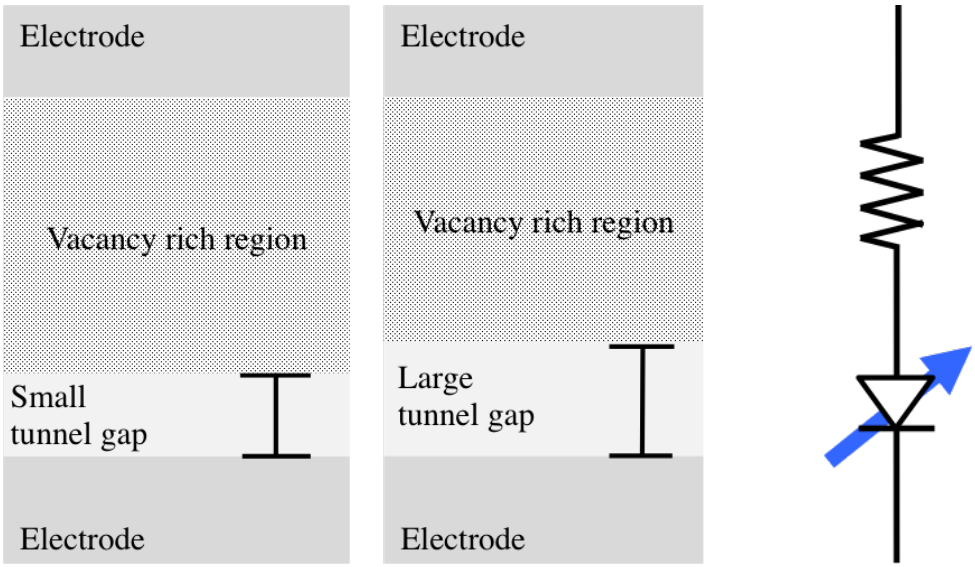


Figure 2.21: Model of series conduction RRAM. The conductance of an RRAM changes from a) a high conductance, small band gap structure to b) a low conductance large tunnel gap state. This behavior can be modeled as a resistor in series with a diode with a variable conductance, as shown in c).

Oxides based on parallel conduction [137], such as TaO and HfO, exhibit a more linear conduction characteristic. In these devices, the source of memristive behavior is the cross-sectional area of the filament, as depicted in Fig. 2.22. As the filament area approaches zero, the thin film behaves like a classic metal-insulator-metal (MIM) diode. As the filament area increases, the device operates as a linear

resistor. The device is therefore modeled as a MIM diode in parallel with a resistor, where the memristive state changes the relative current contribution of the resistive path as compared to the diode path.

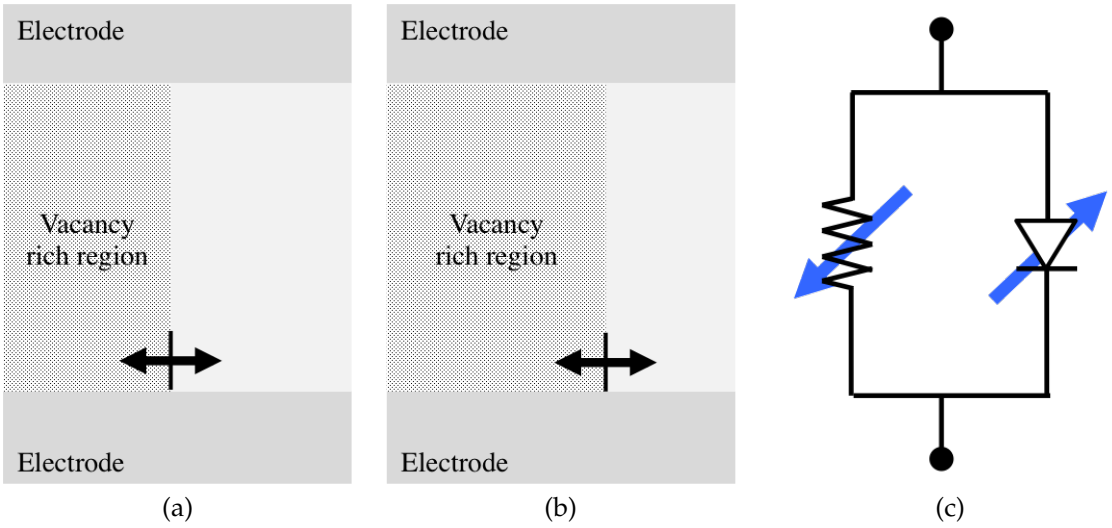


Figure 2.22: Model of parallel conduction RRAM. The conductance of an RRAM changes between a) a more resistive state structure to b) a more nonlinear state. In c), the electrical model contains two parallel paths where the magnitude of the conductance of the resistor and the diode is dependent on the area of the electrode terminal.

Intuitively, a series conduction mechanism is weakly dependent on the area of the device, as an individual filament dominates the conduction process, and only a single filament is required. A technology based on parallel conduction is bounded by the area of the conduction channel. The maximum on-resistance is bounded by the planar area of the thin film.

2.4 Conclusions

Both metal oxide RRAM and STT-MRAM have undergone significant research and development. These memristive technologies exhibit different properties that affect the use of these technologies in memory applications. STT-MRAM is a discrete bipolar memristor with high endurance and state retention, but suffers from a low resistance and resistance ratio. RRAM is a continuous bipolar or unipolar memristor with a high resistance ratio and state retention, but suffers from relatively low endurance. The significant difference in performance of these technologies stems from the physical mechanisms that govern each device. STT-MRAM has evolved from decades of research into magnetic memory devices. Resistive RAM is based on material modulation of metal oxide thin films scaled to small dimensions. As a result of these differences, each technology has a different set of circuit models, a different set of relevant characteristics for memory systems, and are generally considered as competitive technologies in the memory marketplace.

Chapter 3

CMOS and the memory hierarchy

The early portion of the 21st century has been dominated by the collection, analysis, and proliferation of data for insight into natural processes and human endeavors. These systems rely on ever more sophisticated memory circuits and systems to provide efficient storage and access for computational engines. With the evolution of CMOS transistor technology, DRAM and SRAM have proliferated as run time memory for computers, while magnetic disks have served as secondary storage. Modern memory organization and circuitry have evolved to leverage the strengths of these technologies. Resistive memories, poised to replace these technologies, will (at least initially) need to fit into the classic memory organization based on semiconductor memories. The process of replacing SRAM, DRAM, and magnetic disks with resistive memories and the organization of the memory hierarchy and access circuitry are described in this chapter, concluding with a discussion of how

the different memristive technologies can potentially replace semiconductor memory technologies.

3.1 History of memory systems

Modern microprocessor systems originate from Jon Von Neumann's description of the Electronic Discrete Variable Automatic Computer (EDVAC), in "First Draft of a Report on the EDVAC" [139]. This unpublished, unfinished technical report outlined the basic computational structure of a stored program computer. "First: [a computer] will have to perform the elementary operations of arithmetics most frequently... therefore [it is] reasonable that it should contain specialized organs for just these operations." "Second: The logical control of the device, that is the proper sequencing of its operations can be most efficiently carried out by a central control organ." "Third: Any device which is to carry out long and complicated sequences of operations (specifically of calculations) must have a considerable memory." "Fourth: The device must have organs to transfer (numerical or other) information from [an external recording medium] into [its specific parts]. These organs form its input..." Fifth: The device must have organs to transfer (presumably only numerical information) from [its specific parts]. These organs form its output..."

Drawing an analogy to the human body, the design of the EDVAC envisioned a computer system that separated processing, memory, communication, and storage into independent discrete systems. EDVAC, a computer composed of magnetic

tape, diodes, and vacuum tubes, was one of the first binary electronic computers. When discrete transistors and later integrated circuits were developed, these systems adopted the basic structure of the EDVAC. An unintended consequence of adopting this system organization, however, was the "Von Neumann bottleneck."

"There is an old network saying: Bandwidth problems can be cured with money. Latency problems are harder because the speed of light is fixed—you can't bribe God" [140]. In early computer systems, memory access and arithmetic processing developed at approximately the same rate. As IC technology advanced, communication between the memory and the CPU became the primary performance constraint. To alleviate this problem, Intel x86 processors, beginning with the x386 series, supported board-level SRAM cache to enhance performance [141]. The next generation x486 processor included on-die SRAM with an additional off-chip SRAM cache, representing the first cache hierarchy in modern computer systems. Cache memory was introduced and evolved to reduce the access penalty to main memory, easing the Von Neumann bottleneck.

The introduction of cache memory coincided with the "frequency war" era of microprocessors, as illustrated in Fig. 3.1. Early systems began as single contiguous memory with local registers for arithmetic processing, connected by a single main data bus. As transistors shrank and device count per die began to surge, processors began to operate at higher frequencies. Higher on-chip frequencies, enabled

by shorter on-chip interconnects, operated much faster than board level interconnects. In contrast, off-chip interconnect exhibited relatively little change in physical lengths, remaining orders of magnitude longer and slower than on-chip interconnections.

Initially, the cache memory was placed on a separate die on the same board or package as the microprocessor, physically closer to the microprocessor than main memory. As the disparity between processor speed and DRAM access time continued to diverge, additional levels were added to the cache hierarchy (see Figure 3.1). Level one (L1) caches are physically close to or within the microprocessor pipeline and are generally small. Level two (L2), level three (L3), and so on are additional layers of memory that are placed "closer" to the main memory. These cache

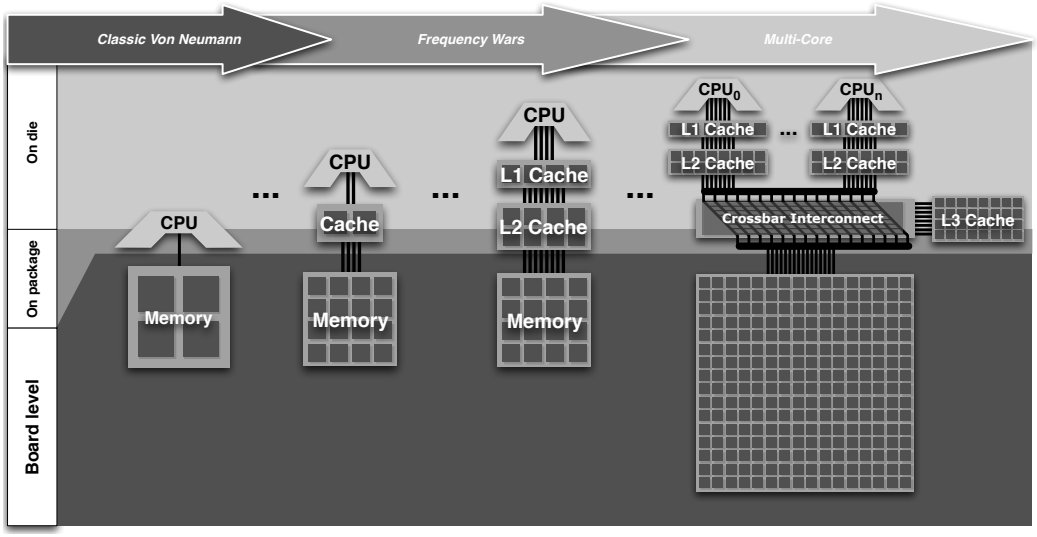


Figure 3.1: Evolution of memory hierarchy with CMOS scaling

memories exhibit both increased capacity and latency since these memories are farther from the microprocessor than the L1 cache memory. To support these cache memories and the increased performance of processor cores, buses also grew wider to provide greater data bandwidth [142–144]. The frequency wars were stymied, however, by increased unpredictability in fabrication and by power limits [145,146].

The modern multicore era arose as a method to increase IC performance, despite frequency limits, introducing novel software and hardware paradigms [145]. For memory systems, this change introduced additional complexity as individual cache memories needed more complex logic to arbitrate buses and data within the cache memory as well as a more striated memory system [147, 148]. Scratchpad memories [149–151] were introduced to complement the traditional memory hierarchy. The increasing complexity and size of new and different applications demanded larger cache memories to operate more efficiently.

The increased transistor count and density has, traditionally, supported greater on-chip cache and main memory capacity as well as increased speed (reduced access time), managing data demands under the constraints of the Von Neuman bottleneck. Both DRAM capacity and on-chip SRAM capacity have grown at an exponential rate, as illustrated in Figure 3.2. DRAM memory capacity increased in size upon the introduction of the one transistor, one capacitor (1T) DRAM in the 1970s through the 2000s. During the 2000s, DRAM production focused on smaller dies with the same nominal capacity of approximately 4 Gb per die to increase process

yield. Systems continued however to increase memory by incorporating more ICs at the board level. The first 8 Gb DRAM die was demonstrated in 2009 with an IC area of 98.1 mm^2 [152]. In contrast, the first 4 Gb die was demonstrated in 2000 with a 650.1 mm^2 area.

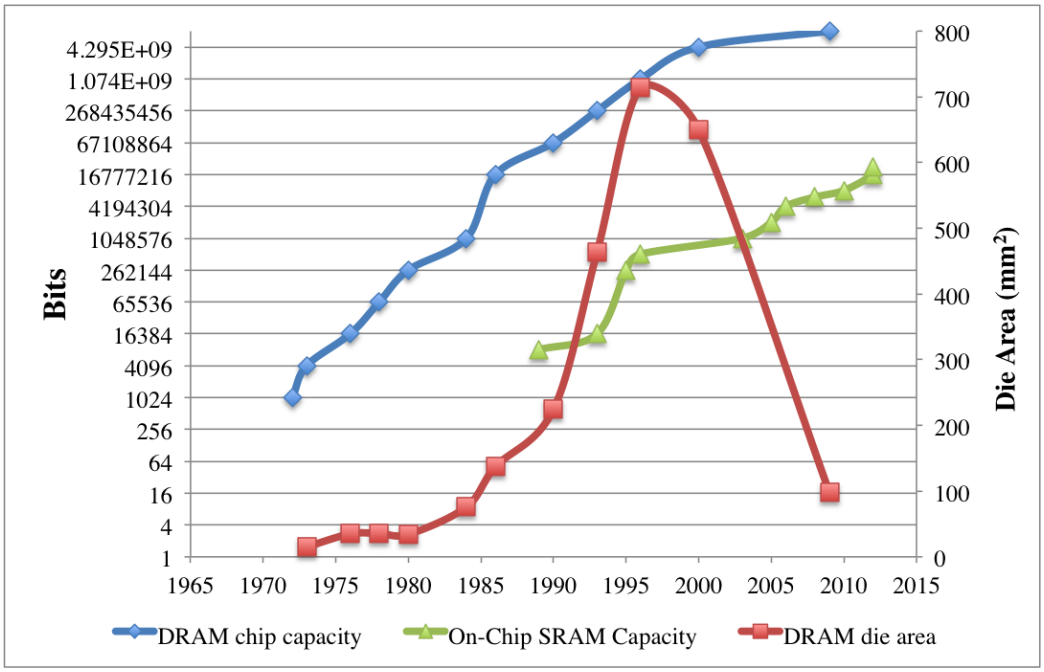


Figure 3.2: Evolution of DRAM and SRAM memory capacity [152–164].

The scaling complexity of DRAM and SRAM, however, are now likely to derail the continuous improvement of memory capacity and performance. From the introduction of on-chip SRAM in 1989 to the mid 2000s, cache memory capacity has doubled approximately every two years. Since 2006, on-chip SRAM capacity has required, on average, 3.3 years to double in size. As discussed in Section 3.2.2, scaling of SRAM has slowed down while DRAM has become extremely difficult to further scale.

3.2 Overview of CMOS Memories

DRAM and SRAM serve as the primary memory in modern computer systems. Flash memory is a non-volatile technology that serves as secondary storage. An overview of each memory technology is presented in this section.

3.2.1 Dynamic random access memory (DRAM)

DRAM, invented by Robert Dennard of IBM [165], served as the work horse of main memory since the introduction of the Intel 1103 in 1971 [166, 167]. A DRAM cell consists of two basic circuit elements, a transistor and a single capacitor, as illustrated in Figure 3.3.

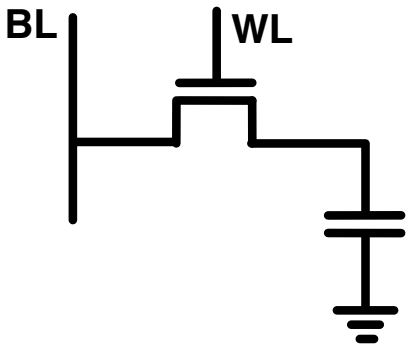


Figure 3.3: DRAM cell.

An individual cell stores a state by storing charge on a capacitor. Writing a logic '1' requires biasing the bitline (BL) to a high voltage and turning on the selection transistor via the word line (WL). The cell capacitor charges until the cell reaches

the bitline voltage. The process for a logic '0' is the same but BL is biased to ground, sinking any charge that may remain on the capacitor.

The process for reading a DRAM cell begins by charging BL to half the supply voltage and connecting the select transistor. Once the select transistor is turned on, the cell either pushes or pulls charge from the bitline, causing the voltage on the line to, respectively, increase or decrease. A high stored state sources charge and a low state sinks charge. This voltage change is compared to a reference voltage to determine the DRAM state. An adjacent line without a connected cell is typically precharged and provides the reference voltage.

Two system level issues arise with DRAM cells. First, a read operation deletes the logic state from the cell, *i.e.*, reads are *destructive*. Reads must be immediately proceeded by a *write back*, which rewrites the data bit back into the cell. Second, transistors are imperfect switches and leak charge from the capacitor over time. Due to this leakage current, cells need to be periodically read and written back to restore the charge on the capacitor, a procedure known as a *refresh*.

3.2.1.1 DRAM scaling issues

These system level complications as well as general fabrication complexities compromise the performance of scaled DRAM. From a system perspective, decreases in cell charge and increases in leakage current require more frequent refresh cycles and error correction overhead. Both of these system overheads decrease the

availability of memory blocks during operation, reducing the performance of the memory.

Fabrication of DRAM devices has also grown in complexity. DRAM capacitors have evolved from classic planar MOS-based circuits to more complex stacked and trench capacitor variants that provide greater capacitor density, and utilize insulator materials with higher dielectric constants [16–18,168]. In a trench-based DRAM cell, the capacitor is produced by etching a vertical hole into the CMOS substrate [169]. In the mid-2000, a 90 nm process requires the height of the trench depth to be roughly ten times the trench diameter. Early devices at the 20 nm node exhibit a ratio of trench height to diameter of approximately 100 [21]. These growing challenges make DRAM scaling more difficult and costly than in previous generations.

3.2.2 Static random access memory (SRAM)

SRAM is a memory structure designed for high speed reads and writes. Unlike DRAM, SRAM cells are connected to the power supply and therefore do not require refresh or write back after a read. A six transistor (6T) SRAM circuit is a storage latch connected to two differential bitlines, as illustrated in Figure 3.4. All modern SRAM cell variants, such as the eight transistor (8T), ten transistor (10T), and twelve transistor (12T) cells, are based on the 6T SRAM cell.

During reads, the wordline (WL) is driven high, connecting the cell to the bitlines (BL and $\overline{\mathrm{BL}}$). Like DRAM, both BL and $\overline{\mathrm{BL}}$ are initially precharged to $\frac{V_{DD}}{2}$. If

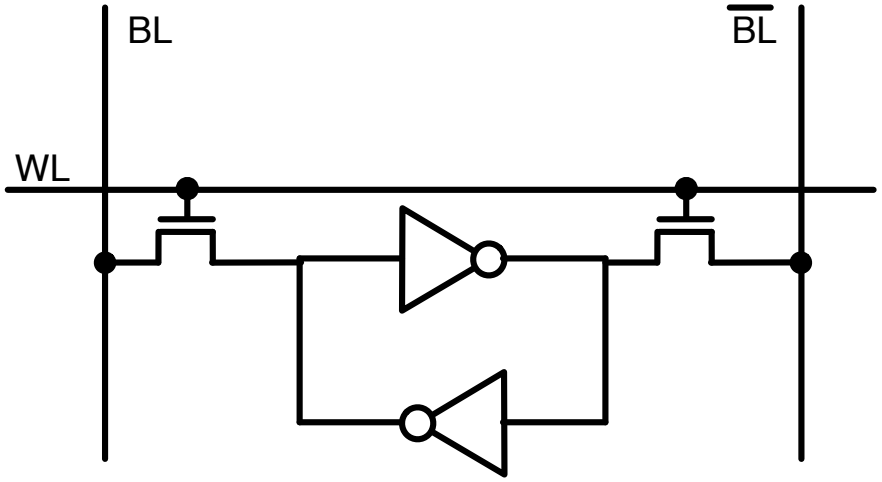


Figure 3.4: Circuit diagram of SRAM memory cell.

the stored state is logic '1,' current is sourced to BL and sunk from $\overline{\mathrm{BL}}$, purturbing the voltages on the two bitlines in opposite directions. This differential signal is more easily detected than in DRAM and thus the sensing process is much faster. In older SRAM, the active connection to power and ground enabled a cell to directly drive logic. For density and speed reasons, modern SRAM uses sense amplifiers rather than directly connecting to logic [170, 171].

At the system level, SRAM is generally different than DRAM. Whereas an individual memory location in DRAM is read infrequently, access patterns for SRAM require near constant availability. This capability is especially important in L1 cache memory and register blocks which are within a microprocessor pipeline and accessed one to twelve times per cycle. As a result, L1 cache memories typically use *multiported* cells, which provide multiple read and write ports per cell, as illustrated in Figure 3.5. A read port (see Figure 3.5a) adds two transistors, a word line ($\mathrm{WRL_n}$),

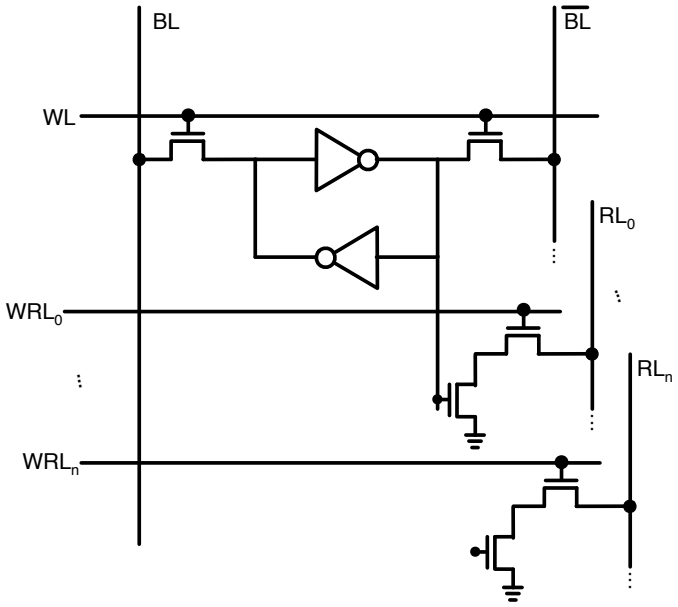
and a bitline ($\mathrm{RL_n}$) to a standard 6T SRAM cell. The $\mathrm{WRL_n}$ connects the bitline to a single transistor, which has the gate terminal connected to the internal node of the latch and the source terminal connected to ground. The internal state stored within the latch controls the impedance of the transistor, which is read from the bitline.

SRAM can also add write-read ports to a cell, as illustrated in Figure 3.5b. These ports are identical to the access ports of a 6T SRAM cell. Additional access ports, however, allow multiple on-going writes to cells within the same memory array.

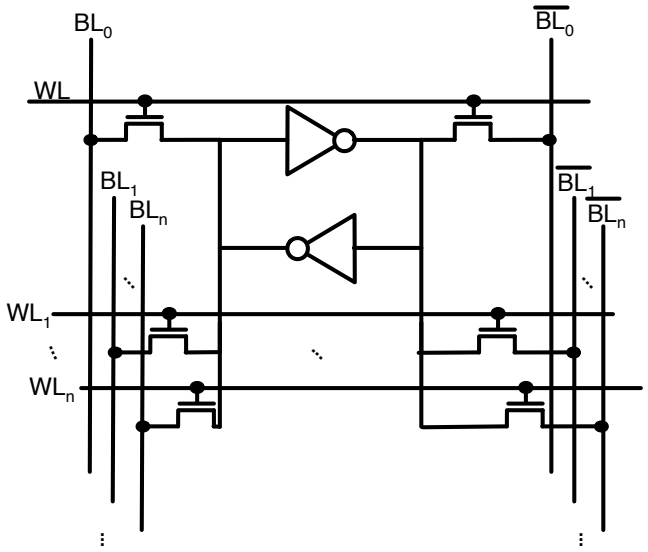
3.2.2.1 SRAM scaling issues

In on-chip cache memories, the memory density is paramount in achieving system performance and power efficiency. As compared to DRAM, the rate of increase in SRAM capacity has slowed. On-chip memory has classically relied on 6T SRAM for cache. During the mid-2000s, more specialized cell variants were introduced to address specific problems with SRAM.

SRAM suffers from several problems as technology is scaled. First, each cell maintains a constant connection between the supply voltage and ground, creating a leakage current path that dissipates static power. To manage leakage currents and to reduce area, SRAM cells are designed with higher thresholds and minimum width transistors. As a result, there is reduced drive strength within the cell, leading to smaller read currents and reduced cell stability. The reduced stability increases the likelihood of accidental writes during reads, and errors due to random



(a)



(b)

Figure 3.5: Circuit diagram of multiported SRAM memory cells: a) SRAM cell with multiple read ports, and b) SRAM cell with multiple read/write ports.

external radiation [172–174].

Process related complications have also reduced the reliability of SRAM. Random variations in fabrication affect the drive strength of individual transistors, causing each cell to produce different read and write currents. Over time, the transistors within an individual cell begin to accumulate charge within the transistor oxide layer. This phenomenon, known as *negative bias temperature instability* (NBTI), causes both PMOS and NMOS threshold voltages to degrade over time [175, 176]. This degradation causes an SRAM cell to become less tolerant to noise voltages on bitlines and power rails, increasing the likelihood of errors.

Several approaches have been introduced to combat these issues. To lower power, power gating [177, 178] and power stepping [179] have been introduced, where the supply voltage of an SRAM block is reduced without compromising the stored data. The stability of the cells is reduced in the low power state since the cache is unavailable during a power down.

To improve cell stability, one approach is to use dedicated ports for reads and writes. An 8T SRAM cell adds a read port to a 6T SRAM cell [180, 181], and dedicates the access port of the original 6T SRAM cell to writes. This method allows the internal latch to be optimized for stability and low leakage current without increasing errors due to read disturbances. One 10T cell variant improves the sensing capability of SRAM by utilizing symmetric read ports to both internal nodes of the latch, generating a differential signal for improved read performance [182]. Other 10T cell variants enable lower (potentially sub-threshold) operating voltages while

maintaining the logic state of the cell [183–185]. Alternatively, the transistors within the cell can be increased in size to enhance drive strength. While these techniques improve stability, they also increase the area of the cells and peripheral circuitry. Due to the additional design complexity and circuit area, the growth rate of on-chip cache memory capacity has declined.

3.2.3 Flash memory

Flash memories are non-volatile, targeting long term storage. Unlike SRAM, flash memory is not connected to a power supply to store a state. Flash also does not require a refresh operation or a write back after a read, key advantages of flash memory technology as compared to DRAM. Additionally, flash memory stores multiple bits of storage within a single device, supporting a greater effective density than DRAM or SRAM.

Flash memory utilizes a device called a floating gate MOS transistor, as illustrated in Figure 3.6. The device behaves similarly to a three terminal NMOS transistor. An additional metal layer, however, is placed between the gate terminal and the transistor channel region. This layer, called the floating gate, enables memory storage within the device. The stored charge on the floating gate changes the threshold voltage of the transistor, changing the channel resistance of the device. The state information is stored and read out through the channel resistance. The floating gate is electrically isolated from any conductors. This property enables

the long non-volatile lifetime of flash memory. Under high electric fields, however, electrons can tunnel through the insulator into the floating gate. Writes in flash memory technologies therefore require large voltages, up to 20 volts [186].

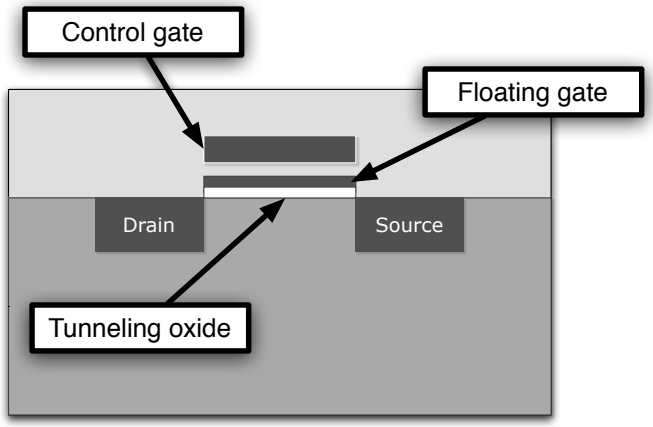


Figure 3.6: Profile view of floating gate transistor.

Flash memory arrays are typically organized into either a NOR or NAND configuration for read out, as illustrated in Figure 3.7. NOR topologies typically provide fast read performance and simple interfacing with digital circuits. NAND topologies exhibit significantly higher density than NOR circuits, but suffer lower read performance and require a dedicated memory controller to interface with standard digital logic. A NOR topology is therefore used in embedded applications to store startup software and operating systems for boot sequences, while NAND is typically used as high density storage [186].

In NOR configurations, the MOS source and drain are connected, respectively, to two parallel lines, the bit line and the source line (see Figure 3.7a). The MOS

gate terminal is connected to a word line that traverses horizontally across the array, perpendicular to the bit line and source line. Enabling the word line connects a single flash transistor to a bit line and source line. The state of the floating gate transistor affects the current sourced from the bit line. The magnitude of the current is evaluated to determine the state of the device. NOR flash memories use channel hot electron (CHE) based injection into the floating gate, a mechanism that requires the floating gate transistor to conduct current through the channel during a write [187]. This process is inefficient at transferring electrons to the floating gate due to an active DC current path to ground through the transistor. As a result, write operations in NOR arrays are typically slower and require greater energy than NAND memories.

NAND configurations connect multiple transistors in series, as illustrated in Figure 3.7b. The key advantage of this approach is that multiple flash transistors are connected in series. Each floating gate transistor within the array avoids bit line contacts at the source and drain terminals. As a result, the size of a storage cell is 2.5x smaller than a NOR topology [186]. The read out is, however, slower due to the high resistance of the series connected transistors. The physical mechanism for writes is Fowler-Nordheim tunneling from the transistor body into the floating gate [187,188]. This mechanism can be controlled by biasing the body and gate contacting the flash transistor. Writes and erase operations therefore occur in parallel.

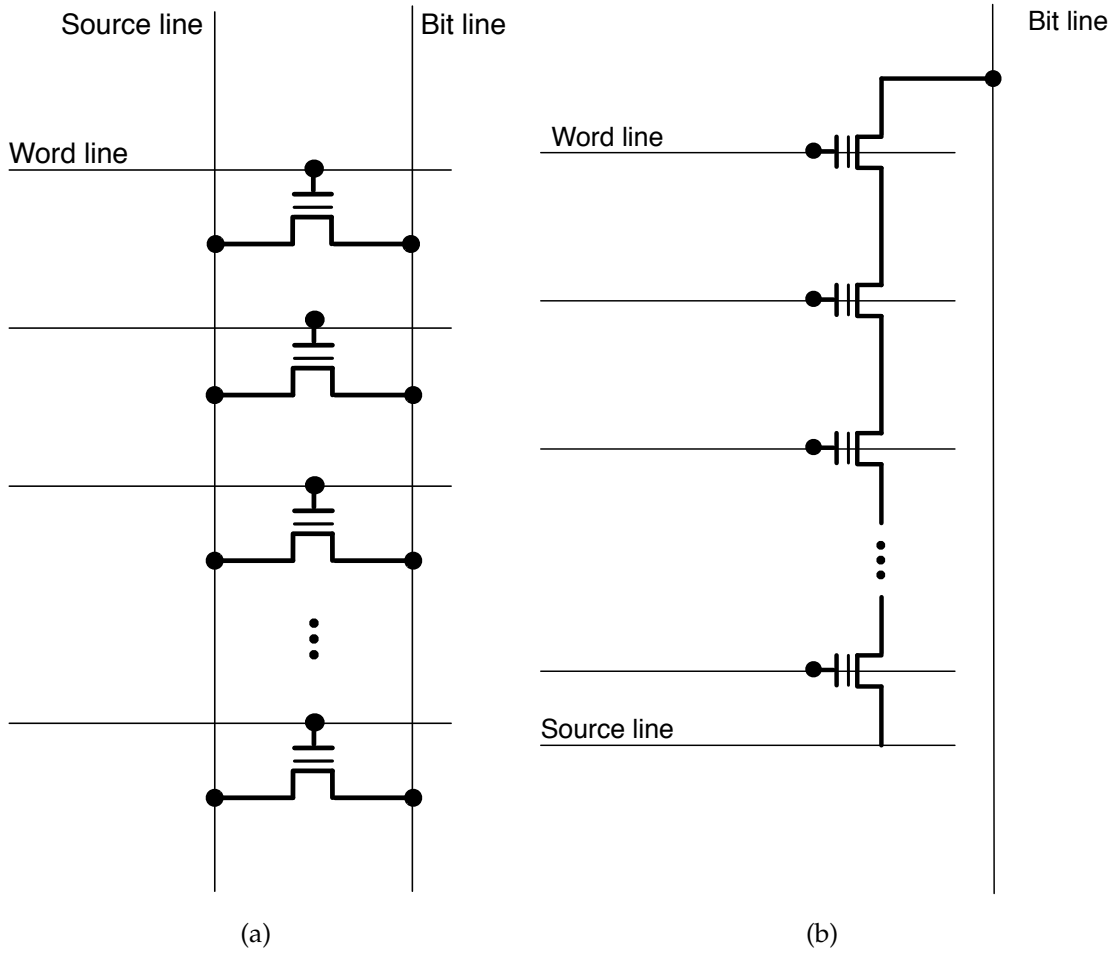


Figure 3.7: Circuit diagram of flash memory array topologies: a) NOR organization, and b) NAND organization.

3.2.3.1 Scaling issues

Several issues complicate scaling flash memories. From a fabrication perspective, the physical structure of a floating gate transistor is more difficult to realize as the position and size of the floating gate are harder to control. Lithographic technology for CMOS devices is currently based on 192 nm wavelength patterning with multiple masks to expose finer layout dimensions [189]. The lack of progress

towards finer lithographic patterning wavelengths has also complicated the fabrication of the floating gate within flash transistors.

The reduced size also degrades the data retention characteristics of the device. Each generation of technology results in fewer electrons to store a state within a single device, resulting in greater sensitivity to charge loss from the floating gate [190]. Flash memory based on 90 nm CMOS requires 50 electrons to shift the threshold voltage of a floating gate transistor [26]. In 25 nm CMOS, the number of electrons is approximately ten [26].

From an operational and reliability perspective, the primary scaling issues with flash memory affect writing and storing to the floating gate within the device. The high electric fields applied for each erase operation damage the tunneling oxide [191]. A modern flash device can withstand thousands of erasures before failure. The thinner oxides, required with deeply scaled technologies, are more susceptible to oxide breakdown, resulting in shorter device retention time and write endurance before failure. Industrial standards have typically held to a ten year retention time with 10,000 writes before failure [187]. This standard has degraded to a one to five year retention time with between 1,000 and 4,000 writes before failure [189].

Additionally, programming disturb errors have become a prominent issue [192]. During a write, cells adjacent to the selected cells are inadvertently biased, causing the state within the unselected cell to change. Random telegraph noise in flash transistors cause the drain current to fluctuate due to traps in the oxide, causing

erroneous reads and writes to a flash transistor [193]. This effect becomes more pronounced with scaling as the relative magnitude of the noise is much larger [193, 194].

Finally, the performance of flash transistors has also begun to decrease. Successive generations of flash memory are subject to increased latency and reduced bandwidth [195]. This behavior occurs both due to the increased resistivity of the interconnect, and the increasingly complex error correction required in deeply scaled flash memory.

3-D NAND technologies have been developed as a potential alternative to classical planar floating gate transistors [196]. This technology requires multiple layers of silicon on a single substrate, where each layer contains a plane of flash transistors. While these scaled technologies provide enhanced planar bit density, stacked layers exhibit temperature sensitivity issues [197, 198], exacerbating the endurance and reliability of flash devices.

3.3 Organization of the memory hierarchy

From a software perspective, memory appears as a single contiguous block that stores data. In hardware, however, hierarchical layers of memory are stacked to lower the average memory access time and power. Modern memory hierarchies have evolved from the Von Neumann structure of main memory, caches, and disk.

Each of these three layers are optimized for different goals. Main memory serves as the primary location of all operating programs and is structured according to the logical organization of the operating system. Cache memory emulates the memory system, but places a small subset of the data close to the microprocessor core, thereby improving performance. Disks store programs offline for future use and are optimized for high storage capacity. As a result, the access patterns of each memory type differ significantly. Typical latencies for each layer within the memory stack is listed in Table 3.1.

Table 3.1: Performance characteristics of modern memories [199]

Cache Level	**Bytes per access**	**Access time**	**Access energy**	**Cost per megabyte** [1]
L1 cache and registers	2 to 8	100's of ps	1 nJ	\$1 to 100
L1, L2 cache	10	1 to 2 ns	1 to 100 nJ	\$1 to 100
L2/L3/off-chip cache	100	5 to 10 ns	10 to 100 nJ	\$1 to 10
DRAM	1000	10 to 100 ns	1 to 10 nJ	\$0.1
Flash	1000	100 to 500 μs	1 μJ (reads) 10 μJ (writes)	\$0.001
Disk	1000	1 to 100 ms	100 to 1000 mJ	\$0.0000001

Beginning from the CPU, the L1 cache is a low latency, high bandwidth memory block responsible for supplying data to the CPU (see Figure 3.8). L1 cache memory and the local registers are accessed when a microprocessor is active. These cache memories are small and optimized for low latency and energy. In general, memory systems place data as close to the CPU as possible. Cache memory access is therefore strongly dependent on the application size and the data set. A smaller

[1]Memory IC cost is highly volatile. Estimating cache cost is dependent on the application. Both consumer desktop microprocessors and ASIC microcontrollers may use 50% of the die area for cache, but exhibit a vastly disparate cost based on the market and design complexity of each IC.

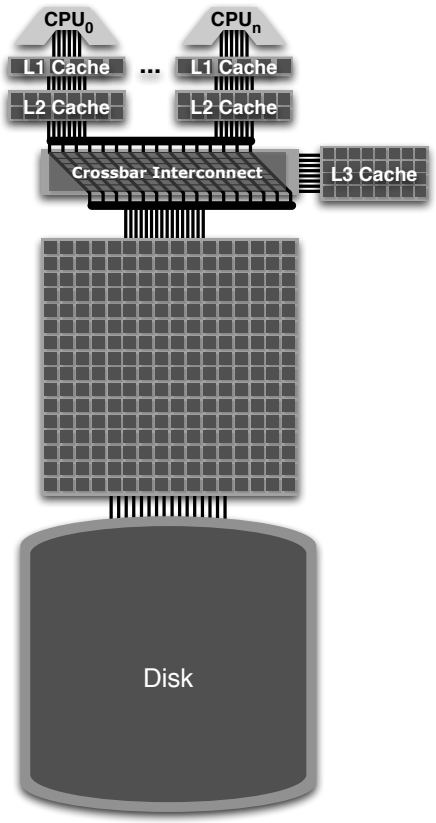


Figure 3.8: Modern memory hierarchy

application may fit entirely into the L2 cache memory, resulting in infrequent accesses to the L3 cache. Larger applications will "spill over" into the L3, causing more frequent accesses, and thus lower average application performance. For data intensive applications, such as multimedia, the effect of the cache is small since the primary system bottlenecks are the bandwidth and latency between the DRAM and the on-chip cache memories.

3.3.1 Hardware organization

The SRAM, DRAM, and flash memory blocks are internally organized as a hierarchy of arrays, as illustrated in Figure 3.9. At the top level of the hierarchy, each array block is composed of multiple memory banks. Each bank typically contains read and write ports with buffers to handle both internal and external bus arbitration and to manage multiple memory accesses. A bank comprises multiple sub-banks interconnected with multiple internal blocks called mats. Within a mat is a local array that directly interfaces with individual memory cells. From a systems perspective, the local array is the fundamental block of the memory subsystem. All of the higher level abstractions (*e.g.*, banks and sub-banks) represent layers of interconnect and decoding to organize and transmit stored data to and from the local array.

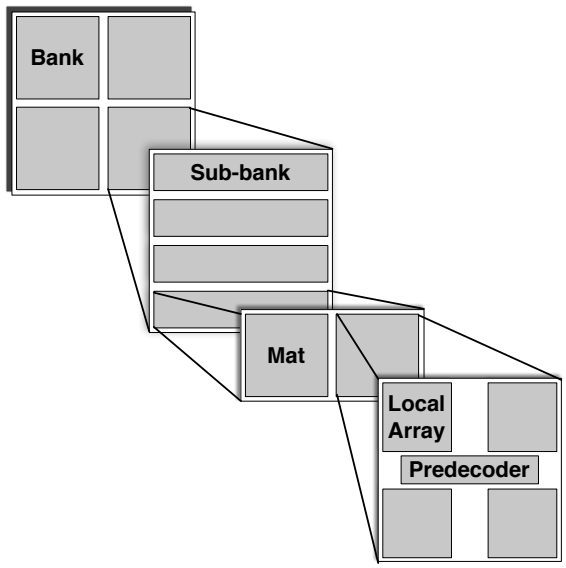


Figure 3.9: Internal hierarchy of local arrays

Two key reasons drive this memory organization as opposed to large monolithic memory arrays. One reason is that reads and writes to different local arrays can operate in parallel, increasing the effective bandwidth of the cache memory. Another reason is that small memory arrays exhibit lower internal parasitic impedances, leading to shorter sensing delay, reduced power, and simplified redundancy mechanisms [200, 201]. A layered approach to memory block structures enhances both the access time and energy for common application classes, while facilitating the large memory space required in modern applications.

3.3.1.1 Structure of a local memory array

A read is initiated by first decoding the data address, accessing a cell within the local array, sensing the voltage or current signal on the metal lines, and transmitting the data out. To support this functionality, a local memory array consists of three basic circuit blocks: row and column decoders, sense amplifiers, and the cell array, as illustrated in Figure 3.10.

3.3.2 Row selection and drivers

Upon arrival of the address to the local array, the remaining address bits are decoded and an individual row is selected. Decoding proceeds by driving a set of address lines ($\mathrm{ADR_n}$) placed from top to bottom along the edge of the array. Each row is permanently encoded to a specific address by selectively connecting the

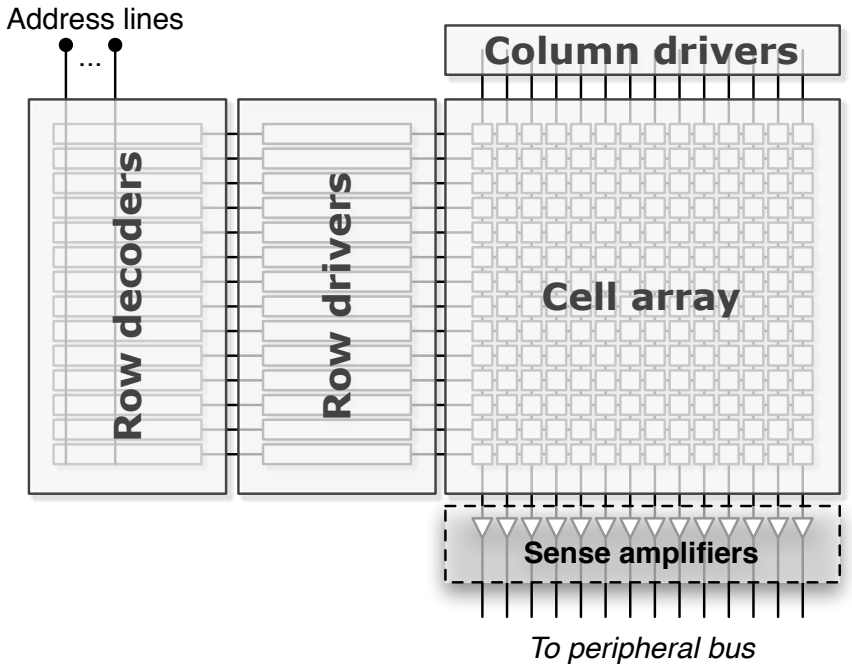


Figure 3.10: Structure of a memory array.

address lines to a NAND gate at the edge of the cell array. An example of a row with address "0111" is illustrated in Figure 3.11. In this case, a NAND gate of row seven ($b0111$) is connected to $\overline{\mathrm{ADR}_3}$, ADR_2, ADR_1, and ADR_0. If all of the inputs are high, the NAND gate registers a logic '0' on the output, which feeds into the row driver circuit. As the impedance of the wordline increases, the drive capability of the row driver must also increase. In these situations, multiple gates with increasing output current are cascaded to improve the drive strength [202, 203].

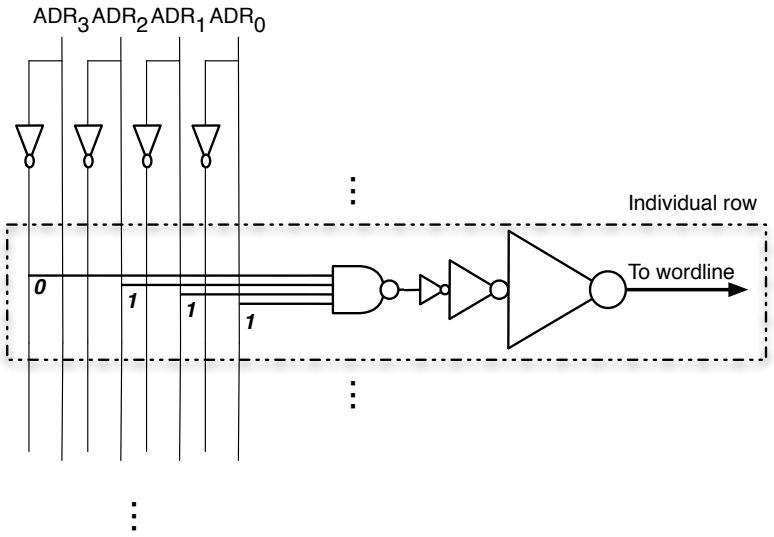


Figure 3.11: Schematic of logic circuit for row access for a four bit address ($b0111$). The address $b0111$ is encoded by connecting the inputs of the NAND gate to $\overline{\mathrm{ADR}_3}$, ADR_2, ADR_1, and ADR_0. The row driver triggers only if all of the inputs are at logic '1'

3.3.2.1 Sense amplifiers

Once the row is selected and the wordline is driven high, all of the cells along the row are connected to a set of bitlines. A sense amplifier is located at the foot of each column and detects the voltage difference on the line to determine the state of the cell connected to the bitline (See Fig. 3.12a). The traditional CMOS sense amplifier is a differential latch, as illustrated in Figure 3.12b. The sense amplifier is initialized by lowering the $SenseEnable$ signal to zero, disconnecting the output terminals (Out and $\overline{Out}$) from the array, and disconnecting the amplifier from ground. Detaching the ground terminal forces both Out and $\overline{Out}$ to the same voltage. The memory cell is connected to the bitlines, producing a voltage difference.

The sense amplifier is reconnected to both ground and the bitlines. In this initial state, the amplifier is unstable. The difference in the bitline voltages causes the circuit to switch to a stable binary state.

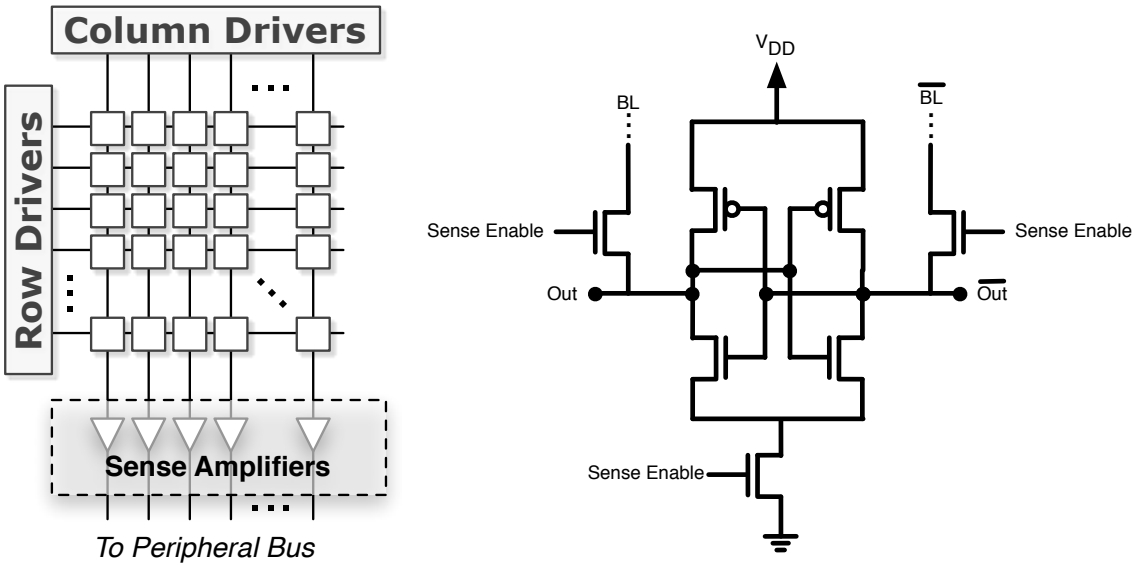


Figure 3.12: Schematic of a local SRAM array: a) the topology of a local array with sense amplifiers, and b) a latch type sense amplifier for column read out.

To read from a read port of an SRAM cell, where there is no differential signal, an alternative sense circuit is required. A dynamic sense amplifier based on a precharge and discharge scheme is typically utilized [199], as illustrated in Figure 3.13. This circuit precharges the bitline to ground and the internal node of the sense amplifier to V_{DD}. Once the sense enable signal is triggered, the bitline begins to charge. If the memory cell is in a high state, the corresponding transistor within the read port is in a low state, draining the charge from the bitline. This charge drain prevents the sense amplifier from changing state. If the memory cell is in the

opposite state, charge remains on the bitline and the sense amplifier switches state.

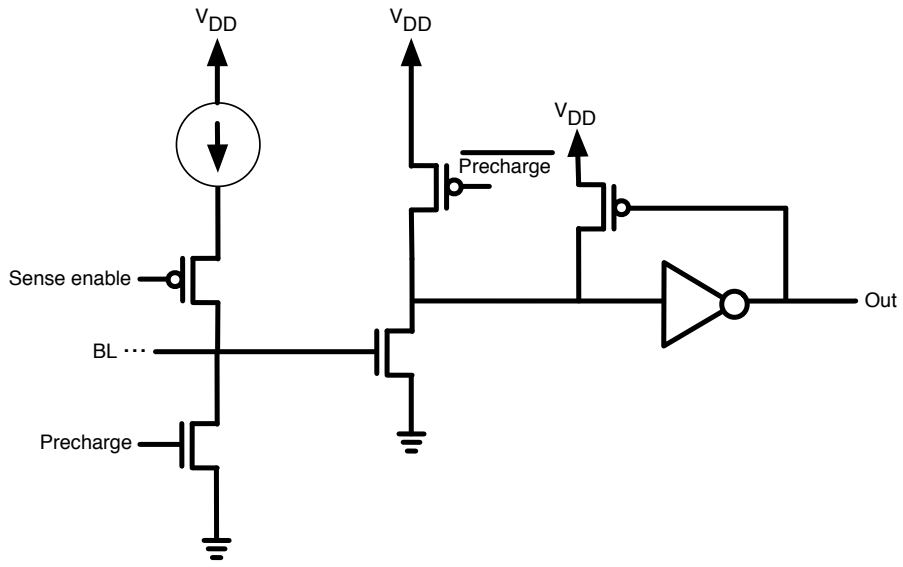


Figure 3.13: Single ended sense amplifiers for dynamic sensing [204].

3.3.2.2 Cell array

The cell array stores data in a physical location, and consists of rows and columns of individual cells, as illustrated in Figure 3.14. The area of a memory cell quadratically affects the area of the array and the impedance driven by the peripheral circuitry. If N is the number of rows and columns in an array, and D is the height and width of a cell, the area of an array is proportional to $(ND)^2$. As the area grows, the length of the wordlines and bitlines also grows, increasing the resistive and capacitive load. Minimizing cell area is therefore of paramount importance. As a result, traditional memory arrays are structured as either high performance memories or high density memories. High density memories minimize cell area and thus

the area of the overall array to improve capacity at the cost of lower performance. High performance memories use larger cells and additional ports to improve read and write performance, increase cell stability, and reduce access conflicts during operations.

In addition to the effect on area, the physical layout of the memory cell constrains the physical structure of the peripheral circuity. Row decoders and sense amplifiers need to be the same physical height and width, respectively, as a memory cell, as illustrated in Figure 3.15. Pitch matching is more complex in multiported memories as each additional port requires a separate set of sense circuits and row

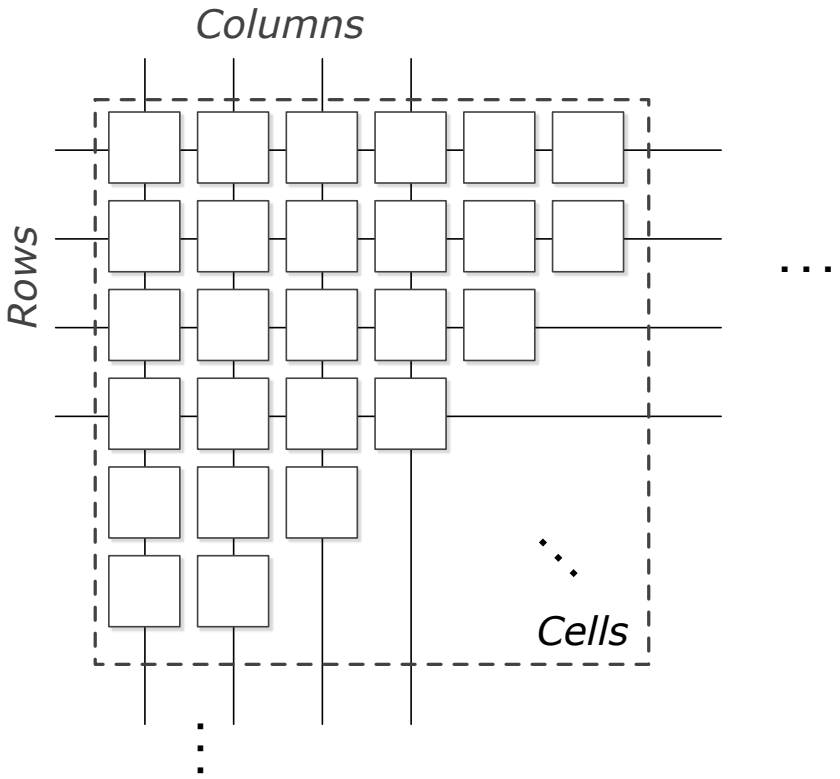


Figure 3.14: Structure of memory cell array.

driver circuits, increasing the overhead of the peripheral circuitry.

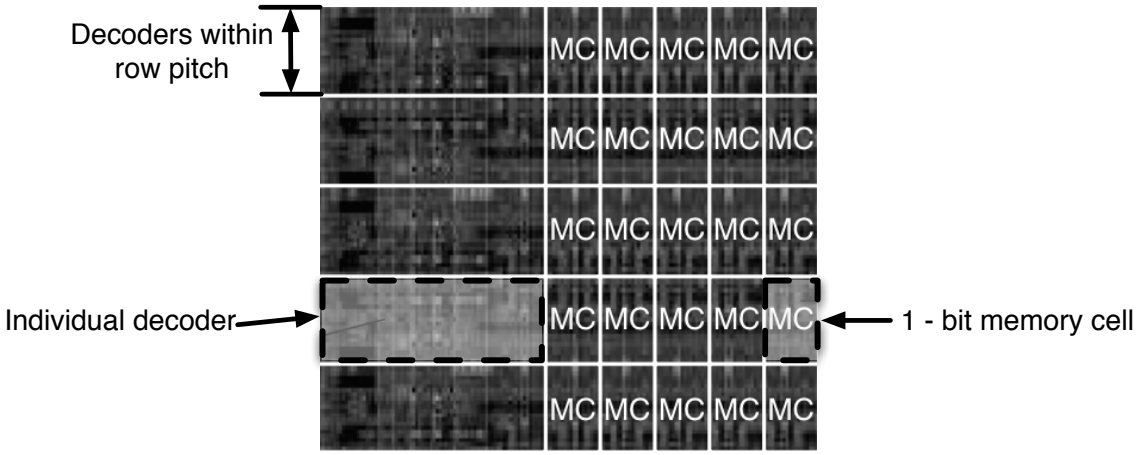


Figure 3.15: Pitch matching of row decoder [199].

These area considerations in memory blocks as well as system requirements affect the tradeoffs among speed, stability, and energy. In SRAM, a shorter array height has been shown to reduce latency due to shorter precharge and discharge times for the bitlines [200]. Reducing the array height may increase the delay as both the driver circuitry and the area of the peripheral decoders, buses, and control logic is greater.

3.4 Resistive memories within the traditional memory hierarchy

As the scaling of traditional CMOS memories becomes more problematic, resistive memory technologies offer a potential path for increased capacity with improved system performance. Despite the important advantages of resistive memories such as non-volatility and high density, each memristor technology introduces entirely new issues to the memory system design process. From a system design perspective, there are two primary differences between charge based semiconductor memories such as DRAM and SRAM, and resistive memory technologies. The first difference is the write endurance. Classic CMOS semiconductor memories exhibit no practical endurance problems. The exception is flash memory, which is limited to disk level applications due to relatively low endurance. A second difference is that resistive memories exhibit highly asymmetric read and write accesses, both in terms of latency and power. Write power and latency are at least an order of magnitude higher than for reads. In contrast, DRAM and SRAM exhibit symmetric read and write latencies and energy per access.

These system level considerations constrain the use of specific memristive technologies to specific layers of the memory hierarchy. Typical parameters for those resistive memory technologies considered here are listed in Table 3.2. CMOS-based

memory characteristics are included for comparison. STT-MRAM exhibits a resistance range similar to CMOS. Both devices exhibit practically infinite endurance, suggesting that cache level replacement of SRAM with STT-MRAM is practical. RRAM with limited endurance is not yet suitable for cache memory but the large tunable range and high density support replacing both DRAM and disk with RRAM. Improving the endurance is, however, an active area of research and may someday enable RRAM for on-chip cache memory [106, 134, 205]. Fine grain control of the RRAM resistance enables multi-bit operation, opening the possibility of using RRAM for high density storage. To realize the potential of resistive memories, circuits and architecture are required to leverage the relative strengths of each technology while managing existing device limitations.

Table 3.2: Characteristics of CMOS and resistive memory technologies

Figure of merit	**Resistive switching oxide (RRAM)**	**Spin torque transfer magnetic tunnel junction (STT-MTJ)**	**Bulk 22nm CMOS transistor**
Density	$4F^2$	$6F^2$ to $30F^2$	$6F^2$
R_{on} *resistance range*	100 Ω to 100 kΩ	2 kΩ to 10 kΩ	1 kΩ to 50 kΩ
Resistance ratio	1 to 10^5	1.5 to 3	10^5
Write latency	100 ps to 100 ms	2 to 20 ns	Circuit dependent
Tunability	Fine grained control	Bistable	Circuit dependent
Endurance	10^8 to 10^12	$> 10^{15}$	Practically infinite

3.5 Conclusions

Memory systems originate from the Von Neumann architecture. This basic structure has evolved to a mult-layer hierarchy consisting of cache memories and

main memory, designed to alleviate the limitations of the Von Neumann bottleneck. These classic CMOS systems utilize local memory arrays as fundamental building blocks with a hierarchy of interconnect and decoding logic to transfer data. Each memory array consists of storage cells and peripheral circuitry to access and store data. The hierarchy of interconnect and decoding logic interfaces each local array to external data buses. With the increasing demands on memory capacity and CPU performance, the memory hierarchy will expand to incorporate additional cache layers, more memory arrays within each cache memory and DRAM block, and a greater number of cells within each local memory array.

As CMOS memories become more difficult to scale, memristive technologies are poised to supplant both SRAM and DRAM. Resistive memories, however, introduce new and different tradeoffs in array structure and organization, changing the process in which memory systems are designed. Replacing CMOS memory with memristor-based memory requires understanding the organization of existing systems, and the strengths and limitations of different and evolving memristive technologies.

Chapter 4

Multistate Register Based on Resistive RAM

The traditional approach of increasing CPU clock frequency has abated due to constraints on power consumption and density. To increase performance with each CMOS generation, thread level parallelism is exploited with multi-core processors [206]. This approach utilizes an increasing number of CMOS transistors to support additional cores on the same die, rather than increase the frequency of a single processor. This larger number of cores, however, has increased static power. Multi-threading is an approach to enhance performance of an individual core by increasing logic utilization [207], without additional static power consumption. Handling each thread, however, requires duplication of resources (e.g., register files, flags, pipeline registers). This added overhead increases the area, power, and complexity of the processor, potentially increasing on-chip signal delays. The thread count is therefore typically limited to two to four threads per core in modern general purpose processors [208].

The high density, nonvolatility, and soft error immunity exhibited by RRAM enables novel tradeoffs in digital circuits, allowing new mechanisms to increase thread count without changing the static power. These tradeoffs support innovative memory structures for novel microarchitectures. In this chapter, a memristive multi-state pipeline register (MPR) is proposed that exploits these properties to enable high throughput computing. The MPR is compatible with existing digital circuits while leveraging RRAM devices to store multiple machine states within a single register. This behavior enables an individual logic pipeline to be densely integrated with memory while retaining state information for multiple independent, on-going operations. The state information for each operation can be stored within a local memory and recalled at a later time, allowing computation to resume without flushing the pipeline.

This functionality is useful in multithreaded processors to store the state of different threads. This situation is demonstrated in the case study of a novel microarchitecture — continuous flow multithreading (CFMT) [209]. It is shown that including an RRAM MPR within the CFMT microarchitecture enhances the performance of a processor, on average, by 40%, while reducing the energy, on average, by 6.5%. The proposed MPR circuit can also be used as a multistate register for applications other than pipeline registers.

Background of RRAM and crosspoint style memories is reviewed in Section 9.2. The operation of the multistate register is presented in Section 4.2. The simulation

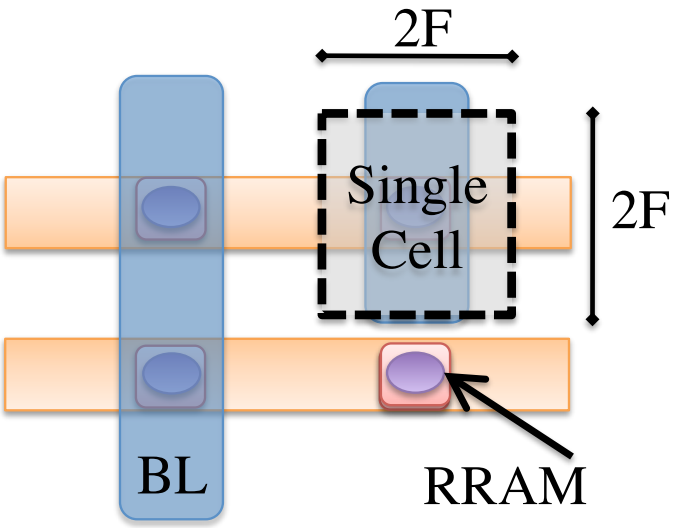
setup and circuit evaluation process are described in Section 4.3. A case study examining the multistate register as a pipeline register within a CPU is presented in Section 4.4, followed by some concluding remarks in Section 4.5.

4.1 Background on Nonlinear RRAM Crosspoint Arrays

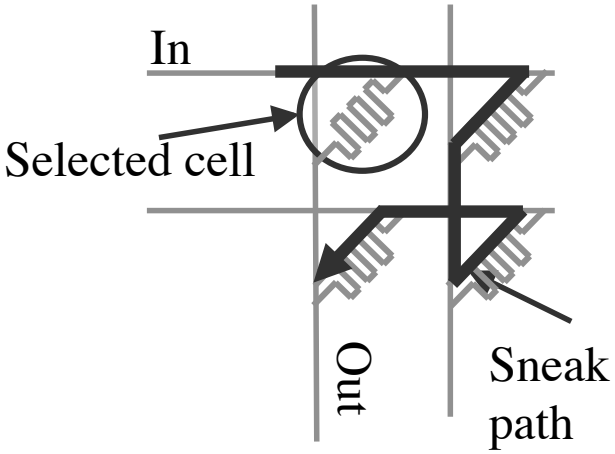
RRAM has the greatest density when utilized in a crosspoint configuration. In this structure, a thin film is sandwiched between two sets of parallel interconnects. Each set of interconnects is orthogonal, allowing any individual memristive device to be selected by biasing one vertical and one horizontal metal line. In this configuration, the circuit density is only limited by the available metal pitch. The structure of a crosspoint is shown in Figure 4.1a.

Crosspoint arrays have the inherent problem of sneak path currents [47], where currents propagate between the two selected lines through unselected memristors. The sneak path phenomenon is illustrated in Figure 4.1b. The nonlinear I-V characteristic of certain memristive devices lessens the sneak path phenomenon [31]. This nonlinearity can be achieved by depositing additional materials above or below the memristive thin film. Depending on the material system used for RRAM, the nonlinearity can result from an insulator to metal transition or a negative differential resistance [31]. From a circuits perspective, the combined device can be modeled

as a pair of cross coupled diodes in series with a memristor, as shown in the inset of Figure 4.2b. Since the rectifying structure requires an additional thin film layer, there is no effect on the area of the crosspoint structure.



(a)



(b)

Figure 4.1: RRAM crosspoint (a) structure, and (b) an example of a parasitic sneak path within a 2 x 2 crosspoint array

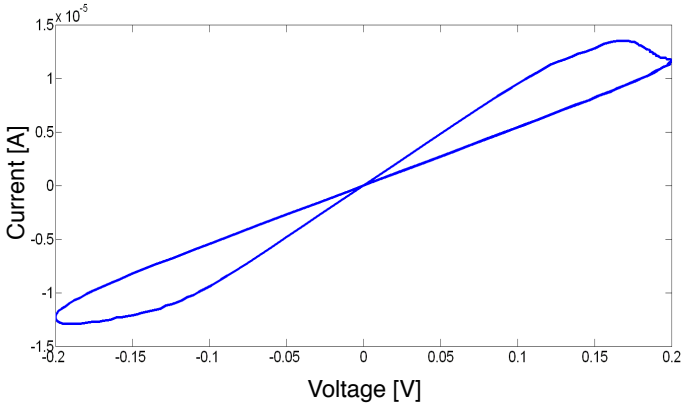
An I-V curve of a memristive device with cross coupled diodes is shown in Figure 4.2a. The device nonlinearity ensures that any unselected rows exhibit a high resistance state several orders of magnitude larger than the memristor resistance. The high resistance of the unselected devices reduces sneak currents and ensures that the leakage power of the array is relatively small. A DC analysis of on and off state memristors within a crosspoint is listed in Table 4.1, where a 4 x 4 crosspoint with RRAM devices is DC biased at 0.8 volts. These RRAM devices exhibit an on/off current ratio of 30. In an unrectified crosspoint, the observed current ratio drops to less than two. The rectified crosspoint displays a current ratio of 28.5, only 5% less than the ideal ratio of an RRAM device. Furthermore, the total power consumption is reduced by almost an order of magnitude.

Table 4.1: Comparison of DC on/off memristor current for 4 x 4 crosspoint array

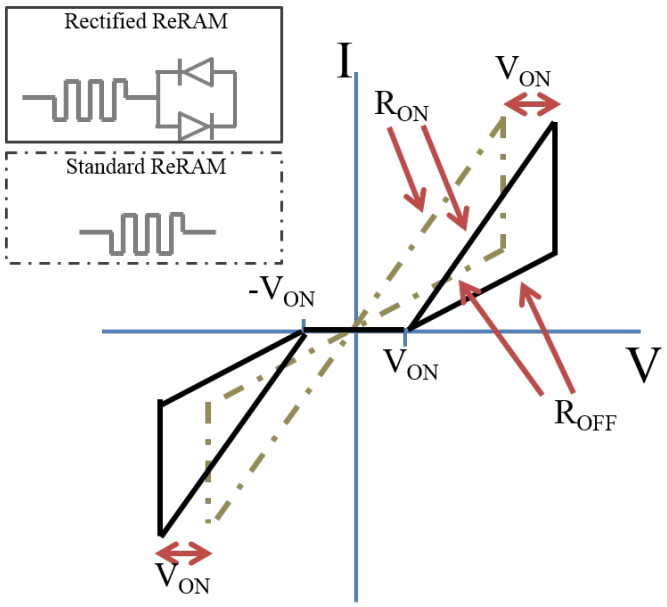
	I_{on} [mA]	I_{off} [mA]	Ratio	Average Active Power [mW]
Unrectified	2.3	0.132	1.7	1.45
Rectified	0.486	0.017	28.5	0.201

4.2 RRAM Multistate Register

The multistate register is a novel circuit used to store multiple bits in a single logic gate. The multistate register is "drop-in" compatible with existing CMOS based flip flops. The element utilizes a clocked CMOS register augmented by additional sense circuitry (SA) and global memristor select (MS) lines. The symbol



(a)



(b)

Figure 4.2: I-V characteristic of a memristor for (a) a ThrEshold Adaptive Memristor (TEAM) [210] model with a 0.2 volt sinusoidal input operating at a frequency of 2 GHz, and (b) resistive devices with and without ideal cross-coupled diodes. The parameters of the TEAM models are listed in Table 4.2. V_{ON} is the on-voltage of the diode, and R_{ON} and R_{OFF} are, respectively, the minimum and maximum resistance of the memristor

and topology of the multistate register are shown in Figure 4.3. Multistate registers can be used as pipeline registers within a processor pipeline, as shown in Figure

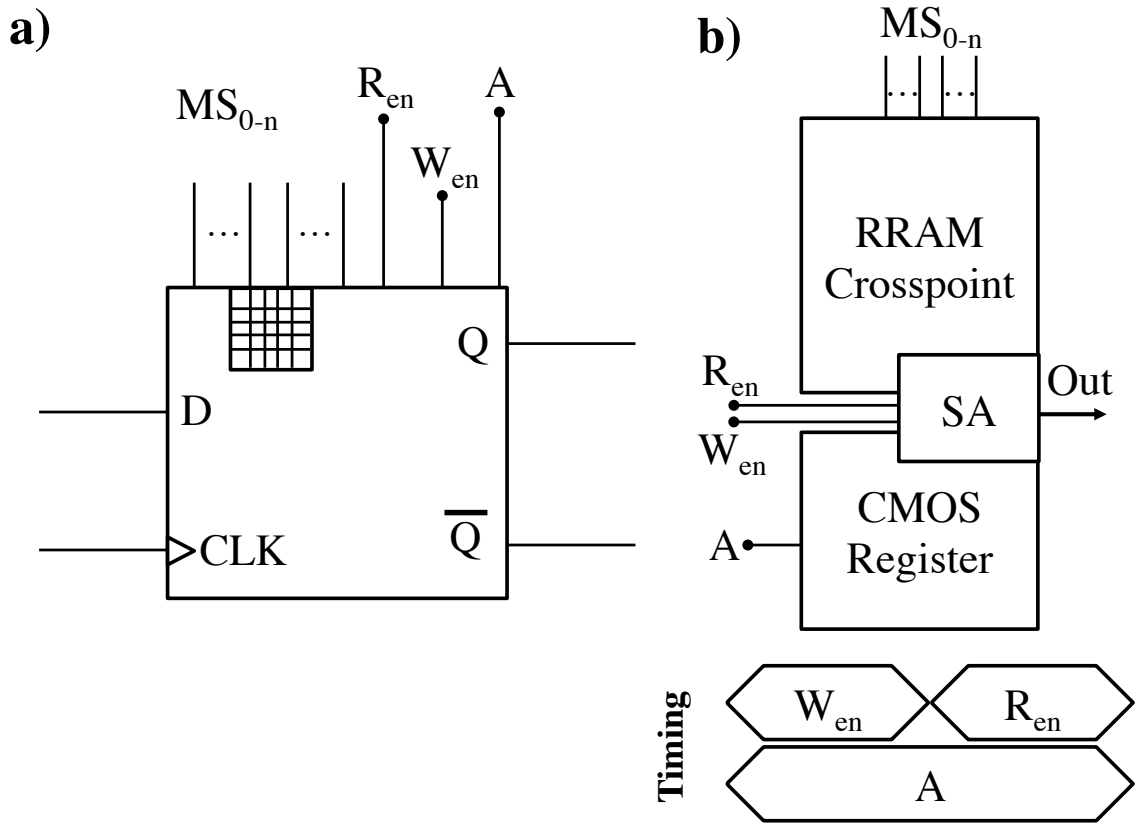


Figure 4.3: Multistate register element. (a) Symbol of the multistate register, and (b) block diagram with control signal timing. The symbol is similar to a standard CMOS D flip flop with the addition of a symbol of the crosspoint array.

4.4 and further explained in Section 4.4.

The MS lines select individual RRAM devices within the crosspoint memory co-located with the CMOS register. A schematic of the proposed RRAM multistate register is shown in Figure 4.5a. The signals W_{en} and R_{en} are global control signals that, respectively, write and read within the local crosspoint memory. Signal A sets the CMOS register into an intermediate state that facilitates writes and reads

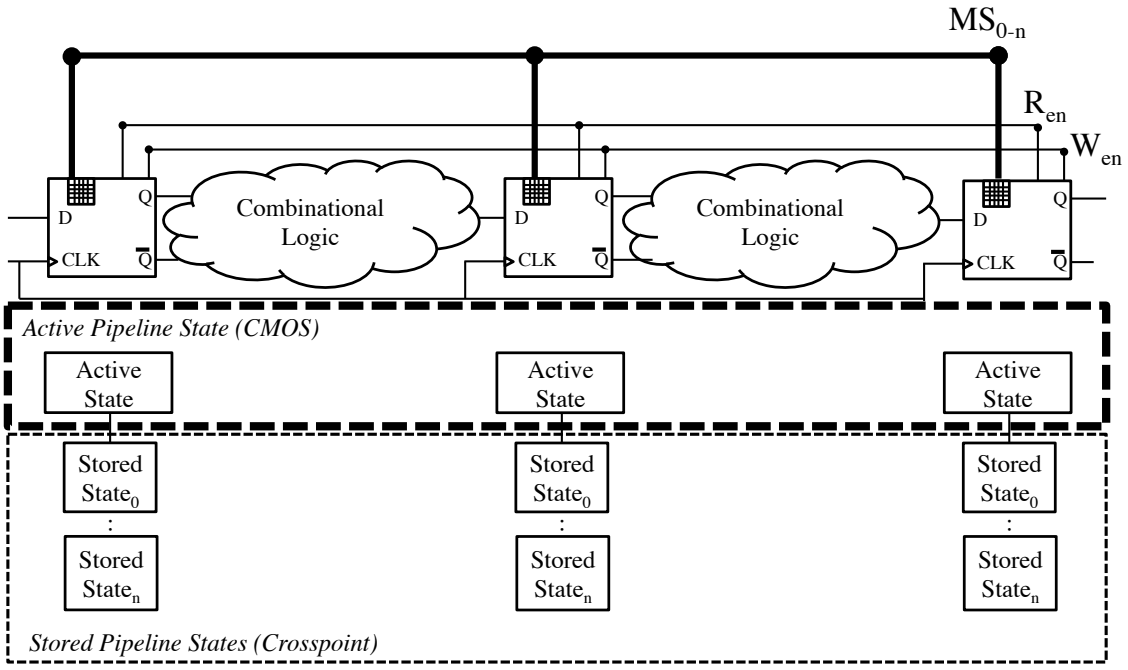


Figure 4.4: Multistate pipeline register (MPR) based pipeline with active and stored pipeline states. The MPR replaces a conventional pipeline register and time multiplexes the stored states.

from the crosspoint. The gates within the CMOS register are reconfigured to provide a built-in sense amplifier for the RRAM crosspoint [204]. The overhead of the additional circuitry (shown in Figure 4.3) is relatively small (see Section 4.3.2).

The multistate register primarily operates as a CMOS register. In this mode, the structure behaves as a standard D flip flop, where a single bit is stored and is active while the idle states are stored within the RRAM crosspoint array. When global control circuitry triggers a change in the pipeline state (e.g., for a pipeline stall or context switch), the circuit stores the current bit of the register and reads out the value of the next active bit from the internal RRAM-based storage. Switching

between active bits consists of two phases. In the first half of the cycle, an RRAM write operation stores the current state of the register. During a write operation, the transmission gate A disconnects the first stage from the following stage, isolating the structure into two latches. The input latch stores the currently evaluated state, while the output latch stores the data of the previous state. Once W_{en} goes high, the input latch drives a pair of multiplexers that write the currently stored state into the RRAM cell selected by the global MS lines. The active devices during the write phase are shown in Figure 4.5b. The write phase may require more than half a cycle depending upon the switching time of the RRAM technology. During the second half of the clock cycle, the new active bit is selected within the resistive crosspoint array and sensed by the output stage of the CMOS D flip flop. During a read operation, the globally selected row is grounded through the common node N_{in}. The voltage on the common line N_{out} is set by the state of the RRAM cell. To bias the RRAM cell, the common line is connected through a PMOS transistor to the supply voltage V_{DD}. The voltage is sensed at the output of M1. If R_{en} is set high, M1 to M5 reconfigure the last inverter stage as a single ended sense amplifier [209], and the crosspoint array is read. The active devices during the read phase are shown in Figure 4.5c.

The physical design of the multistate register can be achieved by two approaches. RRAM devices can be integrated between the first two metals, as illustrated in Figure 4.6a, or the RRAM can be integrated on the middle level metal layers, as shown

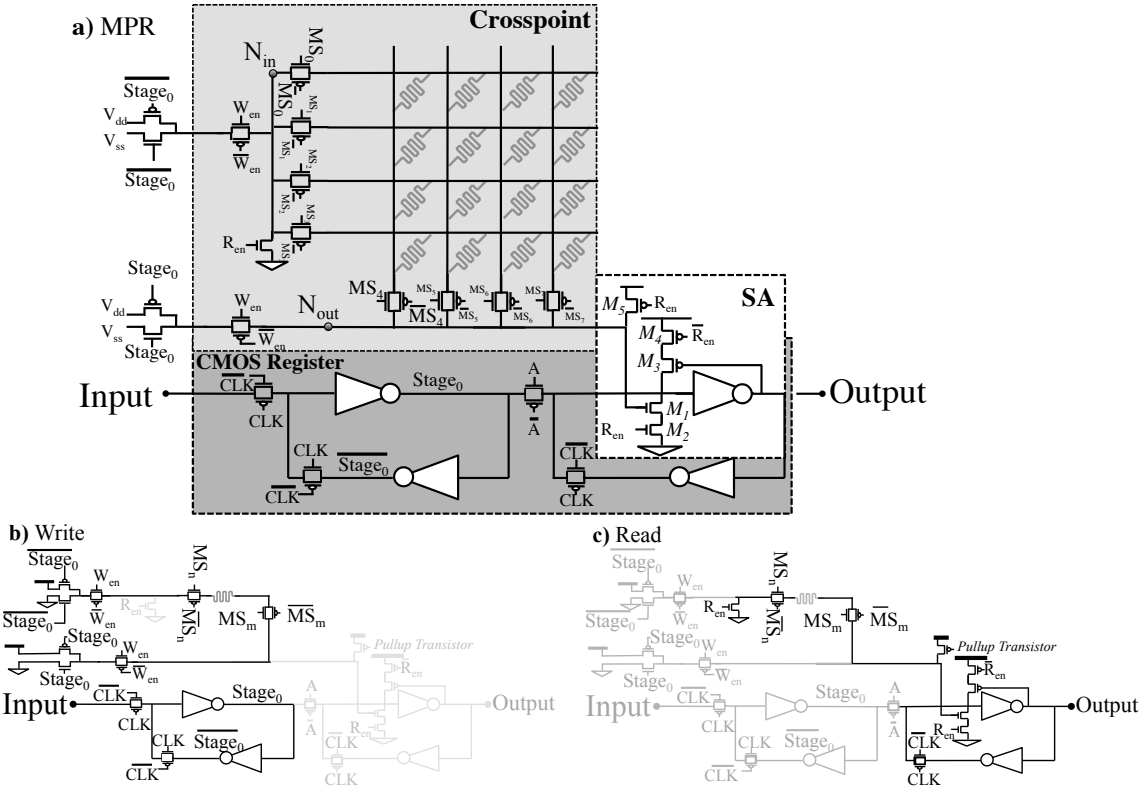


Figure 4.5: Proposed RRAM multistate pipeline register. (a) The complete circuit consists of a RRAM-based crosspoint array above a CMOS-based flip flop, where the second stage (the slave) also behaves as a sense amplifier. The (b) write and (c) read operations of the proposed circuit.

in Figure 4.6b. The middle metal layer approach allows the RRAM to be integrated above the CMOS circuitry, saving area. A standard cell floorplan is shown in Figure 4.7b, where a dedicated track is provided for the RRAM interface circuitry. This dedicated track runs parallel to the CMOS track. The addition of this track wastes area in those cases where multistate registers are sparsely located among the CMOS gates. Additional routing overhead increases the area required to pass signals around the crosspoint array.

The approach illustrated in Figure 4.7a, where the RRAM is integrated on the lower metal layers, requires slightly more area but is compatible with standard cell CMOS layout rules. Fabrication on the lower levels maintains standard routing conventions, where the lower metal layers are dedicated to routing within the gates, and the middle metal layers are used to route among the gates.

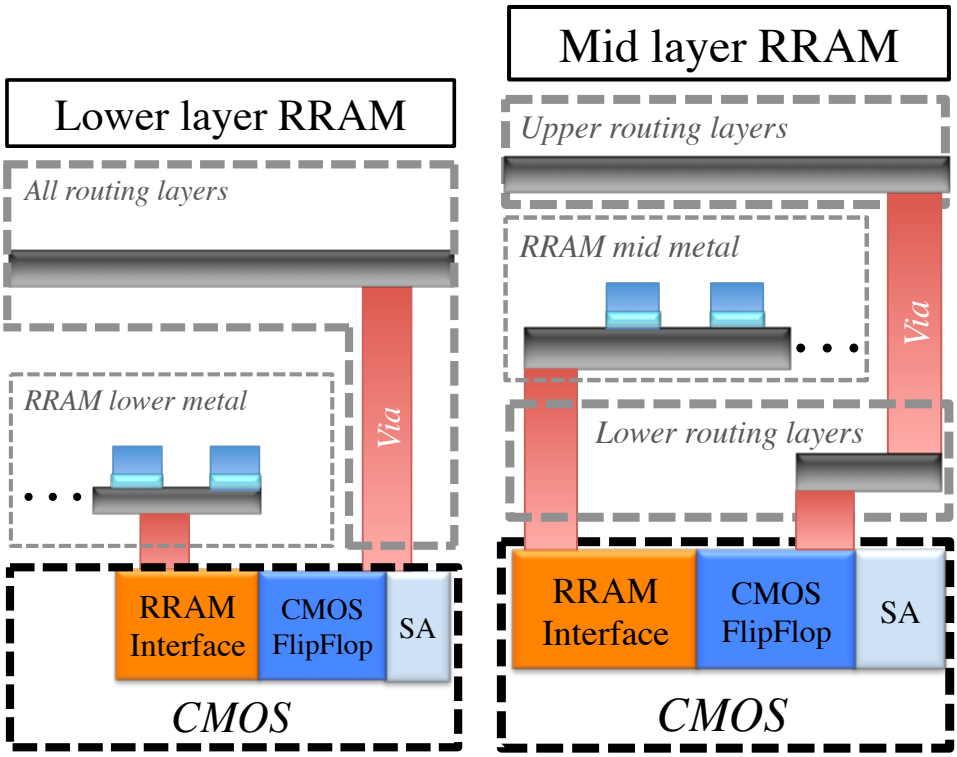


Figure 4.6: Vertical layout of RRAM in MPR circuit for (a) lower level, and (b) mid-layer crosspoint RRAM array

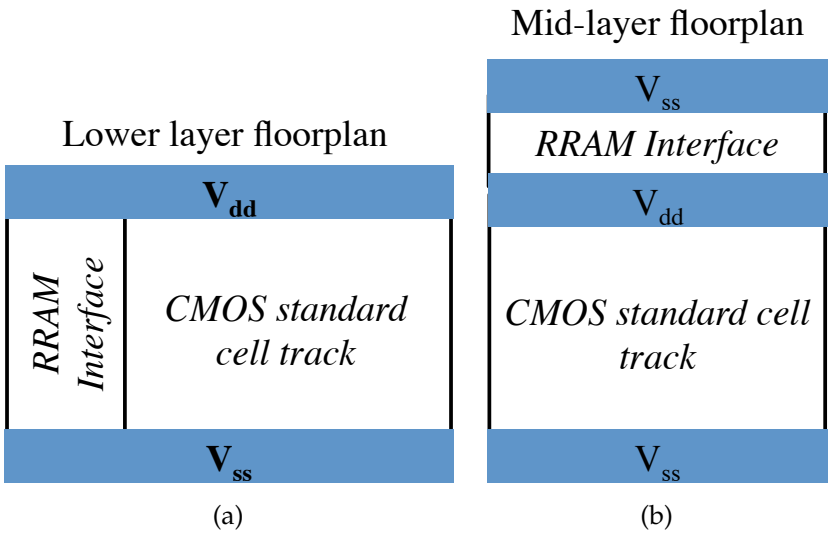


Figure 4.7: Planar floorplan of MPR with lower metal and mid-metal RRAM layers. The RRAM array is not marked in this figure since it is located above the CMOS layer and has a smaller area footprint.

4.3 Simulation Setup and Circuit Evaluation

The multistate register has been evaluated for use within a high performance microprocessor pipeline. The latency, energy, and area of the register are described in this section as well as the sensitivity to process variations.

4.3.1 Latency and Energy

The energy and latency of an MPR are dependent on the parameters of an RRAM device and the CMOS sensing circuitry built into the MPR. The RRAM device is modeled using the TEAM model [210] based on the parameters listed in Table 4.2.

The parameters of the resistive device are chosen to incorporate device nonlinearity into the I-V characteristic, as shown in Figure 4.2b and described in Section 9.2. The multistate register is evaluated across a range of internal cross point sizes (e.g., different number of states per register). The resistance of the device is extracted from [31]. The transistor and cell track sizing information is from the FREEPDK45 Standard Cell Library [211] and scaled to a 22 nm technology. Circuit simulations utilize the 22 nm PTM CMOS transistor model [212]. The RRAM and diode device parameters are listed in Table 4.2. Standard CMOS timing information for the register is listed in Table 4.3. The read operation requires 28.6 ps, equivalent to a 16 GHz clock frequency (the read operation is less than half a clock cycle). Hence, the read operation is relatively fast and does not limit the operation of the multistate register.

The performance of the multistate register is limited by the switching characteristics of the RRAM device. The performance of the multistate register is demonstrated on a 3 GHz CMOS pipeline. To maintain this performance, the desired RRAM devices must be relatively fast [213]. These characteristics are chosen to achieve a target write latency of a 3 GHz CPU. As mentioned in Section 4.2, the RRAM write operation occurs sequentially prior to the read operation. Due to the sequential nature of the multistate register accessing the RRAM array, a half cycle is devoted to the read operation.

Table 4.2: Memristor and diode parameters

R_{on} [kΩ]	0.5
R_{off} [kΩ]	30
k_{on}	-0.021-0.07
k_{off}	0.0021-0.007
α_{on-off}	3
i_{on} [μA]	-1
i_{off} [μA]	1
V_{on} (diode) [V]	0.5
R_{out} (diode) [Î Ī]	1

Table 4.3: Access latency of a 16 bit MPR

Clock to Q [ps]	11.2
Setup Time [ps]	13.2
RRAM Read [ps]	28.6

The energy of the multistate register depends upon the RRAM switching latency, as listed in Table 4.4. $E_{Low-High}$ and $E_{High-Low}$ are the energy required to switch, respectively, to R_{off} and R_{on} for a single device write to the multistate register crosspoint array. Since the switching time of the memristor dominates the delay of a write to the multistate register, $E_{Low-High}$ and $E_{High-Low}$ increase linearly as the switching time increases. Note that the read energy only depends on R_{on} and R_{off} and is therefore constant for different switching times. The read energy, however, depends on the size of the crosspoint array (i.e., the number of RRAM devices), as listed in Table 4.5.

Table 4.4: Write latency and energy of a 16-bit multistate register

Write Time [cycles @ 3 GHz]	0.5	1.5	2.5	3.5	4.5
$E_{Low-High}$ [fJ]	2.24	5.26	8.3	10.49	13.23
$E_{High-Low}$ [fJ]	3.78	10.33	16.89	23.5	30.08

Table 4.5: Read access energy of RRAM

States per Multistate Register	4 States	16 States	64 States
$E_{read,Off}$ [fJ]	1.6	2.2	3.5
$E_{read,On}$ [fJ]	0.33	0.41	0.71

4.3.2 Layout and Physical Area

The layout of the proposed RRAM multistate register is shown in Figure 4.8. The layout of the multistate register is based on 45 nm design rules and scaled to the target technology of 22 nm. The number of RRAM devices within a crosspoint array is scaled from four devices to 64 devices. The MPR is evaluated for both the middle metal and lower metal approaches, as described in Section 4.2. The physical area is listed in Table 4.6.

Table 4.6: MPR area

		Area [μm^2]	Overhead [%]	Overhead per State [%]
	CMOS Register (1 state)	2.8	-	-
Lower Metal	MPR 4 states	5.5	96.2%	24%
	MPR 16 states	6.3	126.5%	8%
	MPR 64 states	8.1	192.5%	3%
Middle Metal	MPR 4 states	3.9	41.3%	10.3%
	MPR 16 states	4.3	54.7%	3.4%
	MPR 64 states	5.2	87.9%	1.4%

The transistors required to access the crosspoint, as shown in Figure 4.8, dominate the area overhead of both the lower metal and middle metal multistate register. Due to the relatively small on-resistance of the RRAM devices, the access transistor needs to be sufficiently large to facilitate a write operation. Additionally, CMOS transmission gates are used to ensure that there is no threshold drop across the

pass transistors. As a result, the area of the crosspoint memory is only a small fraction of the area overhead of the multistate register. Note that alternative RRAM technologies with a higher R_{on} supports smaller transistors and less area. Under these constraints, the most area efficient structure is a 64 bit array, as the overhead per state is, respectively, 0.08 $\mu\mathrm{m}^2$ for the lower metal approach and 3.75 $\mu\mathrm{m}^2$ for the middle metal approach.

As shown, the middle metal register requires less area than a lower metal multistate register. As described in Section 4.2 and depicted in Figure 4.8b, the middle metal register requires an additional track dedicated to the control transistors within the crosspoint array. Positioning the crosspoint array over the register also adds complexity as the upper metal layers can no longer be used to route signals above the multistate register.

4.3.3 Sensitivity and Device Variations

The sense amplifier compares the voltage across a biased RRAM device to a voltage level generated within the circuit. Any voltage above the voltage level produces a logical zero at the output, and any voltage below the voltage level produces a logical one. Similar to digital CMOS circuits, the structure is tolerant to variability in the RRAM resistance. To evaluate the sensitivity of the circuit to variations, the nominal R_{on} is varied from 0.35 to 0.65 $\mathrm{k}\Omega$. This range produces a maximum and minimum change of ± 2 mV in the voltage input of the sense amplifier. For

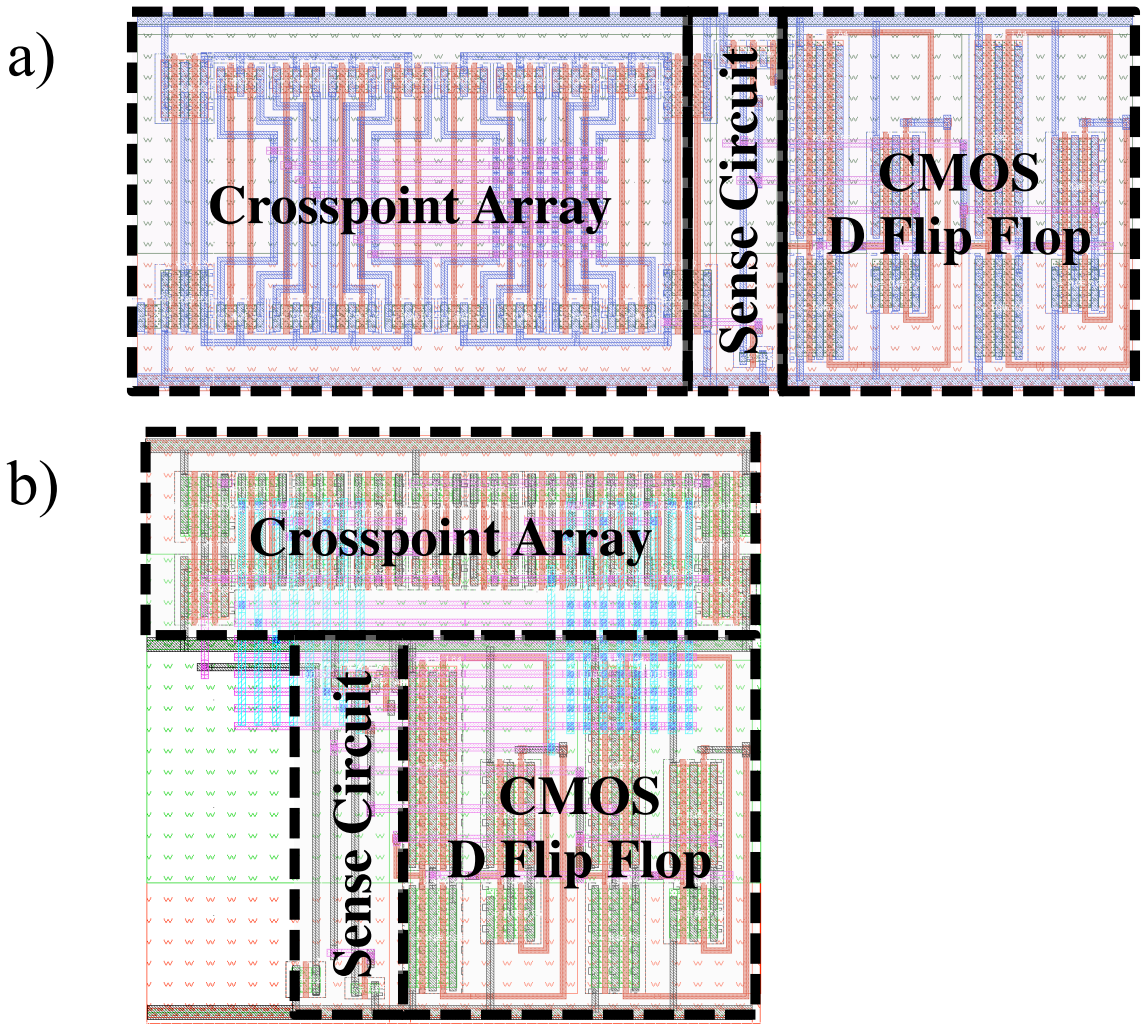


Figure 4.8: Physical layout of 64 state MPR within the crosspoint array on (a) lower metal layers (M1 and M2), and (b) upper metal layers (M2 and M3) above the D flip flop.

$21\ \mathrm{k}\Omega < R_{off} > 39\ \mathrm{k}\Omega$, a voltage ranging from -40 mV to +26 mV is produced. Both ranges represent a 30% variation in the device resistance of R_{on} and R_{off}. In these cases, the correct output state is read out, indicating a high degree of tolerance to variations in the RRAM resistance.

The RRAM circuit can tolerate an R_{on} of up to 12 $\mathrm{k}\Omega$ before the circuit produces

an incorrect output. In a 64 bit multistate register, this behavior corresponds to an increase in the RRAM read delay from 78 ps to 476 ps. With increasing R_{on}, the sense amplifier no longer generates a full range signal at the output, dissipating static energy. Much of this increased delay is due to the device operating near the switching threshold of the sense amplifier.

As R_{off} varies from 30 $\mathrm{k}\Omega$ to 300 $\mathrm{M}\Omega$, the performance of the circuit improves due to two effects. As the resistance increases, the voltage at the sense amplifier input also increases, placing the transistor into a higher bias state, which lowers the delay of the sense amplifier. Additionally, the large resistance of the sensed RRAM device prevents the sense line within the crosspoint array from dissipating charge, maintaining a high voltage at the input of the sense amplifier. Counterintuitively, this effect lowers the delay when R_{off} is greater than 30 $\mathrm{M}\Omega$. Due to the interplay of R_{on} and R_{off}, a delay tradeoff therefore exists between the average resistance of the RRAM technology and the resistive ratio of the device.

4.4 Multistate Registers as Multistate Pipeline Register for Multithread Processors — A Test Case

Replacing CMOS memory (e.g., register file and caches) with non-volatile memristors significantly reduces power consumption. Multithreaded machines can exploit the high density and CMOS compatibility of memristors to store the state of

the in-flight instructions within a CPU with fine granularity. Hence, using memristive technology can dramatically increase the number of threads running within a single core. This approach is demonstrated in this test case, where RRAM multistate registers store the state of multiple threads within a CPU pipeline. In continuous flow multithreading [209], the multistate registers are used as MPRs to store the state of multiple threads. A single thread is active within the pipeline and the instructions from the other threads are stored in the MPRs. The MPRs therefore eliminate the need to flush instructions within the pipeline, significantly improving the performance of the processor.

To exemplify this behavior, the performance and energy of a CFMT processor with the proposed RRAM-based MPRs have been evaluated [214]. To evaluate the performance, the GEM5 simulator [215] is extended to support CFMT. The energy has been evaluated by the McPAT simulator [216]. The simulated processor is a ten stage single scalar ARM processor, where the execution stage operates at the eighth stage. The performance and energy of the CFMT processor are compared to a switch-on-event (SoE) multithreading processor [217], where a thread switch occurs for each long latency instruction (e.g., L1 cache miss, floating point instructions), causing the pipeline to flush. The characteristics of the evaluated processors are listed in Table 4.7. The energy is compared to a 16 thread processor (i.e., with an MPR storing 16 states) which is a sufficient number of threads to achieve the maximum performance for most benchmark applications.

Table 4.7: SoE MT and CFMT processor configurations

	Switch on Event	**RRAM-based CFMT**
Number of pipeline stages	10	
CMOS process	22 nm	
Clock frequency [GHz]	3	
Switch penalty [cycles]	7	1 to 5
L1 read/write latency [cycles]	0	
L1 miss penalty [cycles]	200	
Data L1 cache configuration	32 kB, 4 way set associative	
Instruction L1 cache configuration	32 kB, 4 way set associative	
Branch predictor	Tournament , lshare 18kB/gshare 8kB	

Table 4.8: Performance speedup for different MPR write latencies as compared to switch-on-event multithread processor for CPU SPEC 2006

Benchmark	**MPR Write Latency [clock cycles]**				
	1	**2**	**3**	**4**	**5**
libquantum	1.35	1.28	1.21	1.15	1.09
bwaves	1.22	1.15	1.08	1.04	1
milc	1.47	1.26	1.18	1.11	1.06
zeusmp	1.85	1.59	1.40	1.29	1.21
gromacs	1.53	1.32	1.21	1.17	1.14
leslie3d	1.67	1.48	1.33	1.22	1.15
namd	1.40	1.24	1.15	1.08	1.04
soplex.pds-50	1.35	1.28	1.21	1.16	1.1
lbm	1.5	1.31	1.2	1.12	1.08
bzip2.combined	1.13	1.1	1.08	1.05	1.03
gcc.166	1.35	1.28	1.21	1.15	1.09
gobmk.trevorc	1.3	1.24	1.19	1.14	1.09
h264ref.foreman_baseline	1.06	1.02	1	1	1
GemsFDTD	1.45	1.3	1.18	1.08	1.04
hmmer.nph3	1.18	1.14	1.11	1.07	1.04
soplex.ref	1.7	1.42	1.29	1.19	1.1
gcc.c-typeck	1.33	1.26	1.21	1.15	1.1
gobmk.trevord	1.29	1.23	1.18	1.13	1.08
Average	1.40	1.27	1.19	1.13	1.08

Table 4.9: Energy and area for CFMT test case

	Switch on Event	**RRAM-based CFMT**	**Difference**
Thread switch energy [pJ]	109.9	9.1 @ 1 cycle penalty	-91.7%
		19.1 @ 2 cycle penalty	-82.6%
		29.2 @ 3 cycle penalty	-73.4%
		38.4 @ 4 cycle penalty	-65.1%
		48.2 @ 5 cycle penalty	-56.1%
Processor area [mm^2]	123.276	126.426	2.55%

The performance of the processors is measured by the average number of instructions per clock cycle (IPC), as listed in Table 4.8. The average speedup in performance is 40%. A comparison of the thread switch energy is listed in Table 4.9. The average energy per instruction for various CPU SPEC 2006 benchmarks is listed in Table 4.10, where the average reduction in energy is 6.5%. The area overhead for an 16 thread CFMT as compared to a SoE processor is approximately 2.5%, as listed in Table 4.9.

4.5 Conclusions

Emerging memory technologies, such as RRAM, are more than just a drop-in replacement to existing memory technologies. In this chapter, a RRAM based multistate register is proposed using an embedded array of memristive memory cells within a single flip flop. The multistate register can be used to store additional data that is not conventionally contained within a computational pipeline.

The proposed multistate register is relatively fast due to the physical closeness

Table 4.10: Energy per instruction for different CPU SPEC 2006 benchmark applications

Benchmark	**SoE MT [pJ/inst.]**	**CFMT**				
		RRAM MPR—Write Latency				
		1 cycle [pJ/inst.]	**2 cycles [pJ/inst.]**	**3 cycles [pJ/inst.]**	**4 cycles [pJ/inst.]**	**5 cycles [pJ/inst.]**
libquantum	15.17	14.12	14.29	14.46	14.63	14.80
bwaves	19.63	18.83	19.03	19.25	19.42	19.42
milc	24.51	22.61	23.23	23.47	23.74	24.11
zeusmp	21.10	18.04	18.62	19.19	19.18	19.95
gromacs	30.16	27.94	28.62	29.05	29.23	29.34
leslie3d	27.27	24.72	25.20	25.68	26.08	26.39
namd	22.90	21.42	21.91	22.21	22.50	22.65
soplex.pds-50	17.62	16.52	16.71	16.88	17.03	17.20
lbm	22.54	20.29	20.90	21.36	21.76	21.94
bzip2.combined	21.86	21.44	21.51	21.65	21.65	21.72
gcc.166	19.37	18.32	18.49	18.66	18.83	19.01
gobmk.trevorc	23.05	22.15	22.28	22.71	22.56	22.71
h264ref.foreman_baseline	25.95	25.27	25.35	25.50	25.69	25.76
GemsFDTD	23.89	21.88	22.43	22.99	23.36	23.49
hmmer.nph3	24.27	23.65	23.75	23.84	23.84	24.04
soplex.ref	21.92	19.47	20.04	20.44	20.80	21.17
gcc.c-typeck	19.94	19.16	19.12	19.27	19.43	19.58
gobmk.trevord	22.73	21.71	21.87	22.40	22.25	22.40
Average	22.44	20.97	21.30	21.61	21.78	21.98

of the CMOS and RRAM devices. A 16 bit multistate register requires only 54% additional area as compared to a single state standard register. The multistate register is also relatively low power due to the non-volatility of the resistive devices.

As an example, the proposed multistate register has been applied to a continuous flow multithreading processor, exhibiting a significant performance improvement of 40% as compared to a conventional switch-on-event processor. An RRAM-based MPR therefore enables novel microarchitectures, such as the CFMT. The proposed multistate register is shown to significantly improve performance and reduce

energy with a small area overhead.

Chapter 5

Reducing Switching Latency and Energy in STT-MRAM with Field-Assisted Writing

5.1 Introduction

The performance scaling of modern computing systems is largely constrained by conventional memory technologies. Six transistor (6T) SRAM, which has long been the workhorse of high performance caches, is projected to be replaced by 8T, 10T, and 12T variants to tolerate retention errors, variability, and read disturbance [218]. As a result, SRAM density has not increased commensurately with CMOS scaling.

Emerging resistive memories, which rely on resistivity (rather than charge) to carry information, have the potential to scale to much smaller geometries than charge based memories (e.g., SRAM). The smaller cell area, near-zero leakage power, and enhanced scalability make resistive memories viable alternatives to SRAM and

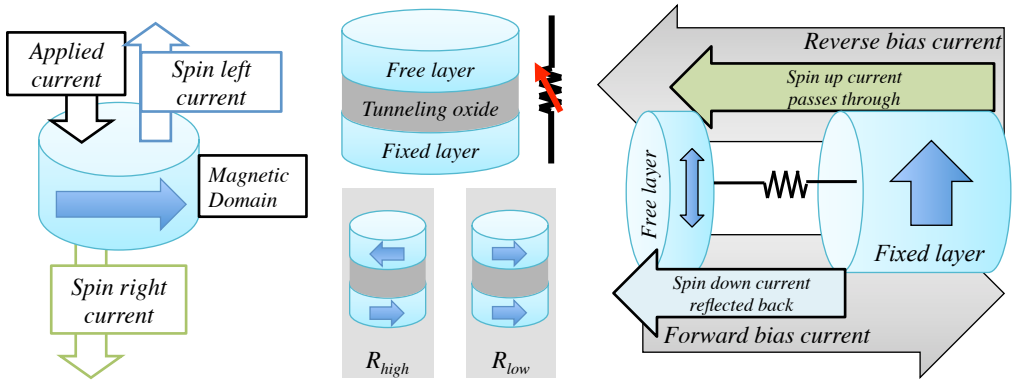
DRAM in next generation memory systems. Among other resistive memories, spin-torque transfer magnetoresistive RAM (STT-MRAM) exhibits low access latency (< 200 ps in 90 nm) [219], densities comparable to DRAM ($8F^2$) [220], and practically unlimited endurance [41]. STT-MRAM is close to becoming a CMOS-compatible universal memory technology. 64 Mb STT-MRAM products have already entered the marketplace [221]. Despite these advantages, STT-MRAM generally suffers long write latency and high write energy, which constrains the use of STT-MRAM to low activity caches (e.g., last level cache).

The storage element in an STT-MRAM cell is a magnetic tunnel junction (MTJ), which is the primary factor limiting the speed of STT-MRAM due to the relatively long switching latency. In addition, the write energy of STT-MRAM is orders of magnitude higher than SRAM. A constant, large amplitude current must be applied during the entire switching period, which dissipates large static power.

To address these issues, an MRAM field-assisted mechanism is proposed to be incorporated into STT-MRAM. The physical topology utilizes an assistive field current to destabilize the MTJ during switching, which reduces the switching latency of STT-MRAM by an order of magnitude, from 6.45 ns to 0.62 ns. The additional energy consumed by the field current can be amortized by applying the field over a row of STT-MRAM cells (along with the wordline), which leads to an 82% reduction in energy per cell. Evaluation of a microprocessor cache system demonstrates a 55% average energy reduction and a 5% speedup as compared to a standard SRAM

cache subsystem. Different from previous work [222] that trades off STT-MRAM retention time for improved write speed and energy, the approach described in this chapter does not require modification of the MTJ structure nor is the data retention time compromised.

The rest of this chapter is organized as follows. Background on STT-MRAM and cell topologies is provided in Section 9.2. The field-assisted writing mechanism is described in Section 5.3. Models of an STT-MRAM cell and array are presented, respectively, in Sections 5.4 and 5.5. Several STT-MRAM cell variants (with and without the applied field) are compared with SRAM within a microprocessor cache system in Section 5.6. Some conclusions are offered in Section 9.5.



(a) (b) (c)

Figure 5.1: Demonstrations of a) the domain dependent polarization effect, b) an MTJ stack, and c) the spin torque transfer effect.

5.2 MTJ Background

5.2.1 MTJ structure and operation

An MTJ is a two terminal resistive element that operates on the principle of spin dependent conduction through magnetic domains [41,49,50]. When applying a current to a magnetic domain, two spin currents (with opposite polarization) are generated across the device due to spin dependent tunneling and reflection, as illustrated in Figure 5.1a. Electrons passing through the domain exhibit a net spin polarization aligned with the magnetic domain, whereas electrons reflecting off the domain have a net spin anti-parallel to the domain.

An MTJ is a stack of two magnetic layers separated by a tunneling oxide, as illustrated in Figure 5.1b. One layer has a fixed magnetization direction, and the other (free layer) can flip between two opposite polarities, one parallel to the fixed layer and the other anti-parallel. When domains in the two layers are aligned (in parallel), electrons passing through both layers will be unimpeded; the MTJ exhibits a low resistance (R_{Low}). When domains in the two layers are anti-parallel, however, an electron will obtain a net polarity in one layer, and enter a layer with the opposite polarity. The electron may reflect off in the second domain. This effect increases the MTJ resistance (R_{High}).

Conventional MRAM circuits use two large orthogonal currents to generate magnetic fields within the free layer. These fields must be sufficiently strong to

induce a torque on the magnetization, which eventually induces a reversal in the polarity of the free layer. STT-MRAMs, however, utilize spin-dependent currents to alter the polarity of the free layer, as illustrated in Figure 5.1c. With reverse bias, current passes through the fixed layer and attains a large net magnetic polarity. Electrons in the STT current transfer angular momentum to the electrons in the free layer, thereby inducing a net torque on the free layer polarity. When the magnitude of the STT current reaches a threshold, the generated torque switches the free layer to a parallel alignment with the fixed layer. The switching mechanism is similar in the forward bias case, except that the free layer is subjected to a reflected spin current with a polarity anti-parallel to the fixed layer. The free layer will therefore switch to an anti-parallel alignment.

5.2.2 MTJ switching dynamics

Spin polarization of electrons incident on a free layer induces a torque on the magnetic polarity. This torque, depicted in Figure 5.2a, is immediately countered by a natural damping torque, which acts to stabilize the magnetic polarity along the long axis of the domain. When the current induced torque is sufficiently large to overcome the damping torque, the domain polarity is aligned with the short axis. At this point, the damping torque switches sides and assists the current induced torque which switches the polarity of the domain.

Note that this switching process is inherently stochastic. Since the current induced torque is parallel or anti-parallel to the resting polarity of the device, the effective torque on the polarity is zero (the cross product of two parallel or anti-parallel torques is zero). If the polarity deviates slightly from a resting position, the cross product becomes non-zero. This deviation is due to thermal fluctuations within the MTJ device. The probability of STT switching is therefore based on the magnitude of the current, bias duration, and ambient temperature [91].

5.2.3 Field-assisted switching

Stochastic switching requires that random thermal fluctuation must be sufficiently large to allow for STT current induced switching. Utilizing a perpendicular magnetic field during the switching process directly addresses this issue. Field-assisted switching requires application of an orthogonally oriented magnetic field in addition to the STT current to reduce the switching latency. The magnetic field torque destabilizes the MTJ polarity towards the short axis, as illustrated in Figure 5.2b. As a result, the spin transfer torque exhibits a larger effective magnitude. This method ensures that the process is less reliant on random thermal fluctuation for switching to occur.

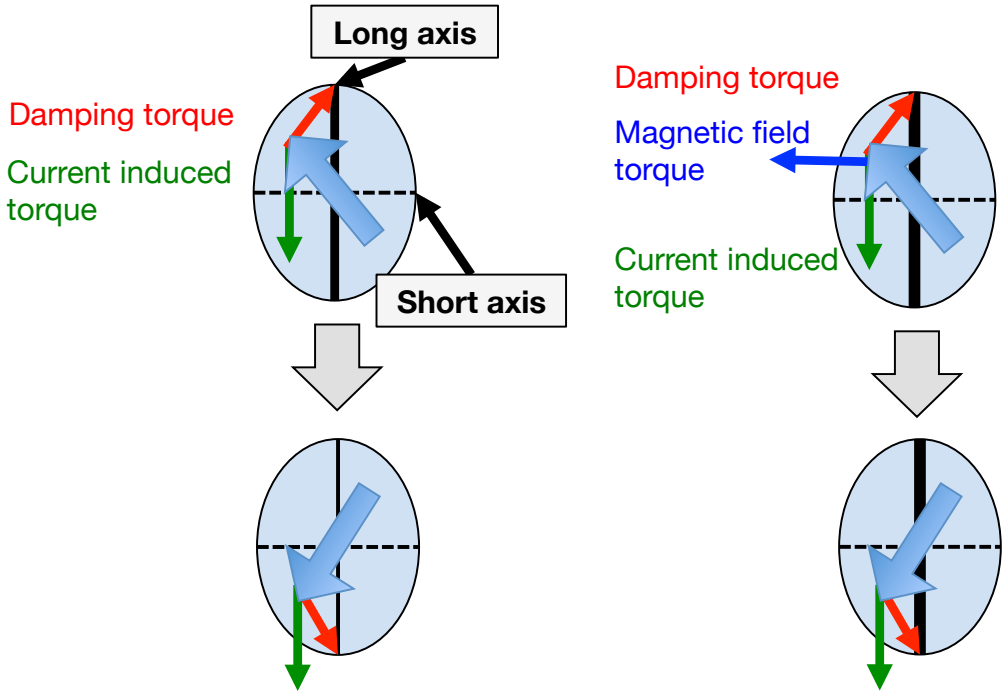


Figure 5.2: Switching dynamics for a) standard STT switching, and b) field-assisted STT switching.

5.2.4 STT-MRAM cell structure

STT-MRAM is CMOS-compatible. A typical one transistor, one MTJ STT-MRAM cell is shown in Figure 5.3. The MTJ serves as a storage element and the resistance represents a single data bit. The access transistor, in series with the MTJ, behaves as a gating element. To read a cell, the wordline (WL) is asserted and the resistance of the MTJ is sensed. To write a cell, the wordline is turned on and the cell is driven by a write current. The direction of the write current determines the value of the bit written into the cell.

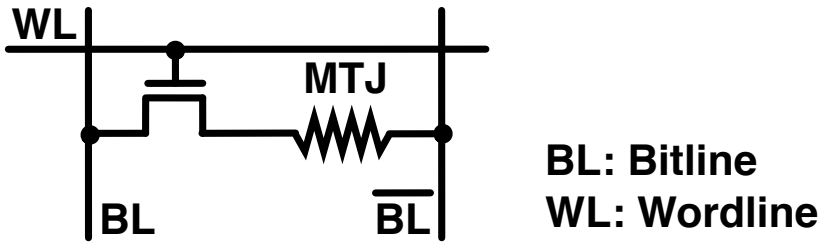
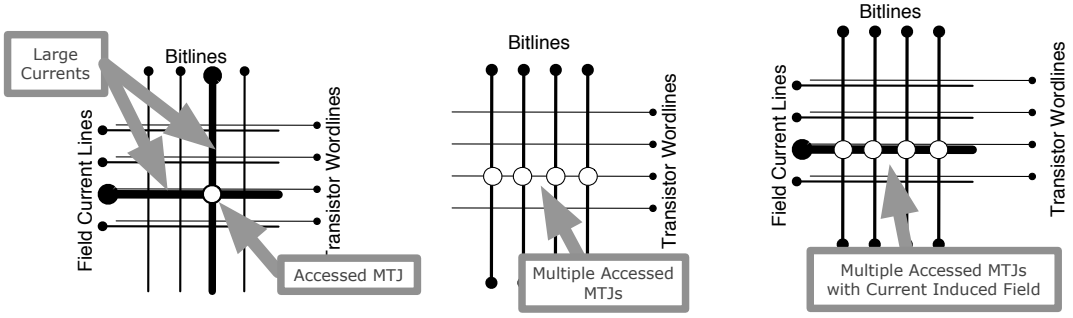


Figure 5.3: A one transistor, one MTJ (1T-1MTJ) STT-MRAM cell.

5.3 Field-assisted STT-MRAM



(a) (b) (c)

Figure 5.4: Current biasing schemes for a) a conventional MRAM, b) an STT-MRAM, and c) the proposed field-assisted STT-MRAM

Since the introduction of the spin torque transfer effect into MTJ switching [41], MRAMs have exclusively used this effect for writing. The STT effect, however, can complement the field-assisted excitation of the magnetic free layer within an MTJ. Classical MRAMs use two perpendicular currents with a single selected MTJ at the intersection to produce a magnetic field that acts on the free layer of an MTJ (see Figure 5.4a). This approach suffers from two key issues: (1) the use of two currents to switch a single bit consumes a large amount of energy, and (2) the MTJs

in adjacent columns and rows are half-selected by the high fields caused by the write currents, constraining the design space to avoid erroneous writes [223].

The STT effect overcomes these problems by using a single current that passes through the MTJ. This technique enables a row of MTJs (along the wordline) to be written in parallel, as illustrated in Figure 5.4b. The direction of the applied current translates into the final state of the MTJs, *i.e.*, a forward bias exclusively sets the device to "0," and a reverse current exclusively sets the device to "1." The switching current is much lower than that of toggle-mode MRAMs, which alleviates the half-select problem. The write latency, however, remains significantly longer than the read latency, and the switching energy is also significantly greater than SRAM. Supplying a sufficiently large write current requires a large access transistor, which reduces the density of the circuit.

The approach proposed herein combines an STT-based current with a field-generating current used in toggle-mode MRAM circuits. In this approach, the field current produces an assistive magnetic field that destabilizes the MTJs across a row. Each MTJ is biased with an STT current that controls the switching direction of the MTJs in each column. Use of a field current in this manner has two beneficial effects: (1) the alignment of the field with respect to the MTJ can destabilize the device, which reduces both the write latency and energy, and (2) the field current is shared across the row, ensuring that the energy consumption of the field current is amortized across all of the cells within a row.

5.3.1 Related work

External magnetic fields are used in conventional MRAM as the primary switching mechanism. This work shows that the superposition of an external magnetic field with local STT currents reduces both the switching latency and energy while removing the issue of half-select disturbance in on-chip, write intensive memories. The use of both a magnetic field and an STT current for switching was demonstrated physically in [224] but explored the context of discrete off-chip memories as a replacement for conventional MRAM switching. The approach in [224] used a nascent STT device and an older CMOS technology. The aggressive sizing and structure are limited to "DRAM-replacement" applications where the design rules facilitate denser cell layouts. In the proposed method presented here, the magnitude of the applied current and size of the memory are used to reduce the switching latency of the MTJ device. For the first time, the combination of the reduced latency and the shared switching current are shown to lower the system level energy consumption of a cache, resulting in both power and performance improvements.

A patent by Andre *et al.* presented a similar structure that utilizes a field current to set the MTJ device to an initial reset state (either R_{on} or R_{off}) prior to writing the device. This method enables the uni-directional cells and diodes to select the individual memory cells [225], which provide cell density advantages appropriate for "DRAM-replacement" memory applications. A reset process, however, requires

the MTJ devices to undergo two switching events for every write, one to switch to a reset state (either R_{on} or R_{off}), and a second switching event to write the correct state for the remaining bits. This process doubles the write latency of an MRAM array. The approach presented in this chapter requires CMOS transistors for bipolar switching and utilizes magnetic fields to enhance the dynamic behavior of the switching process to reduce the energy of a write, while sharing the field current to amortize the energy across multiple columns. The device is not reset to a stable state but rather an additional torque is applied dynamically to enhance the switching process, reducing the overall write latency and enabling use in latency critical application.

5.4 Model of a field-assisted STT-MRAM cell

An individual MTJ is modeled here using the classical Landau-Lifshitz-Gilbert macrospin model with thermal agitation based on a Langvin random field using the M^3 simulator [226]. The MTJ free layer parameters are selected to ensure that the thermal stability factor (Δ) provides ten year retention of the device state ($\Delta = 40$). The MTJ parameters for the resistance and TMR (from ITRS 2011 [51]) are listed in Table 8.1. The critical switching current of the MTJ is dependent on the geometric and material properties of the free layer, permitting the current to be determined from the free layer geometry. The resultant critical current is in agreement with the

switching current targeted by the ITRS [51].

Table 5.1: MTJ parameters

Saturation Magnetization (M_s)	8 x 10^5 A/m
Long axis	70 nm
Short axis	20 nm
Thickness	2.9 nm
$\mathrm{R_{on}}$	5 kΩ
TMR	150%
I_{crit}	39.4 μA

The predictive technology model (PTM) is used to characterize the cell access transistor [227]. A low threshold transistor is used for the selection device and is modeled with a 20% reduction in threshold voltage. The wordline is bootstrapped to $V_{DD}+V_{th}$. The cell transistor width is set to provide a switching current 1.5 times greater than the critical switching current. This width is selected to ensure that the device operates in precessional mode [91], while allowing the access transistor to be small.

Table 5.2: STT-MRAM cell parameters

STT-MRAM Cell Type	Isometric	Minimum	Field-Assisted
Technology	22 nm		
Supply ($\mathrm{V_{DD}}$)	0.8 V		
Nominal switching current	59.1 μA		
STT switching current	75 μA	59.1 μA	66.2 μA
Field line spacing	N/A	N/A	21 nm
Cell length	119 nm	119 nm	161 nm
Cell width	228 nm	175 nm	167 nm

Durlam *et al.* present a classical MRAM cell and memory. Measurements of the field observed by the free layer are demonstrated at a distance of 0.3 $\mu\mathrm{m}$ for a 0.6 $\mu\mathrm{m}$ process. Simple linear scaling of this dimension is not sufficient as the MTJ

dimensions are proportionally larger than a classical MRAM. To compensate, the MTJ dimensions are scaled linearly and the thickness of the MTJ stack is assumed to occupy an additional 10 nm. This thickness is typical of many demonstrated STT-MTJ stacks [228, 229].

The cell layout is based on 45 nm FreePDK design rules and scaled to 22 nm, as shown in Fig 5.5. A spectrum of cell sizes is evaluated for performance. The base cell area is $55.5F^2$. In prior work, the area of a conventional 1T-1MTJ cell is shown to be $49.9F^2$ with the same logic process rules, indicating that the area overhead of the metal line supporting the additional field current is small [230, 231]. This cell has a relatively large cell density as compared to commodity STT-MRAM ($6F^2$ [51]) because the layout design rules originate from a logic process. A standalone memory process with tighter design rules would provide greater density.

Three distinct physical configurations of a 1T-1MTJ memory cell are compared and listed in Table 5.2. The field-assisted STT-MRAM cell (Field-Assisted) is compared with a minimum-sized 1T-1MTJ cell capable of supplying the same nominal switching current (Minimum). The additional metal line devoted to the field current impedes contact sharing and consumes additional area as compared to the nominal cell. The third memory cell (Isometric) has the same total area as the field-assisted cell. Due to extra area consumed by the bit lines above the silicon substrate, the field-assisted cell can use a slightly larger transistor than the nominal cell without affecting cell density, resulting in a slightly larger STT switching current.

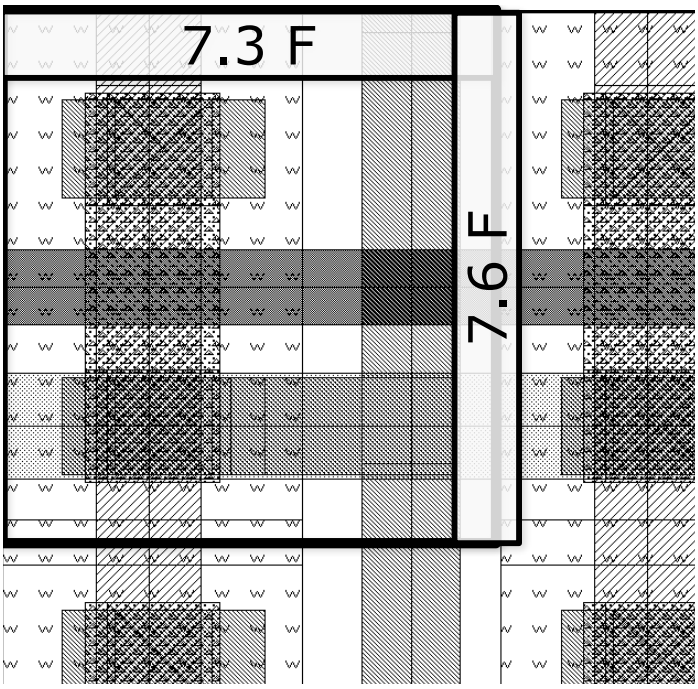


Figure 5.5: Layout of the proposed field-assisted STT-MRAM cell.

The magnetic field through a current loop can be estimated by the Bio-Savart's law [232],

$$B = \frac{\mu_0 I_{field}}{2\pi d}. \tag{5.1}$$

The current through the MTJ induces a spin torque on the free layer, generating a magnetic field that adds linearly to the magnetic field generated by the field current. The magnetic field produced by the STT is assumed to be negligible for two reasons. The STT current is almost two orders of magnitude smaller than the field current, making the field generated by the STT current relatively small. Secondly, the field current is applied to the MTJ before the STT current is applied, ensuring

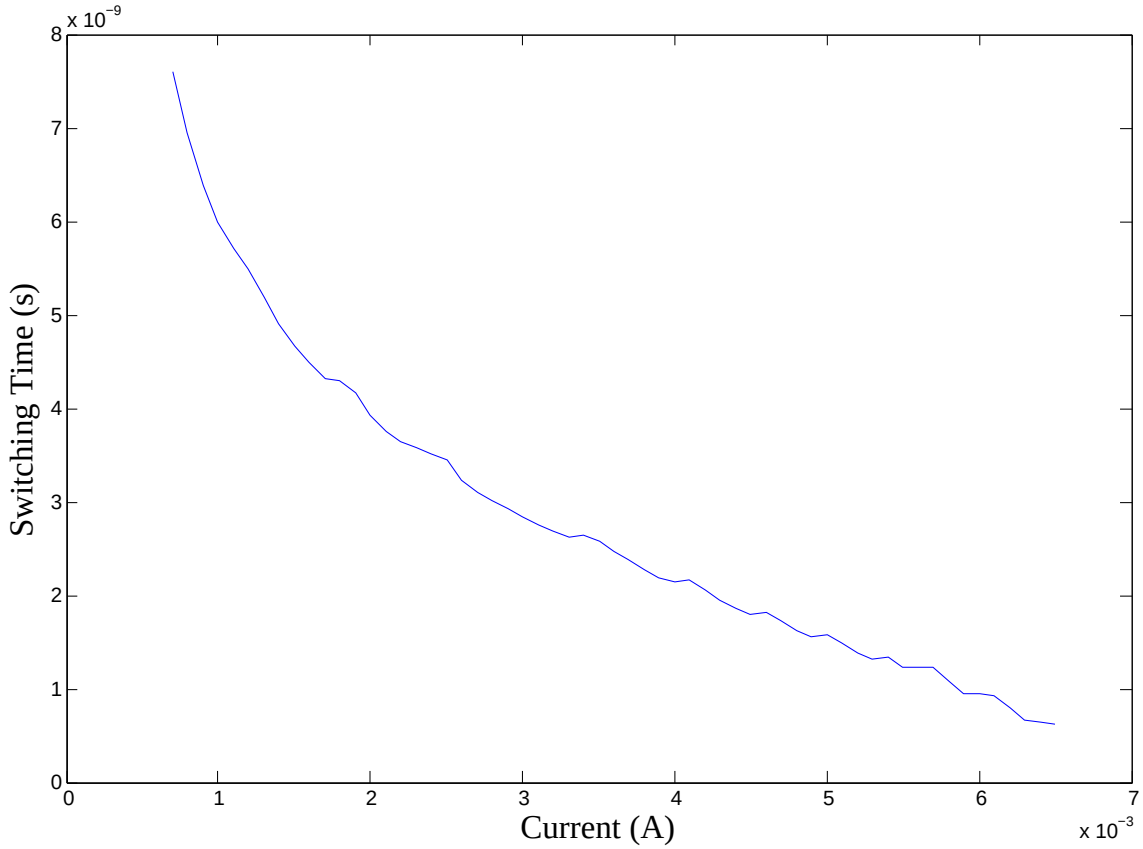


Figure 5.6: Switching latency of a field-assisted classical MRAM cell. The STT switching current is 59.1 μA

that the free layer magnetization is in an unstable state prior to application of the STT current. As a result, the magnetic field of the STT current does not affect the destabilization process.

5.5 Model of an STT MRAM Array

Optimizing the energy consumed by an MRAM array with a field assisted write produces a tradeoff between the size of the array and the current bias when minimizing the switching time of an MTJ. The parasitic impedances of the array, extracted from the cell layout, are listed in Table 5.3 [227].

Table 5.3: Memory array parameters

R_{flcell} (Ω)	0.7
C_{flcell} (aF)	28.8

The array is biased using a field current that traverses the entire row. As the size of the row increases, the energy associated with the field current is amortized across the entire row. The energy associated with the field current is the sum of the dynamic energy to charge the line as well as the static current to generate the magnetic field. Expression (5.2) quantifies this dependance, where R_{flcell} and C_{flcell} describe, respectively, the per cell parasitic resistance and capacitance, N describes the number of cells in a row, R_{access} is the resistance of the access transistor, V_{DD} represents the supply voltage, $t_{switching}$ is the MTJ switching latency, and I_{field} is the generated field current of the line. The dynamic component of the energy is therefore a function of the array width and DC voltage on the bit line during a write.

$$E_{field} = R_{flcell}C_{flcell}N\frac{NR_{flcell}}{NR_{flcell} + R_{access}}V_{DD} + V_{DD}I_{field}(t_{switching}) \qquad (5.2)$$

The energy of the static current is a function of the field current, supply voltage, and switching time of the MTJ. The static component is independent of array size as the supply voltage is constant and the voltage drop is across the peripheral write drivers and the array. The array field current is also constrained by the resistance of the field line,

$$I_{field} R_{flCell} N \leq V_{DD}. \tag{5.3}$$

The energy to switch a single MTJ (E_{switch}) is

$$E_{switch} = I_{STT} V_{DD} t_{switching}, \tag{5.4}$$

where I_{STT} is the spin torque switching current. E_{switch} is therefore only dependent on the switching time of the MTJ. The total energy per bit is

$$E_{total} = E_{switch} + \frac{E_{field}}{N}. \tag{5.5}$$

The switching energy is shown in Figure 5.7. For comparison, the minimum energy to switch an MTJ, as described by (5.4) for a non-field-assisted STT-MRAM cell, is 0.3 pJ per bit. The minimum switching energy of the field-assisted cell is 0.054 pJ per bit with a corresponding switching latency of 618 ps. Due to the bit line resistance, larger rows support a maximum field current at a specific supply voltage.

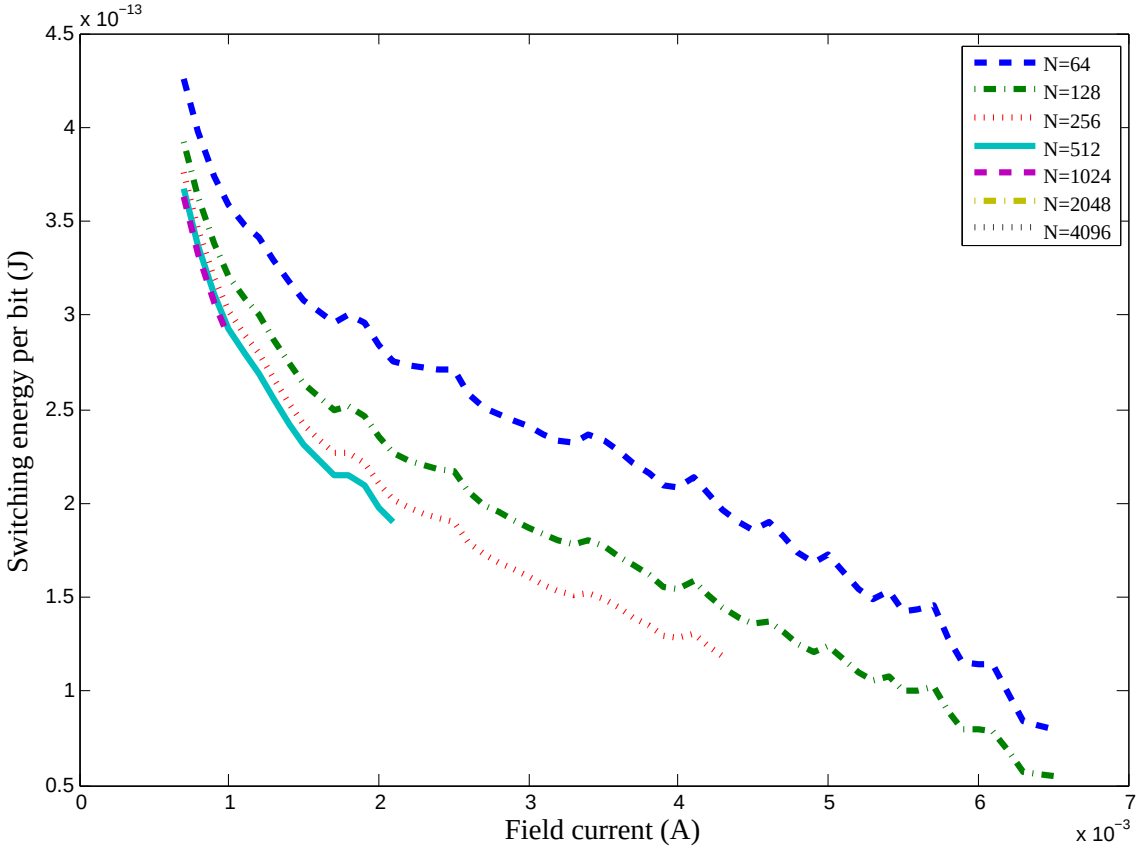


Figure 5.7: Switching energy of a field-assisted classical MRAM cell.

A sufficiently high field cannot be generated to reduce the switching latency of the MTJ, ensuring that the energy consumption is higher than with a shorter row. An optimum row length therefore exists that minimizes the overall switching energy of an array during a write. For the configuration shown in Figure 5.7, the optimum row length is 128 cells.

As illustrated by the figure, increasing the number of cells in a row produces a linear increase in energy consumed per bit. However, as the row length increases,

the maximum current becomes bounded. For latency critical as well as energy critical circuits, the field currents should be maximized for a given row length, and a larger current should be used rather than a longer row, except for small row lengths.

5.5.1 Effects of stochastic switching

As previously described, STT switching is a stochastic operation [221]. While deterministic information is sufficient to determine a suitable design point, practical design methods require that the stochastic nature of the switching process be considered.

Table 5.4: Energy and latency of STT-MRAM cells

Cell Type	Latency (ns)	σ (ns)	90% (ns)	Energy (fJ/bit)
Field-Assisted	0.47	0.481	0.996	93.4
Field-Assisted ($\Delta = 30$)	0.18	0.18	0.38	35.4
Minimum	4.96	1.62	6.65	316.9
Isometric	3.06	0.94	4.10	246.0

The energy and latency for each of the physical memory cells are listed in Table 5.4. Each cell type is evaluated at a row length of 128 with a 6.5 mA field current applied to the device. The field-assisted cell exhibits a significant reduction in energy and latency as compared to the nominal and isometric STT cells. As the field is applied, the switching latency decreases; the standard deviation, however, falls disproportionately. A nominal STT cell exhibits a switching latency of 4.96 ns with a 30% standard deviation. The field-assisted cell exhibits a 0.47 ns latency with 102% standard deviation. Intuitively, as the field is applied, the effect of the

damping torque is diminished and thus the system is more susceptible to thermal fluctuations during switching. This effect causes greater variability in the switching latency.

For comparative purposes, a field-assisted cell with reduced non-volatility is also presented. Unlike the baseline cell, this cell assumes a reduced thermal barrier for the MTJ which lowers the retention time of the MTJ to one day. This combination produces the shortest latency and lowest energy configuration. The combination of a reduced thermal barrier also exhibits no additional variability as compared to the baseline field-assisted cell. In subsequent analysis, however, the baseline cell is designed to ensure that a typical industrial ten year retention time is maintained.

5.6 Cache Evaluation

Although STT-MRAM has been projected to replace SRAM caches in next generation memory applications, the relatively long write latency and high write energy confine the use of STT-MRAM to last level caches. Field-assisted STT-MRAM, however, which significantly improves both the write latency and energy, serves as a viable universal cache candidate. The development of L1 and L2 caches with a field-assisted STT-MRAM is evaluated in this section. SRAM caches and caches using conventional STT-MRAM (without the field-assisted switching mechanism) are treated as a baseline for comparative purposes.

Naive replacement of SRAM arrays and sensing circuitry with STT-MRAM arrays would degrade performance in write critical caches due to the long switching latency, making the comparison unfair. The baseline STT-MRAM (Nominal and Isometric) caches therefore incorporate two state-of-the-art architectural techniques to improve system performance while tolerating write latency. The caches are typically divided into multiple subbanks to increase the parallel throughput of data acccesses and to amortize the cost of the peripheral logic circuitry. *Subbank buffering* [233] adds an SRAM write buffer in front of each cache subbank (Figure 5.8a), which locally buffers on-going writes. When data is stored within a subbank buffer, the H-Tree data bus, which is shared across all of the subarrays, is available to serve the next cache access while the long latency STT-MRAM write is local within the sub-bank. Decoupling the access circuitry and interface bus from the long latency write significantly improves the cache throughput. Additionally, differential writes [234] is a technique commonly used to reduce the write energy. Before a write, the stored data are read and compared to the to-be-written data. Only STT-MRAM cells with different binary states actually switch.

Field-assisted STT-MRAM caches (Figure 5.8b) also employ subbank buffering, but do not incorporate differential writes since all of the STT-MRAM cells in a row are affected by the field. To guarantee a successful STT-MRAM switching process, a checker read is issued after every write. Upon a write failure, a retry write is issued.

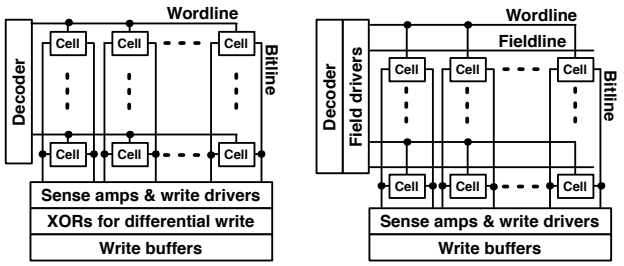


Figure 5.8: Array organization for a) baseline STT-MRAM, and b) field-assisted STT-MRAM.

5.6.1 Simulation Setup

The cycle accurate SESC simulator [235] has been modified to model a multithread (CMT) processor with eight cores and four threads per core operating at 4 GHz. The configuration for the baseline caches and memory subsystem is listed in Table 5.5. The L1 and L2 caches use STT-MRAM with different cell types. CACTI [236] and NVSim [237] are used to estimate the cache energy and access latencies. The cache capacities are maintained the same for both the STT-MRAM and SRAM caches. The estimated cache latencies for these configurations are summarized in Table 5.6. For the baseline STT-MRAM cache configuration, the isometric cells are used for the L1 caches to minimize the MTJ switching latency, and minimum sized STT-MRAM cells are used for L2 to decrease the cache area and read latency. The field-assisted STT-MRAM cache configuration uses the field assisted cells for all of the caches within the hierarchy.

A wide range of parallel workloads have been simulated for each configuration.

Table 5.5: Cache and memory parameters

L1 Caches	
iL1/dL1 size	32kB / 32kB
iL1/dL1 block size	64B / 64B
iL1/dL1 round-trip latency	2 / 2 cycles (uncontended)
iL1/dL1 ports	1 / 1
iL1/dL1 banks	1 / 1
iL1/dL1 MSHR entries	8 / 8
iL1/dL1 associativity	2-way / 2-way
Coherence protocol	MESI
Consistency model	Release consistency
Shared L2 Cache and Main Memory	
Shared L2 cache	4MB, 64B block, 8-way
L2 MSHR entries	64
L2 round-trip latency	20 cycles (uncontended)
Write buffer	64 entries
DRAM subsystem	DDR3-1600 SDRAM
Memory controllers	4

Table 5.6: STT-MRAM cache parameters (cycle: 250 ps)

	Baseline STT	Field-Assisted STT
iL1/dL1 latency	1 cycle	1 cycle
L1s write occupancy	17 cycles	4 cycles
L2 latency	6 cycles	7 cycles
L2 write occupancy	28 cycles	4 cycles

The benchmark suite includes nine software applications, among which three programs are from SPEC OMP2001 [238] and six programs are from SPLASH2 [239]. All workloads are executed in 32 threads on an eight core processor.

5.6.2 System Performance and Energy

The system performance and cache energy are shown in Figures 5.9 and 5.10. All of the comparisons are normalized to the performance of the SRAM caches with the same capacity.

The field-assisted STT-MRAM caches exhibit a slight performance increase as

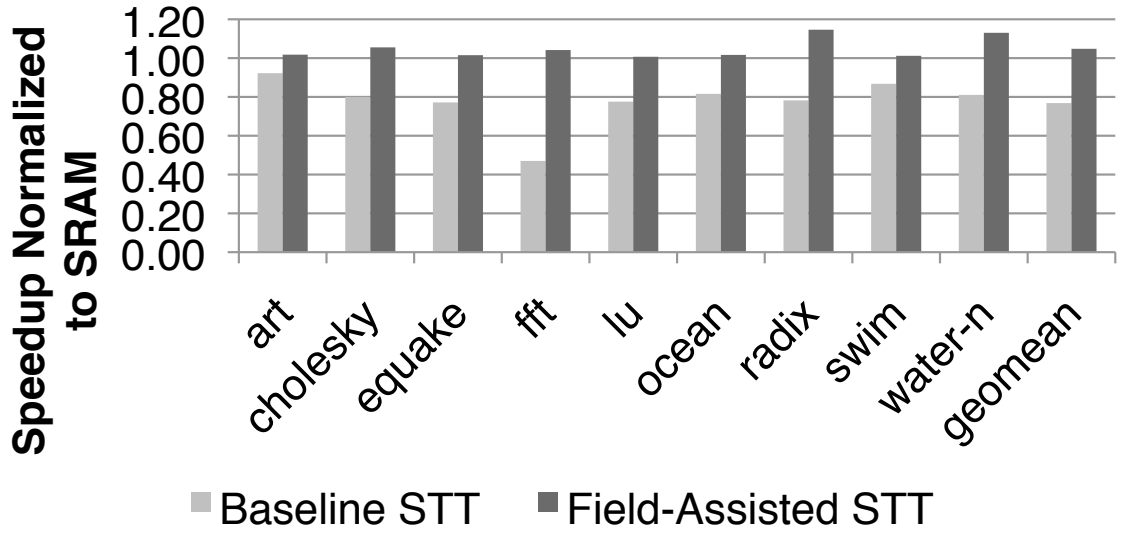


Figure 5.9: System performance of STT-MRAM caches normalized to baseline SRAM caches for each cell type.

compared to the SRAM caches (see Figure 5.9) because the STT-MRAM caches occupy smaller area while maintaining the same capacity, hence benefiting from a shorter wire delay. The baseline STT-MRAM caches exhibit an overall decrease in performance as compared to the baseline SRAM caches due to the long write latency. Despite subbank buffering, the reads can be blocked by writes when there are subbank conflicts.

For these applications, STT-MRAM based caches require less energy (see Figure 5.10). The field-assisted STT-MRAM caches consume slightly higher energy as compared to the baseline STT-MRAM caches due to two reasons: (1) the field current consumes additional energy, and (2) differential writes are applied to the baseline STT-MRAM but not to the field-assisted STT-MRAM. In the application LU, however, the field assisted STT-MRAM caches consume less energy. This behavior

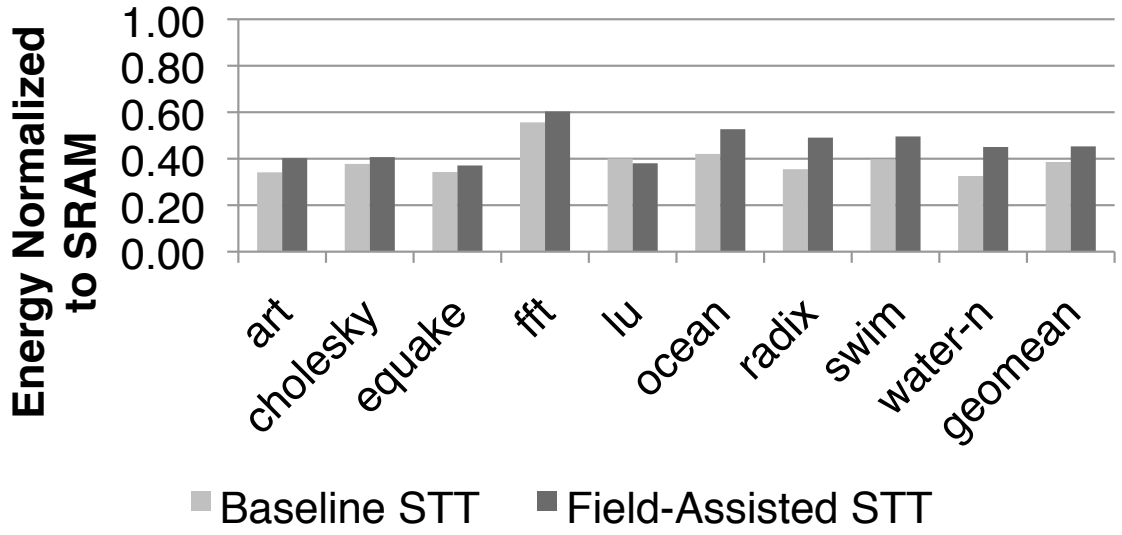


Figure 5.10: Energy of STT-MRAM caches normalized to baseline SRAM caches for baseline and field-assisted cell types.

occurs because LU uses a greater number of bit flips during the write operations. As a result, differential writes have less of an effect on the write energy as compared to other applications using isometric or nominal STT-MRAM cells.

The power dissipated by the benchmarks circuits is depicted in Figure 5.11 for STT-MRAM and SRAM caches. For all of the STT-MRAM caches, the leakage power is less than SRAM. The power dissipated by the read operations is also reduced due to the smaller array area and shorter wires. For the baseline STT-MRAM caches, the power required by the write operations is comparable to the power required by the SRAM writes because the MTJs consume greater switching power but the access time is smaller than the SRAM caches. The field-assisted STT-MRAM caches require higher write power due to the additional field currents applied to each write.

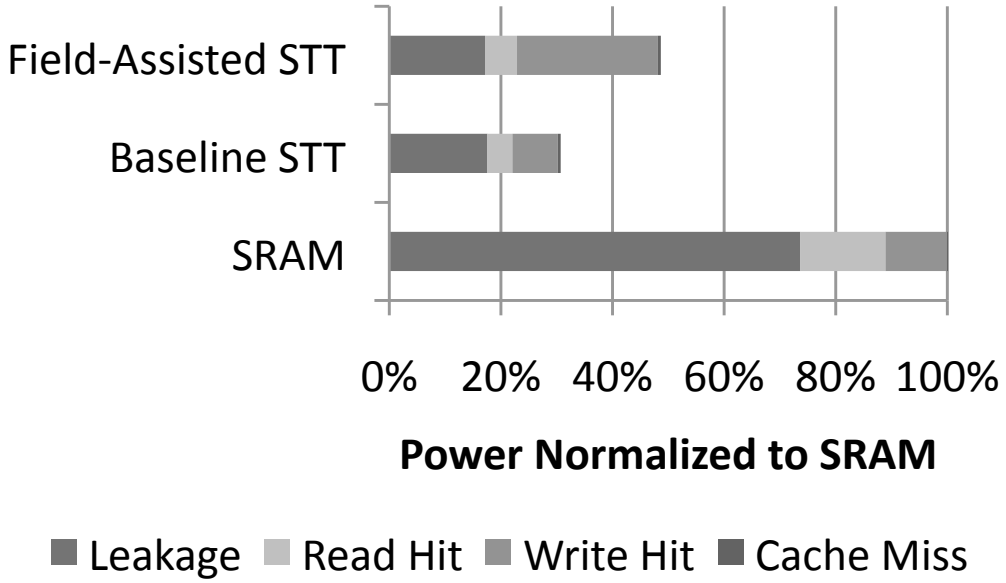


Figure 5.11: Power dissipation of STT-MRAM and SRAM caches.

The field-assisted STT-MRAM caches, however, provide faster write and shorter execution time; hence, the effect of the field currents on the total energy is amortized across the row.

The baseline STT-MRAM caches exhibit a greater than 20% performance penalty to realize an energy savings of approximately 60%. The field-assisted STT-MRAM caches provide a 5% increase in performance with an energy savings of 55% as compared to SRAM, reducing energy while maintaining performance.

5.7 Conclusions

The field-assisted approach utilized in MRAM cells reduces the switching latency of an STT-MTJ. The mechanism of STT switching is reviewed and a field-assisted STT-MRAM is presented. An array model of the switching latency and energy consumption for different field currents and array sizes is also described. It is shown that the per bit switching latency is reduced by a factor of four. If the nonvolatility constraints are relaxed, the overall switching latency is reduced by a factor greater than ten.

Several field-assisted STT-MRAM cells are compared to minimum sized and isometric area based STT-MRAM cells. Each of these cells is evaluated for a variety of applications and compared to standard L1 and L2 SRAM caches. The field-assisted STT-MRAM cache demonstrates a 25% performance improvement as compared to a non-field assisted cache STT-MRAM cache and a 5% improvement as compared to an SRAM cache while reducing overall energy consumption by an average of 55% as compared to an SRAM cache. The reduction in both switching energy and latency support embedded high performance STT-MRAM based cache subsystems, enabling the use of STT-MRAM in upper level caches within high performance microprocessors.

Chapter 6

2T - 1R STT-MRAM Memory Cells for Enhanced On/Off Current Ratio

STT-MRAM is limited by a small on/off resistance ratio. This limitation requires sophisticated read circuitry which leads to greater sensitivity to noise. To address these limitations, two memory cells are proposed that significantly improve the output read ratio. These memory cell variants utilize additional CMOS transistors within the cell to enhance the observed on/off resistance ratio of the MTJ device leading to a shorter read delay. Additional transistors are added in either a gate connected or diode connected manner to the adjacent metal lines that interface with the sense circuitry. Each cell exhibits an order of magnitude increase in the current ratio as compared to a traditional 1T - 1R structure while requiring more area and delivering comparable energy efficiency under high bias. This improvement in current ratio yields a 29% and 81% reduction in memory sensing delay as compared, respectively, to the standard 1T - 1R STT-MRAM memory cell and a 8T-SRAM.

Each cell type is introduced in Section 6.1. Methods for modeling STT-MRAM

arrays along with an evaluation of the sense margin and sense ratio for each cell are presented in Section 6.2. Some conclusions are offered in Section 9.5.

6.1 STT-MTJ Memory Cells

Three basic cell types are described for use in STT-MTJ memories shown in Fig. 6.1. The standard 1T - 1R memory cell is described in Section 6.1.1, followed by the proposed 2T - 1R cell variants in Section 6.1.2, and a discussion of the effects of technology on the memory array write current in Section 6.1.3.

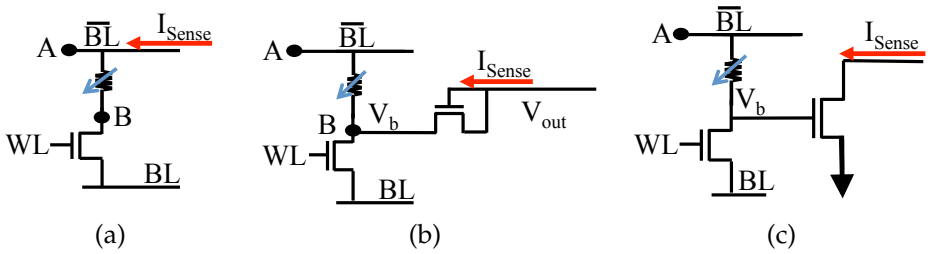


Figure 6.1: Circuit diagram of STT-MTJ memory cells: (a) standard 1T - 1MTJ, (b) 2T - 1MTJ diode cell, and (c) 2T - 1MTJ gate cell.

6.1.1 1T - 1R cell

The 1T - 1R cell, the basic building block of resistive memory arrays (see Fig. 6.1a), must satisfy several design constraints to operate correctly. At full bias, the internal cell transistor and access circuitry must supply sufficiently high current to ensure that the MTJ switches (I_c); however, currents in excess of this amount are

typically required for high speed switching. For reads, a cell current must remain sufficiently below the critical current to mitigate the potential for erroneous writes to the device. Moreover, each transistor isolates a selected memory cell from any peripheral cells to maintain the required sense margin. For this purpose, the read operation biases the access transistor within the linear region; a reverse bias would needlessly reduce the sense margin.

A 1T - 1R cell is the simplest memory cell topology for typical STT-MRAM technologies. The sense margin of the device is observed as a voltage or current proportional to the TMR of the device.

6.1.2 2T - 1R cells

Alternate memory cell topologies utilizing an additional transistor can produce voltage and current amplification without sacrificing immunity to leakage current within the MRAM array.

6.1.2.1 Diode connected transistor read port

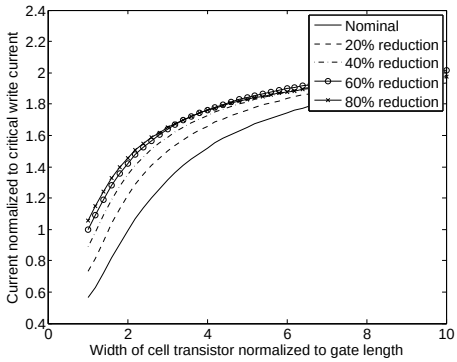
A diode connected transistor incorporated into a memory cell, as shown in Fig. 6.1b, amplifies the voltage of the internal node of the memory cell (node B) to produce a current and voltage signal at the transistor output. The maximum amplification occurs when node B is biased to ensure that the R_{on} and R_{off} states produce a voltage, respectively, above and below the threshold of the transistor.

6.1.2.2 Gate connected transistor read port

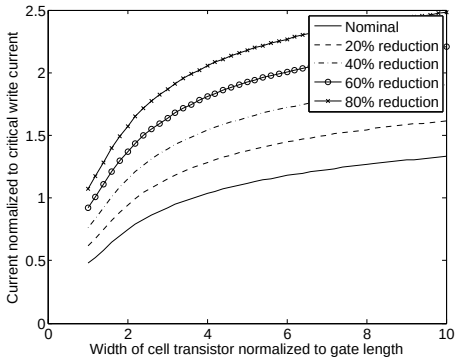
A gate connected memory cell, as shown in Fig. 6.1c, achieves the same amplification as the diode connected transistor and operates at a similar maximum voltage. This topology, however, differs in several key aspects. First, the gate connected transistor is electrically isolated from node B, facilitating the addition of multiple gate connected read ports. Secondly, the source of the transistor is connected to ground, eliminating any source body voltage bias, improving the conductance of the transistor. Thirdly, the output current margin is a function of transistor width which can be increased to improve the sense margin.

6.1.3 Effect of technology on write current

Given that the writes in both the standard 1T - 1R cell and the 2T - 1R memory cells occur in the same manner, a reduction in the threshold voltage can increase the available write current for a given CMOS technology. The source current as a function of transistor device width and threshold voltage is shown in Fig. 6.2. As expected, the write current increases with decreasing threshold voltage. In the reverse write current case, the cell transistor exhibits a threshold drop due to the inability of an NMOS transistor to pass a full voltage swing. Threshold voltage reduction shortens the minimum width of the NMOS transistor needed to provide sufficient current to write to the MTJ. This voltage drop is the primary limitation



(a)



(b)

Figure 6.2: Write current for 1T - 1R cell versus gate length of access transistor and threshold reduction in the CMOS transistor. (a) Forward write current, and (b) reverse write current.

to sourcing a sufficiently high write current in standard CMOS technologies. As shown in Fig. 6.2, a 60% reduction in the threshold current allows a minimum sized device to supply the required 35 μA. Under forward bias, the maximum current is only limited by the total resistance of the write lines and memory cell. A reduction in the threshold voltage has a smaller effect on the on-resistance of the cell transistor within the linear region as compared to the saturation region. A reduced threshold voltage under forward bias therefore exhibits a smaller increase in the transistor

write current than in the reverse bias case. The current saturates to approximately twice I_c despite an increased transistor size and reduced threshold voltage. This result shows that as the threshold voltage is reduced, the array size and not the cell transistor becomes the primary constraint to supplying sufficient write current to the device.

6.2 Memory array model of STT-MRAM

The following section evaluates each of the cell types with respect to current margin and current ratio. The simulation setup is described in Section 6.2.1. The circuit models used to evaluate the current margin and current ratio for each of the memory cells are discussed in Section 6.2.2.

6.2.1 Simulation setup

Each memory cell type is evaluated for size and bias conditions to enhance the sense margin. The evaluation is based on the device parameters listed in Table 8.1 [51]. The MTJ is modeled by (6.3), where the MTJ half bias voltage (V_h) is inferred

Table 6.1: MTJ parameters

$MTJR_{ON}$	5 $k\Omega$
$MTJR_{OFF}$	12.5 $k\Omega$
TMR	150%
I_c	35 μA

from the Slonczewski expression for TMR [80]. This expression describes the 50% bias point at the switching current of an MTJ. SPICE simulations of the MOS transistors are based on the predictive technology model (PTM) at the 22 nm node [227]. The initial circuit characteristics are determined from the procedure described in Appendix 6.3.

The layout of each of the three STT-MRAM cells is depicted in Fig. 6.3, and are based on the FreePDK45 design kit [211]. The cell density for the 1T - 1R, 2T - 1R diode connected, and 2T - 1R gate connected is, respectively, 46.6 F^2, 75.6 F^2, and 101.5 F^2. Note that the size of the memory cells is much larger than a state-of-the-art STT-MRAM, which typically exhibits a $6F^2$ cell area. These smaller area circuits, however, are typically created in a standalone memory process flow where layout regularity and technological focus facilitates the use of more aggressive design rules.

For the physical layout shown in Fig. 6.3, the width of each transistor is 2.2 F. Any transistor width smaller than this dimension does not decrease the area of either 2T - 1R memory cell. This limit is due to minimum sizing rules for contact and transistor spacing; in a practical setting for memories, rules could be tighter to improve density. Design rules specific to memories, however, were not available so design rules tailored for logic circuits are used for these layouts.

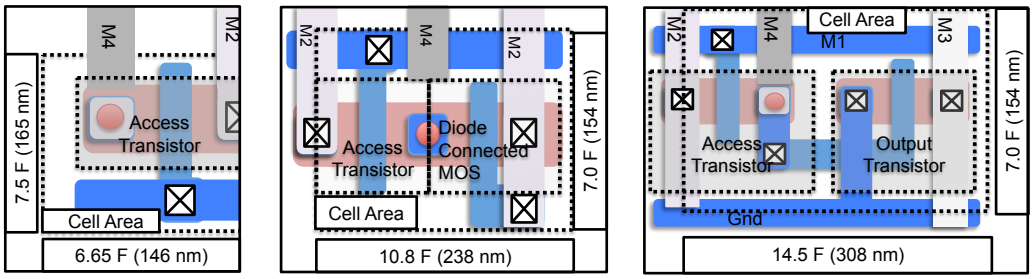


Figure 6.3: Physical layout of STT-MTJ memory cells: (a) standard 1T - 1MTJ, (b) 2T - 1MTJ diode cell, and (c) 2T - 1MTJ gate cell. The physical layout is based on the FreePDK45 where F represents the feature size of the technology [211].

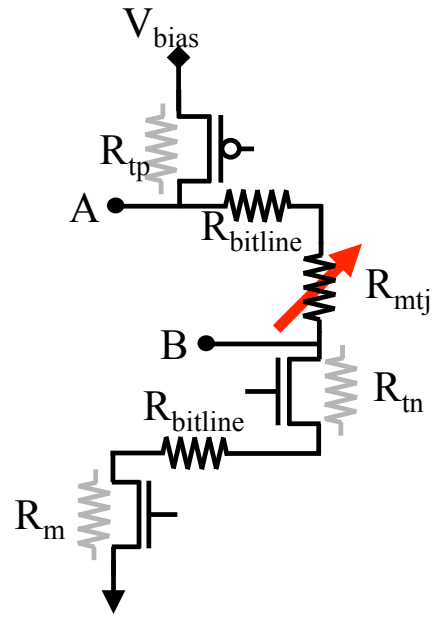
6.2.2 Modeling approach

An STT-MRAM array can be modeled as a voltage divider. The subsequent discussion describes this model and the response of the sense margin and current ratio to the array voltage bias, array size, threshold voltage, and device width of the NMOS transistors within the data array. The current margin is the difference between the off current and the on current for an MTJ cell under bias. The current ratio is the on current divided by the off current for a cell under bias.

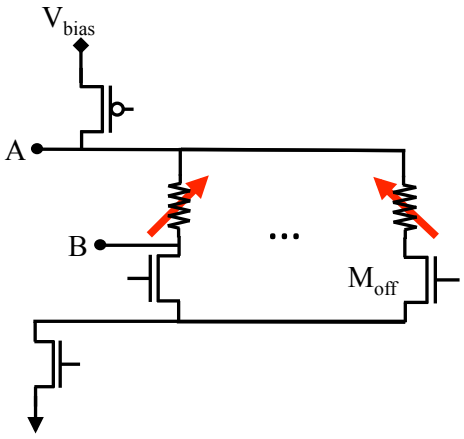
6.2.2.1 1T - 1R data array

The circuit model is shown in Fig. 6.4a, where R_{tp} is the resistance of the PMOS transistor, R_{tc} is the resistance of the NMOS cell access transistor, R_{bl} is the resistance of the bitline, R_{tn} is the column access transistor, and R_{mtj} is the resistance of the STT-MTJ. The sense node A is the observable voltage on the network.

The voltage bias (V_{Bias}) applied to the data array directly controls the magnitude



(a)

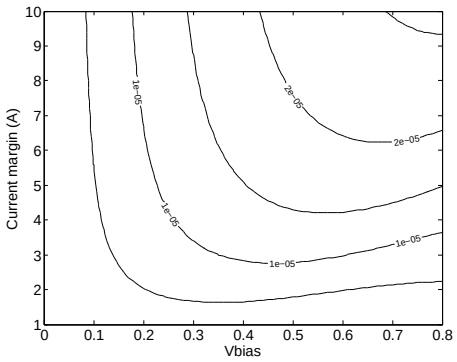


(b)

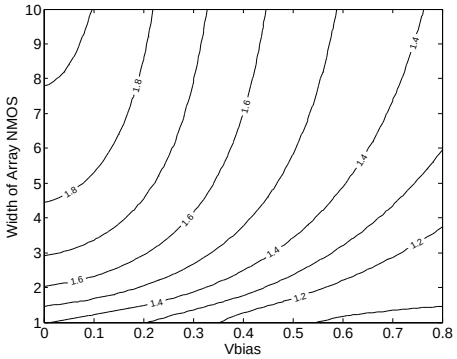
Figure 6.4: Circuit diagram of STT-MRAM array, (a) memory cell sensing model, and (b) data array model.

of the signal detected by the sense circuitry. In the case of the 1T - 1R memory cell, however, there is a diminishing return with a larger bias as the current ratio drops

off.



(a)



(b)

Figure 6.5: Design space of 1T - 1R memory cell at nominal threshold, (a) current margin, and (b) current ratio.

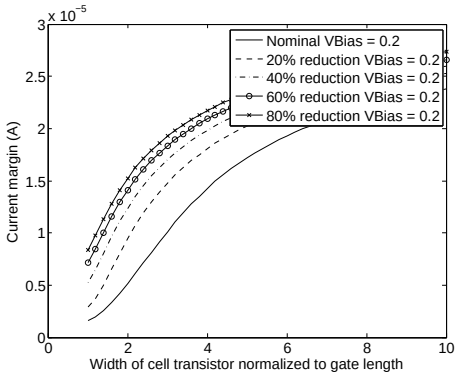
This degradation, shown in Fig 6.5, depicts a contour map of the current margin and current ratio for varying NMOS transistor sizes within the array and voltage biases. The standard 1T - 1R cell exhibits a peak current ratio of approximately two for a 0.1 volt bias. This ratio is 0.6 less than the expected 2.5 predicted by an ideal MTJ device due to the nonlinear voltage drop across the access transistors and the reduction in device TMR with larger voltage bias. Additionally, the voltage

dependence of the TMR ensures that increasing the voltage bias of the array further reduces the current ratio. At full V_{DD}, the peak current ratio drops to approximately 1.4.

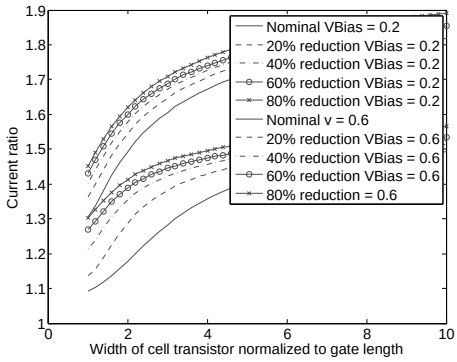
These peak current ratios occur with large access transistor sizes. For smaller NMOS transistor sizes, more practical in high density memories, a reduced current ratio as compared to the peak ratio is noted. At low bias (0.1 to 0.4 volts), the current margin remains relatively constant for increasing NMOS device size. This behavior indicates that low voltage, large transistors are preferable to increase the bias voltage when improving the sense margin of a 1T - 1R array.

The effect of a reduced array threshold voltage for a 1T - 1R cell is shown in Fig. 6.6 for a high voltage bias (0.6 volts) and low voltage bias (0.2 volts) array. Note that the reduction in threshold voltage has a limited effect on the current margin at low bias voltages. The sense margin is approximately invariant with the threshold voltage of the transistors in the data array. In contrast, the current ratio increases by approximately 0.2 over a nominal threshold voltage due to an 80% reduction in threshold voltage. A reduction in the transistor threshold voltage in low voltage, small transistor 1T - 1R memory cells are therefore desirable.

A reduction in the threshold voltage, however, also degrades the isolation of the data array. The decreased current margin and switching ratio for increasing array size are shown in Fig. 6.7. At increasing array size, the current margin and current ratio decrease as the threshold voltage is reduced. This effect occurs despite that



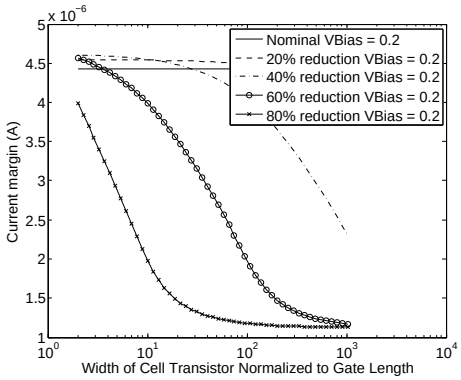
(a)



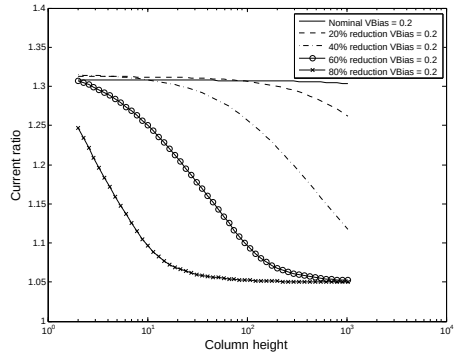
(b)

Figure 6.6: Effect of reduced threshold voltage of 1T - 1R memory cell for increasing size of data array transistors. (a) current margin, and (b) current ratio.

reduced threshold voltages will improve the current ratio and current margin for a single 1T - 1R cell due to the smaller threshold voltages affecting the transistor off-current more than the on-current. The reduced transistor off current results in additional active leakage through the unselected cells, causing both the current ratio and current margin to degrade as a function of array size. As the array size approaches 1,024 rows, the current ratio drops to unity, indicating that the change in resistance of the MTJ is negligible.



(a)

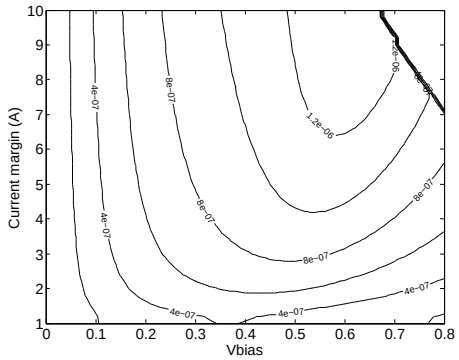


(b)

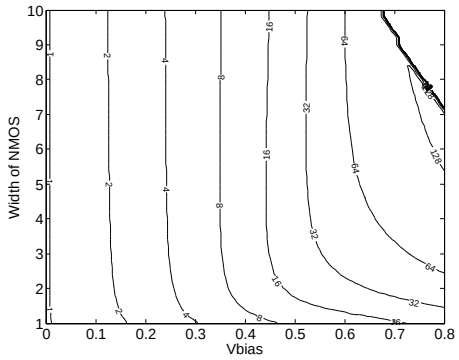
Figure 6.7: Effect of increased data array size on 1T - 1R memory cell at multiple threshold voltages. (a) current margin, and (b) current ratio.

6.2.2.2 2T - 1R diode connected memory cell

Unlike the 1T - 1R cell, where only the signal on the MTJ is sensed, this cell amplifies the internal voltage of the cell. Each of the 2T - 1R cell topologies utilizes an external port to read the cell state. Both cells are connected to the MTJ at node B (see Fig. 6.1b). The increasing voltage difference observed at node B increases the sense margin observed at the output.



(a)



(b)

Figure 6.8: Design space for 2T - 1R diode connected cell at nominal threshold, (a) current margin, and (b) current ratio.

6.2.2.3 Current ratio

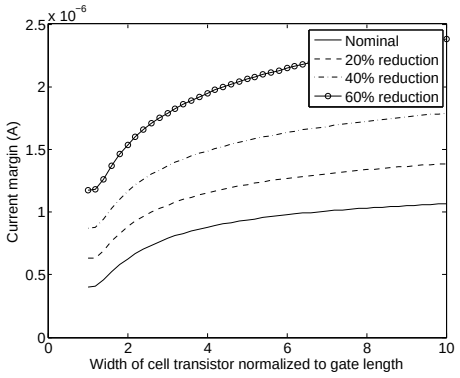
The current ratio of the diode connected memory cell is shown in Fig. 6.8b for a nominal threshold voltage. The array voltage bias has a strong effect on the current ratio. Increasing both the voltage bias and the width of the NMOS transistors within the array increases the current ratio to 151. However, operating at this point is close to writing to the MTJ and provides a small current margin. With a minimum current margin of at least $1\ \mu$A, the maximum achievable current ratio

is 127, $1.4\ \mu\mathrm{A}/11\ \mathrm{nA}$. Note that a relatively high resistance for the read circuitry is assumed as compared to the write driver circuitry to mitigate inadvertent writes within the data array.

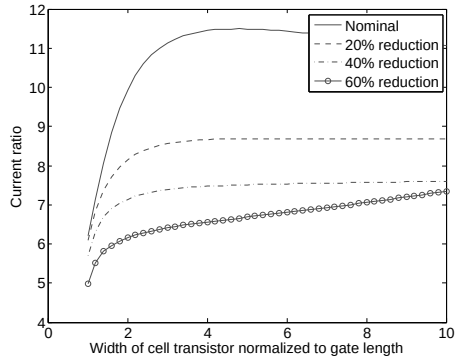
The size of the data array transistor has little effect on the output current ratio or current margin. Only at high bias (above 0.4 volts) does the current margin increase with transistor size. The diode connected cell is therefore more tolerant to transistor variations than the 1T - 1R cell. Additionally, the density advantage can be exploited while producing a higher current ratio. At a relatively small size (2F), the diode connected cell achieves a peak current ratio of 50.3. This advantage, however, is achieved at a low current margin of 0.35 μA. In this case, I_{off} falls at a faster rate than I_{on}; however, the magnitude of both currents decrease, causing a reduction in current margin with an increase in the current ratio.

6.2.2.4 Current margin

The current margin, however, can be further improved by reducing the threshold voltage of the data array transistors, as shown in Fig. 6.9a. A 20% reduction in the threshold voltage doubles the current margin for a minimum sized device. At larger device widths, the current margin is greater than 1 μA while suffering minimal loss in current ratio. A 40% reduction is sufficient to allow a 2F transistor to supply a 1 μA current margin while maintaining a current ratio of 8.



(a)



(b)

Figure 6.9: Effect of reduced threshold voltage on 2T - 1R diode connected cell for increasing size of data array transistors. (a) current margin, and (b) current ratio.

6.2.2.5 Array size

Similar to the 1T - 1R memory cell, there is a penalty associated with reducing the threshold voltage. For a 60% reduction in threshold voltage, an array column height of 1,024 bits produces a drop in current ratio from 6.7x to 3.3x, as shown in Fig. 6.10. Unlike the 1T - 1R memory cell, the current margin is relatively independent of the size of the array and increases by 0.5 μA with a 60% reduction in

threshold voltage. Moreover, for both 20% and 40% reductions in threshold voltages, both the current ratio and current margin remain relatively constant (below 0.2% variation) with increasing array size. The reduced threshold voltage does not affect the voltage signal, as in the case of the 1T - 1R cell.

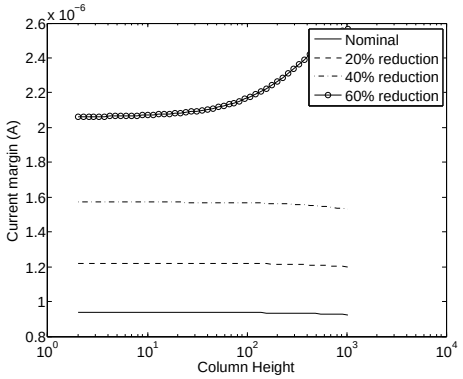
6.2.2.6 Tradeoff between current margin and current ratio

Note the tradeoff between the current ratio and current margin. The diode connected cell alternates between cutoff and saturation. Increasing the internal voltage (node B, shown in Fig 6.4a) increases the gate bias for the on state MTJ and the current through the diode connected transistor. The higher voltage, however, also increases the voltage at node B in the MTJ off state. Since the current in cutoff is an exponential function of the transistor bias, and only quadratic in saturation, the current ratio drops.

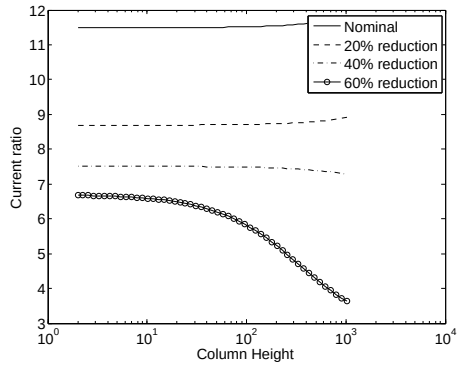
6.2.2.7 2T - 1R gate connected memory cell

As previously mentioned, the external port of the 2T - 1R cells separates the read operation from the write operation. Similar to the diode connected cell, the circuit depends on the internal cell voltage at node B. Correct operation of the 2T - 1R gate connected cell requires the transistor gate voltage to be sufficiently large to switch the transistor between the two MTJ resistive states.

The current ratio and current margin for a gate connected memory cell is shown



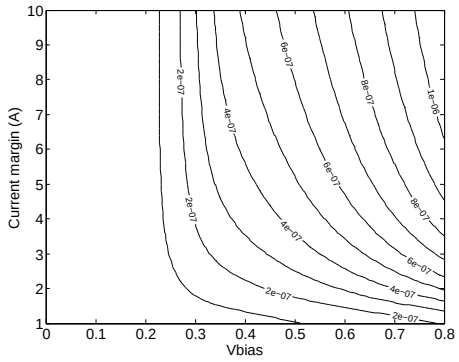
(a)



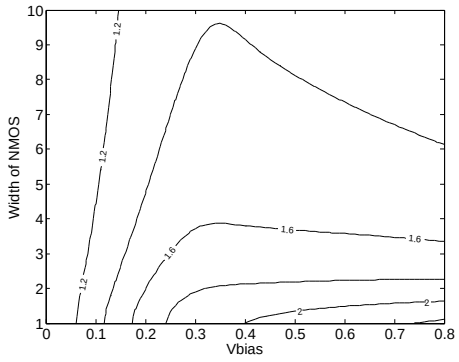
(b)

Figure 6.10: Effect of array size on 2T - 1R diode connected cell for reduced threshold voltage of data array transistors. (a) current margin, and (b) current ratio.

in Fig. 6.11 for varying access transistor sizes and array biases. Unlike the 1T - 1R case, increasing the array bias has a strong effect on the current ratio. For full bias, the gate connected transistor achieves a current ratio of 2.2. Counterintuitively, increasing the cell transistor size reduces the current ratio. This behavior is due to the dependence of the internal voltage of the cell (node B) on the voltage drop across the transistor. A linear reduction in transistor size leads to a reduced voltage drop and voltage change at node B (see Fig 6.4a).



(a)

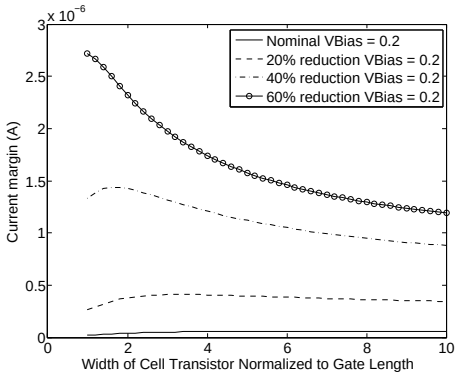


(b)

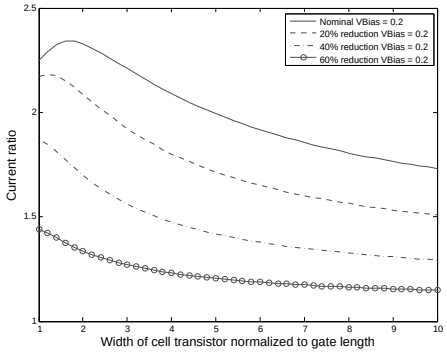
Figure 6.11: Design space of 2T - 1R gate connected cell for nominal threshold 2T - 1R gate connected cell. (a) current margin, and (b) current ratio.

The effect of a reduced threshold voltage is depicted in Fig. 6.12. A reduction in the threshold voltage enhances the current margin. For example, a 40% reduction in the threshold current is sufficient to increase the current margin above 1 μA. This reduction lowers the current ratio from a peak of 1.7 to 1.4 at 2F transistor sizing.

Similar to the 2T - 1R diode cell, the gate connected cell is independent of the array size as a function of technology, as depicted in Fig. 6.13. This characteristic occurs since the gate connected transistor is always grounded at the source terminal



(a)



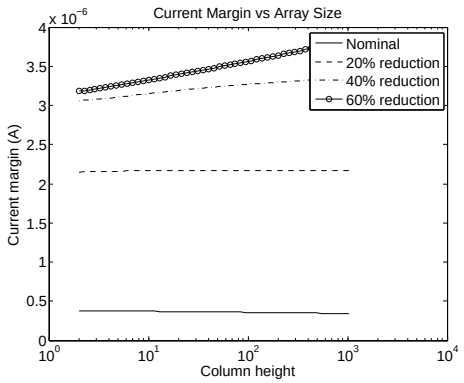
(b)

Figure 6.12: Effect of reduced threshold voltages on 2T - 1R gate connected cell for increasing data array transistor width, (a) current margin, and (b) current ratio.

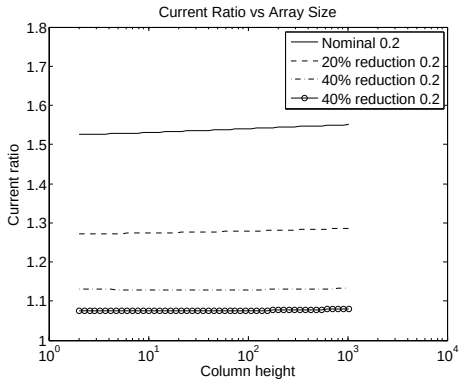
and electrically isolated from the MTJ.

6.2.3 Comparison of current margin and ratio across memory cells

Both the gate connected cell and the diode connected cell improves the current ratio (or margin) observed at the sense circuity by providing additional read ports. The diode connected cell produces the largest increase in current ratio. The current margin of the diode connected cell is limited by the large on-resistance of the diode



(a)



(b)

Figure 6.13: Effect of array size on 2T - 1R gate connected cell for reduced threshold voltages, (a) current margin, and (b) current ratio.

connected transistor. This issue can be addressed, however, by reducing the threshold voltage of the devices within the data array. This reduction in threshold voltage provides additional current at high bias conditions, either decreasing the switching times or improving density through smaller access transistors. Moreover, this reduction does not degrade the current ratio or current margin with increasing array size as with the 1T - 1R memory cell. Intuitively, the target current can be linearly increased by widening the gate connected transistor while maintaining the same

Table 6.2: Single bit access delay (ns)

Number of Bits	SRAM 8T HD RP	SRAM 8T HD WRP	SRAM 8T Logic RP	SRAM 8T Logic WRP	1T-1R	2T-1R Gate	2T-1R Diode
2,048	14.879	14.708	25.793	26.927	3.106	4.200	3.762
1,024	4.471	3.716	7.189	6.754	0.718	1.242	0.969
512	1.537	0.960	2.219	1.721	0.265	0.377	0.295
256	0.626	0.273	0.800	0.466	0.127	0.139	0.111
128	0.306	0.094	0.352	0.145	0.078	0.067	0.057

Table 6.3: Single bit access energy (fJ)

Number of Bits	SRAM 8T HD RP	SRAM 8T HD WRP	SRAM 8T Logic RP	SRAM 8T Logic WRP	1T-1R	2T-1R Gate	2T-1R Diode
2,048	31.182	28.529	31.197	28.561	5.382	50.285	98.113
1,024	19.144	18.243	19.175	18.274	1.081	26.014	39.430
512	12.093	12.014	12.124	12.045	0.568	12.559	17.441
256	6.047	6.235	6.078	6.266	0.370	6.250	7.891
128	2.736	2.973	2.767	3.004	0.284	3.170	3.620

width of the access transistor.

6.2.4 Comparison of SRAM and STT-MRAM memory cells

A comparison of 8T SRAM with STT-MRAM memory cells in terms of read delay, read energy, and physical area are listed, respectively, in Tables 6.2, 6.3, and 6.4. The SRAM read ports (RP) and write-read ports (WRP) are evaluated for both memory specific high density (HD) and logic process (Logic) design rules [240]. Note that the 8T SRAM is selected as the benchmark due to the use of 8T SRAM for high performance caches in sub-45 nm technologies [241]. The wordline energy

Table 6.4: Area comparison

	SRAM 8T HD	SRAM 8T Logic	1T-1R	2T-1R Diode	2T-1R Gate
Cell Height (F)	8	8	7	7	7
Cell Width (F)	31.6	45.4	6.65	10.8	14.5
Density (F^2)	252	363.2	46.55	75.6	101.5

associated with a 8T SRAM cell is negligible as compared to the bitline power. Each memory cell and the associated parasitic impedances are scaled to the 22 nm technology node in the same manner as described in Section 6.2.1. The 8T SRAM read port (RP) is sensed using a standard single-ended inverter sense amplifier [242]. The SRAM write-read port is sensed using a standard dynamic latch sense amplifier [243]. Each of the STT-MRAM cells is sensed using a clamped bitline sense amplifier [171]. The array sizes are typical of an on-chip cache array.

Delay metrics for square array sizes ranging from 128 to 2,048 bits are listed in Table 6.2. STT-MRAM arrays exhibit significantly less delay than the SRAM counterparts. At array sizes of 2,048 cells, the delay of SRAM and STT-MRAM is dominated by the wordline delay. The STT-MRAM has an advantage over SRAM since only one transistor is required to select the cell. Additionally, the reduced length of the wordline further reduces the delay. The write-read port of the 8T SRAM cell is sensed differentially and thus compensates this increased delay. This effect is more clearly observed at the smaller SRAM arrays where the singled-ended read port delays are a factor of three longer than the write-read port read time. Each of the STT-MRAM memory cells are also read in a single ended manner. As compared to the single ended SRAM read port, the delay of each STT-MRAM memory cell type is smaller by a factor of 3.9, 4.6, and 5.37, respectively, for the 1T-1R, 2T-1R gate connected, and 2T - 1R diode connected memory cells. Both the gate and diode connected cells exhibit an area overhead larger then the 1T-1R cell but overcome

this issue through an improved current ratio which reduces the delay.

The energy consumption of each of the cell types is listed in Table 6.3. Each of the cell types exhibits a significant reduction in energy consumption with a smaller data array. The gate connected and diode connected cells plateau at an energy compatible to SRAM arrays at smaller sizes. This behavior is due to the additional bias required to drive the internal node of the cell. The 1T-1R cell does not require an additional bias, enabling more energy efficient reads than the other memory cell types. At larger array sizes, the 2T - 1R cell variants require more energy than the other cell types. The additional area occupied by the logic version of the 8T SRAM cell has little effect on the per bit energy. The word line energy is spread over the length of the row during accesses. This effect can also be observed between the diode and gate connected cells, as the gate connected cell is more energy efficient than the diode connected cell despite the larger area of the diode connected cell.

Between each of the memory types, the SRAM requires longer delays and greater energy than the STT-MRAM memory. In general, the 1T - 1R outperforms SRAM for all array sizes. Both of the 2T - 1R cells require more energy at large array sizes, indicating that each topology is better suited to small active on-chip caches where speed is paramount. At these sizes, the 2T-1R topology exhibits the fastest read operation of any memory cell type at a energy consumption comparable to SRAM.

6.3 Conclusions

Two topologies are proposed to complement the standard 1T - 1R topology commonly used in STT-MTJ based memories. The diode-connected memory cell demonstrates greater than an order of magnitude improvement in the output current on/off ratio. The diode cell, due to the small area and high output current ratio, is therefore the most effective at increasing the current ratio as compared to the other cell topologies. The gate connected cell can, however, be more easily expanded into a multi-port cache structure due to electrical isolation between the internal node of the memory cell and the output port. Furthermore, the current margin of the gate connected cell can be increased irrespective of cell bias by increasing the size of the gate connected transistor. The relative importance of the current margin as compared to the current ratio determines the applicability of each cell for a particular data array. A comparison of each of the memory cells to an 8T SRAM cell shows that these additional cell topologies are advantageous in small area high speed on-chip caches.

Appendix: Parameter Selection

An STT-MRAM array can be modeled as a simple two resistor circuit. This discussion describes this two resistor model and presents expressions to maximize the sense margin. The linear resistor model is applied to produce an initial design for

the 2T - 1R cells.

A.1 Two resistor model

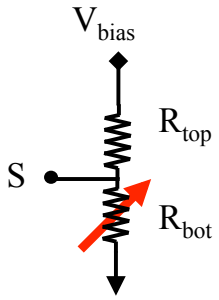


Figure 6.14: Two resistor model of an STT-MRAM array.

A data array can be modeled as a two resistor model where the sense node (node S) is used for sensing, as shown in Fig. 6.14. In this structure, R_{bot} toggles between the two resistance states, $R_{bot_{on}}$ and $R_{bot_{off}}$. In a manner analogous to the TMR of an MTJ, the switching ratio (SR) of R_{bot} is defined as

$$SR \equiv \frac{R_{bot_{off}} - R_{bot_{on}}}{R_{bot_{on}}}. \tag{6.1}$$

The sense margin (ΔV_A) for this structure is the change in the maximum voltage at node S,

$$|\Delta V_A| = V_A\big|_{R_{bot_{off}}} - V_A\big|_{R_{bot_{on}}}. \tag{6.2}$$

This difference produces the largest swing at node S, which in the aforementioned

model represents the voltage detected by the sense circuitry. The maximum change in voltage, *i.e.*, the maximum sense margin [47], occurs under the constraint,

$$R_{top} = \sqrt{(R_{bot_{on}})(R_{bot_{off}})}. \tag{6.3}$$

From (6.3), the voltage sense margin can be expressed as

$$|\Delta V_A| = \frac{\sqrt{1+SR}-1}{\sqrt{1+SR}+1} V_{bias}. \tag{6.4}$$

For current sensing, the sense margin ΔI is the change in current passing through the MTJ, where

$$\begin{aligned} |\Delta I| &= I_{R_{top}}\big|_{R_{bot_{on}}} - I_{R_{top}}\big|_{R_{bot_{off}}} \\ &= \frac{R_{bot_{on}}(SR)V_{bias}}{R_{bot_{on}}^2 + 2R_{bot_{on}}R_{top} + R_{top}^2 + (R_{bot_{on}}^2 + R_{bot_{on}}R_{top})SR}. \end{aligned} \tag{6.5}$$

Intuitively, increasing the voltage through the network increases the voltage sense margin by increasing the voltage drop across the switching resistor. Reducing the resistance of R_{top} monotonically improves the current sense margin through the path. A 2T - 1R data cell, however, produces current through an adjacent read

port. Maximizing the voltage margin at node B in Fig. 6.4a, therefore, produces the largest current ratio and margin.

2T - 1R data array

The voltage margin of the 2T - 1R cell is increased by substituting the resistances illustrated in Fig. 6.4a into R_{top} and R_{bot},

$$R_{top} = R_{tp} + R_{bl} + R_{mtj}, \tag{6.6}$$

$$R_{bot} = R_{bl} + R_{tn} + R_{m}, \tag{6.7}$$

$$SR = \frac{R_{on}TMR}{R_{on} + R_{tp} + R_{bl}}. \tag{6.8}$$

These expressions maximize the voltage difference at node B, the central node within the memory cell (rather than the bitline at node A, as in the case of the 1T - 1R data array). By maximizing the voltage difference at node B, the additional gain produced at the output of both the diode connected and the gate connected cell read ports is greatly increased.

A.2 Design parameter selection

The target MTJ write current is specified by the MTJ technology along with the on-resistance of the write drivers determined from the CMOS technology parameters. The expression,

$$R_{total} = \frac{V_{dd}}{I_c} = R_{mtj} + R_{tp} + R_{tn} + R_m, \tag{6.9}$$

describes the constraint placed by the MTJ write current on the cell size. The size of the transistors is determined from (6.2) - (6.9). These expressions produce the greatest change in output current for both types of 2T - 1R cells. For the 2T - 1R diode connected cell, the output port voltage exceeds the voltage at node B in Fig. 6.4a by the threshold voltage of the diode connected transistor. This higher voltage ensures that the diode switches between the on and off states when the MTJ is, respectively, off and on. A sweep of the bias voltage can be conducted to determine the location where the current ratio is maximum. The gate connected cell does not exhibit this limitation.

Chapter 7

Arithmetic Encoding for Memristive Multi-Bit Storage

A memristive digital memory architecture is proposed herein utilizing the unique analog properties of these devices to compress digital information within a data array. The proposed circuit leverages *a priori* knowledge of a bit sequence for storage. Through use of a compression algorithm with supporting circuitry, the circuit yields the potential to store significantly more bits per cell than a standard multi-bit approach. This system is realized through a memristor driven sensing scheme and an adaptive write circuit that assign a resistance value to a memristive device with fine grain control.

In Section 7.1, background on the proposed compression procedure is described. In Section 7.2, the circuit architecture is reviewed. A description of the data modeling approach for memristive compression is presented in Section 7.3. A discussion of the simulation-based experimental results is presented in Section 7.4. The paper is concluded in Section 7.5.

7.1 Background

Memristive devices can be described as non-volatile resistor-like devices whose conductance is modulated by an applied bias. Since memristors retain a written state when the voltage bias is removed, these devices are useful for low power storage applications. The key feature of this specific type of memristor stems from the continuous "resistance" characteristic. This specific feature enables an encoding based approach. A brief review of the applied coding scheme is provided below.

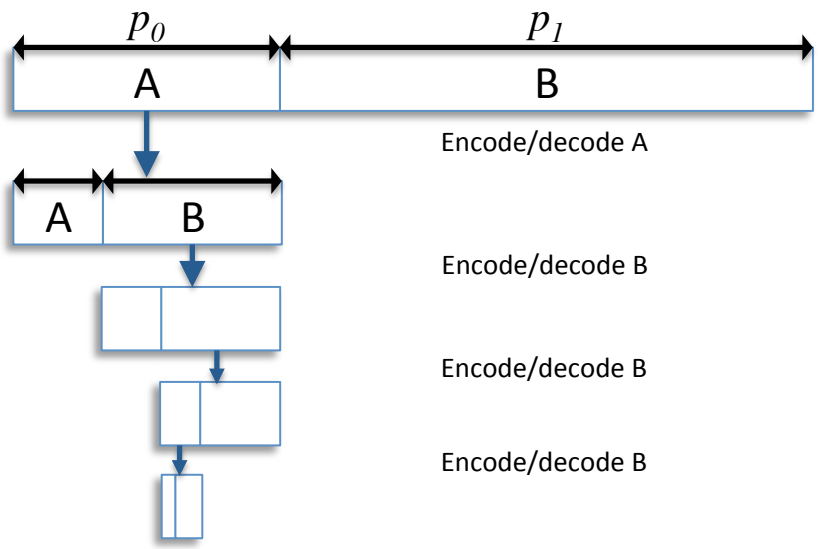


Figure 7.1: Encoding process for a two symbol alphabet and sequence $S_e = ABBA$. The initial interval is divided into sections corresponding to each symbol. The section size is governed by the probability of the symbol. Encoding S_e requires selecting the initial section that corresponds to A, subdividing this section according to the specified probabilities, selecting the subsection associated with C, and continuing the process until all symbols in S_e are encoded

7.1.1 Overview of arithmetic coding

Arithmetic coding is a long standing method for compressing data [244]. The procedure relies on assessing the probability of certain values within a data stream to create an encoding model that favors more frequent values. A string of bits is represented by a single compressed value. The continuum of potentially compressed values can be encoded into a memristor due to the inherent continuous resistance characteristic of the device, improving the storage density of a memristor over standard multilevel approaches. The encoding process relies on mapping an uncompressed sequence to a fractional value within the interval $[0, 1)$ which is related to a particular resistance within the resistive range of a memristive device. A probability model of a sequence informs the coding mechanism which encodes the data to a target resistance value.

Arithmetic coding uses a finite, non-empty set of elements A, designated as an *alphabet*. Each element $\{a_0, a_1, ..., a_k\}$ in the set, known as a *symbol*, represents a possible value within the data sequence being compressed A *sequence* is a series $S = s_n$ such that $\{s_n \epsilon A, \forall s_n \epsilon S\}$; this series represents an uncompressed data stream of symbols from the alphabet. A model $P = \{p_0, ...p_k\}$ associates each a_k in the alphabet with a probability p_k where $\sum_{i=0}^{k} p_i = 1$.

For example, consider the arbitrary sequence $S_e = ABBA$ for $A_e = \{A, B\}$, and the probability model $P_e = \{\frac{1}{4}, \frac{3}{4}\}$. This probabilistic model is defined according

to the frequency of symbols within the sequence. The interval $[0, 1)$ is divided into subintervals, each corresponding to a symbol in A_e. The length of each interval is equal to the probability associated with the corresponding symbol, as shown in Figure 7.1. The first detected symbol is A; the interval $[0, \frac{1}{4})$ represents the first symbol in the sequence. Any value within the interval is sufficient to encode the first bit of the sequence. To encode the second symbol, the interval $[0, \frac{1}{4})$ is again divided according to the probability model (see Figure 7.1). For the next symbol in the sequence (B), the interval $[\frac{1}{16}, \frac{1}{4})$ is selected which represents the top $\frac{1}{4}$ of the previous interval. A value within this interval encodes the first two symbols of the sequence. The process continues until all symbols in the sequence are encoded into a single value. The final interval for this example sequence is $[\frac{28}{256}, \frac{37}{256})$. Intuitively, the final interval is unique to the sequence S_e as other symbols would lead to different intermediate intervals. Selecting the value $\frac{32.5}{256}$ is, therefore, sufficient to encode the entire sequence.

The decoding process begins with the selected value ($\frac{32.5}{256}$) and the starting interval $[0, 1)$. In a manner similar to the encoding process, the interval is partitioned according to the probability model. The selected value lies in the region of the interval corresponding to the symbol A. From this information, the first symbol in the sequence is decoded as A. Continuing to the second symbol, the interval $[0, \frac{1}{4})$ is selected and partitioned. The selected value occurs in the top $\frac{3}{4}$ of the interval $[0, \frac{1}{4})$ and corresponds to the symbol B, permitting the second symbol to be decoded.

The process continues until the full sequence is retrieved. Through this process, a full data sequence can be reduced to a single fractional value without any loss of information.

These fractional values and the corresponding intervals are mapped to either a voltage or current by biasing a memristive device. The precise mapping mechanism is described in the following sections.

7.2 Memristive multi-bit encoding

The goal of memristive compressive storage is to map a binary sequence to a fractional value using two symbol arithmetic encoding and store the value within a memristive data cell. A continuous set of encoded fractional values is mapped to the continuous resistive characteristic of a memristive device. The design of these circuits is predicated on two basic memristive building blocks, a resistive divider with adjustable memristors, and a memristor data cell containing the stored data. Circuits to both write and read a memristive data cell within the proposed encoding scheme are described in the following section.

7.2.1 Decoding and read circuitry

The process of reading and decoding a data cell proceeds in a manner consistent with the compression process, as illustrated by the circuit shown in Figure 7.2.

V_{bottom} and V_{top} begin with, respectively, the initial interval of V_{min} and V_{max}; these voltages correspond to the arithmetic coding interval $[0, 1)$. Electrically, these levels are the maximum and minimum voltage biases that correspond, respectively, to the memristor states, R_{off} and R_{on}. The memristor values correspond to a probability model $P = \{p_0, 1 - p_0\}$ for a two symbol alphabet $A = \{0, 1\}$, where each memristor assumes the resistance value $p_k R_{off}$ for the two encoded symbols. When a read occurs, the voltage bias applied to the memristive data cell is set below the memristor threshold voltage. The current generated by this circuit is mirrored to a comparator. Within the first interval, a voltage divider generates the initial comparison voltage V_{rn}, where n represents the symbol being decoded (see Figure 7.2). The result of the comparison operation is stored in a shift register. Following this operation, if the result is logic 1, V_{bottom} is set to V_{rn}, otherwise V_{top} is set to V_{rn}. Setting V_{top} and V_{bottom} in this manner is the same procedure through which an arithmetic coding interval is selected for a given sequence. The voltage divider, with resistances set according to the probability of each symbol, generates the appropriate comparison threshold. This process continues until a maximum number of bits has been decoded. The initial voltage of the biased cell is stored within the sample and hold circuit, shown in Figure 7.2, to prevent writing to the memristor during an on-going read operation. The total number of bits stored per memristor is limited by the minimum distinguishable voltage at the output of the comparator. This limit is specified at design time and assumes that external decoding mechanisms detect

when a specific output vector generates more bits than the noise level allows, and truncates the bit vector to the number of bits that can be stored within the cell..

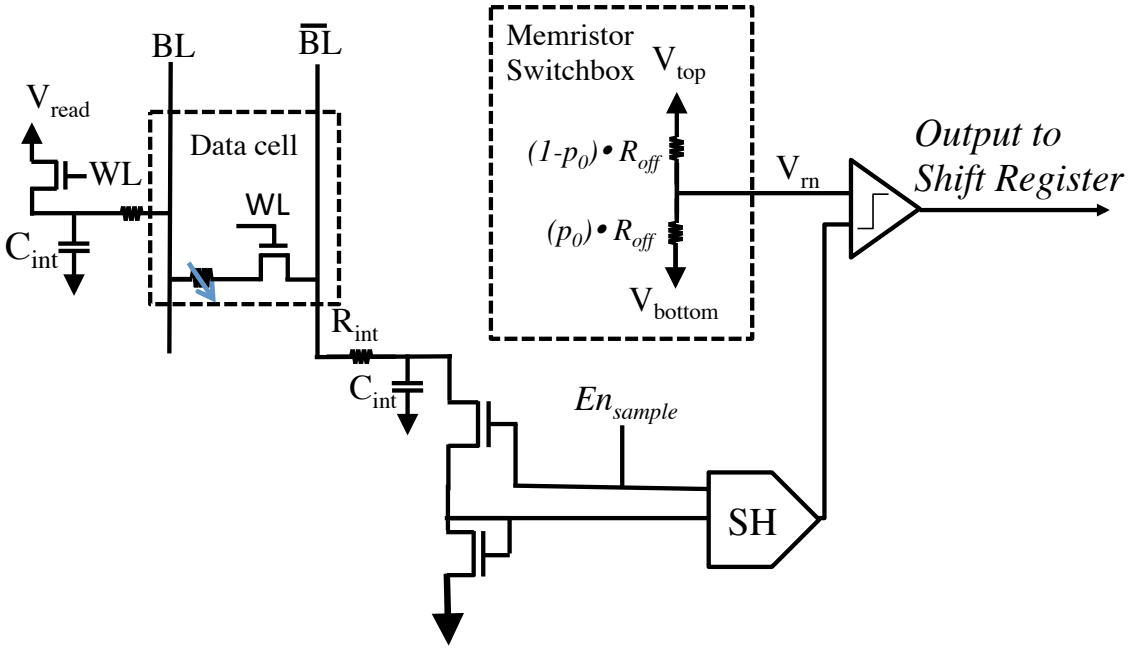


Figure 7.2: Circuitry for reading an encoded value from a memristive data cell. Each read operation begins by selecting the cell in the data array which is compared against a reference voltage. The comparison is stored in a shift register at the end of each interim read operation. Depending upon the result of the comparison, either V_{top} or V_{bottom} is set to V_{rn}

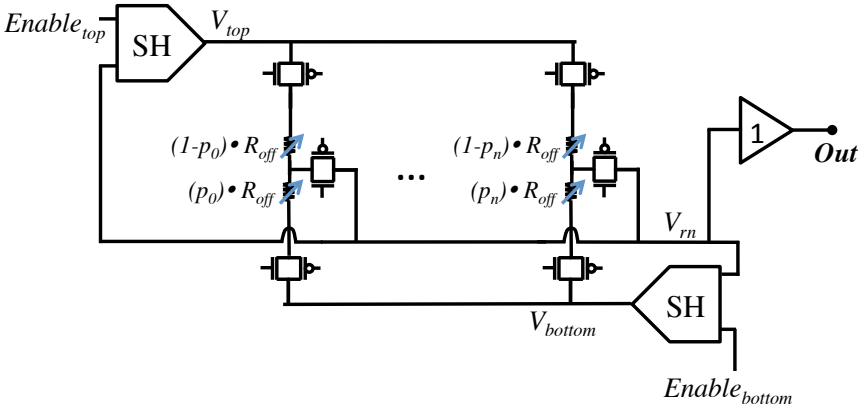
Voltage divider switchbox: To properly modulate V_{top} and V_{bottom}, a circuit is required to both generate V_{rn} and to store intermediate values during the decoding process. This objective is accomplished by the voltage switchbox shown in Figure 7.3a. Each sample and hold circuit drives a single pair of resistors. The resistance values of each pair correspond to the probabilities associated with a particular bit stream (*e.g.*, $p_0 = 0.1, 0.2...$). During the decoding process, V_{max} and V_{min} are applied, respectively, to V_{top} or V_{bottom}. The pair of resistors that correspond to the

selected branch is switched on; the voltage division across the resistors gives rise to the threshold voltage V_{rn}. The sample and hold circuit stores the current value of V_{rn}. If the readout voltage is greater than the threshold voltage, the bottom sample and hold circuit is switched, otherwise the top sample and hold circuit is triggered. Switching the sample and hold circuit generates a new value for either V_{top} or V_{bottom}, producing the next threshold voltage V_{rn}. This process continues until the stored sequence is decoded.

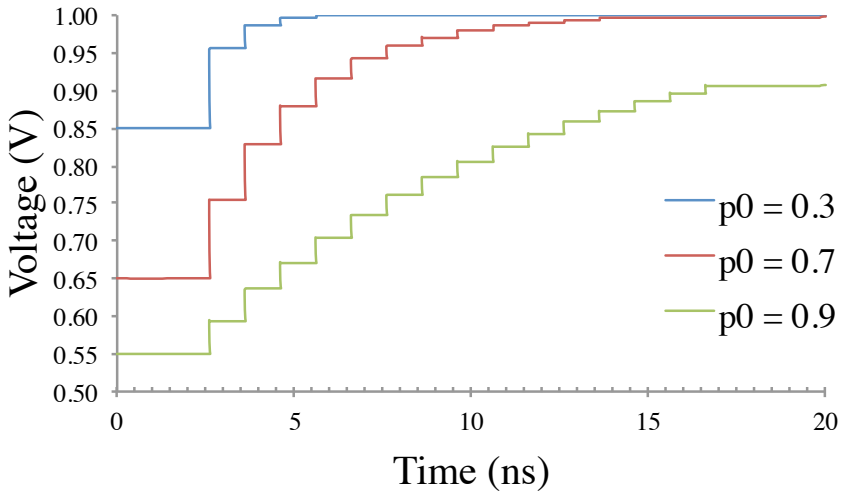
The operation of the circuit is illustrated in Figure 7.3b. This graph depicts the voltage divider switchbox for an input bit stream containing only ones. A larger probability (p_0) indicates that ones are more prevalent in the input bit stream then zeros. Storing this specific sequence as a voltage level is more effective when the circuit is configured to a probability of 0.9 rather than the other two cases, resulting in a larger detectable difference between voltage levels. A larger detectable voltage level illustrates the process in which arithmetic encoding can be used to improve the storage density of a memristive device as compared to a traditional multi-bit approach.

7.2.2 Encoding and write circuitry

A variable-length data sequence is encoded into a single memristor by the circuity shown in Figure 7.4. The write operation occurs in three steps. First, the data being written, transmitted to the array in a pre-encoded state, creates a reference



(a)



(b)

Figure 7.3: Voltage divider switchbox. (a) Circuitry for generating threshold voltages for both encoding and decoding circuitry. V_{bottom} and V_{top} are initially set to, respectively, V_{min} and V_{max}. If $Enable_{top}$ is set high, the sample and hold corresponding to V_{top} is set to the threshold voltage V_{rn}; the same is true for $Enable_{top}$. (b) Switchbox output for a bit stream of ones these for cases where the switchbox is set to zero probability

voltage using the switchbox. Afterwards, the word line of the selected cell is biased high. Once En_0 and En_1 are switched on, the reference voltage is compared to the voltage generated at the output of the array. Given the bidirectional nature

of memristive devices, this initial comparison, carried out by the write direction comparator shown in Figure 7.4, determines which direction to apply the bias to perform the write operation. This process establishes whether an increase of decrease of the initial device resistance achieves the target value.

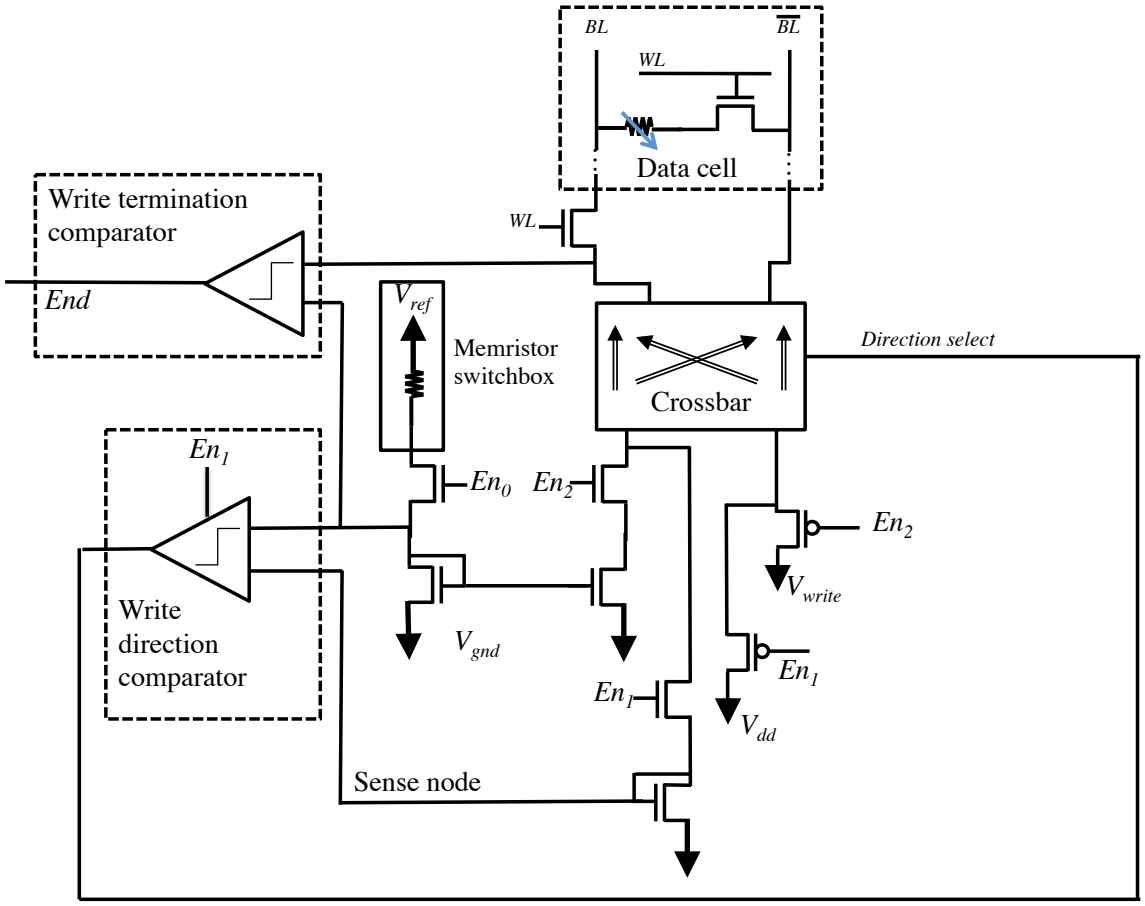


Figure 7.4: The adaptive write circuit. An initial read of the selected data cell determines the direction required to write the device (En_1). This signal is relayed to the crossbar which selects the direction of the device. A fixed current is applied to both the reference switchbox and the data array (En_2). The write termination comparator indicates whether the state has been written. This event occurs when the voltage across the current mirror transistors is the same, fixing the equal currents and memristor resistance.

After this initial read, the second stage applies a voltage to the selected cell by raising the voltage on En_2. Applying a voltage to the memristor device for a prolonged period changes the device resistance. The write termination comparator continuously compares the two voltages, as indicated in Figure 7.4. The End signal is pulled low once the memristive device has been correctly written to the target resistance. Drift in the resistance, which occurs at the termination of a write operation, is a source of noise in the circuit. Note that the linear memristor model utilized in this analysis is known to be inaccurate [132]; however, the write procedure adaptively adjusts the target resistance to any write based on the electrical resistance of the device.

7.3 Improvements In Bit Density

Encoding a fraction to a continuous memristor is only limited by the granularity at which the resistance can be changed, and the ability to distinguish values during read and write operations. For these operations, noise in the circuit as well as resistive drift governs the maximum number of bits that can be stored within a memristor.

$$V_{min} \geq 2(V_n + V_{drift} + V_{SH}). \tag{7.1}$$

Equation (7.1) describes the minimum distinguishable voltage V_{min} within a memristive sensing operation. V_{drift} specifies the maximum voltage caused by resistive

drift from the write operation, and V_{SH} represents the cumulative error from each of the sample and hold circuits. V_{drift} is due to the delayed termination of the write operation. For example, assume a change in resistance between 10 $\mathrm{K}\Omega$ to 100 $\mathrm{K}\Omega$ corresponds to an output voltage swing between 0 to 1 volts. If 25 mV of circuit noise is seen at the sensing circuitry and a 1 $\mathrm{K}\Omega$ drift gives rise to a 25 mV error, the minimum distinguishable voltage would be 100 mV. Resistive drift and circuit noise are dependent on the circuit topology and resistive state of the device. The low resistance states drift more than the high resistance states due to the higher currents during the write procedure [245]. The sample and hold circuitry contributes three sources of error: the pedestal error associated with the sampling of a voltage level, the resistive drift caused by sampling during a read, and the droop rate of the hold state [246]. All three sources of error have a direct effect on the minimum distinguishable voltage.

For a simple two symbol alphabet, the disparity, a measure of the relative probability of symbols within an alphabet, is

$$disparity = abs((1 - p_0) - p_0) = abs(1 - 2p_0). \tag{7.2}$$

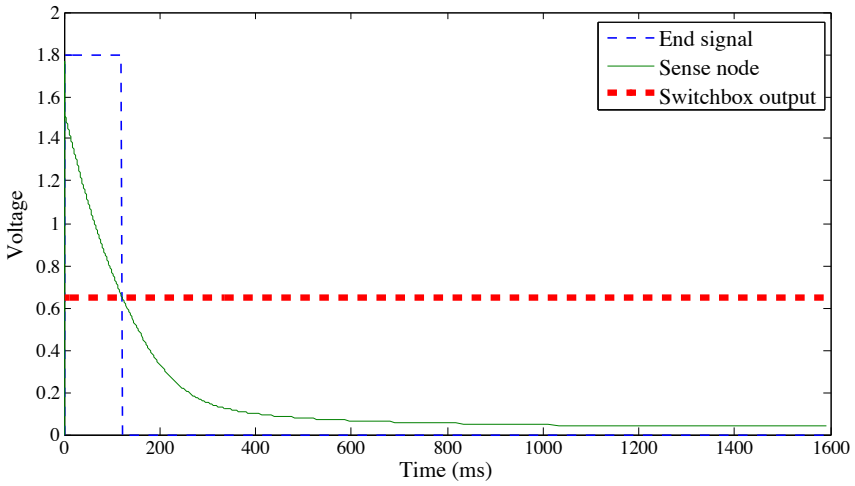
This metric describes the compression characteristics of a particular input bit stream. The storage capability of an array of memristive data cells can be characterized by this metric.

Table 7.1: Memristor model parameters [131]

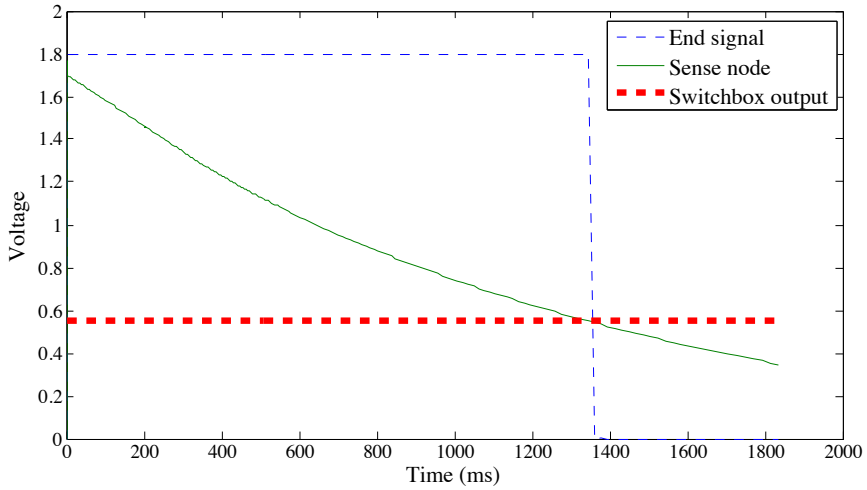
$\mathrm{u_v}$	1×10^{-14}	$\mathrm{m^2V^{-1}s^{-1}}$
$\mathrm{R_{off}}$	38	$\mathrm{k\Omega}$
$\mathrm{R_{on}}$	100	$\mathrm{k\Omega}$
D	10	nm
$\mathrm{V_{th}}$	1	V

7.4 Experimental Evaluation

The proposed circuit architecture has been evaluated using a 1.8 volt, 180 nm CMOS technology. The memristor behavior is modeled by a linear VerilogA model [247]. This model corresponds to (1) and (2). Device parameters are from [131], and listed in Table 7.1. For the purposes of this analysis, an ideal sample and hold circuit is assumed. The effects of non-idealities are modeled by the parameter V_{min}, as described in (7.1). The data stream is modeled as a random binary sequence. The arithmetic coding algorithm, applied to this sequence to determine the average improvement in bit density, is a function of the probability characteristics of the data stream and theï£¡ tolerable noise. For simplicity, the probability is determined from the average occurrence of the symbols ($A = \{0, 1\}$) within the sequence.



(a)



(b)

Figure 7.5: Adaptive write circuitry for target voltage levels (a) 650 mV, and (b) 550 mV. The End signal is pulled to ground when the device resistance has crossed the target threshold.

7.4.1 Circuit simulation

A simulation of the write circuitry is shown in Figure 7.5, which demonstrates that a memristive device adaptively switches to the target voltage. As the memristor resistance changes, the voltage on the *Sense node* converges to the voltage specified by the switchbox. The End signal is pulled to ground when the memristive device surpasses the target voltage. A key limitation of this adaptive circuit is the wide range over which the device switching speed can vary. Switching from R_{on} to the voltage level shown in Figure 7.5a requires approximately 100 ns; however, switching to the level shown in Figure 7.5b requires more than 1.3 s. The adaptive scheme has a one-to-one correspondence between a write bias voltage and a memristive state. In this adaptive scheme, higher resistance states correspond to lower write bias voltages. As a result, switching to a higher resistance state causes the switching process to require more time than if a full voltage bias is applied to the circuit.

The maximum voltage range delivered to the memristor varies between 500 mV (V_{min}) and 980 mV (V_{max}). This range considers the voltage drop across the access transistors and the adaptive current mirror (which is utilized during write operations).

The resistive drift of the device during this process is shown to be negligibly small. This small drift is due to the slow switching speed observed in the device,

which is on the order of milliseconds. The total peak V_{drift} is observed to be 0.3 μV, comparable to the thermal noise generated by a memristor in the on state. Resistive drift is therefore neglected.

7.4.2 Bit density

The minimum noise level determines the storage density as a function of the data disparity. The bit density is illustrated in Figure 7.6 and listed in Table 7.2. The case of no disparity models a traditional multi-bit approach, where the voltage range is divided equally by the minimum increment in observable voltage (V_{min}). For this comparison, the voltage drop across the access devices for a traditional multi-bit approach is assumed to be the same as the encoded approach.

An improvement in storage density over a traditional approach is seen for all cases, however, only a marginal improvement is noted for those data sets with a disparity less then 0.5. The average bit storage density can, however, be improved by a factor of 7.6 for high noise, high disparity data sets. The overall improvement in storage density is dependent on the relative frequency of the different sequences.

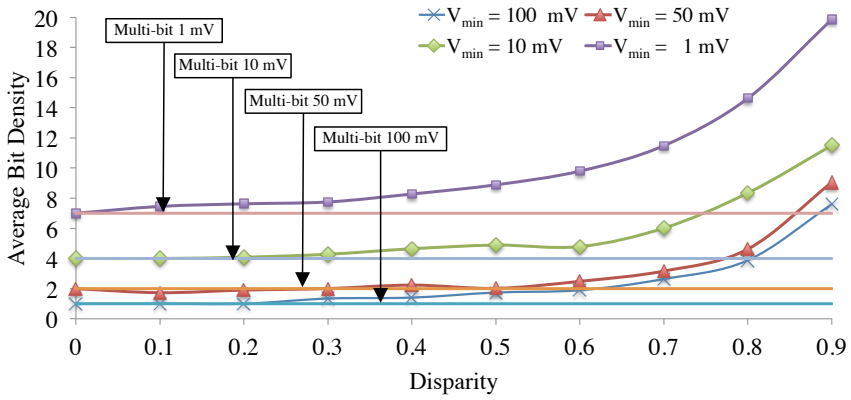


Figure 7.6: Improvement in bit density versus disparity for increasing V_{min}

Table 7.2: Average Bit Density vs V_{min}

Disparity	100 mV	50 mV	10 mV	1 mV
0	1	2	4	7
0.1	1	1.7309	4.0015	7.4546
0.2	1	1.9029	4.0855	7.6279
0.3	1.344	2.0048	4.2803	7.7496
0.4	1.4055	2.2359	4.6454	8.2768
0.5	1.742	2.0182	4.8872	8.8869
0.6	1.8945	2.4858	4.7909	9.7958
0.7	2.6363	3.1652	6.0143	11.483
0.8	3.8793	4.6234	8.3456	14.622
0.9	7.6021	9.0552	11.498	19.867

7.5 Conclusions

A circuit architecture is presented which supports arithmetic encoding of data within memristive data cells. Novel read and write circuits are described that support fine grain control and detection of the memristor device resistance. The encoding procedure exhibits storage density improvements of 7.6x for a specific data set.

Chapter 8

Sub-Crosspoint RRAM Decoding for Improved Area Efficiency

For high density applications, physical area is of paramount importance. A standard approach in semiconductor memory is to place the access circuitry, such as the decoders and sense amplifiers, peripherally around the memory cells. The RRAM devices, however, are integrated into the metal layers without using the silicon area beneath the array.

Two topologies are proposed that integrate RRAM within the intermediate metal layers, where the decode circuits are placed beneath the array (which is called here, sub-crosspoint decoding). The peripheral row and column decode circuits are integrated beneath the crosspoint array by introducing crosspoint gaps, and by vertically and horizontally staggering contacts to the rows and columns. A topology

where only the row decode circuitry is placed underneath the array exhibits 38.6% reduction in area for a single array with a 21.6% improvement in array efficiency. A second topology, with sub-crosspoint placement of both the row and column decoders, reduces the area of large RRAM crosspoint arrays by 27.1% and improves area efficiency to nearly 100%.

Background on crosspoint memories are reviewed in Section 8.1. The physical topology of the RRAM crosspoint array is described in Section 8.2. The proposed topology is evaluated and compared to standard peripheral approaches in Section 8.3, and some conclusions are offered in Section 9.5.

8.1 Nonlinear crosspoint array

RRAM and other memristive devices have been proposed for use in crosspoint arrays. Crosspoints arrays achieve a high cell density by integrating an RRAM device at the intersection of perpendicular metal lines on adjacent metal layers, as described in Chapter 2.

An individual bit is selected by biasing a row and grounding a column within an array. Selecting a single row produces a voltage drop across the unselected rows and columns. This effect produces parasitic sneak currents that propagate through unselected cells, causing a degradation in sense margin and an increase in power consumption [46]. These currents prohibit the use of RRAM-only crosspoints in all

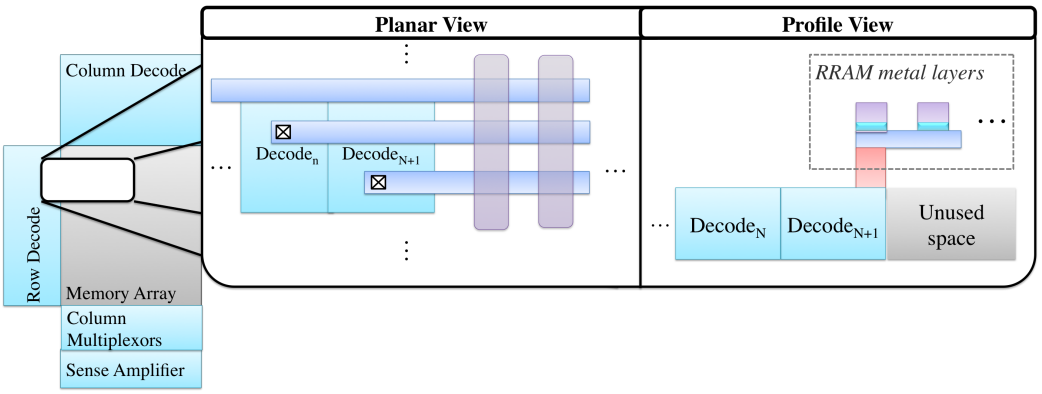


Figure 8.1: Planar and profile view of peripheral RRAM crosspoint array.

but the smallest arrays [47]. Larger arrays utilize a selector device (*e.g.*, a tunneling barrier) in series with the RRAM to ensure that only a small (leakage) current is passed through the unselected rows [48]. Unlike traditional CMOS memories, crosspoint memories also need to be bit addressable. Only a single bit can be written into a crosspoint array during a write operation due to the resistive load of the bit lines in large arrays. This characteristic requires additional area for the peripheral circuitry.

8.1.1 Related work

Recently, Liu *et al.* [248] demonstrated a vertically integrated RRAM crosspoint memory that integrates the peripheral access circuitry beneath the RRAM array. The approach places the column and row segmenting circuitry as well as the driver circuitry beneath the array, while placing the column decode circuitry peripheral to the array. The topology proposed here avoids bit line segmentation and integrates

both the column and decode circuitry beneath the array and is compatible with the approach described in [248]. Expressions are provided in this paper to size the array according to the physical dimensions of the decoder to ensure that the decode circuitry is beneath the RRAM crosspoint array. Niu *et al.* [249] provide an area and cost model that supports placing the driver circuits beneath the crosspoint array.

8.2 Physical design of RRAM crosspoint array

The area efficiency of a memory is the portion of the IC composed of the memory cells as compared to the total area of the memory system including all of the peripheral circuitry. Memories typically exhibit array efficiencies ranging from 30% to 40% for deeply scaled technologies. Only a fraction of the total die area is therefore dedicated to data storage. Higher array efficiencies increase memory capacity without additional die area.

CMOS memory arrays rely on pitch matching of the peripheral circuits, such as the decoders and sense amplifiers, to the width of the corresponding row or column. The height of a row decoder is equivalent to the height of a cell. In a minimum sized RRAM technology, however, the dimensions are significantly smaller (see Fig 2.5), making pitch matching difficult. The height of a crosspoint cell is 2F to 3F but the minimum height and length of a transistor is generally more than 3F before considering interconnect. The peripheral decode circuitry is therefore placed

with staggered interconnect to drive an individual row or column, increasing the area of an array, as illustrated in Fig. 8.1.

RRAM crosspoint arrays, however, are fabricated within the metal layers and do not utilize the silicon area beneath the memory array. The proposed topologies embed the decoding circuitry beneath the crosspoint array to reduce area, thereby increasing the area efficiency.

The key idea of the proposed topologies is to place the decode circuitry within a grid, beneath the crosspoint array, and to stagger the contacts to ensure that each decode block connects to a single row, as illustrated in Fig. 8.2. Intuitively, the height of a decoder can be hidden across multiple rows and the width of the decoder can be hidden beneath columns. This structure creates a grid of sub-crosspoint decoders beneath the crosspoint array.

Furthermore, gaps are introduced into the array interconnect to improve area efficiency. Despite the slight reduction in cell density, the overall area of an array is reduced. These gaps are strategically introduced into the array to facilitate access to the columns for column decoding with minimal effect on the physical area.

The decoding circuit used for both topologies is described in Section 8.2.1 The proposed topology for sub-crosspoint row decoding is described in Section 8.2.2 followed by the topology for sub-crosspoint row and column decoding in Section 8.2.3.

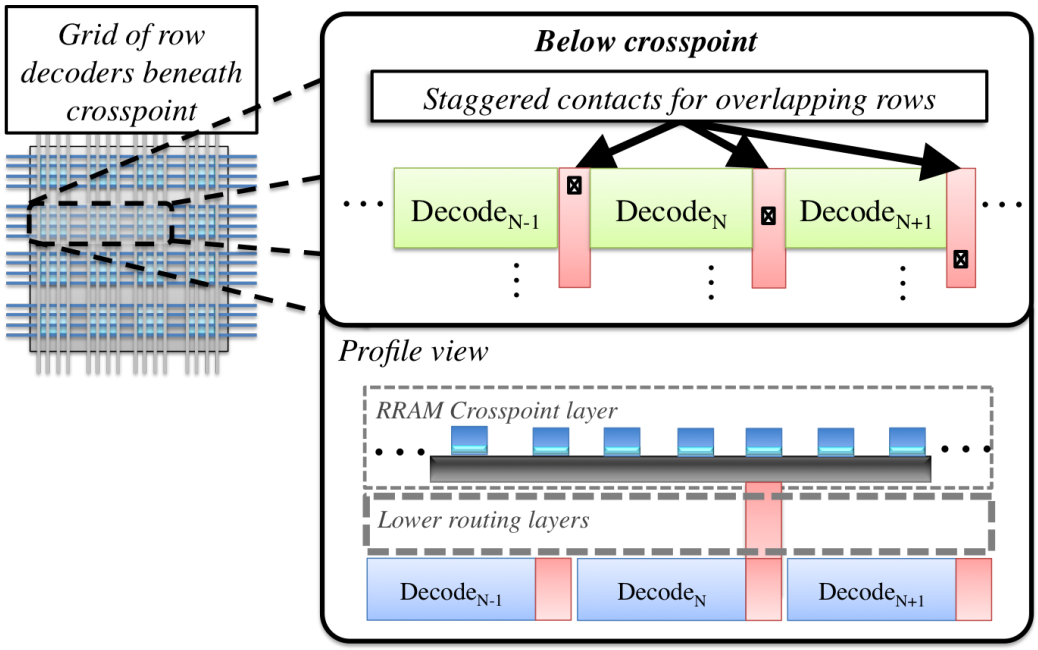


Figure 8.2: Planar and profile view of proposed RRAM sub-crosspoint row decoders.

8.2.1 NOR decoder circuit

A NOR-style decoder [199], commonly used in DRAM circuits, provides row and column decoding for both topologies and is shown in Fig. 8.3. The decoder is modified for resistive memories. The selection transistors, required for address decoding, are shown on the left. The driver circuitry for reads and writes are shown on the right. If all of the inputs are low, indicating a match, the decoding node is pulled high. If either R_{en}, W_{en_h}, or W_{en_l} is enabled, the address is valid and the row or column is driven. The write enable signals (W_{en_h} and W_{en_l}) are connected to the high and low voltage drivers to enable the bidirectional writes necessary for bipolar RRAM devices.

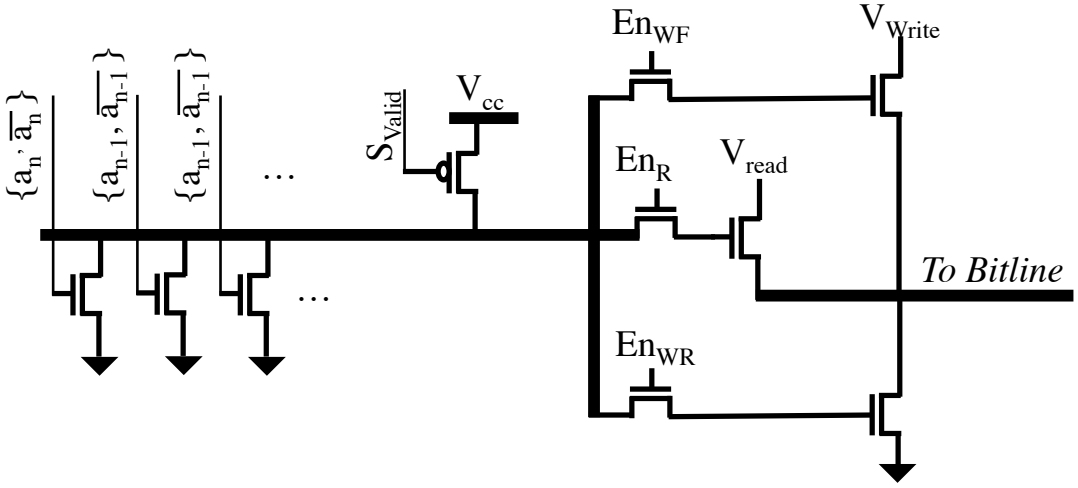


Figure 8.3: Decoder circuit

8.2.2 Sub-crosspoint row decoding

A sub-crosspoint row decoder is constrained by the physical size of the minimum sized transistor, size of the output driver, metal routing for the power and ground lines, and the number of columns and rows within the array. The transistor size and metal lines constrain the height of the decoder. The number of columns or rows determines the number of transistors required for decoding, which, in addition to the size of the output driver, determines the width of the decoder.

The height of a decoder is amortized across multiple rows, as illustrated in Fig 8.2. If the height of a decoder is H, $\frac{H}{k}$ decoders are placed side-by-side beneath a crosspoint array, where k is the minimum metal pitch of the technology. For binary decoding, the number of rows is a power of two. The number of rows (N_{row}) also indicates the number of decoders placed in parallel, as each decoder connects to

a single row. Placing decoders horizontally beneath a crosspoint array allows the global predecoding circuitry to drive a column of decoders, as depicted in Fig. 8.2.

The number of predecode bits is

$$N_{r\_pdec} = \left\lceil log_2(\frac{H_{sub} + H_{routing}}{H_{c\_pitch}}) \right\rceil, \tag{8.1}$$

where H_{dec} and $H_{routing}$ are, respectively, the height of the decode and routing lines normalized to the feature size of the technology. Hence, $2^{N_{r\_pdec}}$ is the number of rows required to "hide" a row decoder. The physical width of a row is

$$W_{row} = 2^{N_{r\_pdec}} W_{r\_sub}, \tag{8.2}$$

where

$$W_{r\_sub} = W_{drive} + 2W_{tr} log_2(N_{row}) - N_{r\_pdec}. \tag{8.3}$$

W_{r_sub} is the width of a single row decoder, W_{drive} is the width of a row driver circuit, N_{row} is the number of rows within an array, and W_{tr} is the width of the selection transistor within the decoder. This expression provides the minimum width of a row. The number of columns to maximize the density of this approach is $\frac{W_{row}}{W_{c\_pitch}}$.

8.2.3 Sub-crosspoint row and column decoding

Placing both the row and column decode circuitry beneath the crosspoint array creates an interdependence between the number of rows and columns. The rows of an RRAM crosspoint array are located directly above the silicon, permitting access from beneath. The row plane of the crosspoint, however, blocks access to the column plane of the crosspoint from beneath. A gap is therefore introduced between the rows to enable the sub-crosspoint decoder to communicate with the column rows. The gap between individual rows provides access to the crosspoint columns using the same metal layer as the row layer.

The physical topology of the sub-crosspoint decoder for both columns and rows is shown in Fig. 8.4. A decode sub-block consists of a co-located row and column decoder. The sub-blocks are oriented in a grid pattern beneath the crosspoint array. Contacts are staggered horizontally for rows and vertically for columns, as illustrated in Fig. 8.4. The column decoder is placed below the row decoder to share the power rails with the row decoder, and to ensure that the shape of a row and column decode sub-block is as close as possible to a square.

Completely hiding a sub-block requires $2^{N_{r\_pdec}}$ rows. If the same methodology is applied to a column decoder, $2^{N_{c\_pdec}}$ columns are required. The number of rows and columns of an array is $2^{N_{r\_pdec}+N_{c\_pdec}}$, ensuring that the array has an equal number of rows and columns.

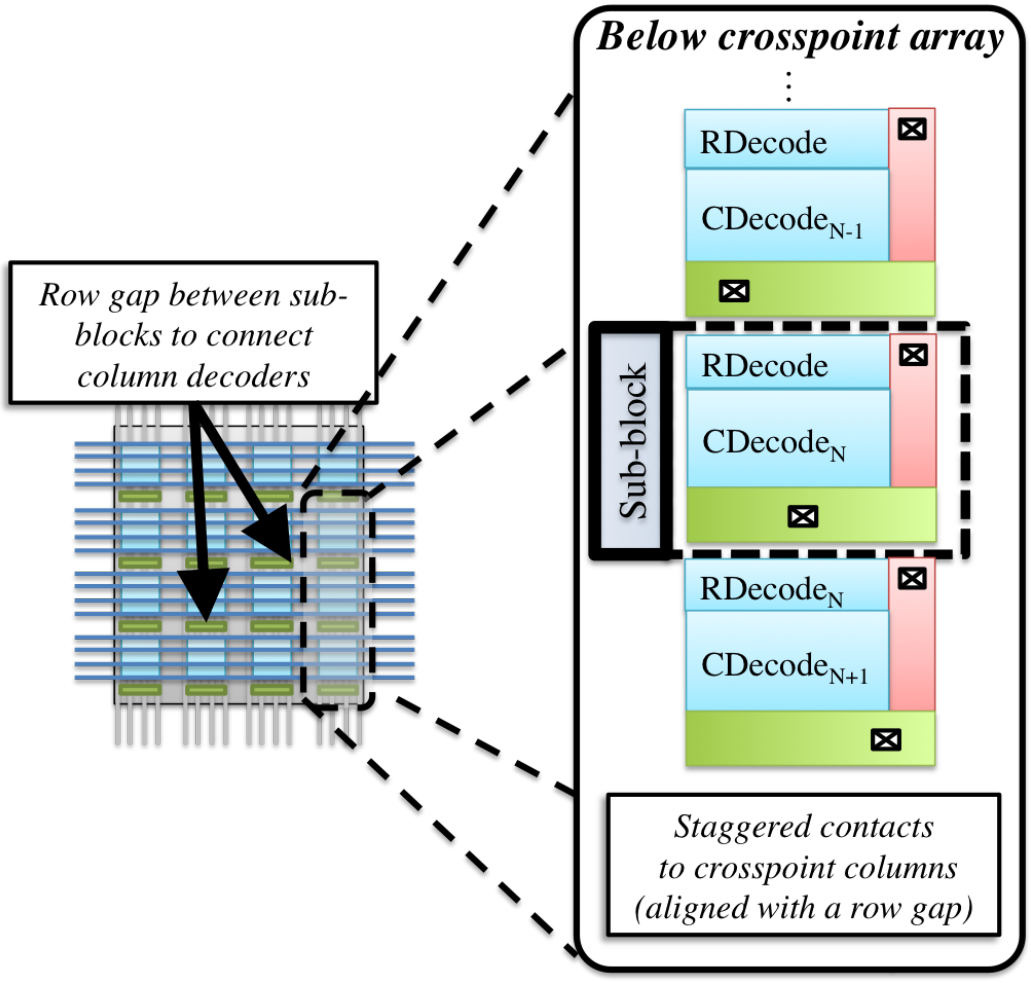


Figure 8.4: Physical topology of sub-crosspoint row and column decoders.

Expression (8.1) is used to determine the number of row predecode bits based on the height of a sub-block. The required number of column predecode bits is

$$N_{c\_pdec} = \lceil log_2(\frac{W_{r\_sub}}{W_{c\_pitch}}) \rceil, \tag{8.4}$$

where

$$W_{r\_sub} = W_{drive} + 2W_{tr}N_{r\_pdec}. \tag{8.5}$$

The height of an array is $2^{N_{c\_pdec}}(2^{N_{r\_pdec}}H_{c\_pitch}+1)$. The width is $2^{N_{r\_pdec}}2^{N_{c\_pdec}}W_{c\_pitch}$.

Assuming the same pitch for both the crosspoint rows and columns, the total area of a memory array with sub-crosspoint decoder circuitry is

$$A = 2^{N_{c\_pdec}+N_{r\_pdec}}H^2_{c\_pitch}(2^{N_{r\_pdec}} + 1). \tag{8.6}$$

Note that (8.1) and (8.4) contain integer ceiling functions, ensuring that the decode circuitry occupies less planar area than the crosspoint array. This constraint as well as the 2^n growth of the columns and rows with predecode bits produces unused space beneath the sub-block. This space can be utilized to increase the write drivers and reduce the resistive load. Note that these expressions produce a unique array size that is ultimately dependent on the height of the sub-block. The array is therefore no longer a function of the number of rows, as in row-only sub-crosspoint

decoding.

8.3 Evaluation

Sub-crosspoint decoding is evaluated and compared to standard peripheral approaches in the following section. Note that this evaluation only considers the array efficiency of individual memory arrays and does not consider the global logic, decoders, and routing. The cell layout is based on 45 nm FreePDK design rules and is scaled to a feature size of 22 nm. It is assumed that an additional intermediate metal layer is available in 22 nm technology (see Table 8.1). The layout of the decoding circuitry is constrained to the first two metal layers (see Figs. 8.5 and 8.6).

The RRAM parameters are based on [137] and scaled to 33 nm (3F is the minimum metal pitch for the local and intermediate metal layers and thus defined the geometry of an RRAM device). A Simmons tunnel barrier model [108] is used to simulate the selector device. *In lieu* of high voltage transistor models, the write driver is sized according to a 0.9 volt, 22 nm PTM model and scaled to provide double the current required to apply 3 volts to an RRAM device in the on state. No more than 10% of the resistive load is due to the bit lines to ensure that at least 3 volts are dropped across the RRAM device during a write. The area of the peripheral sense amplifiers and column multiplexors is modeled using the methodology provided in CACTI [201].

Table 8.1: Parameters

Feature size (F)	22 nm
Metal pitch	3F
Routing metal layers	1 to 2
Crosspoint metal layers	3 to 4
R_{on}	34.9 kΩ
RRAM write voltage	3 V
Tunnel barrier thickness	1.15 nm
Tunnel barrier bandgap	0.6 eV

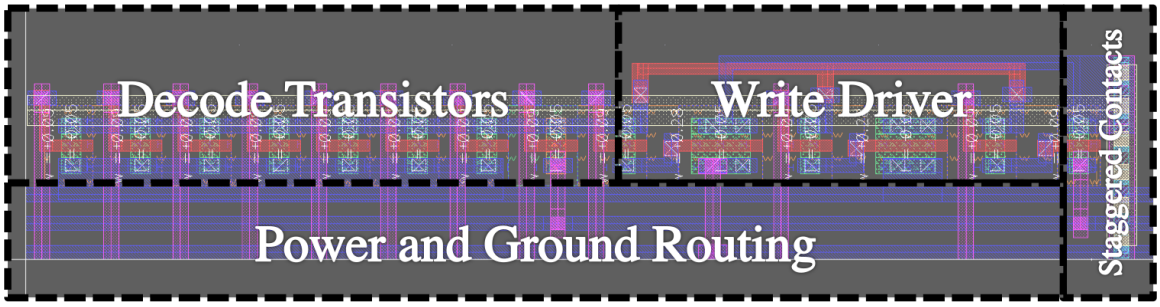


Figure 8.5: Layout of row decoder in 45 nm CMOS.

A comparison of the peripheral approach with the sub-crosspoint row decode topology for a rectangular array is listed in Table 8.2. A rectangular array integrates as many columns as possible. For smaller array sizes, the proposed topology reduces the overall area by 38.6% and improves area efficiency by more than 20%. For large arrays with 2,048 rows, the area advantage decreases to 6.4% as the area of the array dominates the structure.

The area of a traditional square array with an equal number of rows and columns is listed in Table 8.3. Similar to a rectangular array, the physical area for smaller arrays exhibits an improvement of 36.0% and 16.8%, respectively, for area and area

Table 8.2: Comparison of rectangular arrays

	Peripheral array			**Sub-crosspoint row decoder array**		
Number of rows	Number of column	Area (μm^2)	Array efficiency	Area (μm^2)	Area efficiency	Area reduction
128	167	298.3	35.7%	183.3	57.3%	38.6%
256	197	590.8	42.9%	430.4	58.0%	27.2%
512	228	1,206.3	48.5%	992.2	58.2%	17.8%
1,024	258	2,518.7	52.5%	2,244.0	58.3%	10.9%
2,048	288	5,352.2	55.0%	5,009.5	58.4%	6.4%

Table 8.3: Comparison of square arrays

	Peripheral array		**Sub-crosspoint row decoder array**			
Number of rows and columns	Area (μm^2)	Area efficiency	Area (μm^2)	Area efficiency	Sub-crosspoint unused space	Area reduction
128	249.2	31.8%	159.5	48.6%	-	36.0%
256	705.3	44.9%	493.9	62.2%	77.1	30.0%
512	2,109.9	59.1%	1,622.9	74.5%	660.9	23.1%
1,024	6,760.0	72.3%	5,657.8	84.0%	3,477.4	16.3%
2,048	2,3168.3	82.6%	20,707.2	90.5%	15,833.5	10.6%

efficiency. While the reduction in area follows a similar trend in rectangular arrays, the proposed approach maintains an area efficiency advantage unlike with rectangular arrays. This approach also demonstrates that sub-crosspoint row decoding utilizes only a small portion of the area under a crosspoint array for larger array sizes. At 2,048 columns and rows, 76% of the area under a crosspoint array is unused, permitting additional peripheral logic to be placed under the array to improve the area efficiency of larger size arrays.

The column decode circuitry can be integrated beneath the crosspoint array in the manner described in Section 8.2. The sub-block decoder is shown in Fig. 8.6. As

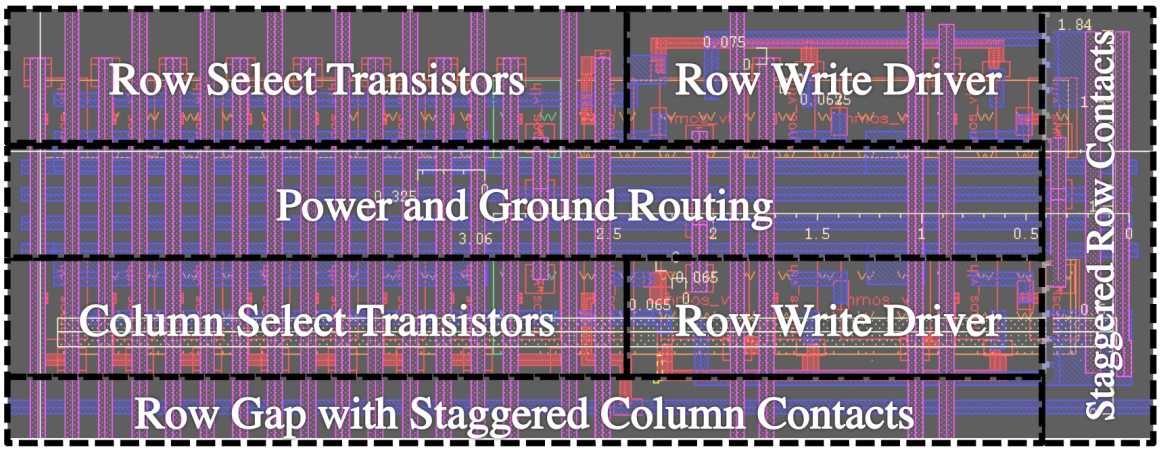


Figure 8.6: Layout of decoder sub-block in 45 nm

Table 8.4: Sub-crosspoint row and column decoder, square array

Number of rows and columns	Area μm^2	Area efficiency	Area reduction
2,048	19,412.4	99.9996%	27.1%

listed in Table 8.4, the array efficiency of this approach is nearly 100%. Moreover, a 27% reduction in area is produced.

8.3.1 Implications of sub-crosspoint decoder on array size

Sub-crosspoint row decoding is best applied to smaller RRAM arrays, where the array size and peripheral circuitry are comparable. For square arrays greater than 512 x 512, the unused space beneath the array is at least 31.3% of the array area and grows as high as 76% in 2,048 x 2,048 arrays. For an array size of 256 x 256, the unused area is 11% with a 17.3% improvement in area efficiency.

The sub-crosspoint row and column decoding approach presented here presents an inflection point at an array size of 2,048 x 2,048 that achieves near 100% area efficiency. A smaller number of rows and columns exhibits lower area efficiency. While sub-crosspoint topologies produce smaller arrays than the standard peripheral approach, the reduction in area efficiency degrades the storage capacity of an individual memory. Arrays larger than 2,048 x 2,048 are dominated by the area of the crosspoint array and result in a negligible improvement in array efficiency. While additional sub-blocks are required to decode larger arrays, the area of the memory cells is larger than the area of the sub-blocks. Thus, additional unused area is available beneath the array, although with increased resistive and capacitive impedances within the crosspoint array.

8.4 Conclusions

Two physical topologies for sub-crosspoint decoding of an RRAM based memory are demonstrated. The two approaches are sub-crosspoint row decoding, and sub-crosspoint row and column decoding. Expressions are provided for both topologies to size a crosspoint array as well as the column and row decode circuitry. Sub-crosspoint row decoding reduces area by up to 38.6% over the standard peripheral

approach, with an improvement in area efficiency of 21.6% for small 128 x 128 arrays. For large 2,048 x 2,048 square arrays, area is reduced by 10.6% with a corresponding improvement in area efficiency of 8.0%. Sub-crosspoint row and column decoding reduces the RRAM crosspoint area by 27.1% and improves area efficiency to nearly 100%.

Chapter 9

STT-MRAM Based Multistate Register

9.1 Introduction

Resistive memories are poised for integration into standard CMOS processes, providing novel performance, power, and reliability capabilities. Resistive memory exhibits qualities that are amenable to a variety of different applications. Metal-oxide resistive RAM (RRAM) is suitable as a DRAM and flash memory replacement due to the multi-bit capability and high density [27, 250]. While device technology remains an active area of research, state-of-the-art RRAM endurance is on the order of 10^{12}, still too low for high performance microprocessor applications [251, 252], and requires high voltages (1.8 to 10 volts) for sub-nanosecond writes [253].

Spin torque transfer MRAM (STT-MRAM) is a resistive memory with CMOS compatible voltages and practically infinite endurance [254]. The challenges of STT-MRAM, however, have been providing reliable, fast switching while overcoming

the relatively small off/on resistance ratio ($\frac{R_{off}}{R_{on}}$) [231].

In this paper, STT-MRAM is integrated into a non-volatile multistate pipeline register (MPR) [255]. Physical MRAM phenomena, such as field assisted switching [256–258], reduced non-volatility [222], and device structures are evaluated. An STT-MRAM-based circuit is compared to an RRAM-based multistate register in terms of area, power, and speed, and applied at both the gate level and within a multi-threaded microprocessor using multistate registers as pipeline registers [209]. A STT-MRAM-based MPR circuit is demonstrated that achieves a 487 ps write latency, consistent with in-core operation.

Background on resistive memories is provided in Section 9.2. The proposed multistate register is described in Section 9.3. Device and circuit evaluation is discussed in Section 9.4, followed by some conclusions in Section 9.5.

9.2 Background

Different resistive memory technologies share a set of common characteristics. Electrically, the devices behave as two terminal resistors, where a sufficiently large current or voltage bias changes the steady state resistance. Physically, each device is typically fabricated by depositing one or more thin films in series with an intermetal via on the back end of a CMOS process.

9.2.1 MRAM

MRAM has several traits that differ from other resistive memories. Unlike RRAM and other memristive technologies, MRAM is a bistable resistive memory, capable of existing in either a maximum (R_{OFF}) or minimum (R_{ON}) resistance state. MRAM also exhibits a high-to-low resistance ratio on the order of two to three, much smaller than the 10^2 to 10^6 ratio exhibited by RRAM devices [27].

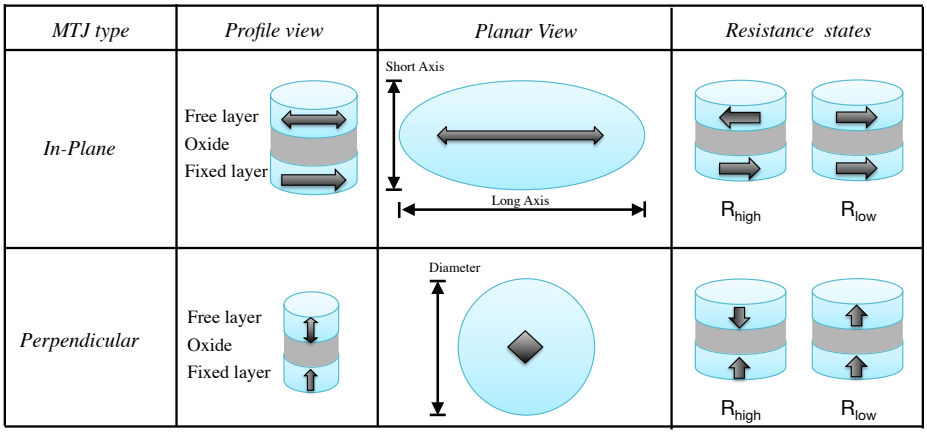


Figure 9.1: Device structure and magnetic orientation of in-plane and perpendicular MTJs

These differences arise due to the physical switching and conductance mechanisms exhibited by MRAM. MRAM devices operate by switching the orientation of a magnetic domain (free layer) with respect to a magnetic reference domain (fixed layer). This switching process does not alter the crystalline structure of the material and therefore avoids the physical degradation exhibited by RRAM. Consider the in-plane MTJ illustrated in Fig. 9.1. If the domains are in the opposite orientation, the device resistance is R_{OFF}. Domains orientated in the same direction exhibit a low

resistance (R_{ON}). Applying a bias to the device produces a force on the magnetic orientation of the domain. A sufficiently high bias will switch the device state. The polarity of the bias controls the final resistance, *i.e.*, a positive bias switches the device to a high resistance, and a negative bias switches the device to a low resistance. Switching a domain, however, is a stochastic process that is strongly dependent on temperature.

9.2.1.0.1 Perpendicular versus in-plane MTJ structure STT-MRAM devices can also be configured in either an in-plane or perpendicular-to-plane structure, as illustrated in Fig. 9.1. The key difference between these two device types stems from the steady state orientation of the magnetic domains. In-plane MTJs (IMTJ) are oriented with the plane of the thin film whereas perpendicular MTJs (PMTJ) are oriented orthogonal to the film. Structurally, in-plane MTJs are patterned in an oval shape to control the relative direction of the device magnetization. The magnetization of a perpendicular-to-plane device is controlled by the crystalline structure of the magnetic layer and material interfaces [251]. These devices are circular unlike in-plane MTJs.

The oval structure of an in-plane MTJ creates a demagnetization field within the device that is caused by the non-uniform dimensions of the device. This field partially opposes the switching process. Perpendicular MTJ are symmetric, avoiding

the effects of the demagnetisation field. PMTJs also exhibit higher thermal stability than in-plane MTJs, and require lower switching currents. PMTJ fabrication, however, is more difficult than IMTJs due to the complex material stack needed to control the crystallographic orientation of the thin film [52].

9.2.1.0.2 Physical mechanisms to enhance STT switching In first generation MRAM, current induced magnetic fields switch the individual devices [259]. Magnetic fields, however, are difficult to control and exhibit scaling issues such as half select disturbance [223, 260] and high write energy . Augmenting an STT-MRAM with a magnetic field [224, 225] has been proposed to improve device switching speeds [257, 258, 261] to enhance system performance in on-chip caches [256].

Additionally, the ten year state retention constraint commonly used in commodity MRAM and DRAM [222, 251] is unnecessary for in-core microprocessor applications. By reducing the volume of the switching layer of the MRAM device, the stability of the device degrades, shortening the retention time. This reduced stability also lowers the energy and latency of the switching process, enabling greater performance in STT-MRAM based caches [262].

9.3 Circuit Design

A multistate flip flop with STT-MRAM is proposed for in-pipeline registers. The circuit is based on [255] with modifications to support STT-MRAM. A brief

overview of circuit operation is provided in this section.

The multistate flip flop primarily operates as a digital CMOS register within a pipeline which switches data from a local scratchpad memory when triggered by a global signal, as illustrated in Fig. 9.2. The modified circuit structure is illustrated in Fig. 9.3. In this state, the circuit operates in a CMOS mode without interfacing to the STT-MRAM. Upon application of the enable signal, the master stage of the flip flop writes data into a cell within the scratchpad memory (see Fig. 9.3). After the write completes, a second MTJ is selected from the scratchpad, and is passed to the slave stage of the CMOS register.

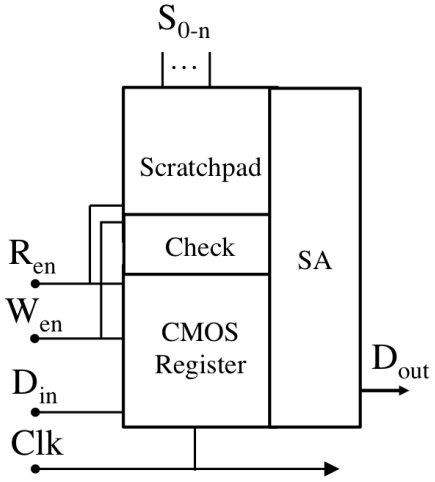


Figure 9.2: STT-MPR flip flop. During normal operation, the MPR operates as a digital register. When triggered, a MTJ is selected from the scratchpad by $\mathrm{S_n}$. The write enable signal $\mathrm{W_{en}}$ is set high, and the data in $\mathrm{Stage_0}$ is written to the scratchpad. A local check circuit ensures a successful write. A different MTJ is selected from the scratchpad. The $\mathrm{R_{en}}$ signal is set high and the CMOS register is reconfigured into a sense amplifier to read the selected MTJ.

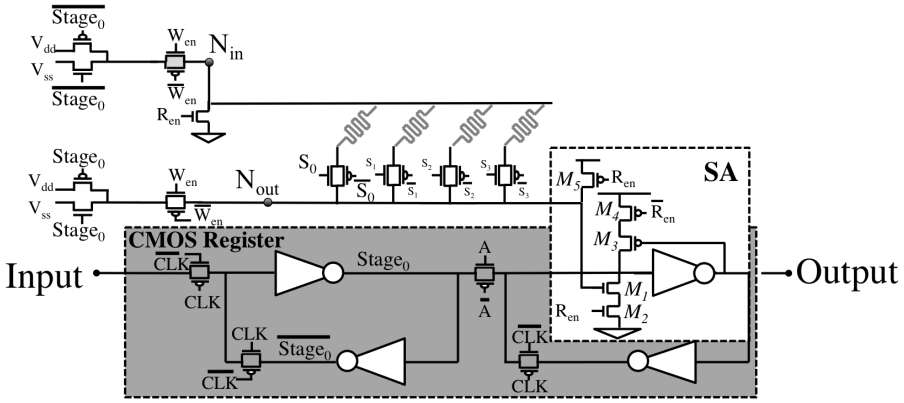


Figure 9.3: Circuit diagram of STT-MPR flip flop.

9.3.0.0.3 Field switching Field assisted switching [224] applies a large perpendicularly oriented magnetic field across an MTJ. The applied magnetic field destabilizes the MTJ, decoupling the switching process from random thermal perturbations, enabling the device to switch faster. To generate the required magnetic field, a current of up to several milliampheres is placed adjacent to the MTJ device. The energy can be reduced by sharing this current among multiple MTJs [256].

9.3.0.0.4 Compensating for stochastic switching The switching process of an MTJ is a random process, resulting in a finite probability of incomplete switching [263]. To compensate for this issue, a checker read is included to ensure that the target value is written. An additional read step is performed after the write. The output of the read is compared to the stored state in the master stage with an XOR gate. If the output of the XOR gate is high, the write is attempted again until the correct result is written. While writes are triggered globally, only a small fraction

of local registers need to be rewritten due to incomplete writes. Additional logic is therefore included to bypass the global write enable signal to ensure that only unsuccessfully switched registers are rewritten. The circuit is shown in Fig. 9.4.

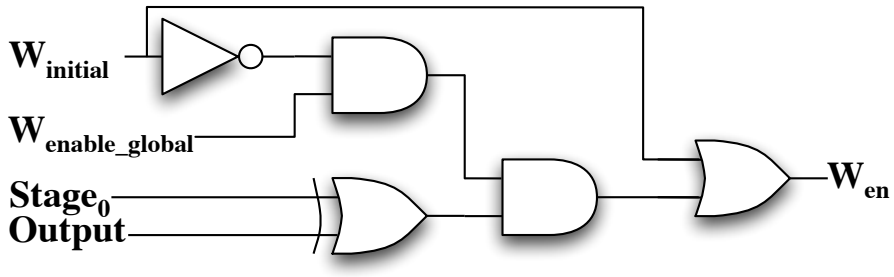


Figure 9.4: Check logic for STT-MPR flip flop.

Each switching event x can be treated as a Bernoulli trial [264]. The number of switching attempts (N) is therefore a binomial random variable. The probability of switching p_{sw} is determined from the switching probability of an STT-MTJ for an applied write pulse of specific duration. The probability of switching on the n^{th} attempt is

$$P(x_1 = 0 ... x_{n-1} = 0, x_n = 1) = p_{sw}(1 - p_{sw})^{n-1}. \tag{9.1}$$

The total number of switching attempts is therefore the expected value of N,

$$\begin{aligned} E[N] &= \sum_{n=0}^{\infty} n p_{sw}(1 - p_{sw})^{n-1} \\ &= 1/p_{sw}. \end{aligned} \tag{9.2}$$

The average switching latency t_{sw_avg} of a device is therefore

$$t_{sw\_avg} = \frac{t_{pw}}{p_{sw}}, \tag{9.3}$$

where t_{pw} is the length of the current pulse applied to the MTJ.

9.3.0.0.5 Architectural concerns In the continuous flow multithreaded architecture (CFMT) [209], an idle state is stored within an MPR scratchpad during a DRAM memory access. An MPR based multithreaded microprocessor therefore needs to retain an MTJ state for $S * t_{DRAM}$, where S is the number of stored states. The latency of a DRAM access (t_{DRAM}) is approximately 50 nanoseconds [199], but varies based on the configuration, application, and queueing policy for read accesses in the memory subsystem. Much of the architectural speedup provided by CFMT is achievable with 16 scratchpad states [255]. A retention time of one microsecond is therefore the lowest possible delay without causing errors.

9.4 Setup and evaluation of circuit

The circuit simulations are based on the predictive technology model for 22 nm CMOS [227]. The physical area is based on FreePDK45 scaled to the 22 nm technology [211]. The MTJ device parameters are based on ITRS [51] for the in-plane

device and [265] for the perpendicular MTJ. MTJ switching for in-plane and perpendicular devices utilizes a macrospin LLG solver with the M^3 simulator [226]. Parameters for each MTJ variant are listed in Table 9.1 with the LLG simulation parameters listed in Table 9.2. The thermal stability of each MTJ is varied by reducing the thickness of the free layer [222, 266].

Table 9.1: MTJ parameters

	In-Plane	Perpendicular
Saturation Magnetization (M_s)	8 x $10^5 A/\mathrm{m}$	12.6 x $10^5 A/\mathrm{m}$
Long axis	70 nm	40 nm
Short axis	20 nm	40 nm
Thickness	2.9 nm	1.3 nm
$\mathrm{R_{on}}$	5 kΩ	5 kΩ
TMR	150%	127%
$I_{switching}$	61.5 μA	80 μA
Field line spacing	21 nm	21 nm

Table 9.2: LLG Simulation parameters

γ	1.76x10^{11} $\frac{\mathrm{rad}}{\mathrm{s-T}}$
α	0.01 (in-plane) / 0.027 (perpendicular)
Temperature	350 K
Time step	0.25 ps
Initial angle (θ_0)	2 °

9.4.1 Physical area

The area of the STT-MPR consists of the CMOS register, scratchpad memory, sense circuit, and check logic (see Table 9.3). The additional area required for the write checking logic is greater than a typical digital CMOS register. The scratchpad requires an additional transistor per MTJ, exhibiting linear growth with additional stored states. These additional constraints result in a greater area overhead than

an RRAM based MPR [255]. Note that the in-plane and perpendicular MTJs are integrated within metal vias, and therefore either technology does not effect the overall area.

The field based STT-MPR requires greater area than a non-field assisted register due to the additional field lines. The area is therefore larger, and the overhead per state shrinks at a slower rate than a non-field assisted STT-MPR.

Table 9.3: Area of STT-MPR configurations

	Area(μm)	Overhead	Overhead per state
Register	2.0	-	-
Check Logic	3.0	152.8%	-
STT 4	7.2	264.3%	66.1%
STT 4+8	8.1	310.4%	25.9%
STT 4+16	8.9	350.8%	17.5%
STT 4+32	10.5	429.2%	11.9%
STT 4+64	13.7	591.5%	8.7%
STT 4 Field	7.5	279.8%	70.0%
STT 4+8 Field	11.8	495.0%	41.2%
STT 4+16 Field	14.5	632.4%	31.6%
STT 4+32 Field	19.9	907.2%	25.2%
STT 4+64 Field	30.8	1456.9%	21.4%

9.4.2 STT-MTJ read latency and energy

The read energy and latency of the STT-MPR are listed in Table 9.4. The read energy and latency are dominated by the junction capacitances and resistance of the MTJ and select transistors. The read delay also increases as the number of MTJs increases. This trend differs from an RRAM based MPR where the resistance of the RRAM device dominates the impedance observed by the sense circuit, and is therefore mostly independent of the number of storage devices. The smaller TMR

of the PMTJ (127%) as compared to the IMTJ (150%) results in a higher delay (up to 1 ps) and energy (up to 0.5 fJ) for MTJ reads.

Table 9.4: MTJ scratchpad read delay and energy

	In-plane MTJ		Perpendicular MTJ	
MTJ count	Read delay (ps)	Read energy (fJ)	Read delay (ps)	Read energy (fJ)
STT 4	95.3	2.9	96.3	3.4
STT 4+8	119.8	3.8	121.1	4.3
STT 4+16	143.7	4.8	145.2	5.3
STT 4+32	192.2	6.8	193.9	7.2
STT 4+64	290.2	10.7	290.9	11.2

9.4.3 MTJ write latency and energy

By reducing the device retention time and applying a magnetic field, the MTJ switching latency is reduced, as illustrated in Figure 9.5. Reducing the retention time without an applied field reduces the latency of the device to approximately 2.1 nanoseconds. A 6.5 mA field current achieves a switching latency of 1.1 nanoseconds with a retention time of 100 seconds. Applying high fields at retention times below 100 seconds, however, causes the MTJ to oscillate between the high and low states during writes. This oscillation may be avoided by terminating the field immediately after a write. This precise timing, however, is impractical when sharing a field current among multiple STT-MPR circuits; further reductions in the switching latency is therefore not possible without local detection of switching.

The switching latency of perpendicular MTJs with reduced retention time and applied fields is illustrated in Figure 9.6. Unlike in-plane MTJs, magnetic fields have a less pronounced effect on the switching latency of perpendicular MTJs. A 1.6

mA field assistance exhibits switching latencies similar to a non-field assisted MTJ. Moreover, all field currents demonstrate little change over a non-field assisted perpendicular MTJ at lower retention times. The reduced effectiveness of the field can be explained by the lack of a demagnitization field within the PMTJ. Field assisted switching, however, can reduce the switching latency for longer retention times, indicating that field assistance is more effective in array-based memory structures. PMTJs with reduced retention times and zero field current exhibit sub-nanosecond switching (505 picoseconds) and are therefore advantageous for in-core memory applications.

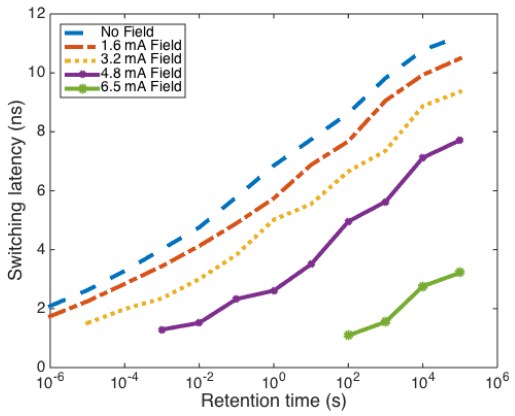


Figure 9.5: Retention time characteristics of an in-plane MTJ with reduced retention time and current induced magnetic fields. The switching latency is reported for p_{sw} = 0.95

The write energy of several MTJ configurations are listed in Table 9.5. These device configurations provide retention times of one microsecond and one millisecond to describe, respectively, aggressive and worst case timing conditions. The energy consumption of an STT multistate register is greater than an RRAM MPR.

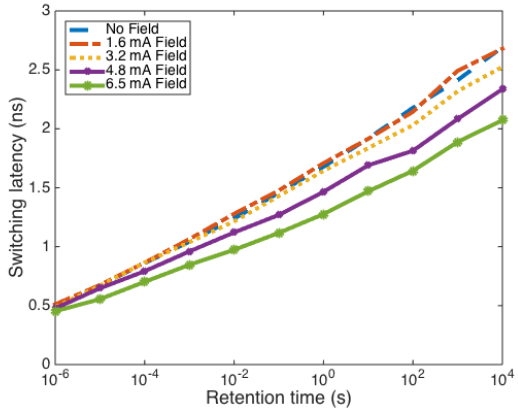


Figure 9.6: Retention time characteristics of a perpendicular MTJ with reduced retention time and current induced magnetic fields. The switching latency is reported for $p_{sw} = 0.95$

This higher energy is due to the much lower resistance of MTJ devices as compared to RRAM devices. While field currents lower the MTJ switching latency, the relatively high resistance of the field line reduces the number of gates that can share a field current, increasing the energy consumption of the register.

Table 9.5: STT Multistate Register Write Energy for MTJ Configurations

		In-plane MTJ			Perpendicular MTJ		
	Number of gates	Retention	Write energy	Latency @ 2 Sigma	Retention	Write energy	Latency @ 2 Sigma
No Field	-	1.00E-06	1.02E-13	2.09E-09	1.00E-06	3.23E-14	5.05E-10
		1.00E-03	1.97E-13	4.02E-09	1.00E-03	6.67E-14	1.04E-09
Field 1.6 mA	33	1.00E-06	1.57E-13	1.71E-09	1.00E-06	5.40E-13	5.11E-10
		1.00E-03	3.11E-13	3.44E-09	1.00E-03	1.13E-13	1.06E-09
Field 3.2 mA	12	1.00E-06	3.90E-13	1.51E-09	1.00E-06	1.32E-13	4.89E-10
		1.00E-03	6.10E-13	2.34E-09	1.00E-03	2.84E-13	1.04E-09
Field 4.8 mA	6	-	-	-	1.00E-06	3.64E-13	4.76E-10
		1.00E-03	9.70E-13	1.29E-09	1.00E-03	7.34E-13	9.60E-10
Field 6.5 mA	2	-	-	-	1.00E-06	1.26E-12	4.52E-10
		1.00E+02	3.05E-12	1.09E-09	1.00E-03	2.37E-12	8.47E-10

9.5 Conclusions

An STT-MTJ based MPR circuit is proposed for use in an in-core microprocessor pipeline and compared with an RRAM based MPR. Lowered thermal stability and current induced magnetic fields are explored to reduce switching latency. These effects are evaluated for both in-plane and perpendicular MTJs, demonstrating a significant reduction in the write latency. Field assisted switching is shown to exhibit a much stronger effect for IMTJs than PMTJs. PMTJs with reduced retention times can achieve sub-nanosecond switching without the application of a nearby magnetic field. Both MTJ technologies are capable of switching latencies on the order of one nanosecond, enabling switching delays compatible with in-pipeline microprocessor blocks. While additional write power (an increase of at least 10.5 fJ) and area (an increase of at least 168%) are required as compared to an RRAM based MPR, the infinite write endurance and voltage compatibility of STT-MRAM is more amenable to the high frequency operation required by in-pipeline memory structures and should therefore be considered as an alternative to RRAM for in-core microprocessor.

Chapter 10

Conclusions

The cadence of semiconductor technology development has echoed through many industries, enabling exponential computing advancements and novel computing platforms. Computing has relied on the availability of inexpensive, low power, high capacity memory, both to manage system challenges like the Von Neumann bottleneck and to increase the size and scope of computing. As the generational improvements of mainstream CMOS memories begin to slow, memristors are a class of technologies capable of supporting future performance and capacity goals while offering opportunities for novel circuit functions.

Memristor technologies have progressed through a period of rapid development, with MRAM and metal-oxide RRAM being two of the most prominent forms. Much of the current research has focused on fabrication challenges. Multiple material systems have been evaluated and explored for RRAM, such as HfO, TiO, TaO, AlO, and SiO as well as other forms of oxide-based technologies. Research in this

domain has focused on increasing density, improving endurance, controlling variability, and reducing write and read time and energy. MRAM research has focused on improving yield, optimizing device characteristics, and improving cell density. Increasing the on-off resistance ratio while reducing switching delay and write currents has been a key challenge.

These objectives directly affect the performance of modern memory systems. The density of a memory system determines the impedance of a memory array, placing a fundamental limit on capacity. The write energy affects the viability of a memory device for certain applications. Variability and resistance ratios impact the read performance of an array.

The research described in this dissertation considers these challenges from physical topology and circuit design levels. Field assisted writes of MRAM devices specifically address the read and write energy of MRAM caches, significantly reducing energy and latency while enabling MRAM for latency critical applications such as in-core cache. This concept is extended to STT-MPR, which enables the use of MRAM in pipeline registers. STT-MRAM memory cells enhance the low resistance ratio of MRAM to improve sense margins, providing a mechanism to enhance read performance. Arithmetic encoding introduces compression storage into an individual cell, improving the effective density of an array. Sub-crosspoint decoding similarly reorganizes the array circuitry to both increase both array efficiency and memory array capacity.

The research described in this dissertation also proposes novel circuits enabled by resistive memories. The memristive pipeline register integrates memristors within a high performance microprocessor, supporting greater functionality through faster thread switching and higher system performance. An MTJ-based inductor explores the novel use of MTJs to achieve high density on-chip inductance.

The topics described in this dissertation address some of the core concerns of resistive memories while exploring novel applications for these devices. Modern memristive circuits and systems promise to revolutionize memory systems. The results described in this dissertation support the development of next generation computing systems.

Chapter 11

Future Work

RRAM and flash memories are considered competing memory devices. Combining these two technologies, however, may solve the pressing issue of variational effects in semiconductor analog circuits. Analog circuits are typically relegated to highly mature process technologies that cannot achieve the performance of leading edge processes. Despite the use of mature technologies, these circuits are subject to overdesign to achieve acceptable yield. The combination of these effects leads to a "fixed bandwidth-accuracy-power tradeoff which is set by technology constants" [267]. Moreover, performance characteristics often degrade in final products. Analog-to-digital converters (ADCs), for example, are designed with a target bit resolution but the actual circuit will often support a much lower resolution due to noise and process variations.

RRAM coupled with flash memory has the potential to avoid these problems and achieve circuits with low variations. This capability is due to two fundamental characteristics of RRAM and flash transistors: 1) the conductance/transconductance

is tunable, and 2) the change in conductance/transconductance is invariant below a threshold voltage.

Analog circuits based on memristors and memristive transistors, *i.e.*, three terminal switches with modifiable gain and channel conductance, are suggested as a remedy to the issues of variation in modern analog ICs, potentially achieving better performance while consuming less power. Circuits based on these devices can be fabricated with a configurable circuit topology, programmed after fabrication, and operated below the memristor threshold voltage. The device programmability enables two key features currently unrealizable in high performance analog circuits. Variations can be detected and corrected during the post-fabrication test process, enabling more aggressive transistor sizing to produce higher performance circuits while dissipating less power. General circuit structures can be programmed after device fabrication, enabling field programmable analog circuits.

These two research directions are presented in this chapter for further investigation. Variation reduction with tunable analog circuits is proposed in Section 11.2. Enhanceing the circuit structure, interconnect design, and related design methodologies required for memristor-based high performance programmable analog circuits are suggested in Section 11.3. A summary of this chapter is provided in Section 11.4.

11.1 Variations in Analog Circuits

The analog design process requires an understanding of variational effects on transistor and interconnect performance. Global variations can affect large portions of an IC. Local variations can cause parameter mismatch among individual transistors in close proximity.

Local variations produce changes in transistor process parameters that are typically controlled by changing the physical area of a device. Channel carrier concentration, gate oxide thickness, sheet resistance, and free carrier mobility exhibit a variance,

$$\sigma_p^2 \propto \frac{1}{LW}, \tag{11.1}$$

where L and W are, respectively, the transistor length and width, and p represents the process parameter of interest [268]. Based on geometric parameters, the variance of electrical parameters (e of a device) can be modeled by

$$\sigma_e^2 = \sum_i (\frac{\delta e}{\delta p_i})^2 \sigma_{p_i}^2, \tag{11.2}$$

where p_i indicates the process parameter i [268]. Those electrical parameters most relevant to an analog circuit are the MOS drain current, transconductance, input voltage, and output conductance.

Addressing these variations is typically handled by employing sizing and patterning techniques. As the size of a transistor increases, the potential mismatch between local components typically decreases [268]. Placing devices in close proximity is therefore a common technique for improving analog circuit performance [268]. Active and passive devices are patterned with symmetric geometries to minimize any nonuniformities in component performance. Note that scaled technologies, due to increased variability with smaller lithographic patterning, are more sensitive to device mismatch. Consequently, analog circuits typically utilize much larger transistors than required by the design rules of the technology, forfeiting much of the performance enhancements derived from CMOS scaling.

11.2 Variation reduction in tunable analog circuits

Three basic amplifier circuits form the basis of many analog circuits. Memristive programmable versions of the common source, common drain, and differential pair amplifier circuits are shown in Figure 11.1. Both memristors and memristive transistors can be programed by controlling the power rail and certain internal nodes within the amplifiers. For the common source amplifier, programming the memristor consists of controlling the power rail, and grounding the central node and all other contacts within the amplifier. Each terminal can be controlled with high voltage (HV) transistors configured with multiplexers connected to external voltage

sources. Note that it is important to control the parasitic capacitance introduced by the high voltage transistors. The capacitance of the HV transistors can be reduced by biasing the bulk contact during normal circuit operation. This method requires additional wells for the HV transistors. The amplifier transistor is programmed in a similar manner by controlling each device contact with HV transistors. The same steps can be used to program the common drain and differential pair amplifiers.

11.3 Programmable analog circuits

To provide field programmability, a regular structure is required where several programmable memristors and transistors are incorporated within a configurable interconnect. For example, a programmable transistor (PT) can be implemented as an array of regular sized transistors (RST), as illustrated in Figure 11.2. RSTs are switched on to increase the conductance of a PT. Each set of RSTs is connected to the same gate, drain, and source contacts with different body contacts for programmability. A high conductance transistor (HCT) is introduced within an array of RSTs to control the parasitic junction capacitance of the array. Shutting off the HCT separates one set of RSTs from subsequent sets of RSTs in the array, isolating the capacitive load of the RST blocks. Similarly, arrays of memristors and RSTs can be used as programmable resistors and capacitors.

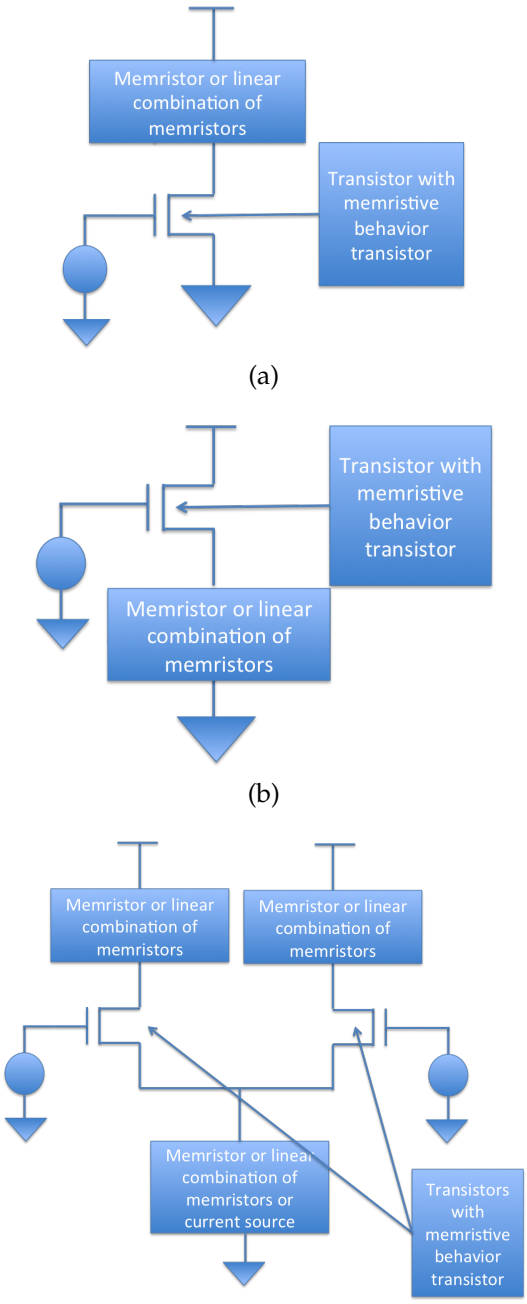


Figure 11.1: Circuit diagram of a) common source, b) common drain, and c) differential pair amplifiers with tunable components.

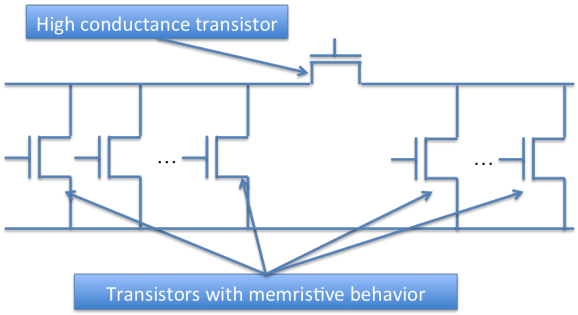


Figure 11.2: Programmable transistor array.

11.3.1 Support circuitry and methodology

Programmable analog circuits require support circuitry to program, test, and reprogram devices. A feedback loop based write circuit similar to the circuit described in Chapter 7 can be used to program specific memristor values based on a reference voltage or current. Some RRAM material systems also exhibit thermal degradation over time in high temperature environments [269, 270]. Periodic circuit refresh of device parameters may be required to maintain specific values in embedded applications.

Supporting a wide range of transistor characteristics while achieving a high degree of variation mitigation requires methodologies to size programmable memristors and transistor arrays and the related interconnect. Tradeoffs need to be made that consider the number of HCT devices, reduction in capacitance, and size of the HCTs.

An interconnect topology is required to interconnect these transistor and device arrays into an effective amplifier circuit. Two configurable layers of interconnect

are required. A local interconnect layer connects blocks of memristor devices and transistor arrays to create individual amplifier circuits. A global interconnect layer interconnects multiple amplifier blocks. The topology of the interconnect strongly affects the load characteristics, introducing a tradeoff between the parasitic interconnect impedance and the performance of the configured amplifier. As more devices are included within the local analog blocks, the interconnect load on the PTs and memristors increases. The parasitic interconnect impedances of the global network also introduce fan out problems, potentially degrading the performance of the amplifier.

11.3.2 Alternatives to flash

Other transistor technologies with tunable behavior have also been proposed [271]. Resistive switching has been demonstrated in bipolar transistors [271, 272]. In these technologies, a memristor layer is placed between the emitter and the base within the bipolar transistor. Modulating the memristor resistance changes the effective tunnel barrier thickness at the base-emitter junction. This method results in tunable current gain within the device which can support the proposed programmable analog circuitry. These technologies are based on the same set of conductance and fabrication mechanisms as RRAM. While these devices are still nascent, further development provides an avenue for enhanced scalability than is currently achievable with floating gate transistors.

11.4 Summary

RRAM technology has provided an avenue for circuit programmability that was previously only possible with discrete components. Two key research avenues are discussed for further investigation. Analog circuits that can be tuned after fabrication to remove variational effects are discussed, exhibiting the potential to use more aggressively sized transistors at scaled technology nodes to achieve higher performance. Programmable analog circuits are proposed for real-time changes to computing requirements. The development of these two research avenues may provide a radically different process for analog circuit design, enabling higher performance than existing approaches. Furthermore, high performance programmable analog circuits supports circuit synthesis, potentially enabling design automation of analog circuits as is commonly achieved with digital circuits. Future research on the development of circuits, physical layout, and methodologies to leverage the programmability of memristors for analog circuitry is of general importance to high performance computing.

Bibliography

[1] L. Chua, "Memristor — The Missing Circuit Element," *IEEE Transactions on Circuit Theory*, Vol. 18, No. 5, pp. 507–519, September 1971.

[2] D. B. Strukov, G. S. Snider, D. R. Stewart, and R. S. Williams, "The Missing Memristor Found," *Nature*, Vol. 453, No. 7191, pp. 80–83, October 2008.

[3] T. Prodromakis, C. Toumazou, and L. Chua, "Two Centuries of Memristors," *Nature Materials*, Vol. 11, No. 6, pp. 478–481, June 2012.

[4] W. Thomson, "On the Electro-Dynamic Qualities of Metals: — Effects of Magnetization on the Electric Conductivity of Nickel and of Iron," *Proceedings of the Royal Society of London*, Vol. 8, pp. 546–550, January 1856.

[5] M. Faraday, "Experimental Researches in Electricity," *Philosophical Transactions of the Royal Society of London*, Vol. 122, pp. 125–162, January 1832.

[6] H. Davy, "Researches on the Oxymuriatic Acid, Its Nature and Combinations; And on the Elements of the Muriatic Acid. With Some Experiments on Sulphur and Phosphorus, Made in the Laboratory of the Royal Institution," *Philosophical Transactions of the Royal Society of London*, pp. 231–257, January 1810.

[7] L. O. Chua and S. M. Kang, "Memristive Devices and Systems," *Proceedings of the IEEE*, Vol. 64, No. 2, pp. 209–223, February 1976.

[8] M. Di Ventra, Y. V. Pershin, and L. O. Chua, "Circuit Elements with Memory: Memristors, Memcapacitors, and Meminductors," *Proceedings of the IEEE*, Vol. 97, No. 10, pp. 1717–1724, October 2009.

[9] L. Chua, "Resistance Switching Memories are Memristors," *Applied Physics A*, Vol. 102, No. 4, pp. 765–783, March 2011.

[10] A. Beck, J. G. Bednorz, C. Gerber, C. Rossel, and D. Widmer, "Reproducible Switching Effect in Thin Oxide Films for Memory Applications," *Applied Physics Letters*, Vol. 77, No. 1, pp. 139–141, June 2000.

[11] C. Rossel, G. I. Meijer, D. Bremaud, and D. Widmer, "Electrical Current Distribution Across a Metal–Insulator–Metal Structure During Bistable Switching," *Journal of Applied Physics*, Vol. 90, No. 6, pp. 2892–2898, September 2001.

[12] X. Duan, Y. Huang, and C. M. Lieber, "Nonvolatile Memory and Programmable Logic from Molecule-Gated Nanowires," *Nano Letters*, Vol. 2, No. 5, pp. 487–490, April 2002.

[13] P. Van Der Sluis, "Non-Volatile Memory Cells Based On $Zn_xCd_{1-x}S$ Ferroelectric Schottky Diodes," *Applied Physics Letters*, Vol. 82, No. 23, pp. 4089–4091, June 2003.

[14] H. Kim, M. P. Sah, and S. P. Adhikari, "Pinched Hysteresis Loops is the Fingerprint of Memristive Devices," *arXiv preprint arXiv:1202.2437*, February 2012.

[15] B. Mouttet, "Memresistors and Non-Memristive Zero-Crossing Hysteresis Curves," *arXiv Condensed Matter — Mesoscale and Nanoscale Physics*, February 2012.

[16] J. H. Comfort, "DRAM Technology: Outlook and Challenges," *Proceedings of the IEEE International Conference on VLSI and CAD*, pp. 182–186, October 1999.

[17] K. Kim, "Technology for Sub-50nm DRAM and NAND Flash Manufacturing," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 323–326, December 2005.

[18] K. Kim and G. Jeong, "Memory Technologies for Sub-40nm Node," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 27–30, December 2007.

[19] Semiconductor Industry Association, *The International Technology Roadmap for Semiconductors*, 2001.

[20] Semiconductor Industry Association, *The International Technology Roadmap for Semiconductors*, 2013.

[21] Ki. Kim, U-In Chung, Y. Park, J Lee, J Yeo, and D. Kim, "Extending the DRAM and FLASH Memory Technologies to 10nm and Beyond," *Proceedings of the SPIE Advanced Lithography Conference*. International Society for Optics and Photonics, pp. 832605–832605, February 2012.

[22] P. Bai *et al.*, "A 65nm Logic Technology Featuring 35nm Gate Lengths, Enhanced Channel Strain, 8 Cu Interconnect Layers, Low-k ILD and 0.57 μm^2 SRAM Cell," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 657–660, December 2004.

[23] E. Karl *et al.*, "A 4.6 GHz 162 Mb SRAM Design in 22 nm Tri-Gate CMOS Technology with Integrated Read and Write Assist Circuitry," *IEEE Journal of Solid-State Circuits*, Vol. 48, No. 1, pp. 150–158, January 2013.

[24] E. Karl *et al.*, "A 4.6 GHz 162Mb SRAM Design in 22nm Tri-Gate CMOS Technology with Integrated Active V MIN-Enhancing Assist Circuitry," *Proceedings of the IEEE International Solid-State Circuits Conference*, pp. 230–232, February 2012.

[25] T. Piazza, H. Jiang, P. Hammarlund, and R. Singhal, "Technology Insight: Intel (r) Next Generation Microarchitecture Code Name Haswell," *Intel Developer Forum*, 2012.

[26] G. Atwood, "Current and Emerging Memory Technology Landscape," *Presented at Flash Memory Summit*, 2011.

[27] H. S. Wong *et al.*, "Metal–Oxide RRAM," *Proceedings of the IEEE*, Vol. 100, No. 6, pp. 1951–1970, June 2012.

[28] K. Kim *et al.*, "32x32 Crossbar Array Resistive Memory Composed of a Stacked Schottky Diode and Unipolar Resistive Memory," *Advanced Functional Materials*, Vol. 23, No. 11, pp. 1440–1449, October 2013.

[29] O. Bass, A. Fish, and D. Naveh, "A Memristor as Multi-Bit Memory: Feasibility Analysis.," *Radioengineering*, Vol. 24, No. 2, June 2015.

[30] X. A. Tran *et al.*, "Self-Rectifying and Forming-Free Unipolar HfO_x Based-High Performance RRAM Built by Fab-Available Materials," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 31.2.1–31.2.4, December 2011.

[31] J. J. Yang *et al.*, "Engineering Nonlinearity into Memristors for Passive Crossbar Applications," *Applied Physics Letters*, Vol. 100, No. 11, pp. 113501–113501–4, March 2012.

[32] Y. Li *et al.*, "Bipolar One Diode-One Resistor Integration for High-Density Resistive Memory Applications," *Nanoscale*, Vol. 5, pp. 4785–4789, April 2013.

[33] J. H. Nickel *et al.*, "Memristor Structures for High Scalability: Non-Linear and Symmetric Devices Utilizing Fabrication Friendly Materials and Processes," *Microelectronic Engineering*, Vol. 103, pp. 66 – 69, March 2013.

[34] J. Woo *et al.*, "Selector-Less RRAM with Non-Linearity of Device for Cross-Point Array Applications," *Microelectronic Engineering*, Vol. 109, pp. 360 – 363, September 2013.

[35] C. W. Smullen *et al.*, "Relaxing Non-Volatility for Fast and Energy-Efficient STT-RAM Caches," *Proceedings of the IEEE/ACM International Symposium on High Performance Computer Architecture*, pp. 50–61, February 2011.

[36] D. Lelmini, F. Nardi, C. Cagli, and A. L. Lacaita, "Size-Dependent Retention Time in NiO-Based Resistive-Switching Memories," *IEEE Electron Device Letters*, Vol. 31, No. 4, pp. 353–355, April 2010.

[37] U. Russo, D. Ielmini, C. Cagli, and A. L. Lacaita, "Self-Accelerated Thermal Dissolution Model for Reset Programming in Unipolar Resistive-Switching Memory (RRAM) Devices," *IEEE Transactions on Electron Devices*, Vol. 56, No. 2, pp. 193–200, February 2009.

[38] A. Chen, S. Haddad, and Y. Wu, "A Temperature-Accelerated Method to Evaluate Data Retention of Resistive Switching Nonvolatile Memory," *IEEE Electron Device Letters*, Vol. 29, No. 1, pp. 38–40, January 2008.

[39] J. Akerman *et al.*, "Demonstrated Reliability of 4-Mb MRAM," *IEEE Transactions on Device and Materials Reliability*, Vol. 4, No. 3, pp. 428–435, September 2004.

[40] Q. Xia, J. J. Yang, W. Wu, X. Li, and R. S. Williams, "Self-Aligned Memristor Cross-point Arrays Fabricated with One Nanoimprint Lithography Step," *Nano Letters*, Vol. 10, No. 8, pp. 2909–2914, June 2010.

[41] M. Hosomi *et al.*, "A Novel Nonvolatile Memory with Spin Torque Transfer Magnetization Switching: Spin-RAM," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 459–462, December 2005.

[42] Q. Xia *et al*, "Memristor—CMOS Hybrid Integrated Circuits for Reconfigurable Logic," *Nano Letters*, Vol. 9, No. 10, pp. 3640–3645, September 2009.

[43] E. Chen *et al.*, "Advances and Future Prospects of Spin-Transfer Torque Random Access Memory," *IEEE Transactions on Magnetics*, Vol. 46, No. 6, pp. 1873–1878, June 2010.

[44] M. Meier *et al.*, "Resistively Switching Pt/Spin-On Glass/Ag Nanocells for Non-Volatile Memories Fabricated with UV Nanoimprint Lithography," *Microelectronic Engineering*, Vol. 86, No. 4, pp. 1060–1062, April 2009.

[45] S. Y. Chou, P. R. Krauss, and P. J. Renstrom, "Nanoimprint Lithography," *Journal of Vacuum Science & Technology B*, Vol. 14, No. 6, pp. 4129–4133, November 1996.

[46] M. A. Zidan *et al.*, "Memristor-Based Memory: The Sneak Paths Problem and Solutions," *Microelectronics Journal*, Vol. 44, No. 2, pp. 176 – 183, February 2013.

[47] S. Shin, K. Kim, and S. Kang, "Data-Dependent Statistical Memory Model for Passive Array of Memristive Devices," *IEEE Transactions on Circuits and Systems II: Express Briefs*, Vol. 57, No. 12, pp. 986 –990, December 2010.

[48] Y. Deng *et al.*, "RRAM Crossbar Array with Cell Selection Device: A Device and Circuit Interaction Study," *IEEE Transactions on Electron Devices*, Vol. 60, No. 2, pp. 719–726, February 2013.

[49] T. Kishi *et al.*, "Lower-Current and Fast Switching of a Perpendicular TMR for High Speed and High Density Spin-Transfer-Torque MRAM," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 1–4, January 2008.

[50] T. Kawahara *et al.*, "2 Mb SPRAM (Spin-Transfer Torque RAM) with Bit-by-Bit Bi-Directional Current Write and Parallelizing-Direction Current Read," *IEEE Journal of Solid-State Circuits*, Vol. 43, No. 1, pp. 109–120, January 2008.

[51] The ITRS Technology Working Groups, *International Technology Roadmap for Semiconductors (ITRS)*, `http://public.itrs.net`.

[52] H. Sato *et al.*, "Perpendicular-Anisotropy CoFeB-MgO Magnetic Tunnel Junctions with a MgO/CoFeB/Ta/CoFeB/MgO Recording Structure," *Applied Physics Letters*, Vol. 101, No. 2, pp. 022414–022414–2, July 2012.

[53] W. Kim *et al.*, "Extended Scalability of Perpendicular STT-MRAM Towards Sub-20nm MTJ Node," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 24.1.1–24.1.4, December 2011.

[54] K. T. Nam *et al.*, "Switching Properties in Spin Transfer Torque MRAM with Sub-50nm MTJ Size," *Proceedings of the Non-Volatile Memory Technology Symposium*, pp. 49–51, November 2006.

[55] S. Ikeda *et al.*, "Tunnel Magnetoresistance of 604% at 300 K by Suppression of Ta Diffusion in CoFeB/MgO/CoFeB Pseudo-Spin-Valves Annealed at High Temperature," *Applied Physics Letters*, Vol. 93, No. 8, pp. 082508–1—082508–3, August 2008.

[56] W. Kim *et al.*, "Fabrication and Magnetoresistance of Tunnel Junctions Using Half-Metallic Fe_3O_4," *Journal of Applied Physics*, Vol. 93, No. 10, pp. 8032–8034, April 2003.

[57] K. Inomata *et al.*, "Large Tunneling Magnetoresistance at Room Temperature Using a Heusler Alloy with the B_2 Structure," *Japanese Journal of Applied Physics*, Vol. 42, No. 4B, pp. L419, April 2003.

[58] P. M Tedrow and R. Meservey, "Spin-Dependent Tunneling into Ferromagnetic Nickel," *Physical Review Letters*, Vol. 26, No. 4, pp. 192, January 1971.

[59] M. Julliere, "Tunneling Between Ferromagnetic Films," *Physics Letters A*, Vol. 54, No. 3, pp. 225–226, September 1975.

[60] M. N. Baibich *et al.*, "Giant Magnetoresistance of (001)Fe/(001)Cr Magnetic Superlattices," *Physical Review Letters*, Vol. 61, pp. 2472–2475, November 1988.

[61] G. Binasch, P. Grünberg, F. Saurenbach, and W. Zinn, "Enhanced Magnetoresistance in Layered Magnetic Structures with Antiferromagnetic Interlayer Exchange," *Physical Review B*, Vol. 39, pp. 4828–4830, March 1989.

[62] A. Fert, "Nobel Lecture: Origin, Development, and Future of Spintronics," *Reviews of Modern Physics*, Vol. 80, pp. 1517–1530, December 2008.

[63] P. A. Grünberg, "Nobel Lecture: From Spin Waves to Giant Magnetoresistance and Beyond," *Reviews of Modern Physics*, Vol. 80, pp. 1531–1540, December 2008.

[64] S. S. P. Parkin *et al.*, "Giant Tunnelling Magnetoresistance at Room Temperature with MgO (100) Tunnel Barriers," *Nature Materials*, Vol. 3, No. 12, pp. 862–867, December 2004.

[65] S. Yuasa, A. Fukushima, H. Kubota, Y. Suzuki, and K. Ando, "Giant Tunneling Magnetoresistance Up to 410% at Room Temperature in Fully Epitaxial Co/MgO/Co Magnetic Tunnel Junctions with BCC Co (001) Electrodes," *Applied Physics Letters*, Vol. 89, No. 4, pp. 042505–042505, July 2006.

[66] W. J Gallagher and S. S. P. Parkin, "Development of the Magnetic Tunnel Junction MRAM at IBM: From First Junctions to a 16-Mb MRAM Demonstrator Chip," *IBM Journal of Research and Development*, Vol. 50, No. 1, pp. 5–23, January 2006.

[67] S. Tehrani *et al.*, "Progress and Outlook for MRAM Technology," *IEEE Transactions on Magnetics*, Vol. 35, No. 5, pp. 2814–2819, Sepember 1999.

[68] R. Desikan *et al.*, "On-Chip MRAM as a High-Bandwidth, Low-Latency Replacement for DRAM Physical Memories," Technical Report, Department of Computer Science, University of Texas at Austin, September 2002.

[69] D. D Tang and Y. Lee, *Magnetic Memory: Fundamentals and Technology*, Cambridge University Press, 2010.

[70] N. Sakimura *et al.*, "A 512kb Cross-Point Cell MRAM," *Proceedings of the IEEE International Solid State Circuits Conference*, pp. 278–279, February 2003.

[71] Y. Asao *et al.*, "Design and Process Integration for High-Density, High-Speed, and Low-Power $6F^2$ Cross Point MRAM Cell," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 571–574, December 2004.

[72] M. K Bhattacharyya and J. A Brug, "MRAM Device Using Magnetic Field Bias to Suppress Inadvertent Switching of Half-Selected Memory Cells," U. S. Patent No. 6,097,626, August 1, 2000.

[73] H. Fujiwara, S. Y. Wang, and M. Sun, "Critical-Field Curves for Switching Toggle Mode Magnetoresistance Random Access Memory Devices," *Journal of Applied Physics*, Vol. 97, No. 10, pp. 10P507–1—10P507–5, May 2005.

[74] W. J. Gallagher *et al.*, "Recent Advances in MRAM Technology," *Proceedings of the IEEE VLSI-TSA International Symposium on VLSI Technology*, pp. 72–73, June 2005.

[75] D. C. Worledge, "Spin Flop Switching for Magnetic Random Access Memory," *Applied Physics Letters*, Vol. 84, No. 22, pp. 4559–4561, May 2004.

[76] B. N Engel *et al.*, "A 4-Mb Toggle MRAM Based on a Novel Bit and Switching Method," *IEEE Transactions on Magnetics*, Vol. 41, No. 1, pp. 132–136, January 2005.

[77] S. Wolf *et al.*, "The Promise of Nanomagnetics and Spintronics for Future Logic and Universal Memory," *Proceedings of the IEEE*, Vol. 98, No. 12, pp. 2155–2168, September 2010.

[78] G. Sun, X. Dong, Y. Xie, J. Li, and Y. Chen, "A Novel Architecture of the 3D Stacked MRAM L2 Cache for CMPs," *Proceedings of the IEEE International Symposium on High Performance Computer Architecture*, pp. 239–249, February 2009.

[79] E. Y. Tsymbal, O. N. Mryasov, and P. R. LeClair, "Spin-Dependent Tunnelling in Magnetic Tunnel Junctions," *Journal of Physics: Condensed Matter*, Vol. 15, No. 4, pp. R109, February 2003.

[80] J. C. Slonczewski, "Conductance and Exchange Coupling of Two Ferromagnets Separated by a Tunneling Barrier," *Physical Review B*, Vol. 39, pp. 6995–7002, April 1989.

[81] J. A. Kim, "Spin-Torque Oscillators," *Solid State Physics*, Vol. 63, pp. 217, March 2012.

[82] D. Houssameddine *et al.*, "Spin-Torque Oscillator Using a Perpendicular Polarizer and a Planar Free Layer," *Nature Materials*, Vol. 6, No. 6, pp. 447–453, April 2007.

[83] S. Bonetti *et al.*, "Spin Torque Oscillator Frequency Versus Magnetic Field Angle: The Prospect of Operation Beyond 65 GHz," *Applied Physics Letters*, Vol. 94, No. 10, pp. 102507, March 2009.

[84] J. C Slonczewski, "Conductance and Exchange Coupling of Two Ferromagnets Separated by a Tunneling Barrier," *Physical Review B*, Vol. 39, No. 10, pp. 6995, April 1989.

[85] A. Brataas, A. D. Kent, and H. Ohno, "Current-Induced Torques in Magnetic Materials," *Nature Materials*, Vol. 11, No. 5, pp. 372–381, April 2012.

[86] T. L. Gilbert, "A Phenomenological Theory of Damping in Ferromagnetic Materials," *IEEE Transactions on Magnetics*, Vol. 40, No. 6, pp. 3443–3449, November 2004.

[87] L. Berger, "Emission of Spin Waves by a Magnetic Multilayer Traversed by a Current," *Physical Review B*, Vol. 54, pp. 9353–9358, October 1996.

[88] J. C. Slonczewski, "Current-Driven Excitation of Magnetic Multilayers," *Journal of Magnetism and Magnetic Materials*, Vol. 159, No. 12, pp. L1–L7, June 1996.

[89] A. Aharoni, "Thermal Agitation of Single Domain Particles," *Physical Review*, Vol. 135, No. 2A, pp. A447, July 1964.

[90] J. Z. Sun, "Spin Angular Momentum Transfer in Current-Perpendicular Nanomagnetic Junctions," *IBM Journal of Research and Development*, Vol. 50, No. 1, pp. 81–100, January 2006.

[91] Z. Diao *et al.*, "Spin-Transfer Torque Switching in Magnetic Tunnel Junctions and Spin-Transfer Torque Random Access Memory," *Journal of Physics: Condensed Matter*, Vol. 19, No. 16, pp. 165209, April 2007.

[92] R. H. Koch, J. A. Katine, and J. Z. Sun, "Time-Resolved Reversal of Spin-Transfer Switching in a Nanomagnet," *Physical Review Letters*, Vol. 92, pp. 088302–1—088302–4, February 2004.

[93] J. Fidler and T. Schrefl, "Micromagnetic Modelling - the Current State of the Art," *Journal of Physics D: Applied Physics*, Vol. 33, No. 15, pp. R135, August 2000.

[94] W. Park, I. J. Hwang, T. Kim, K. J. Lee, and Y. K. Kim, "Anomalous Switching in Submicrometer Magnetic Tunnel Junction Arrays Rrising from Magnetic Vortex and Domain Wall Pinning," *Journal of Applied Physics*, Vol. 96, No. 3, pp. 1748–1750, August 2004.

[95] G. Tatara, H. Kohno, and J. Shibata, "Microscopic Approach to Current-Driven Domain Wall Dynamics," *Physics Reports*, Vol. 468, No. 6, pp. 213 – 301, 2008.

[96] J. Shi, S. Tehrani, T. Zhu, Y. F. Zheng, and J.-G. Zhu, "Magnetization Vortices and Anomalous Switching in Patterned NiFeCo Submicron Arrays," *Applied Physics Letters*, Vol. 74, No. 17, pp. 2525–2527, April 1999.

[97] Y. Guo, T. Min, and P. Wang, "Vortex Magnetic Random Access Memory," U. S. Patent No. 7,072,208, July 4, 2006.

[98] V. S. Pribiag *et al.*, "Magnetic Vortex Oscillator Driven by DC Spin-Polarized Current," *Nature Physics*, Vol. 3, No. 7, pp. 498–503, April 2007.

[99] S. Bohlens *et al.*, "Current Controlled Random-Access Memory Based on Magnetic Vortex Handedness," *Applied Physics Letters*, Vol. 93, No. 14, pp. 142508, October 2008.

[100] K. J. Lee, W. Park, and T. Kim, "Kink-Free Design of Submicrometer MRAM Cell," *IEEE Transactions on Magnetics*, Vol. 39, No. 5, pp. 2842–2844, September 2003.

[101] T. Min *et al.*, "Study of Intermediate Magnetization States in Deep Submicrometer MRAM Cells," *IEEE Transactions on Magnetics*, Vol. 41, No. 10, pp. 2664–2666, October 2005.

[102] J. Li *et al.*, "Modeling of Failure Probability and Statistical Design of Spin-Torque Transfer Magnetic Random Access Memory (STT-MRAM) Array for Yield Enhancement," *Proceedings of the ACM/IEEE Design Automation Conference*, pp. 278–283, June 2008.

[103] M. El Baraji *et al.*, "Dynamic Compact Model of Thermally Assisted Switching Magnetic Tunnel Junctions," *Journal of Applied Physics*, Vol. 106, No. 12, pp. 123906–1—123906–6, December 2009.

[104] D. Wang *et al.*, "70% TMR at Room Temperature for SDT Sandwich Junctions with CoFeB as Free and Reference Layers," *IEEE Transactions on Magnetics*, Vol. 40, No. 4, pp. 2269–2271, July 2004.

[105] Y. Wu, S. Yu, X. Guan, and H. P. Wong, "Recent Progress of Resistive Switching Random Access Memory (RRAM)," *Proceedings of the IEEE International Silicon Nanoelectronics Workshop*, pp. 1–4, June 2012.

[106] J. J. Yang *et al.*, "High Switching Endurance in TaO_x Memristive Devices," *Applied Physics Letters*, Vol. 97, No. 23, pp. 232102–1—232102–3, December 2010.

[107] J. F. Gibbons and W. E. Beadle, "Switching Properties of Thin NiO Films," *Solid-State Electronics*, Vol. 7, No. 11, pp. 785 – 790, November 1964.

[108] J. G. Simmons and R. R. Verderber, "New Conduction and Reversible Memory Phenomena in Thin Insulating Films," *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, Vol. 301, No. 1464, pp. 77–102, October 1967.

[109] F. Argall, "Switching Phenomena in Titanium Oxide Thin Films," *Solid-State Electronics*, Vol. 11, No. 5, pp. 535 – 541, May 1968.

[110] G. Dearnaley, D. V. Morgan, and A. M. Stoneham, "A Model for Filament Growth and Switching in Amorphous Oxide Films," *Journal of Non-Crystalline Solids*, Vol. 4, No. 1, pp. 593 – 612, April 1970.

[111] W. W. Zhuang *et al.*, "Novel Colossal Magnetoresistive Thin Film Nonvolatile Resistance Random Access Memory (RRAM)," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 193–196, December 2002.

[112] I. Valov *et al.*, "Nanobatteries in Redox-Based Resistive Switches Require Extension of Memristor Theory," *Nature Communications*, Vol. 4, pp. 1771, April 2013.

[113] H. Jung *et al.*, "Electrical Characteristics of an Ultrathin (1.6 nm) TaO_xN_y Gate Dielectric," *Applied Physics Letters*, Vol. 76, No. 24, pp. 3630–3631, April 2000.

[114] S. A. Campbell *et al.*, "MOSFET Transistors Fabricated with High Permitivity TiO_2 Dielectrics," *IEEE Transactions on Electron Devices*, Vol. 44, No. 1, pp. 104–109, January 1997.

[115] L. Kang *et al.*, "Electrical Characteristics of Highly Reliable Ultrathin Hafnium Oxide Gate Dielectric," *IEEE Electron Device Letters*, Vol. 21, No. 4, pp. 181–183, April 2000.

[116] J. H. Lee *et al.*, "Mass Production Worthy $HfO_2/Al_2/O_3/$ Laminate Capacitor Technology Using Hf Liquid Precursor for Sub-100 nm DRAMs," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 221–224, December 2002.

[117] R. B. Van Dover *et al.*, "Advanced Dielectrics for Gate Oxide, DRAM and RF Capacitors," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 823–826, December 1998.

[118] I. G. Baek *et al.*, "Highly Scalable Nonvolatile Resistive Memory Using Simple Binary Oxide Driven by Asymmetric Unipolar Voltage Pulses," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 587–590, December 2004.

[119] F. Gomez-Marlasca, N. Ghenzi, M. J. Rozenberg, and P. Levy, "Understanding Electroforming in Bipolar Resistive Switching Oxides," *Applied Physics Letters*, Vol. 98, No. 4, pp. 042901, January 2011.

[120] J. Y. Son and Y. Shin, "Direct Observation of Conducting Filaments on Resistive Switching of NiO Thin Films," *Applied Physics Letters*, Vol. 92, No. 22, pp. 222106–222106, June 2008.

[121] U. Russo *et al.*, "Conductive-Filament Switching Analysis and Self-Accelerated Thermal Dissolution Model for Reset in NiO-based RRAM," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 775–778, December 2007.

[122] D. Ielmini, "Modeling the Universal Set/Reset Characteristics of Bipolar RRAM by Field-and Temperature-Driven Filament Growth," *IEEE Transactions on Electron Devices*, Vol. 58, No. 12, pp. 4309–4317, December 2011.

[123] X. Cao *et al.*, "Forming-Free Colossal Resistive Switching Effect in Rare-Earth-Oxide Gd_2O_3 Films for Memristor Applications," *Journal of Applied Physics*, Vol. 106, No. 7, pp. 073723–1—073723–5, October 2009.

[124] Z. Fang *et al.*, "$HfO_x/TiO_x/HfO_x/TiO_x$ Multilayer-Based Forming-Free RRAM Devices With Excellent Uniformity," *IEEE Electron Device Letters*, Vol. 32, No. 4, pp. 566–568, April 2011.

[125] B. Govoreanu *et al.*, "$10x10\ nm^2\ HfO_x$ Crossbar Resistive RAM with Excellent Performance Reliability and Low-Energy Operation," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 31.6.1–31.6.4, December 2011.

[126] N. Raghavan, A. Fantini, R. Degraeve, P. J. Roussel, L. Goux, B. Govoreanu, D. J. Wouters, G. Groeseneken, and M. Jurczak, "Statistical Insight into Controlled Forming and Forming Free Stacks for HfO_x {RRAM}," *Microelectronic Engineering*, Vol. 109, No. 1, pp. 177 – 181, September 2013.

[127] G. Bersuker *et al.*, "Metal Oxide Resistive Memory Switching Mechanism Based on Conductive Filament Properties," *Journal of Applied Physics*, Vol. 110, No. 12, pp. 124518–1—124518–12, December 2011.

[128] K. M. Kim, D. S. Jeong, and C. S. Hwang, "Nanofilamentary Resistive Switching in Binary Oxide System; A Review on the Present Status and Outlook," *Nanotechnology*, Vol. 22, No. 25, pp. 254002–1–254002–17, June 2011.

[129] I. Valov, R. Waser, J. R. Jameson, and M. N. Kozicki, "Electrochemical Metallization Memories - Fundamentals, Applications, Prospects," *Nanotechnology*, Vol. 22, No. 25, pp. 254003, June 2011.

[130] Y. Sato *et al.*, "Consideration of Switching Mechanism of Binary Metal Oxide Resistive Junctions Using a Thermal Reaction Model," *Applied Physics Letters*, Vol. 90, No. 3, pp. 033503, January 2007.

[131] D. Strukov, G. Snider, D. Stewart, and R. Williams, "The Missing Memristor Found," *Nature*, Vol. 453, No. 7191, pp. 80–83, May 2008.

[132] J. J. Yang, M. D. Pickett, X. Li, D. A. Ohlberg, D. R. Stewart, and R. S. Williams, "Memristive Switching Mechanism for Metal/Oxide/Metal Nanodevices," *Nature Nanotechnology*, Vol. 3, No. 7, pp. 429–433, January 2008.

[133] R. R. Das, P. Bhattacharya, W. Perez, R. S. Katiyar, and A. S. Bhalla, "Leakage Current Characteristics of Laser-Ablated $SrBi_2Nb_2O_9$ Thin Films," *Applied Physics Letters*, Vol. 81, No. 5, pp. 880–882, July 2002.

[134] C. H. Cheng, C. Y. Tsai, A. Chin, and F. S. Yeh, "High Performance Ultra-Low Energy RRAM with Good Retention and Endurance," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 19–4, December 2010.

[135] R. Waser and M. Aono, "Nanoionics-Based Resistive Switching Memories," *Nature Materials*, Vol. 6, No. 11, pp. 833–840, June 2007.

[136] H. P. Wolf and D. R. Young, "Charge Storage Structure for Nonvolatile Memories," U. S. Patent No. 4,717,943, January 5, 1988.

[137] J. Strachan *et al.*, "State Dynamics and Modeling of Tantalum Oxide Memristors," *IEEE Transactions on Electron Devices*, Vol. 60, No. 7, pp. 2194–2202, July 2013.

[138] J. J. Yang, F. Miao, M. D. Pickett, D. A. A. Ohlberg, D. R. Stewart, C. N. Lau, and R. S. Williams, "The Mechanism of Electroforming of Metal Oxide Memristive Switches," *Nanotechnology*, Vol. 20, No. 21, pp. 215201, May 2009.

[139] J. Von Neumann, "First Draft of a Report on the EDVAC," *IEEE Annals of the History of Computing*, Vol. 15, No. 4, pp. 27–75, 1993.

[140] D. A Patterson, "Latency Lags Bandwith," *Communications of the ACM*, Vol. 47, No. 10, pp. 71–75, October 2004.

[141] Intel Corporation, *Intel386 SL Microprocessor SuperSet Data Book*, 1992.

[142] D. Burger, J. R. Goodman, and A. Kägi, "Memory Bandwidth Limitations of Future Microprocessors," *Proceedings of the IEEE/ACM International Symposium on Computer Architecture*, pp. 78–89, November 1996.

[143] M. D. Hill and M. R. Marty, "Amdahl's Law in the Multicore Era," *Computer*, Vol. 41, No. 7, pp. 33–38, July 2008.

[144] B. Rogers *et al.*, "Scaling the Bandwidth Wall: Challenges in and Avenues for CMP Scaling," *Proceedings of the IEEE/ACM International Symposium on Computer Architecture*, pp. 371–382, November 2009.

[145] J. Held, J. Bautista, and S. Koehl, "From a Few Cores to Many: A Tera-Scale Computing Research Overview," *Intel White Paper*, 2006.

[146] V. Venkatachalam and M. Franz, "Power Reduction Techniques for Microprocessor Systems," *ACM Computer Survey*, Vol. 37, No. 3, pp. 195–237, Sept. 2005.

[147] D. J. Sorin, M. D. Hill, and D. A. Wood, "A Primer on Memory Consistency and Cache Coherence," *Synthesis Lectures on Computer Architecture*, Vol. 6, No. 3, pp. 1–212, November 2011.

[148] C. Kim, D. Burger, and S. W. Keckler, "An Adaptive, Non-Uniform Cache Structure for Wire-Delay Dominated On-Chip Caches," *Proceedings of the IEEE/ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, pp. 211–222, October 2002.

[149] R. Banakar *et al.*, "Scratchpad Memory: Design Alternative for Cache On-Chip Memory in Embedded Systems," *Proceedings of the ACM International Symposium on Hardware/Software Codesign*, pp. 73–78, May 2002.

[150] D. C. Pham *et al.*, "Overview of the Architecture, Circuit Design, and Physical Implementation of a First-Generation Cell Processor," *IEEE Journal of Solid-State Circuits*, Vol. 41, No. 1, pp. 179–196, January 2006.

[151] D. Patterson, "The Top 10 Innovations in the New NVIDIA Fermi Architecture, and the Top 3 Next Challenges," *NVIDIA Whitepaper*, 2009.

[152] U. Kang *et al.*, "8Gb 3D DDR3 DRAM using Through-Silicon-Via Technology," *Proceedings of the IEEE International Solid-State Circuits Conference*, pp. 130–131,131a, February 2009.

[153] K. N. Kim *et al.*, "Highly Manufacturable and High Performance SDR/DDR 4 Gb DRAM," *Proceedings of the IEEE Symposium on VLSI Technology*, pp. 7–8, June 2001.

[154] K. Itoh *et al.*, "A High-Speed 16-kbit n-MOS Random-Access Memory," *IEEE Journal of Solid-State Circuits*, Vol. 11, No. 5, pp. 585–590, October 1976.

[155] R. A. Abbott, W. M. Regitz, and J. A. Karp, "A 4K MOS Dynamic Random-Access Memory," *IEEE Journal of Solid-State Circuits*, Vol. 8, No. 5, pp. 292–298, October 1973.

[156] K. Itoh *et al.*, "A High-Speed 16K-bit NMOS RAM," *Proceedings of the IEEE International Solid-State Circuits Conference*, Vol. 19, pp. 140–141, February 1976.

[157] E. Arai and N. Ieda, "A 64-kbit Dynamic MOS RAM," *IEEE Journal of Solid-State Circuits*, Vol. 13, No. 3, pp. 333–338, June 1978.

[158] T. Mano *et al.*, "A Fault-Tolerant 256K RAM Fabricated with Molybdenum-Polysilicon Technology," *IEEE Journal of Solid-State Circuits*, Vol. 15, No. 5, pp. 865–872, October 1980.

[159] S. Suzuki *et al.*, "A 128K Word x 8 Bit Dynamic RAM," *IEEE Journal of Solid-State Circuits*, Vol. 19, No. 5, pp. 624–627, October 1984.

[160] T. Furuyama *et al.*, "An Experimental 4-Mbit CMOS DRAM," *IEEE Journal of Solid-State Circuits*, Vol. 21, No. 5, pp. 605–611, October 1986.

[161] K. Nakagawa, M. Taguchi, and T. Ema, "Fabrication of 64M DRAM with I-Line Phase-Shift Lithography," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 817–820, December 1990.

[162] G. Kitsukawa *et al.*, "256-Mb DRAM Circuit Technologies for File Applications," *IEEE Journal of Solid-State Circuits*, Vol. 28, No. 11, pp. 1105–1113, November 1993.

[163] M. Horiguchi *et al.*, "An Experimental 220 MHz 1 Gb DRAM," *Proceedings of the IEEE International Solid-State Circuits Conference*, pp. 252–253, February 1995.

[164] H. S. Jeong *et al.*, "Highly Manufacturable 4 Gb DRAM Using 0.11 μm DRAM Technology," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 353–356, December 2000.

[165] R. H. Dennard, "Field-Effect Transistor Memory," U. S. Patent No 3,387,286, June 4, 1968.

[166] G. E. Moore, "Intel: Memories and the Microprocessor," *Daedalus*, Vol. 125, No. 2, pp. 55–80, April 1996.

[167] G. C. Feth, "Memories are Bigger, Faster — and Cheaper," *IEEE Spectrum*, Vol. 10, No. 11, pp. 28–35, November 1973.

[168] C. Hwang, "Nanotechnology Enables a New Memory Growth Model," *Proceedings of the IEEE*, Vol. 91, No. 11, pp. 1765–1771, November 2003.

[169] P. K. Chatterjee, S. Malhi, and W. F. Richardson, "DRAM Cell with Trench Capacitor and Vertical Channel in Substrate," U. S. Patent No. 5,208,657, May 4, 1993.

[170] D. T. Wong *et al.*, "An 11-ns $8K \times 18$ CMOS Static RAM with $0.5 - \mu; m$ Devices," *IEEE Journal of Solid-State Circuits*, Vol. 23, No. 5, pp. 1095–1103, October 1988.

[171] T. N. Blalock. and R. C. Jaeger, "A High-Speed Clamped Bit-Line Current-Mode Sense Amplifier," *IEEE Journal of Solid-State Circuits*, Vol. 26, No. 4, pp. 542 –548, April 1991.

[172] N. Seifert *et al.*, "Radiation-Induced Soft Error Rates of Advanced CMOS Bulk Devices," *Proceedings of the IEEE International Reliability Physics Symposium*, pp. 217–225, March 2006.

[173] R. C. Baumann, "Radiation-Induced Soft Errors in Advanced Semiconductor Technologies," *IEEE Transactions on Device and Materials Reliability*, Vol. 5, No. 3, pp. 305–316, September 2005.

[174] F. X. Ruckerbauer and G. Georgakos, "Soft Error Rates in 65nm SRAMs–Analysis of New Phenomena," *Proceedings of the IEEE International On-Line Testing Symposium*, pp. 203–204, July 2007.

[175] C. E. Blat, E. H. Nicollian, and E. H. Poindexter, "Mechanism of Negative-Bias-Temperature Instability," *Journal of Applied Physics*, Vol. 69, No. 3, pp. 1712–1720, February 1991.

[176] V. Reddy *et al.*, "Impact of Negative Bias Temperature Instability on Digital Circuit Reliability," *Microelectronics Reliability*, Vol. 45, No. 1, pp. 31–38, January 2005.

[177] Z. Hu *et al.*, "Microarchitectural Techniques for Power Gating of Execution Units," *Proceedings of the ACM International Symposium on Low Power Electronics and Design*, pp. 32–37, July 2004.

[178] Y. Wang, S. Roy, and N. Ranganathan, "Run-Time Power-Gating in Caches of GPUs for Leakage Energy Savings," *Proceedings of the ACM International Conference on Design, Automation and Test in Europe*, pp. 300–303, March 2012.

[179] K. Flautner, N. S. Kim, S. Martin, D. Blaauw, and T. Mudge, "Drowsy Caches: Simple Techniques for Reducing Leakage Power," *Proceedings of the IEEE/ACM International Symposium on Computer Architecture*, pp. 148–157, 2002.

[180] M. F. Chang *et al.*, "A Differential Data-Aware Power-Supplied (D^2AP) 8T SRAM Cell With Expanded Write/Read Stabilities for Lower VDDmin Applications," *IEEE Journal of Solid-State Circuits*, Vol. 45, No. 6, pp. 1234–1245, June 2010.

[181] L. Chang *et al.*, "An 8T-SRAM for Variability Tolerance and Low-Voltage Operation in High-Performance Caches," *IEEE Journal of Solid-State Circuits*, Vol. 43, No. 4, pp. 956–963, April 2008.

[182] T. Song, S. Kim, K. Lim, and J. Laskar, "Fully-Gated Ground 10T-SRAM Bitcell in 45 nm SOI Technology," *Electronics Letters*, Vol. 46, No. 7, pp. 515–516, April 2010.

[183] M. F. Chang, Y. C. Chen, and C. F. Chen, "A 0.45-V 300-MHz 10T Flowthrough SRAM With Expanded Write/Read Stability and Speed-Area-Wise Array for Sub-0.5-V Chips," *IEEE Transactions on Circuits and Systems II: Express Briefs*, Vol. 57, No. 12, pp. 980–985, December 2010.

[184] C. H. Lo and S. Y. Huang, "P-P-N Based 10T SRAM Cell for Low-Leakage and Resilient Subthreshold Operation," *IEEE Journal of Solid-State Circuits*, Vol. 46, No. 3, pp. 695–704, March 2011.

[185] D. Kim *et al.*, "A 1.85fW/bit Ultra Low Leakage 10T SRAM with Speed Compensation Scheme," *Proceedings of the IEEE International Symposium on Circuits and Systems*, pp. 69–72, May 2011.

[186] *Toshiba NAND Flash Applications Design Guide*, April 2003.

[187] R. Bez, E. Camerlenghi, A. Modelli, and A. Visconti, "Introduction to Flash Memory," *Proceedings of the IEEE*, Vol. 91, No. 4, pp. 489–502, April 2003.

[188] M. Lenzlinger and E. H. Snow, "Fowler-Nordheim Tunneling into Thermally Grown SiO_2," *Journal of Applied Physics*, Vol. 40, pp. 278–283, Jan. 1969.

[189] Y. Koh, "NAND Flash Scaling Beyond 20nm," *Proceedings of the IEEE International Memory Workshop*, pp. 1–3, May 2009.

[190] D. Kang, K. Lee, S. Seo, S. Kim, J. Lee, D. Bae, D. H. Li, Y. Hwang, and H Shin, "Generation Dependence of Retention Characteristics in Extremely Scaled NAND Flash Memory," *IEEE Electron Device Letters*, Vol. 34, No. 9, pp. 1139–1141, September 2013.

[191] H. Shim, M. Cho, K. Ahn, G. Bae, and S. Park, "Novel Integration Technologies for Improving Reliability in NAND Flash Memory," *Proceedings of the IEEE International Symposium on Circuits and Systems*, pp. 424–427, May 2012.

[192] J. Lee, S. Hur, and J. Choi, "Effects of Floating-Gate Interference on NAND Flash Memory Cell Operation," *IEEE Electron Device Letters*, Vol. 23, No. 5, pp. 264–266, May 2002.

[193] K. Fukuda, Y. Shimizu, K. Amemiya, M. Kamoshida, and C. Hu, "Random Telegraph Noise in Flash Memories - Model and Technology Scaling," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 169–172, December 2007.

[194] A. Ghetti, C.M. Compagnoni, F. Biancardi, A.L. Lacaita, S. Beltrami, L. Chiavarone, A.S. Spinelli, and A. Visconti, "Scaling Trends for Random Telegraph Noise in Deca-Nanometer Flash Memories," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 1–4, December 2008.

[195] L. M. Grupp, J. D. Davis, and S. Swanson, "The Bleak Future of NAND Flash Memory," *Proceedings of the 10th USENIX Conference on File and Storage Technologies*, pp. 2–2, 2012.

[196] S. M. Jung *et al.*, "Three Dimensionally Stacked NAND Flash Memory Technology Using Stacking Single Crystal Si Layers on ILD and TANOS Structure for Beyond 30nm Node," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 1–4, December 2006.

[197] G. Van der Plas *et al.*, "Design Issues and Considerations for Low-Cost 3-D TSV IC Technology," *IEEE Journal of Solid-State Circuits*, Vol. 46, No. 1, pp. 293–307, January 2011.

[198] A. Rahman and R. Reif, "Thermal Analysis of Three-Dimensional (3-D) Integrated Circuits (ICs)," *Proceedings of the IEEE International Interconnect Technology Conference*, pp. 157–159, June 2001.

[199] B. Jacob, S. S. Ng, and D. Wang, *Memory Systems: Cache, DRAM, Disk*, Morgan Kaufmann, 2010.

[200] B. S. Amrutur and M. A. Horowitz, "Speed and Power Scaling of SRAM's," *IEEE Journal of Solid-State Circuits*, Vol. 35, No. 2, pp. 175–185, February 2000.

[201] Hewlett-Packard Western Research Laboratory, *CACTI 3.0: An Integrated Cache, Timing, Power, and Area Model*, 2001.

[202] B. S. Cherkauer and E. G. Friedman, "A Unified Design Methodology for CMOS Tapered Buffers," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, Vol. 3, No. 1, pp. 99–111, March 1995.

[203] B. S. Amrutur and M. A. Horowitz, "Fast Low-Power Decoders for RAMs," *IEEE Journal of Solid-State Circuits*, Vol. 36, No. 10, pp. 1506–1515, October 2001.

[204] K. Pagiamtzis and A. Sheikholeslami, "Content-Addressable Memory (CAM) Circuits and Architectures: A Tutorial and Survey," *IEEE Journal of Solid-State Circuits*, Vol. 41, No. 3, pp. 712–727, March 2006.

[205] M. Lee *et al.*, "A Fast, High-Endurance and Scalable Non-Volatile Memory Device Made from Asymmetric Ta_2O_{5-x}/TaO_{2-x} Bilayer Structures," *Nature Materials*, Vol. 10, No. 8, pp. 625–630, July 2011.

[206] J. Li and J. F. Martinez, "Power Performance Implications of Thread-Level Parallelism on Chip Multiprocessors," *Proceedings of the IEEE International Symposium on Performance Analysis of Systems and Software*, pp. 124–134, March 2005.

[207] D. M. Tullsen, S. J. Eggers, and H. M. Levy, "Simultaneous Multithreading: Maximizing On-chip Parallelism," *Proceedings of the IEEE/ACM International Symposium on Computer Architecture*, pp. 392–403, June 1995.

[208] "Intel Ivy Bridge Specifications," http://ark.intel.com/.

[209] S. Kvatinsky, Y. Nacson, Y. Etsion, E. Friedman, A. Kolodny, and U. Weiser, "Memristor-Based Multithreading," *IEEE Computer Architecture Letters*, Vol. 13, No. 1, pp. 41–44, January 2014.

[210] S. Kvatinsky, E. G. Friedman, A. Kolodny, and U. C. Weiser, "TEAM: ThrEshold Adaptive Memristor Model," *IEEE Transactions on Circuits and Systems I: Regular Papers*, Vol. 60, No. 1, pp. 211–221, January 2013.

[211] *FreePDK45 User Guide*, April 2011, http://www.eda.ncsu.edu/wiki/FreePDK45.

[212] NIMO Group, "Predictive Technology Model (PTM)," Available online: http://www.eas.asu.edu/~ptm, Arizona State University.

[213] S. Kvatinsky, E. Friedman, A. Kolodny, and U. Weiser, "The Desired Memristor for Circuit Designers," *IEEE Circuits and Systems Magazine*, Vol. 13, No. 2, pp. 17–22, Second Quarter 2013.

[214] S. Kvatinsky, Y. H. Nacson, R. Patel, Y. Etsion, E. G. Friedman, A. Kolodny, and U. C. Weiser, "Multithreading with Emerging Technologies — Dense Integration of Memory within Logic," (in submission).

[215] "The gem5 Simulator System," May 2012, http://www.m5sim.org/.

[216] S. Li, J. H. Ahn, R. D. Strong, J. B. Brockman, D. M. Tullsen, and N. P. Jouppi, "McPAT: An Integrated Power, Area, and Timing Modeling Framework for Multicore and Manycore Architectures," *Proceedings of the IEEE/ACM International Symposium on Computer Architecture*, pp. 469–480, June 2009.

[217] M. Farrens and A. R. Pleszkun, "Strategies for Achieving Improved Processor Throughput," *Proceedings of the IEEE/ACM International Symposium on Computer Architecture*, pp. 362–369, June 1991.

[218] L. Chang *et al.*, "An 8T-SRAM for Variability Tolerance and Low-Voltage Operation in High-Performance Caches," *IEEE Journal of Solid-State Circuits*, Vol. 43, No. 4, pp. 956–963, April 2008.

[219] W. Zhao, C. Chappert, and P. Mazoyer, "Spin Transfer Torque (STT) MRAM-Based Runtime Reconfiguration FPGA Circuit," *ACM Transactions on Embedded Computing Systems*, Vol. 9, No. 2, pp. 14:1–14:16, October 2009.

[220] S. Chung *et al.*, "Fully Integrated 54nm STT-RAM with the Smallest Bit Cell Dimension for High Density Memory Application," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 12.7.1–12.7.4, December 2010.

[221] J. M. Slaughter *et al.*, "High Density ST-MRAM Technology," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 29.3.1–29.3.4, December 2012.

[222] C. W. Smullen *et al.*, "Relaxing Non-Volatility for Fast and Energy-Efficient STT-RAM Caches," *Proceedings of the IEEE International Symposium on High Performance Computer Architecture*, pp. 50–61, June 2011.

[223] R. P. Robertazzi, D. C. Worledge, and J. Nowak, "Investigations of Half and Full Select Disturb Rates in a Toggle Magnetic Random Access Memory," *Applied Physics Letters*, Vol. 92, No. 19, pp. 192510—192510–3, May 2008.

[224] W. C. Jeong *et al.*, "Highly Scalable MRAM Using Field Assisted Current Induced Switching," *Proceedings of the IEEE Symposium on VLSI Technology*, pp. 184–185, June 2005.

[225] T. Andre *et al.*, "Structures and Methods for a Field-Reset Spin-Torque MRAM," U. S. Patent 8,228,715, July 24, 2012.

[226] C. K. A. Mewes and T. Mewes, "M^3 Micromagnetic Simulator," `www.bama.ua.edu/~tmewes/Mcube/Mcube.shtml`.

[227] W. Zhao and Y. Cao, "New Generation of Predictive Technology Model for Sub-45 nm Early Design Exploration," *IEEE Transactions on Electron Devices*, Vol. 53, No. 11, pp. 2816–2823, November 2006.

[228] H. Zhao *et al*, "Low Writing Energy and Sub-Nanosecond Spin Torque Transfer Switching of In-Plane Magnetic Tunnel Junction for Spin Torque Transfer Random Access Memory," *Journal of Applied Physics*, Vol. 109, No. 7, pp. 07C720, April 2011.

[229] S. Ikeda *et al*, "A Perpendicular-Anisotropy CoFeB MgO Magnetic Tunnel Junction," *Nature Materials*, Vol. 9, No. 9, pp. 721–724, September 2010.

[230] R. Patel, E. Ipek, and E. G. Friedman, "STT-MRAM Memory Cells with Enhanced On/Off Ratio," *Proceedings of the IEEE International System-on-Chip Conference*, pp. 148–152, September 2012.

[231] R. Patel, E. Ipek, and E. G. Friedman, "2T-1R STT-MRAM Memory Cells for Enhanced On/Off Current Ratio," *Microelectronics Journal*, Vol. 45, No. 2, pp. 133 – 143, February 2014.

[232] F. T. Ulaby, E. Michielssen, and U. Ravaioli, *Fundamentals of Applied Electromagnetics*, Prentice Hall, 2010.

[233] X. Guo, E. Ipek, and T. Soyata, "Resistive Computation: Avoiding the Power Wall with Low-leakage, STT-MRAM Based Computing," *Proceedings of the*

IEEE/ACM International Symposium on Computer Architecture, pp. 371–382, June 2010.

[234] B. C. Lee, E. Ipek, O. Mutlu, and D. Burger, "Architecting Phase Change Memory as a Scalable DRAM Alternative," *Proceedings of the IEEE/ACM International Symposium on Computer Architecture*, pp. 2–13, June 2009.

[235] J. Renau *et al.*, "SESC: SuperESCalar Simulator," January 2005, `http://sesc.sourceforge.net`.

[236] N. Muralimanohar, R. Balasubramonian, and N. Jouppi, "Optimizing NUCA Organizations and Wiring Alternatives for Large Caches With CACTI 6.0," *Proceedings of the IEEE/ACM International Symposium on Microarchitecture*, pp. 3–14, December 2007.

[237] X. Dong, C. Xu, Y. Xie, and N. P. Jouppi, "NVSim: A Circuit-Level Performance, Energy, and Area Model for Emerging Nonvolatile Memory," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 31, No. 7, pp. 994–1007, July 2012.

[238] `http://www.spec.org/omp2001/`.

[239] S. C. Woo, M. Ohara, E. Torrie, J. P. Singh, and A. Gupta, "The SPLASH-2 Programs: Characterization and Methodological Considerations," *Proceedings of the IEEE/ACM International Symposium on Computer Architecture*, pp. 24–36, June 1995.

[240] K. Nii, Y. Tsukamoto, T. Yoshizawa, S. Imaolka, and H. Makino, "A 90 nm Dual-Port SRAM with 2.04 μm^2 8T-Thin Cell Using Dynamically-Controlled Column Bias Scheme," *Proceedings of the IEEE Solid-State Circuits Conference*, Vol. 1, pp. 508–543, February 2004.

[241] R. Kumar and G. Hinton, "A Family of 45nm IA Processors," *Proceedings of the IEEE International Solid-State Circuits Conference*, pp. 58–59, February 2009.

[242] S. Cosemans, W. Dehaene, and F. Catthoor, "A Low-Power Embedded SRAM for Wireless Applications," *IEEE Journal of Solid-State Circuits*, Vol. 42, No. 7, pp. 1607–1617, July 2007.

[243] A. Hajimiri and R. Heald, "Design Issues in Cross-Coupled Inverter Sense Amplifiers," *Proceedings of the IEEE International Symposium on Circuits and Systems*, Vol. 2, pp. 149–152, May 1998.

[244] A. Said, "Introduction to Arithmetic Coding - Theory and Practice," Technical Report HPL-2004-76, HP Laboratories, April 2004.

[245] M. D. Pickett *et al.*, "Switching Dynamics in Titanium Dioxide Memristive Devices," *Journal of Applied Physics*, Vol. 106, No. 7, pp. 074508, October 2009.

[246] D. Johns and K. W. Martin, *Analog Integrated Circuit Design*, John Wiley & Sons, 1997.

[247] A. G. Radwan, M. A. Zidan, and K. N. Salama, "On the Mathematical Modeling of Memristors," *Proceedings of the IEEE International Conference on Microelectronics*, pp. 284–287, December 2010.

[248] T. Liu *et al.*, "A $130.7 - mm^2$ 2-Layer 32-Gb ReRAM Memory Device in 24-nm Technology," *IEEE Journal of Solid-State Circuits*, Vol. 49, No. 1, pp. 140–153, January 2014.

[249] D. Niu *et al.*, "Design of Cross-Point Metal-Oxide ReRAM Emphasizing Reliability and Cost," *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pp. 17–23, November 2013.

[250] M. Zangeneh and A. Joshi, "Design and Optimization of Nonvolatile Multibit 1T1R Resistive RAM," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, Vol. 22, No. 8, pp. 1815–1828, August 2014.

[251] J. G. Zhu, "Magnetoresistive Random Access Memory: The Path to Competitiveness and Scalability," *Proceedings of the IEEE*, Vol. 96, No. 11, pp. 1786–1798, November 2008.

[252] Y. B. Kim *et al.*, "Bi-Layered RRAM with Unlimited Endurance and Extremely Uniform Switching," *Proceedings of the International Symposium on VLSI Technology and Circuits*, pp. 52–53, June 2011.

[253] A. Torrezan, J. P. Strachan, G. Medeiros-Ribeiro, and R. S. Williams, "Sub-Nanosecond Switching of a Tantalum Oxide Memristor," *Nanotechnology*, Vol. 22, No. 48, pp. 485203, 2011.

[254] C. Chappert, A. Fert, and F. N. Van Dau, "The Emergence of Spin Electronics in Data Storage," *Nature Materials*, Vol. 6, No. 11, pp. 813–823, November 2007.

[255] R. Patel, S. Kvatinsky, E. G. Friedman, and A. Kolodny, "Multistate Register Based on Resistive RAM," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 2014, in press.

[256] R. Patel, X. Guo, Q. Guo, E. Ipek, and E. G. Friedman, "Reducing Switching Latency and Energy in STT-MRAM Caches With Field-Assisted Writing," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 2015, in press.

[257] X. Cao, W. Zhu, R. Lamberton, and K. Gao, "Field Assisted Switching of a Magnetic Memory Element," US Patent 8,422,277, April 16, 2013.

[258] X. Wang *et al.*, "Magnetic Field Assisted STRAM Cells," US Patent 8,400,825, March 19, 2013.

[259] B. N. Engel *et. al*, "A 4-Mb Toggle MRAM Based On a Novel Bit and Switching Method," *IEEE Transactions on Magnetics*, Vol. 41, No. 1, pp. 132–136, January 2005.

[260] T. M. Maffitt, J. K. DeBrosse, J. A. Gabric, E. T. Gow M. C. Lamorey, J. S. Parenteau, D. R. Willmott, M. A. Wood, and W. J. Gallagher, "Design Considerations for MRAM," *IBM Journal of Research and Development*, Vol. 50, No. 1, pp. 25–39, January 2006.

[261] Y. Ding, "Method and System for Using a Pulsed Field to Assist Spin Transfer Induced Switching of Magnetic Memory Elements," US Patent 7,502,249, March 9, 2009.

[262] C. Smullen *et al.*, "The STeTSiMS STT-RAM Simulation and Modeling System," *Proceedings of the IEEE International Conference on Computer-Aided Design*, pp. 318–325, November 2011.

[263] J. Li *et al.*, "Design Paradigm for Robust Spin-Torque Transfer Magnetic RAM (STT MRAM) from Circuit/Architecture Perspective," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, Vol. 18, No. 12, pp. 1710–1723, December 2010.

[264] P. Z. Peebles, *Probability, Random Variables, and Random Signal Principles Third Edition*, McGraw-Hill New York, 1993.

[265] S. Ikeda *et al.*, "A Perpendicular-Anisotropy CoFeB–MgO Magnetic Tunnel Junction," *Nature Materials*, Vol. 9, No. 9, pp. 721–724, September 2010.

[266] H. Sato, M. Yamanouchi, K. Miura, S. Ikeda, R. Koizumi, F. Matsukura, and H. Ohno, "CoFeB Thickness Dependence of Thermal Stability Factor in CoFeB/MgO Perpendicular Magnetic Tunnel Junctions," *IEEE Magnetics Letters*, Vol. 3, pp. 3000204–3000204, April 2012.

[267] P.R. Kinget, "Device Mismatch and Tradeoffs in the Design of Analog Circuits," *IEEE Journal of Solid-State Circuits*, Vol. 40, No. 6, pp. 1212–1224, June 2005.

[268] P.G. Drennan and C.C. McAndrew, "Understanding MOSFET Mismatch for Analog Design," *IEEE Journal of Solid-State Circuits*, Vol. 38, No. 3, pp. 450–456, March 2003.

[269] J. Park *et al.*, "Investigation of State Stability of Low-Resistance State in Resistive Memory," *IEEE Electron Device Letters*, Vol. 31, No. 5, pp. 485–487, May 2010.

[270] K. Jung *et al.*, "Resistance Switching Characteristics in Li-Doped NiO," *Journal of Applied Physics*, Vol. 103, No. 3, pp. 034504–1—034504–4, 2008.

[271] E. Yalon, A. Gavrilov, S. Cohen, D. Mistele, B. Meyler, J. Salzman, and D. Ritter, "Resistive Switching in HfO_2 Probed by a Metal—Insulator—Semiconductor Bipolar Transistor," *IEEE Electron Device Letters*, Vol. 33, No. 1, pp. 11–13, January 2012.

[272] M. K. Hota, C. Mukherjee, T. Das, and C. K. Maiti, "Bipolar Resistive Switching in $Al/HfO_2/In_{0.53}Ga_{0.47}As$ MIS Structures," *ECS Journal of Solid State Science and Technology*, Vol. 1, No. 6, pp. N149–N152, October 2012.