A STATISTICAL STT-RAM DESIGN VIEW AND ROBUST DESIGNS AT SCALED TECHNOLOGIES

by

Yaojun Zhang

B.S. Microelectronics, Shanghai Jiaotong University, 2008

M.S. Electrical Engineering, University of Pittsburgh, 2010

Submitted to the Graduate Faculty of

the Swanson School of Engineering in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2017

UNIVERSITY OF PITTSBURGH

SWANSON SCHOOL OF ENGINEERING

This dissertation was presented

by

Yaojun Zhang

It was defended on

November 19, 2016

and approved by

Yiran Chen, Ph.D., Associate Professor, Department of Electrical and Computer Engineering

Hai Li, Ph.D., Associate Professor, Department of Electrical and Computer Engineering

Ching-Chung Li, Ph.D., Professor, Department of Electrical and Computer Engineering

Ervin Sejdic, Ph.D., Assistant Professor, Department of Electrical and Computer Engineering

Mingui Sun, Ph.D., Professor, Department of Neurological Surgery

Dissertation Advisors: Yiran Chen, Ph.D., Associate Professor, Department of Electrical and Computer Engineering,

Co-Advisor, Hai Li, Ph.D., Associate Professor, Department of Electrical and Computer Engineering

A STATISTICAL STT-RAM DESIGN VIEW AND ROBUST DESIGNS AT SCALED TECHNOLOGIES

Yaojun Zhang, PhD

University of Pittsburgh, 2017

Rapidly increased demands for memory in electronic industry and the significant technical scaling challenges of all conventional memory technologies motivated the researches on the next generation memory technology. As one promising candidate, spin-transfer torque random access memory (STT-RAM) features fast access time, high density, non-volatility, and good CMOS process compatibility. In recent years, many researches have been conducted to improve the storage density and enhance the scalability of STT-RAM, such as reducing the write current and switching time of magnetic tunneling junction (MTJ) devices. In parallel with these efforts, the continuous increasing of tunnel magneto-resistance(TMR) ratio of the MTJ inspires the development of multi-level cell (MLC) STT-RAM, which allows multiple data bits be stored in a single memory cell. Two types of MLC STT-RAM cells, namely, parallel MLC and series MLC, were also proposed. However, like all other nano-scale devices, the performance and reliability of STT-RAM cells are severely affected by process variations, intrinsic device operating uncertainties and environmental fluctuations. The storage margin of a MLC STT-RAM cell, i.e., the distinction between the lowest and highest resistance states, is partitioned into multiple segments for multi-level data representation. As a result, the performance and reliability of MLC STT-RAM cells become more sensitive to the MOS and MTJ device variations and the thermal-induced randomness of MTJ switching.

In this work, we systematically analyze the impacts of CMOS and MTJ process variations and MTJ resistance switching randomness that induced by intrinsic thermal fluctuations. Then, we analyzed the extension of STT-RAM cell behaviors from SLC (single-level-cell) to MLC (multi-level-cell). With the detail analysis study of STT-RAM cells, we proposed several error reduction

design, such as ADAMS structure, and FA-STT structure. In which, *ADAMS* can be dynamically configured between the high-reliable (HR) mode and the high-capacity (HC) mode upon the real-time system requirement: For the performance and reliability critical applications, ADAMS switches to HR mode. For the capacity critical applications, ADAMS switches to HC mode. The ADAMS cell is broken into two "1T1J" cells that can work independently, offering the similar performance and reliability to conventional STT-RAM design.

TABLE OF CONTENTS

PREFACE xiv

1.0 INTRODUCTION 1

2.0 PRELIMINARY 3

2.1 STT-RAM Basics 3

2.2 Process Variations 3

2.3 Thermal Fluctuation in MTJ switching 5

3.0 SINGLE-LEVEL CELL OPERATION ANALYSIS 7

3.1 Write Errors of an STT-RAM cell 7

3.1.1 Persistent Errors 7

3.1.1.1 Geometry Variations of Transistor and MTJ 7

3.1.1.2 Fluctuation of Magnetic Anisotropy 9

3.1.2 Quantitative Analysis on Persistent Write Errors 10

3.1.3 Non-Persistent Errors 14

3.1.3.1 Thermal Fluctuations 14

3.1.3.2 Temperature Dependency 18

3.1.4 Statistical Write Error Rate Analysis 20

3.1.5 Array Level Analysis 21

3.2 Read Errors of an STT-RAM cell 23

3.2.1 Persistent Error: Sensing Errors 23

3.2.2 Non-Persistent Error: Read Disturbance 25

3.2.3 Read Error Rate Analysis 25

3.2.4 Reading Analysis of a STT-RAM Array 26

3.3 STT-RAM Design Space Exploration of Reliability Optimization. 28
3.3.1 Oxide Layer Thickness Design Specification 28
3.3.2 Word-line Override Designs 31
3.4 STT-RAM Cell Design Optimization Flow 32
4.0 MULTI-LEVEL CELL OPERATION ANALYSIS 34
4.1 Variability Sources in MLC STT-RAM Designs 34
4.1.1 Process Variations in MLC 35
4.1.2 Thermal Fluctuations 35
4.2 Readability Analysis of MLC MTJs 36
4.2.1 Nominal Analysis of the Readability of MLC MTJs 36
4.2.2 Statistical Analysis of the Readability of MLC MTJs 38
4.2.2.1 Optimization of Parallel MLC MTJs 38
4.2.2.2 Optimization of Series MLC MTJs 40
4.3 Writability Analysis of MLC MTJs 41
4.3.1 Write Mechanism of MLC STT-RAM Cells 41
4.3.2 Impacts of Thermal Fluctuations 42
4.3.3 Write Operations of Parallel MLC MTJs 43
4.3.4 Write Operations of Series MLC MTJs 46
5.0 DIFFERENTIAL SENSING SCHEME TO IMPROVE THE READ PERFORMANCE OF STT-RAM 48
5.1 motivation 48
5.2 ADAMS Technology 48
5.2.1 Regular Differential Sensing Scheme (RDAMS) 49
5.2.2 Asymmetric Differential Cell Structure (ADAMS) 50
5.2.3 Read and Write Robustness of ADAMS 50
5.2.3.1 Read robustness 50
5.2.3.2 Write robustness 51
5.2.4 Asymmetric SenAmp and Latch Design 51
5.2.4.1 Asymmetric SenAmp 51
5.2.4.2 Asymmetric Latch 53

5.2.5 Reconfigurable Scheme STT-RAM 54
5.3 ADAMS Design Optimization and Analysis 55
5.3.1 Write Operation Analysis 55
5.3.1.1 Asymmetric Write Analysis 55
5.3.1.2 Definition of Write Error Rate 56
5.3.1.3 Write Optimization of ADAMS 58
5.3.2 Read Operation Analysis 60
5.3.2.1 Read Reliability Analysis 60
5.3.2.2 Read Latency Analysis 64
6.0 OTHER PROPOSED STT-RAM IMPROVEMENT WORKS 66
6.1 Basic Concept of FA-STT 66
6.2 FA-STT Read Scheme 68
6.2.1 Self-reference Sensing Scheme in FA-STT 68
6.2.2 Read Operation Analysis 70
6.2.2.1 Read disturbance 70
6.2.2.2 Sensing margin 72
6.3 FA-STT Write Scheme 73
6.3.1 Field-assisted MTJ Switching 73
6.3.2 Write Performance Evaluation 74
6.3.3 Write Error Rate 76
6.4 Layout Design Consideration 77
6.5 GSHE Spin Logic Structure 78
6.5.1 Basic Logic Functions 78
6.5.2 GSHE Logic Operation Scheme 80
6.6 Diode-GSHE Structure 82
6.6.1 Sneak Path Issues 82
6.6.2 Proposed Diode-GSHE Structure 83
6.7 Case Study 84
6.7.1 Full Adder Design 84
6.7.2 Experimental Results 86

7.0 CONCLUSION 89
BIBLIOGRAPHY 90

LIST OF TABLES

1 Summary of Device Parameters 8

2 MTJ Write Current Distribution Under Process Variations 10

3 Summary of Variation Contribution [34] 12

4 Summary of Device Parameters 56

5 Design Parameters 69

6 Comparison of write error rates under 10ns write period. 76

7 Control Signal of Diode-GSHE Structure 83

8 Summary of GSHE MTJ Parameters 86

9 Comparison of Full Adders between CMOS Circuit and Proposed Diode-GSHE Circuit. 87

LIST OF FIGURES

1 MTJ Structure (a) Anti-parallel (high resistance state). (b) Parallel (low resistance state). (c) 1T1J STT-RAM cell structure. 4

2 Examples of the driving strength distribution of the NMOS transistor in the STT-RAM cell: (a) 1→0. (b) 0→1. 13

3 (a) Switching current vs. Switching time mean. (b) Switching time mean vs. SDMR (Switching time standard deviation/Mean Ratio). 15

4 Perpendicular MTJ. (a) Switching current vs. Switching time mean. (b) Switching time mean vs. SDMR. 17

5 (a) MTJ Critical Switching Current vs. Switching Time under Varying Temperature, (b)Threshold Switching Time against Temperature. 18

6 (a) Error Rate for 10ns Write Pulse Width, (b) Error Rate for 20ns Write Pulse Width, (c) 1% and 0.1% error rate of writing '1'. 19

7 In-plane and perpendicular STT-RAM write error rate comparison under 10ns write pulse width. 21

8 Transistor channel length distribution map for a STT-RAM array. 22

9 Probability of Sensing Error and Read Disturbance under different read current. $T_{read} = 5ns$. 24

10 Sense amplifier design. 26

11 Probability of Sensing Error and Read Disturbance in a STT-RAM array. 27

12 Resistance states and resistance difference changes with oxide layer thickness. . . . 28

13 Sensing error rate and disturbance error rate when oxide layer thickness varies. . . . 29

14 (a) NMOS driving ability varies with oxide layer thickness. (b) NMOS driving ability varies with transistor channel width. 29
15 Write error rate under different oxide layer thicknesses. 30
16 Comparison between original design and override design in writing '1'. 31
17 Precess Variation Aware STT-RAM Design Flow. 32
18 Four state resistance distributions of (a) Parallel MLC MTJ and (b) Series MLC MTJ, optimized by nominal design method. 38
19 (a) Error Rate vs. R_2/R_1 Ratio Sweep, (b)Error Rate vs. Resistance of Hard Domain Sweep. 40
20 Switching properties of the two domains for a parallel MLC MTJ. (a) switching time vs. switching current. (b) switching time standard deviation vs. switching current. 43
21 Writing error rate in parallel MLC STT-RAM cell at T_w = 10ns. Notes: The total error rate is not necessarily equal to the sum of incomplete error and overwrite error, which are the errors overwriting the hard domain or incurring the incomplete soft domain flipping, respectively. 45
22 (a)Writing error rate in a parallel MLC STT-RAM cell at different T_w, Threshold current distributions of resistance state trasitions for the parallel MLC MTJ.(b) Dependent transitions. (c) Independent transitions. 45
23 (a)Writing error rate in a series MLC STT-RAM cell at different T_w, Threshold current distributions of resistance state transitions for the series MLC MTJ.(b) Dependent transitions. (c) Independent transitions. 46
24 Structure of (a) RDAMS. (b) ADAMS. 49
25 (a) 3D view of RDAMS. (b) Layout of RDAMS. (c) 3D view of ADAMS. (d) Layout of ADAMS.(e) layout of 1T1J. 50
26 (a) Asymmetric sense amplifier (SenAmp) design. (b) Simulation results of SenAmp Out signal at different corner cases. 52
27 (a) Circuit of Asymmetric Latch. (b) Asymmetric Latch Output Results. 53
28 Reconfigurability of ADAMS. Mode = 0: High-reliable (HR) mode; Mode = 1: High-capacity (HC) mode. 54

29 (a) Switching current vs. Inverse of switching Time. (b) Switching time mean vs Standard deviation and mean ratio (SDMR). 55
30 MTJ switching current vs. NMOS transistor size. (a) P-cell. (b) C-cell. 56
31 STT-RAM writing state. (a) 1T1J. (b) RDAMS. (c) ADAMS 57
32 Write error rate at 10ns write pulse width. 58
33 Write error rates of the RDAMS and ADAMS cells when the write pulse width is set to (a) 10ns; (b) 8ns; (c) 5ns; and (d) 3ns. 59
34 Example of BL voltages distribution of a 1T1J cell. 60
35 STT-RAM reading state. (a) 1T1J. (b) RDAMS. (c) ADAMS 61
36 Sensing errors and disturbance errors of different cell structures. (a) Without redundancy. (b) With 3% redundancy. 63
37 (a) Latency distribution of SenAmps. (b) SenAmp latency, latch latency and total read latency of the ADAMS cell. 65
38 (a) 3D view of FA-STT scheme. (b) MTJ intermediate resistance state generation. . 67
39 (a) Self-reference circuit design. (b) MTJ resistance during read operation. 68
40 (a) Intermediate state generation. (b) Read disturbance of intermediate state. 70
41 (a) MTJ resistance changes in reading '0'. (b) MTJ resistance changes in reading '1'. 71
42 MTJ resistance change under different magnetic field applying speed. 71
43 (a) Sensing margin distributions. (b) Memory yields under different sensing margins. 72
44 (a) The mean of MTJ switching time vs. the magnetic field. (b) The SDMR of MTJ switching time vs. the magnetic field. 73
45 The motion behavior of MTJ free layer magnetization: (a) the standard STT-RAM $1 \rightarrow 0$; (b) FA-STT $1 \rightarrow 0$; and (c) FA-STT $0 \rightarrow 1$. 75
46 The write time distributions. 75
47 3D View of External Metal Placing. 78
48 Examples of Basic Logic Functions. (a) Serial Connection, (b) Parallel Connection. 79
49 (a) Circuit of Three-stage Operation Scheme, (b) Control Signal Diagram. 81
50 An example of a real case where current sneaks through undesired paths. 82
51 Proposed Diode-GSHE Structure. 83
52 Example of Diode-GSHE Based Full Adder. 85

53 N-bit Adder Structure basd on 1-bit Adder. 85
54 Dynamic Power Consumption Under 22nm, 34nm, and 45nm tech nodes. 86

PREFACE

Among many people who helped me with this work, I first thank my advisor, Dr. Yiran Chen, for his relentless support throughout the entire duration of my graduate research, which forms the foundation of this dissertation. It was him who invited me to his excellent research group in which I initiated my first research project and have been actively participated during my PhD program. His instructive advice helped me to build my research experiences from ground up and follow the right direction since then. His strong enthusiasm motivates me to concentrate on my high performance computing research. Without his help, I could have never done this work.

Second, I would like to thank Dr. Hai Li, who has co-advised my research work for over five years of my graduate study. Her encouragement at the early stage of my work made me feel warm and helped me through the hard times. It was from her words I gained the confidence to pursue a PhD degree. Her patient guidance and directions not only helped me to conquer the difficulties I have experienced in my research work but also equipped me with valuable capabilities necessary for conducting research. From her, I have learned many useful techniques including presentation/reasoning skills, academic paper writing, research idea formulating, etc.

I also thank Professor Ching-Chung Li, Professor Ervin Sejdic and Professor Mingui Sun for being on my program committee and giving me constructive advice on this dissertation. I highly appreciate their time spent on reviewing the dissertation.

1.0 INTRODUCTION

Conventional memory technologies, i.e., SRAM, DRAM, and Flash, have achieved a remarkable success in modern electronic industry. As the semiconductor fabrication technology approaches 20nm range, the disadvantages of those technologies has become more and more prominent, i.e., the high leakage power of SRAM and DRAM, the poor endurance performance of NAND Flash, and the generally degraded device reliability. Hence, the research on emerging memory technologies have been triggered to look for alternative process scaling paths. As a promising candidate, spin-transfer torque random access memory (STT-RAM) aims the embedded memory and on-chip cache applications [27, 36, 41]. In an STT-RAM cell, data is stored as the resistance states of a magnetic tunneling junction (MTJ) device [8]. Compared to other competing technologies such as Phase-Change RAM (PCRAM), Resistive RAM (RRAM) and Ferromagnetic RAM (FeRAM), STT-RAM offers faster (nanoseconds) read access time, better CMOS process compatibility, as well as the common properties such as zero standby power, small memory cell size, and good scalability etc. [25].

As technology scales, the STT-RAM density and power consumption improve, followed by the increased process variations. The impacts of the process variations on STT-RAM cell designs, including the MOS transistor device variations, MTJ geometry and resistance variations, have been analyzed by [33, 17]. Meanwhile, the intrinsic device operating uncertainties of STT-RAM, i.e., the thermal fluctuation in the MTJ switching, is aggravated when the working temperature varies in a large range, which was also analyzed in [22]. In previous work, pure CMOS device process variation aware statistical analysis method with the consideration of the MTJ geometry variations is done in [33, 17]. And [22] has proposed some combined circuit and magnetic-level STT-RAM model that can simulate the interaction between MOS transistor and MTJ without taking into account process variations. In our work, we systematically analyze the impacts of both the

device parameter fluctuations of MTJ and transistors, and intrinsic MTJ operating uncertainties on the performances and the reliabilities of STT-RAM cells. In this work, we quantitatively study the influences of thermal fluctuation and process variation on the MTJ switching performance, and extended it from Single level cell (SLC) to multi-level cell (MLC). In Multi-level cell (MLC) STT-RAM, two MLC STT-RAM structures (parallel and serial) are analyzed. Also, by leveraging our proposed STT-RAM cell model, we establish a statistical design flow that can optimize both the persistent and non-persistent errors in STT-RAM design. Finally, two error reduction design and one improved device structure are introduced to improving the existing challenges in STT-RAM technology.

The rest of the paper is organized as follows: We briefly introduce preliminary background on STT-RAM and its variation resource in Chapter 2. In Chapter 3, we start with presenting the analysis of operation errors in single level cell (SLC) STT-RAM . Then, based on the understanding of SLC, multi-level cell (MLC) STT-RAM analysis will be demonstrated in Chapter 4. In Chapter 5, we will give a novel differential sensing design called ADAMS to reduce the read error of STT-RAM. Besides that, we will also present several other error reduction design in 6 And last is our conclusion in Chapter 7.

2.0 PRELIMINARY

2.1 STT-RAM BASICS

Spin-transfer torque random access memory (STT-MRAM) uses magnetic tunneling junction (MTJ) devices to store the information. A MTJ has two ferromagnetic layers (FL) and one oxide barrier layer (BL). The resistance of MTJ depends on the relative magnetization directions (MDs) of the two FLs. When their MDs are parallel or anti-parallel, the MTJ is in its low or high resistance state, as illustrated in Fig. 1. R_h and R_l are usually used to denote the high and the low MTJ resistance, respectively. Tunneling magneto-resistance (TMR) is defined as $(R_h - R_l)/R_l$, which presents the distinction between the two resistance states.

In a MTJ, the MD of one FL (reference layer) is pinned while the one of the other FL (free layer) can be flipped by applying a polarized write current though the MTJ. For example, the switching from low resistance state ("0") to high resistance state ("1") can be realized by applying a current from B to A, as shown in Fig. 1. A larger write current can shorten the MTJ switching time by paying the additional memory cell area overhead: In the popular "1T1J" (one-transistor-one-MTJ) cell structure (see Fig. 1(c)), the MTJ write current is supplied by the NMOS transistor. Increasing the write current requires a larger NMOS transistor. Also, the increased write current raises the breakdown possibility of the MTJ device.

2.2 PROCESS VARIATIONS

The CMOS process variations that contribute to the variability of the driving strength of the NMOS transistor in an "1T1J" STT-RAM cell structure include random dopant fluctuations (RDFs), line-

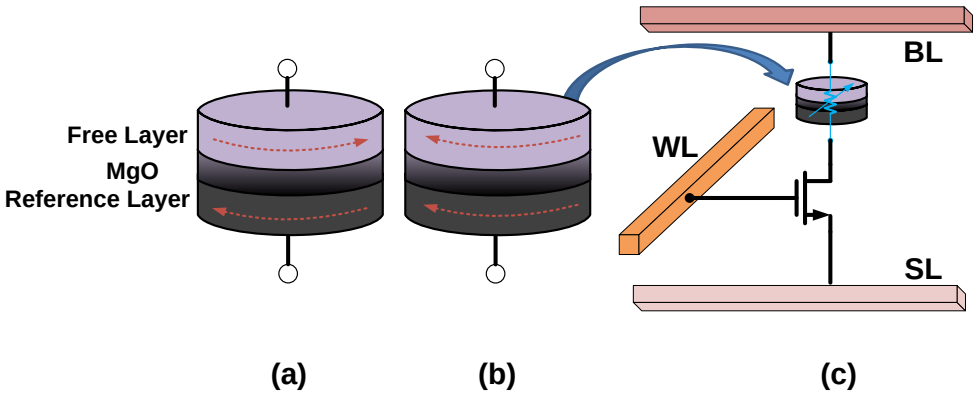

Figure 1: MTJ Structure (a) Anti-parallel (high resistance state). (b) Parallel (low resistance state). (c) 1T1J STT-RAM cell structure.

edge roughness (LER), shallow-trench isolation (STI) stress, and the geometry variations of transistor channel length/width. Besides the geometry variations, most of the CMOS process variations are reflected as the threshold voltage deviations. The random variation of the threshold voltage is prominent in the scaled CMOS technology and can severely affect circuit stability and performance. It is known that the relative deviations of MOS transistor parameters reduce when the transistor size increases.

CMOS process variations affect not only the driving strength of the MOS transistor but also its equivalent resistance. The relative deviations of MOS transistor parameters reduce when the transistor size increases.

The major sources of MTJ device variations include: 1) MTJ shape variations; 2) MgO thickness variations; and 3) normally distributed localized fluctuation of magnetic anisotropy $K = M_s \cdot H_k$ [25]. The first two factors cause the variations of the MTJ resistance and the MTJ switching current by changing the bias conditions of the NMOS transistor. The third factor is the intrinsic variation of magnetic material that affects the MTJ switching threshold current density (Eq. 2.1) and the magnetization stability barrier height (Eq. 2.2) [25].

$$J_{C0} = (\frac{2e}{\hbar})(\frac{\alpha}{\eta})(t_F M_s)(H_k \pm H_{ext} + 2\pi M_s) \tag{2.1}$$

$$\Delta = \frac{K_u V}{k_B T} = \frac{M_s H_k V cos^2(\theta)}{k_B T} \tag{2.2}$$

Here, the switching threshold current density J_{C0} is the minimal current density that causes the MTJ resistance flipping in the absence of any external magnetic field at 0K; e is the electron charge; α is the damping constant; M_s is the saturation magnetization; t_F is the thickness of the free layer; $\hbar$ is the reduced planck's constant; H_k is the effective anisotropy field including magneto crystalline anisotropy and shape anisotropy; H_{ext} is the external field; η is the spin transfer efficiency; T is working temperature; K_B is Boltzmann constant; and V is MTJ element volume.

2.3 THERMAL FLUCTUATION IN MTJ SWITCHING

Device variations are introduced by the uncertainties during the manufacturing process. After the device is fabricated, the device parameters are fixed and their impacts on the circuit performance are deterministic. Besides the device variations of MOS transistor and MTJ, the MTJ switching performance is also affected by the intrinsic thermal fluctuations. In general, the impact of thermal fluctuations can be modeled by the thermal induced random field h_{fluc} in stochastic Landau-Lifshitz-Gilbert (LLG) equation (Eq. 2.3) [8, 2, 9] as

$$\frac{d\vec{m}}{dt} = -\vec{m} \times (\vec{h}_{eff} + \vec{h}_{fluc}) + \alpha \vec{m} \times (\vec{m} \times (\vec{h}_{eff} + \vec{h}_{fluc})) + \frac{\vec{T}_{norm}}{M_s} \tag{2.3}$$

Where $\overrightarrow{m}$ is the normalized magnetization vector. Time t is normalized by γM_s; γ is the gyromagnetic ratio and M_s is the magnetization saturation. $\overrightarrow{h}_{eff} = \frac{\overrightarrow{H_{eff}}}{M_s}$ is the normalized effective magnetic field. $\overrightarrow{h}_{fluc}$ is the normalized thermal agitation fluctuating field at finite temperature which represent the thermal fluctuation. α is the LLG damping parameter. $\overrightarrow{T}_{norm} = \frac{\overrightarrow{T}}{M_s V}$ is the spin torque term with units of magnetic field. And the net spin torque $\overrightarrow{T}$ can be obtained through microscopic quantum electronic spin transport model. Under the intrinsic thermal fluctuations, the MTJ switching time becomes unrepeatable and follows a distribution. As we shall show in the next Section, this distribution is also affected by the MTJ and NMOS transistor device variations and causes the asymmetric STT-RAM cell switching at two switching directions.

3.0 SINGLE-LEVEL CELL OPERATION ANALYSIS

3.1 WRITE ERRORS OF AN STT-RAM CELL

STT-RAM errors mainly include two types – operational error and retention error. In this paper, we mainly focus on the the operational error as normally the STT-RAM is designed with very high retention time to cover the concerned storage time span, e.g., 10 years. Based on the occurrence behaviors, operational errors of an STT-RAM cell can be further divided into two types: persistent error and non-persistent error. In memory design, persistent errors denote the errors that happen deterministically and can be repeated after the chip is fabricated. On the contrary, non-persistent errors denote the transient failures incurred by intermittent events and cannot be repeated deterministically.

3.1.1 Persistent Errors

The persistent error in STT-RAM write is referred to as the errors incurred by insufficient MTJ write current and MTJ switching threshold current variation, which are induced by the process variations of the NMOS transistor and the MTJ, respectively.

3.1.1.1 Geometry Variations of Transistor and MTJ Without considering any power rail bounces, when programming an STT-RAM cell, the write current through the MTJ is mainly determined by the size of the NMOS transistor and the MTJ resistance. The first order approximation of the MTJ write current deviation generated from the process variations W (transistor channel width), L (transistor channel length), V_{th} (threshold voltage), and R_{MTJ} (equivalent resistance of

Table 1: Summary of Device Parameters

Device	Parameters	Mean	Std. Dev.
Transistor	Channel Length L	45nm	2.25nm
	Channel Width W	design dependent	2.25nm
	Threshold Voltage V_{th}	0.466V	δV_{th0}=30mV
MTJ	MgO Thickness τ	2.2nm	2% of mean
	Cross Section A	$40 \times 90nm^2$	5% of mean
	Perpendicular CS A_P	$45 \times 45nm^2$	
	Low Resistance R_l	2000Ω	
	High Resistance R_h	4500Ω	

MTJ) can be expressed as:

$$\begin{aligned}(\sigma I_{MTJ})^2 = &(\tfrac{\sigma I_{MTJ}}{\sigma W}\,|_{W=W_0}\sigma W)^2 \\ &+ (\tfrac{\sigma I_{MTJ}}{\sigma L}\,|_{L=L_0}\sigma L)^2 \\ &+ (\tfrac{\sigma I_{MTJ}}{\sigma V_{th}}\,|_{V_{th}=V_{th0}}\sigma V_{th})^2 \\ &+ (\tfrac{\sigma I_{MTJ}}{\sigma R_{MTJ}}\,|_{R_{MTJ}=R_{MTJ0}}\sigma R_{MTJ})^2.\end{aligned} \tag{3.1}$$

Here W_0, L_0 and V_{th0} are the nominal values of NMOS transistor width, length and threshold voltage, respectively. The standard variation of the threshold voltage V_{th} decreases when the transistor size increases, say, $\sigma V_{th} \propto 1/\sqrt{WL}$. In this work, we select PTM 45nm technology as our reference technology node in the simulations. Assuming a high-performance NMOS transistor is used, σV_{th0} is set to 30mV with the mean of channel length $L_0 = 45$ nm [37]. The standard deviations of W and L (σW and σL) are both set to 5% of the minimal transistor length (= $45nm$). The details of the parameters adopted in our simulations are summarized in TABLE 1.

The MTJ resistance $R_{MTJ} \propto e^{\tau}/A$, where τ is the tunneling oxide thickness and A is the MTJ surface area. The variations of both τ and A follow Gaussian distributions [17]. $\Delta V_{MTJ} = I_{MTJ} \cdot R_{MTJ}$ is the voltage drop across the MTJ where I_{MTJ} is the current through the MTJ. Hence, $V_{ds} = V_{dd} - \Delta V_{MTJ}$ is a function of I_{MTJ}.

Based on the recent experimental results in [7], in our simulations, we choose the nominal values of R_L and R_H ,or R_{L0} and R_{H0} as 2000Ω and 4500Ω, respectively. We also assume that the standard deviations of τ and A are 2% or 5% of their means [17], as shown in TABLE 1.

The MTJ size are modeled by the equations from [40] as:

$$H_K = M_S(N_b - N_a). \tag{3.2}$$

$$N_a = \frac{4\pi}{m^2 - 1}[\frac{m}{\sqrt{m^2 - 1}} ln(m + \sqrt{m^2 - 1}) - 1]. \tag{3.3}$$

$$N_b = 2\pi - \frac{N_a}{2}. \tag{3.4}$$

$$m = \frac{a}{b}. \tag{3.5}$$

Here a and b are the length and width of the MTJ nanopillar. N_a and N_b are the demagnetization factor along the longer a-axis and shorter b-axis, respectively. In a perpendicular MTJ, there is no shape anisotropy since $a = b, N_a = N_b$.

Meanwhile, we assume the variations of MTJ and CMOS devices are independent because these two types of devices are fabricated at different layers with different processes.

3.1.1.2 Fluctuation of Magnetic Anisotropy

Different from CMOS device variations and MTJ geometry variations that directly affecting MTJ write current, localized fluctuation of MTJ magnetic anisotropy results in the variations of switching threshold current density J_{C0}. In the concerned MTJ switching time range (from a few ns to hundreds ns), our magnetic model shows that the fluctuation of MTJ magnetic anisotropy causes a standard deviation of the MTJ switching threshold current density about 2% of its nominal value.

Table 2: MTJ Write Current Distribution Under Process Variations

Transistor Size	$V_{ds}(V)$	Nominal $I_{MTJ}(\mu A)$	$0\rightarrow1 Standard Deviation(\mu A)$			$0\rightarrow1 Standard Deviation/Mean$		
			MOS only	MTJ only	Both	MOS only	MTJ only	Both
90nm	0.8498	75.12	7.53	1.01	7.61	10.03%	1.35%	10.13%
180nm	0.7685	115.7	10.61	3.12	11.11	9.17%	2.70%	9.60%
270nm	0.7201	139.9	11.63	4.84	12.87	8.31%	3.46%	9.20%
360nm	0.6877	156.1	12.46	5.71	14.02	7.98%	3.66%	8.98%
450nm	0.6643	167.8	12.71	7.25	14.72	7.64%	4.32%	8.77%
540nm	0.6465	176.7	12.77	8.25	15.20	7.23%	4.67%	8.60%
630nm	0.6323	183.8	12.83	9.02	15.68	6.98%	4.91%	8.53%
720nm	0.6208	189.6	12.93	9.61	16.10	6.82%	5.07%	8.49%
Transistor Size	$V_{ds}(V)$	Nominal $I_{MTJ}(\mu A)$	$1\rightarrow0 Standard Deviation(\mu A)$			$1\rightarrow0 Standard Deviation/Mean$		
			MOS only	MTJ only	Both	MOS only	MTJ only	Both
90nm	0.5629	97.15	9.08	0.39	9.09	9.35%	0.40%	9.36%
180nm	0.2893	157.9	10.27	1.37	10.35	6.50%	0.87%	6.55%
270nm	0.1914	179.7	9.64	4.07	10.42	5.36%	2.26%	5.80%
360nm	0.1431	190.4	8.42	6.37	10.46	3.73%	2.86%	4.42%
450nm	0.1143	196.8	7.18	7.75	10.57	3.65%	3.94%	5.20%
540nm	0.0952	201.1	3.90	10.03	10.23	1.48%	4.99%	5.37%
630nm	0.0815	204.1	2.84	10.96	11.31	1.39%	5.37%	5.54%
720nm	0.0713	206.4	2.77	11.53	11.85	1.34%	5.59%	5.74%

3.1.2 Quantitative Analysis on Persistent Write Errors

We perform Monte-Carlo simulations to quantitatively study the persistent write errors in STT-RAM cell design with PTM 45nm technology [3]. A Verilog-A MTJ model was created for process variation analysis and the assumptions of the process variations are listed in TABLE 1. All simulations were conducted under Cadence Spectre Analog environment.

Three scenarios are simulated to study the impacts of different process variation sources on the driving ability of the NMOS transistor in STT-RAM cells with different transistor sizes, including:

1. Case 1 (MOS variation only): Assuming no MTJ geometry variations and only NMOS transistor process variations are considered;
2. Case 2 (MTJ variation only): Assuming no NMOS transistor process variations and only MTJ geometry variations are considered;
3. Case 3 (Both Variations): Both MTJ and NMOS transistor process variations are considered.

TABLE 2 summarizes our simulation results. For every cases, $V_{dd} = 1.0V$. Both MTJ switching directions (parallel to anti-parallel, or '0→1' and anti-parallel to parallel, or '1→0') are simulated because the NMOS transistor has different biasing conditions at these two switching directions. For every simulated transistor size, 1000 Monte-Carlo simulations are conducted.

In "MOS variation only" case, when the MTJ switches from '0' to '1', the NMOS transistor always works at its saturation region. Increasing transistor width W reduces the NMOS transistor resistance as well as the V_{ds}. However, the reduction of V_{ds} is very moderate even all the coefficients corresponding to each transistor process variations in Eq. (3.1) become larger. It leads to a larger standard deviation of MTJ write current even though the variations of V_{th} decreases. In the case that MTJ switches from '1' to '0', the NMOS transistor works at saturation region first when its width is small. However, following the increase of the channel width, NMOS transistor will change its working region from saturation to linear. V_{ds} reduces very sharply (even possibly below V_{th}), as shown in TABLE 2. Combining with the decrease of σV_{th}, the coefficients of transistor process variations in Eq. (3.1) reduce when the transistor width increases.

In "MTJ variation only" case, the coefficient of MTJ variation in Eq. (3.1) always increases when transistor size (and hence, I_{MTJ}) increases. Moreover, because of the higher I_{MTJ}, a larger MTJ write current variation is induced by MTJ variations in 1→0' switching compared to '0→1' switching under the same NMOS transistor size. Due to the same reason (and also the reduction of σV_{th}), the MTJ variation induced MTJ write current deviation becomes more prominent when the NMOS transistor size becomes larger.

When both the MTJ and NMOS transistor variations are considered, the contributions of different device variations to the MTJ driving current are mainly represented by the following four terms in Eq. (3.1) as [34]:

$$S_1 = (\frac{\partial I}{\partial W})^2 \cdot \sigma_W^2, S_2 = (\frac{\partial I}{\partial L})^2 \cdot \sigma_L^2,$$
$$S_3 = (\frac{\partial I}{\partial R})^2 \cdot \sigma_R^2, S_4 = (\frac{\partial I}{\partial v_{th}})^2 \cdot \sigma_{v_{th}}^2. \tag{3.6}$$

Table 3: Summary of Variation Contribution [34]

	Variation	Monoto	$W \to \infty$
"0"	S_1	$\downarrow$	$S_1 \to 0$
	S_2	$\nearrow\searrow$	$S_2 \to 0$
	S_3	$\uparrow$	$\max S_3$
	S_4	$\nearrow\searrow$	$S_4 \to 0$
"1"	S_1	$\downarrow$	$S_1 \to 0$
	S_2	$\uparrow$	$\max S_2$
	S_3	$\uparrow$	$\max S_3$
	S_4	$\nearrow\searrow$	$S_4 \to 0$

Based on short-channel BSIM model [34], the MTJ driving current supplied by a NMOS transistor working in saturation region can be calculated by:

$$I = \frac{\beta}{1 + \frac{1}{v_{sat}L}(V_{dd} - IR)} \cdot \left[(V_{dd} - V_{th})(V_{dd} - IR) - \frac{a}{2}(V_{dd} - IR)^2\right]. \tag{3.7}$$

Here $\beta = \mu_0 C_{ox}\frac{W}{L}$, μ_0 is electron mobility, C_{ox} is gate oxide capacitance per unit area, a is body-effect coefficient, and v_{sat} is carrier velocity saturation.

TABLE 3 shows the changing trends of S_1 to S_4 at both switching directions when the transistor channel width W increases. For each S_i ($i = 1 \sim 4$) that do not monotonically changes when W increases, a larger S_i corresponds to more contribution to the MTJ driving current variation. The limits of each S_i when W is approaching infinite are also listed in TABLE 3. It clearly shows that the residual values of S_1–S_4 at '0→1' switching is larger than that at '1→0' switching when $W \to \infty$. In other words, '0→1' switching suffers from a larger MTJ driving current variation than '1→0' switching when the NMOS transistor is large.

Furthermore, the mean of the MTJ write current of '0→1' switching is always lower than that of '1→0' switching at all simulated transistor sizes. Therefore, the STDR (standard deviation vs. mean ratio) of the MTJ switching time of '0→1' switching is always larger than that of '1→0' switching.

As also shown in TABLE 2, following the increase of the NMOS transistor size, the ratio between the means of the MTJ write currents at both switching directions, i.e., $I_{MTJ,mean}^{0\to1}/I_{MTJ,mean}^{1\to0}$, decreases. It is because that the driving ability of the NMOS transistor quickly saturates when V_{gs} reduces. However, the ratio between the standard deviations of the MTJ write currents, i.e., $\sigma_{I_{MTJ}}^{0\to1}/\sigma_{I_{MTJ}}^{1\to0}$, slightly increases when the NMOS transistor size grows. These two trends indicate the aggravation of STT-RAM cell switching asymmetry when the NMOS transistor size increases.

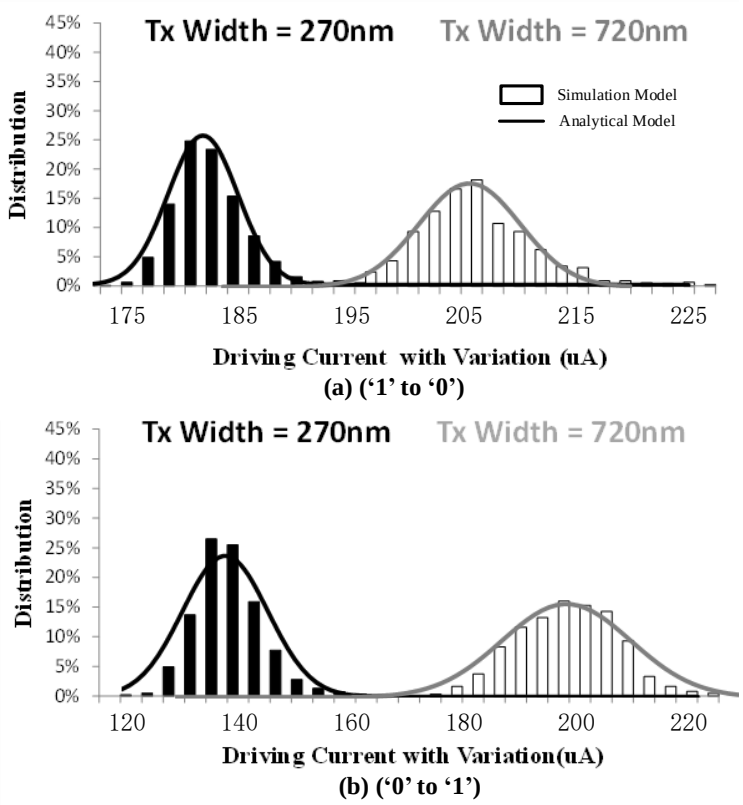

Figure 2: Examples of the driving strength distribution of the NMOS transistor in the STT-RAM cell: (a) 1→0. (b) 0→1.

We note that the analytical expression in Eq. (3.1) is able to provide reasonable estimation on the distribution of the MTJ write current by assuming the MTJ write current follows Gaussian distribution. The results of Monte-Carlo simulation and analytical estimation of the MTJ write current distributions for the NMOS transistor with W = 270nm and 720nm, respectively, are compared in Fig. 2.

Without considering thermal fluctuations, the MTJ write current I_{MTJ} must be larger than the critical MTJ switching current I_C to ensure a successful write. However, thermal fluctuation induced operational randomness makes this statement invalid. In the next section, we will discuss the impact of thermal functions on the write reliability of STT-RAM cells.

3.1.3 Non-Persistent Errors

The critical MTJ switching current at both switching directions, i.e., $I_{C,0\rightarrow1}$ and $I_{C,1\rightarrow0}$, are affected by thermal fluctuations. Thermal fluctuation is a purely random process that cannot be deterministically repeated, and induces non-persistent errors in STT-RAM operations.

3.1.3.1 Thermal Fluctuations Our simulation results of the MTJ switching current vs. the mean and the SDMR of the MTJ switching time are depicted in Fig. 3. The original device parameters are extracted from a 40nm×90nm elliptical MTJ device and have been carefully scaled to the 45nm technology. The results of both switching directions are included.

Since the switching process of a MTJ can be categorized into three working regions based on its switching time range, different fitting equations are generated for each time range as follows: For a long switching time ($> 10ns$):

$$I_{C1}(t_w) = I_{C0}(1 - (1/\Delta)ln(t_w/\tau_0)). \tag{3.8}$$

Here, t_w is switching time; τ_0 is relaxation time.

For an ultra-short switching time ($< 3ns$):

$$I_{C3}(t_w) = I_{C0} + Cln(\pi/2\theta). \tag{3.9}$$

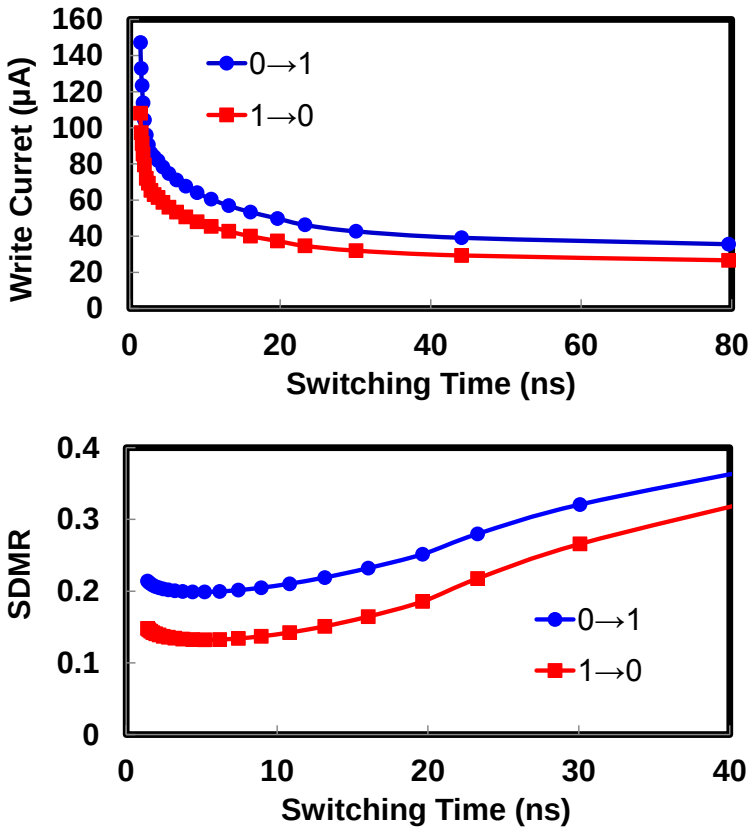

Figure 3: (a) Switching current vs. Switching time mean. (b) Switching time mean vs. SDMR (Switching time standard deviation/Mean Ratio).

Here C is a fitting parameter, θ is the initial angle between the magnetization vector and the easy axis, n is a fitting parameters.

When the MTJ switching time is in the intermediate region ($3ns < t_w < 10ns$), a dynamic reversal that combines the precessional and thermally activated switching occurs [8]. Based on the simulation results of our macro-magnetic model, we derive a fitting function of the critical MTJ switching current I_{C2} for this time range as:

$$\begin{aligned} I_{C2}(t_w) &= 30(I_{C3}(3n) - I_{C1}(10n))/t_w \\ &+(10I_{C3}(3n) - 3I_{C1}(10n))/7. \end{aligned} \tag{3.10}$$

Fig. 3(a) shows the simulation results of the means of the MTJ switching current and the nominal switching time in both '1→0' (red) and '0→1' (blue) switching's using the same MTJ configuration in the previous simulations. Thermal fluctuation influences the MTJ magnetic switching

process and causes the variations of MTJ switching time. When MTJ is operating in a relatively long time region ($> 10ns$), thermal fluctuation is dominated by the thermal component of internal energy; when MTJ working in a short time region ($< 10ns$), thermal fluctuation is dominated by the thermally active initial angle of procession [37].

Under a certain threshold write current, the MTJ write latency is not fixed but suffers from the thermal fluctuation induced variations. This uncertainty may cause unsuccessful writes if the MTJ device fails to switch before the write pulse is removed. Fig. 3(b) shows the distribution of MTJ switching time at both '1→0' and '0→1' switching's. The distinction between the means of MTJ switching time at two switching directions with the same switching current can be explained as the asymmetric impacts of tunneling spin polarization P and follows:

$$\frac{J_{C0}^{0\to1}}{J_{C0}^{1\to0}} = \frac{1+P^2}{1-P^2}. \tag{3.11}$$

Here $J_{C0}^{0\to1}$ and $J_{C0}^{1\to0}$ denotes the MTJ switching threshold current density at the switching of '0→1' and '1→0', respectively.

The difference in the standard deviations of the MTJ switching time at two switching directions, however, is caused by the asymmetric influences of thermal agitation fluctuating field $\overrightarrow{h}_{fluc}$. A larger MTJ switching time deviation is observed in '0→1' switching than '1→0' switching.

We found when the MTJ works at a long switching time range (>40ns, or switched by a low current), the standard deviation of the MTJ switching time for both switching directions are high. Following the decrease of the MTJ switching time, the standard deviation of the MTJ switching time reduces first and then raises again. It is due to the reduced thermal impacts and the increased impact of the spin torque term $\overrightarrow{T}_{norm}$ on MTJ switching under a high switching current. In general, when the nominal MTJ switching time decreases, its standard deviation decreases first and then increases. The minimal SDMR of the MTJ switching time occurs around $t_w = 10ns$.

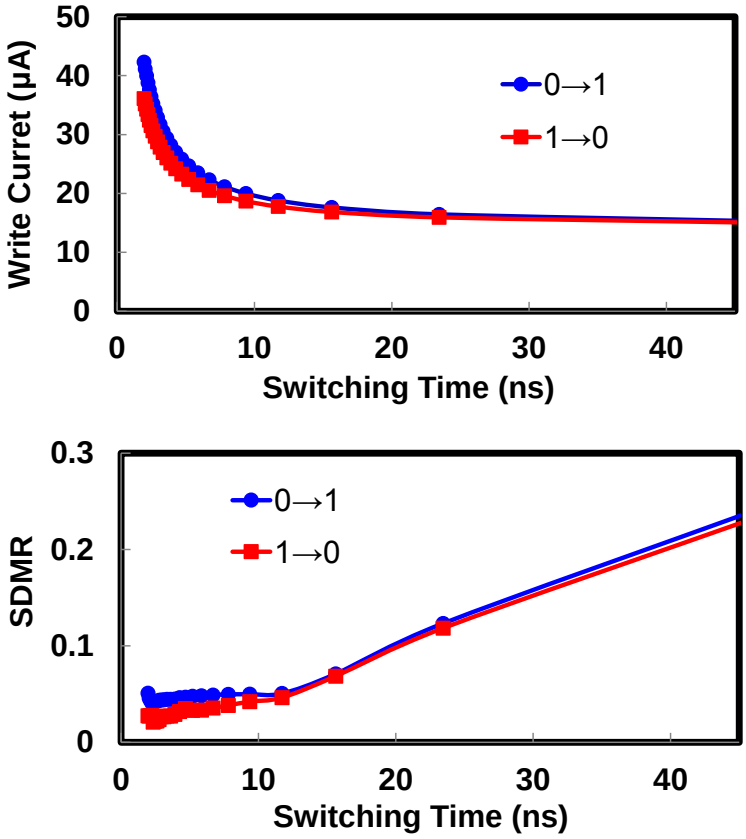

Figure 4: Perpendicular MTJ. (a) Switching current vs. Switching time mean. (b) Switching time mean vs. SDMR.

As aforementioned, PMTJ has a lower switching threshold current density than in-plane MTJ. Similar to Fig. 3, the simulation results of the nominal switching current and the SDMR of the switching time for a 65nm×65nm PMTJ are illustrated in Fig. 4(a) and Fig. 4(b), respectively. Here the size of the PMTJ is adopted from [7], which does not choose the minimal pitch of the technology node due to other circuit design concerns. Compared to in-plane MTJ, PMTJ significantly reduces the requirement of switching current due to the smaller switching threshold current density. The switching current difference between writing '1' and writing '0' also becomes smaller, indicating that PMTJ has a more symmetric switching performance. However, writing '1' ('0→1', blue line) still requires a larger current than writing '0' ('1→0', red line). On the other hand, PMTJ comes with a much smaller switching time variation though its changing trend is the same as that of in-plane MTJ. In general, the SDMRs of the switching time of PMTJ at both MTJ switching directions are very close: writing '1' has a slightly larger switching time variation then writing '0'

when the write current is small due to the asymmetric thermal effect on perpendicular anisotropy. Nonetheless, compared to in-plane MTJ, PMTJ has a better balanced switching performance at different directions.

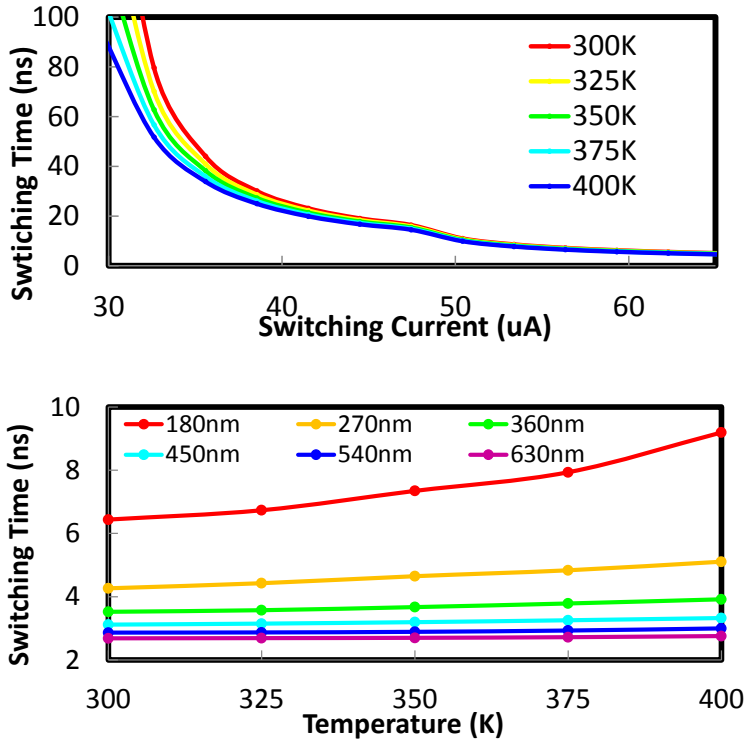

Figure 5: (a) MTJ Critical Switching Current vs. Switching Time under Varying Temperature, (b)Threshold Switching Time against Temperature.

3.1.3.2 Temperature Dependency The switching performance of a MTJ improves when working temperature raises. Higher temperature degrades the magnetization stability barrier height (Eq. 2.2) and reduces the critical MTJ switching current and/or the switching time. Fig. 5(a) shows the relationship between the critical MTJ switching current and the switching time under different temperatures for the adopted PMTJ. The impacts of temperature variations are more significant in long working time region: the thermal impact on the MTJ switching performance is more prominent when the MTJ switching current is low, compared to the impact of spin-torque.

We also simulated the temperature sensitivity of the nominal switching time of the MTJ driven by the NMOS transistor with different sizes, as shown in Fig. 5(b). Only the mean values of the switching performances are analyzed with temperature variation. The MTJ switching time

increases when the temperature raises. Since the driving ability of NMOS transistors becomes worse when operating in a high temperature environment, the result actually indicates that the improvement of MTJ magnetic switching performance cannot compensate the driving ability loss of the NMOS transistor when the working temperature increases.

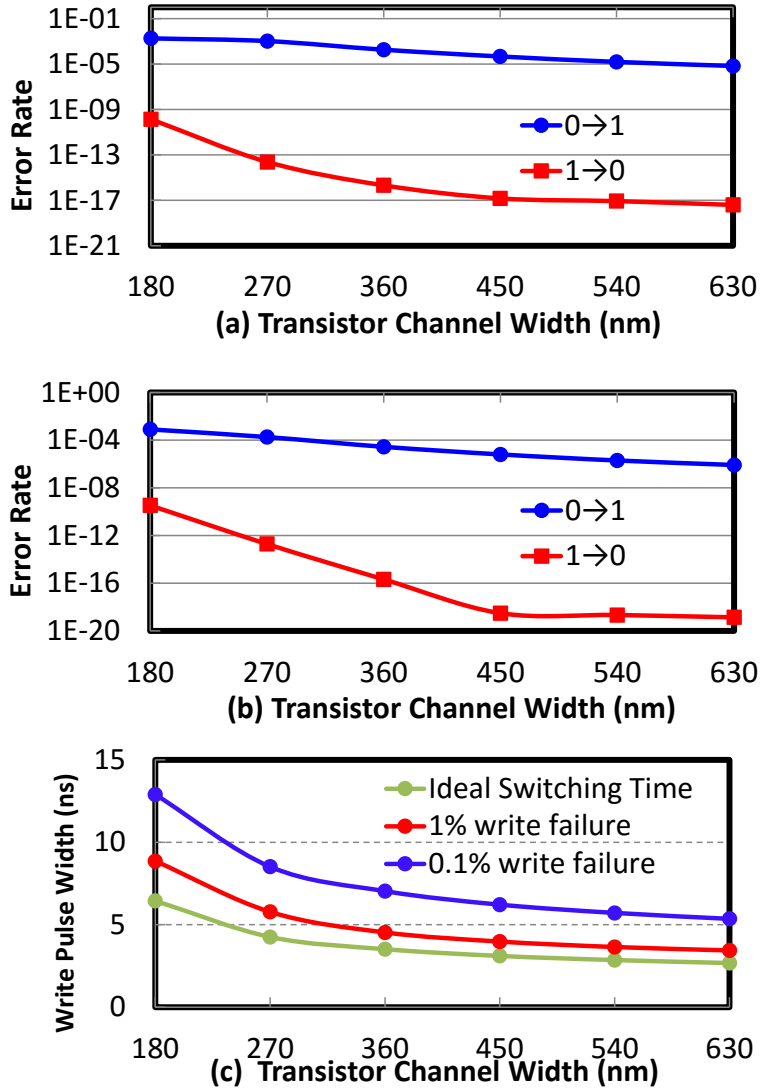

Figure 6: (a) Error Rate for 10ns Write Pulse Width, (b) Error Rate for 20ns Write Pulse Width, (c) 1% and 0.1% error rate of writing '1'.

3.1.4 Statistical Write Error Rate Analysis

The write error rate of an STT-RAM cell can be defined as the probability that the write access to the STT-RAM cell cannot complete within a certain write pulse width. Thus, a Monte-Carlo simulation is conducted by generating 1,000 STT-RAM cell driving ability samples (reflecting the persistent errors) and 1,000 MTJ switching time sampling for thermal fluctuation simulations (modeling the non-persistent errors) on each sample of the STT-RAM cell driving ability.

Fig. 6(a) and Fig. 6(b) shows our simulation results of STT-RAM cell write error rates for both writing "1" and "0" at 300K, when the write pulse width is set at 10ns and 20ns, respectively. Except for the ambient temperature, all other aforementioned variation sources, including the device variations of NMOS transistor and MTJ and the thermal fluctuations are taken into account in our simulations. Increasing the transistor size can effectively suppress write error rate by raising the MTJ write current. Due to the asymmetric cell structure, the NMOS transistor provides less current to the MTJ during '0→1' switching than '1→0' switching. However, '0→1' switching requires higher MTJ switching current than '1→0' switching, and becomes the limiting factor of write error rate. The effectiveness of sizing up the NMOS transistor for error rate reduction degrades when the transistor size is large because the NMOS driving ability becomes saturated due to the reduced V_{ds}.

Fig. 6(c) shows the required write pulse width (MTJ switching time) for the write error rates of 1% and 0.1% when the NMOS transistor size varies. For comparison purpose, the ideal results based on the nominal device parameters without considering thermal fluctuations are also presented. Significant differences are observed between the ideal and the actual performance of the MTJ: the required write pulse width when the variations are considered can be multiple times longer than the ideal result, depending on the targeted error rate.

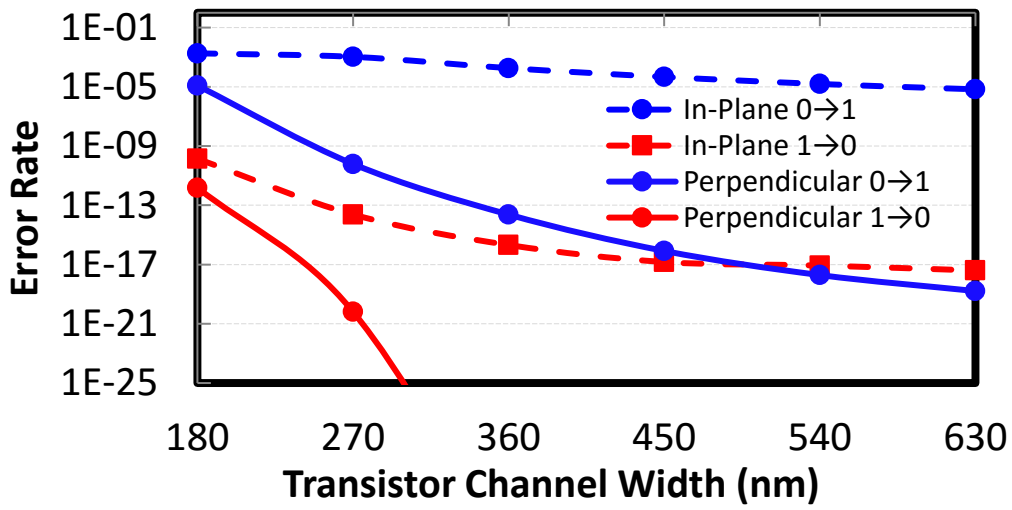

Figure 7: In-plane and perpendicular STT-RAM write error rate comparison under 10ns write pulse width.

We also simulated write error rate of perpendicular STT-RAM cells. Fig. 7 shows write error rates of both in-plane STT-RAM and perpendicular STT-RAM under a 10ns write pulse width. Since the required switching current of perpendicular STT-RAM cell is much less than that of in-plane STT-RAM cell, under the same transistor size, the write error rate of perpendicular STT-RAM is much smaller than the one of in-plane STT-RAM. To maintain a certain level write error rate, perpendicular STT-RAM can achieve a much higher cell density than in-plane STT-RAM.

3.1.5 Array Level Analysis

Variabilities in STT-RAM cell, e.g., geometry variations of transistor and MTJ size, occurs in both random and systematic sources. Systematic variations usually demonstrate strong spatial correlations, that means the neighbour cell variation are much smaller than two cells far apart.

In this section, we use VARIUS to generate distributions of variabilities of STT-RAM array [26] with spatial correlations. Both inter-die and intra-die variations are considered. Particularly, the inter-die variation is reflected as the fluctuation of the mean value of the variability ($\mu_(die)$) while the intra-die variation is shown as the standard deviation ($\sigma_(die)$) which includes all the parameters that affected by process variation, i.e. σ_W, σ_L and σ_R. ρ is the spatial correlation coefficient which decreases when the distance between two cells increases. Furthermore, parameter

ϕ defines the maximum distance where two cell can correlate. Cells that distance between each other is longer than ϕ are assumed to have no correlations. The correlation range is radius of the die, when ϕ is 0.5 as in our simulation, only the cell at the center is affected by the whole die.

We repeatly ran VARIUS to generate a $1k \times 1k$ array by using statistic tool R. The parameter set including (W, L, and $R_(MTJ)$) of each cell in the array follows intra-die and inter-die variations, and these variations are assumed to follow Gaussian distribution.

As an example, Fig. 8 shows two generated sample sets of transistor channel length distribution map and histogram for a STT-RAM with $\sigma_L = 0.05 * L, and \mu_(die) = 0.02 * L$. The values of transistor channel length are represented by the color lightness: lighter color indicates longer transistor length. For example, area A has the shortest transistor channel length, which behaves a strongest driving ability, on the other hand, longest channel length happens in area B, correspondingly, area B has the worst driving ability.

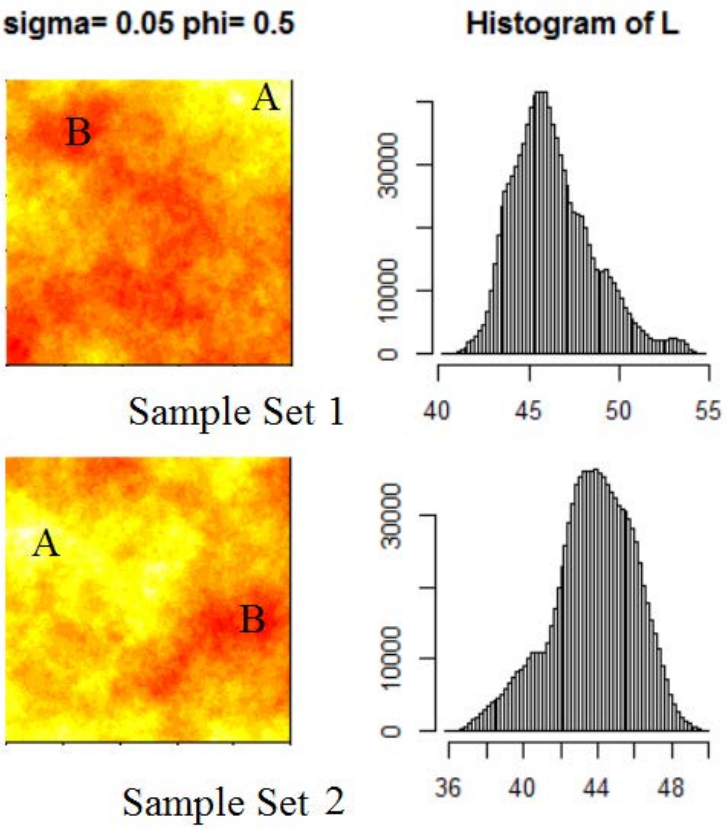

Figure 8: Transistor channel length distribution map for a STT-RAM array.

To systematically calculate the array error rate, we assume that the power supply that applied to each cell is the same. Thus the error rate will only be affected by the same resources, persistent and non-persistent error as describe above. Also using transistor channel length as an example, in single cell analysis we assume that the standard deviation of 45nm is 2.25nm as shown in Table 1. Although 2% inter-die variation and spatial correlation is considered in the simulation, based on the histogram also shows in Fig. 8, the standard deviation is still 2.25nm. Since all the parameter has the same mean, and standard deviation, we can easily conclude that writing error rate will maintain the same as single cell analysis.

3.2 READ ERRORS OF AN STT-RAM CELL

Read operations of STT-RAM are also affected by both persistent and non-persistent variations. On the one hand, process variations of peripheral circuit (e.g., sense amplifier) and variation of equivalent resistance of NMOS transistor and MTJ affect the sensing margin of STT-RAM; On the other hand, thermal fluctuation will cause the MTJ resistance switches when read voltage/current is applied. Such a non-persistent error that randomly occurs in read operations is usually referred to as read disturbance. As a result, read errors of STT-RAM can be classified into two kinds of errors: sensing error which is persistent error and read disturbance error which is non-persistent error.

3.2.1 Persistent Error: Sensing Errors

In traditional current-sensing STT-RAM read scheme, for instance, a read current I_{read} is injected into the memory cell. The generated bit-line voltage is then compared to a reference voltage to read out the MTJ resistance state. The generated sense margin, which can be measured by the voltage difference between the bit-line voltage and the reference voltage, is proportional to $I_{read} \cdot R_L \cdot TMR$. Certain sense margin must be maintained in STT-RAM read operations to overcome the device mismatch in the sense amplifier and keep the sensing errors at a minimum level.

When I_{read} is small, the generated sense margin of STT-RAM will be very limited if the MTJ resistance and/or TMR is fixed. The degraded sense margin may incur sensing errors if the device

variation of sense amplifier is large. Since the process variations of CMOS technology become more and more severe when manufacturing technology scales, readability may replace the write failure to serve as the limiting factor of STT-RAm design reliability. It is necessary to conduct a detailed analysis on the robustness degradation of the STT-RAM read operations and explore the optimization of MTJ scaling from the readability perspective.

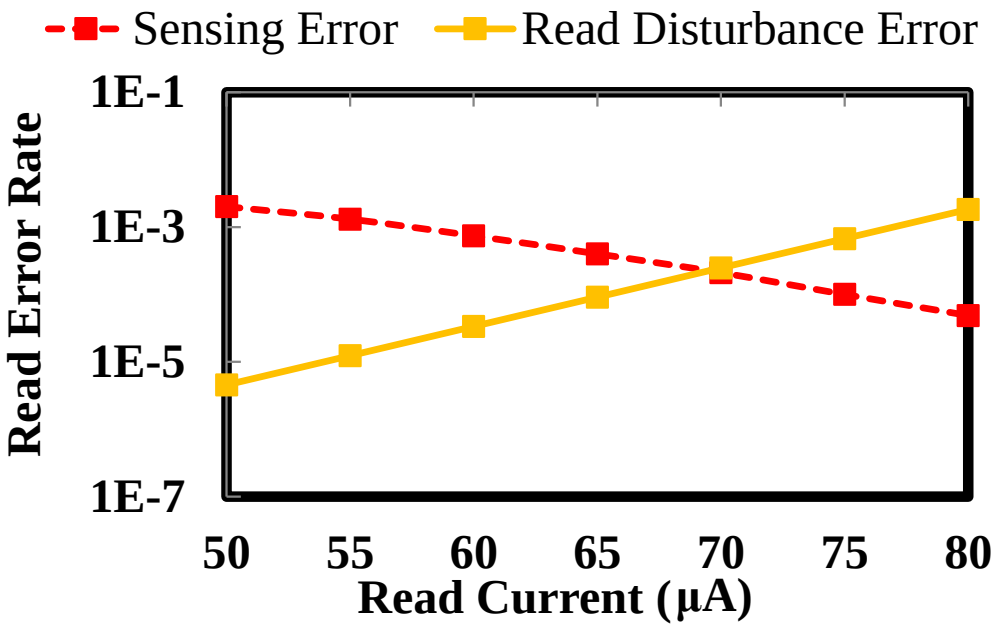

Figure 9: Probability of Sensing Error and Read Disturbance under different read current. $T_{read} = 5ns$.

We define the sense margin as the voltage difference actually generated on the two inputs of the sense amplifier. A large sensing margin generally implies a low sensing error rate. Because of process variations, the sense margin observed by the sense amplifier must be large enough to overcome the device mismatch in the sense amplifier.

The sensing errors occur when the voltage difference on the inputs of the sense amplifier cannot overcome the device mismatch of the circuit. The red line in Fig. 9 shows the sensing error rates of an in-plane STT-RAM cell when changing I_{read}. Here the device variations of both MTJ and NMOS transistor are included in our simulation. The adopted device parameters are shown in TABLE. 1. Following the increase of I_{read}, the sensing error rate reduces rapidly. It is because with the same R_L and TMR, increasing the I_{read} will raise the sensing margin, or say, $I_{read} \cdot \Delta R$.

3.2.2 Non-Persistent Error: Read Disturbance

The resistance state of the MTJ may be flipped by the read current. Since the read current is usually small, the MTJ switching performance in STT-RAM read operations can be modeled by Eq. (3.8). The switching probability of the MTJ, hence, can be approximated by:

$$P_{sw} = 1 - exp\{-\frac{T_w}{\tau_0} exp[-\frac{1}{\Delta}(1 - I_{read}/I_{c0})]\}. \tag{3.12}$$

Eq. (3.12) clearly shows that the MTJ switching probability is a function of the critical switching current I_{c0}, the switching time τ_p, and the applied current I_{read}. Fig. 9 also shows the simulated STT-RAM cell read disturbance rate under different read currents (the yellow line). The read disturbance quickly increases when I_{read} raises. Note that here the read current is applied for 5ns.

3.2.3 Read Error Rate Analysis

It is obvious that the probability of STT-RAM read disturbance and sensing errors follow an opposite trend during STT-RAM design optimization: On the one hand, when increasing the read current or read lantency, sensing error will reduce due to the enlarged sensing margin or more robust sensing process; On the other hand, increasing the read current or read latency will also raise the occurrence probability of read disturbance. Hence, it is possible to find an optimal point that can achieve the minimum total read error rate. In general, read error rate of an STT-RAM cell can be expressed as:

$$P(Re_e) = P(Sen_e) + P(Dis_e) - P(Sen_e) \times P(Dis_e). \tag{3.13}$$

Here $P(Re_e)$, $P(Sen_e)$, and $P(Dis_e)$ represent the probability of total read error rate, sensing error rate, and read disturbance rate, respectively.

In Fig. 9, the optimum read current that achieves the minimum total read error rate (2.1×10^{-4}) is $70\mu A$. Deviating from this optimum value will quickly raise either the sensing error rate or read disturbance rate.

Note that this conclusion is valid only for the sensing time of 5ns, which is the minimum sensing time that is required to charge the sense amplifier for a read current larger than 50uA. Reducing the sensing time will cause a higher requirement of sensing current.

3.2.4 Reading Analysis of a STT-RAM Array

Same as array level writing operation simulation, we generated a array using statistic tool R. To demonstrate the impacts of sensing margin and variation, we used a basic and popular sense amplifier design in STT-RAM arrays as shown in Fig. 10 which is shared by each column of 1k bit cells. Only conventional sensing scheme is adopted. We note that, the performance and reliability can always be further improved by a better SA design. The Sense amplifier we used here was tuned for best possible performance in typical process corner by sizing of the transistors. We also assume that the sense amplifier is placed very close to the array to reduce the affect of routing delays. Thus, for each parameter of transistor width, length and threshold voltage, a 1009×1009 array is generated, we pick a 1009×1000 matrix among the array, and using the $1K \times 1K$ numbers as our sample array, and the rest $9 \times 1K$ represents the parameter ratio of a sense amplifier. Since every column has its own sensing reference, the reference should be adjustable to have the optimize value for its own column instead of using $\frac{R_h+R_l}{2}$ for the whole array.

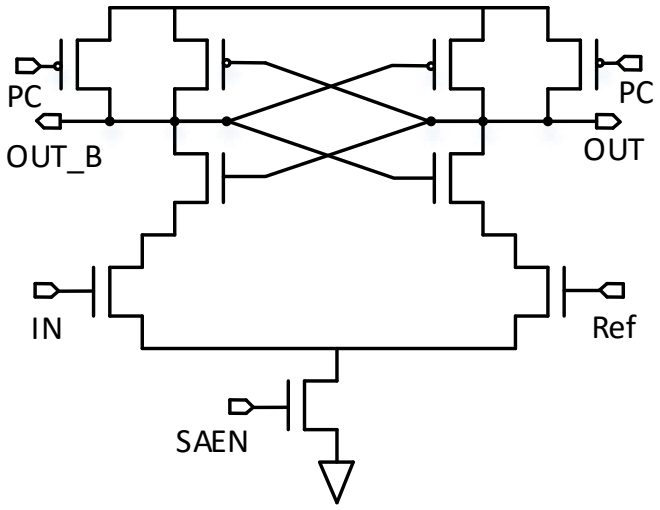

Figure 10: Sense amplifier design.

A Monte Carlo simulation that can read out every bit of the entire array has been developed to systematically analysis the read error rate for STT-RAM array. To accurately model a random noise that may cause a mismatch in sense amplifier, we applied a random noise voltage (from -0.1V to 0.1V) to the output node of the sense amplifier as shown in Fig. 10. Since the simulation determine

a success read or a read failure based on whether readout result is same as the value stored in the cell, it is very difficult to differential a sensing error that is caused by a not enough sensing margin, or mismatch by noise, thus, we count both these read failure as sensing error in here. Since it is impossible to run through all the bit cells, we also assume that each cell that its sensing margin is large enough, i.e. $\leq 30mV$, will always perform a successful reading. We accumulated the results and calculated the final read error rate based on all the roles above.

Fig. 11 shows one Monte Carlo simulation results of a STT-RAM array that generate as above. Compare with read error in the single cell analysis, the read error is higher when the read current is small, however when the read current is increasing, the error rate is largely reduced. It is obvious that with small sensing margin, the error is also increased by the mismatch and noise of the sense amplifier. On the other hand, when sensing margin increases, and effect on amplifiers are reduced. The read error rapidly reduce since every column is compared with its own reference. Especially when spatial correlation are token into account, most of cells in each column are biased from the same direction compare with typical value. The results can be further improved by optimizing the distribution of sense amplifier connections.

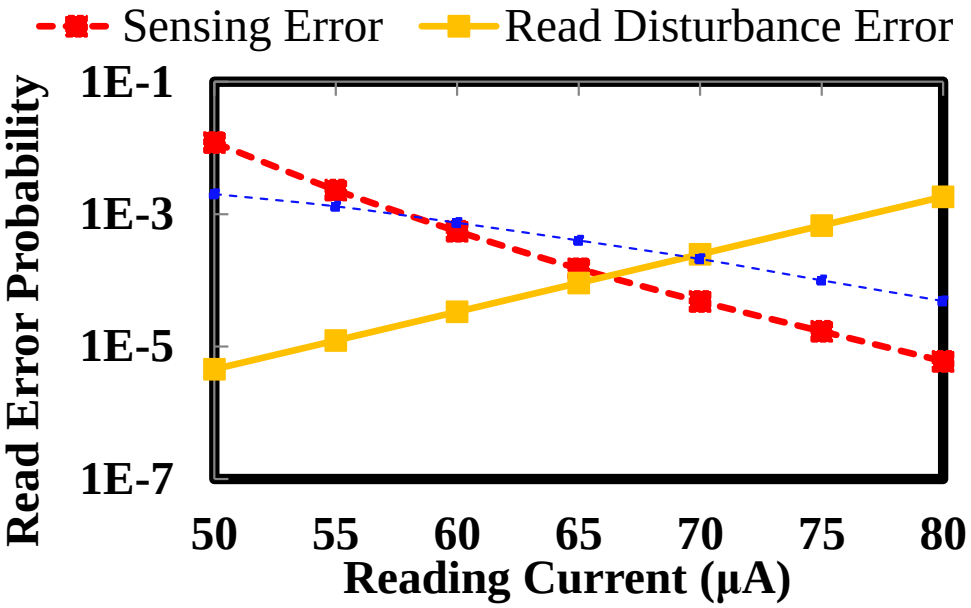

Figure 11: Probability of Sensing Error and Read Disturbance in a STT-RAM array.

3.3 STT-RAM DESIGN SPACE EXPLORATION OF RELIABILITY OPTIMIZATION.

3.3.1 Oxide Layer Thickness Design Specification

Increasing the sensing margin can enhance the read reliability of STT-RAM. As aforementioned, the sensing margin is a product of read current and MTJ resistance difference. Sec. 3.2 concludes that the read current cannot be greatly increased when read disturbance is taken into account. Hence, a more viable way to enhance the sensing margin is increasing the MTJ resistance difference.

One approach to increase the MTJ resistance difference is to raise the MTJ resistance value (i.e., R_{high} and R_{low}) while still maintaining the similar TMR by increasing the thickness of oxide layer. This method may reduce the write current applied to the MTJ during write operation and harm the write reliability of the STT-RAM cell (see Section 3.1). In addition, the TMR of the MTJ will slightly change with the thickness of oxide layer. Nonetheless, it has been proved that such a TMR degradation can be controlled within a small range [20]. To analyze the potential benefit of optimizing the thickness of the oxide layer in STT-RAM readability enhancement, we performed the relevant simulations by sweeping the thickness of the oxide layer from 2nm to 3nm. The corresponding TMR keeps above 100%.

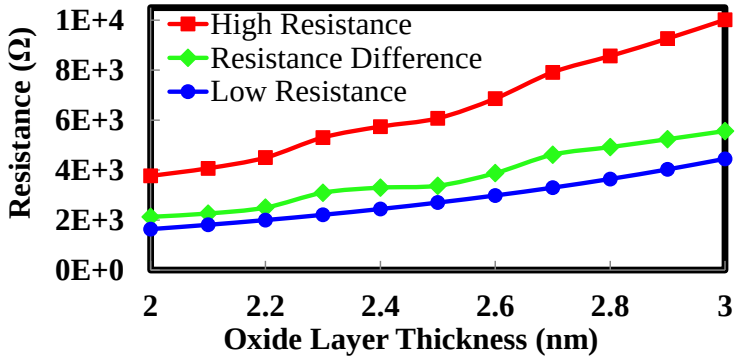

Figure 12: Resistance states and resistance difference changes with oxide layer thickness.

Fig. 12 shows the changes of the high and the low resistance states, and the resistance difference of the MTJ when oxide layer thickness varies. When the oxide layer thickness increases from 2nm to 3nm, the MTJ resistance can vary up to 2.78×. The resistance difference keeps increasing, well-controlled TMR degradation [38] leads to more than doubled sensing margin.

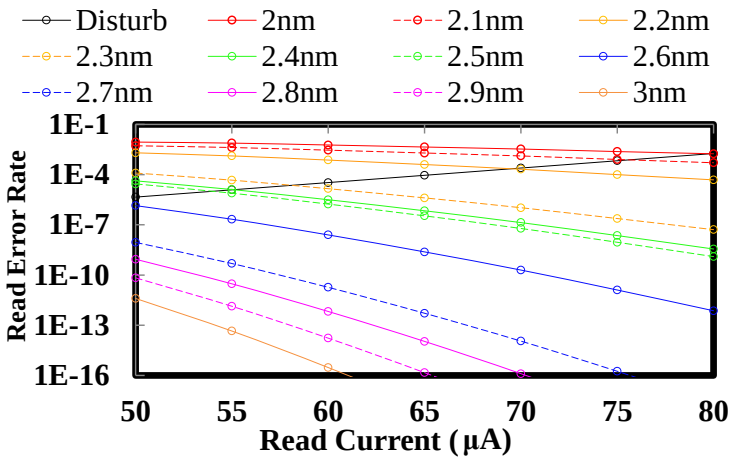

Figure 13: Sensing error rate and disturbance error rate when oxide layer thickness varies.

Although the standard deviation of the oxide layer thickness variation is smaller (2%) than that of other horizontal process variations (5%), the impact of oxide layer thickness variation on MTJ resistance is still significant because of the exponential relation between these two parameters. Fig. 13 depicts both sensing error rate and read disturbance error rate of an STT-RAM cell when the oxide layer thickness varies. Note that the read disturbance error rate is determined by the amplitude of the read current and independent on the oxide layer thickness. As a comparison, the sensing error rate is greatly reduced by increasing the oxide layer thickness, which leads to the improved MTJ resistance difference. As the process variation induced MTJ resistance variability keeps almost the same, the improved MTJ resistance difference generates larger sensing margin.

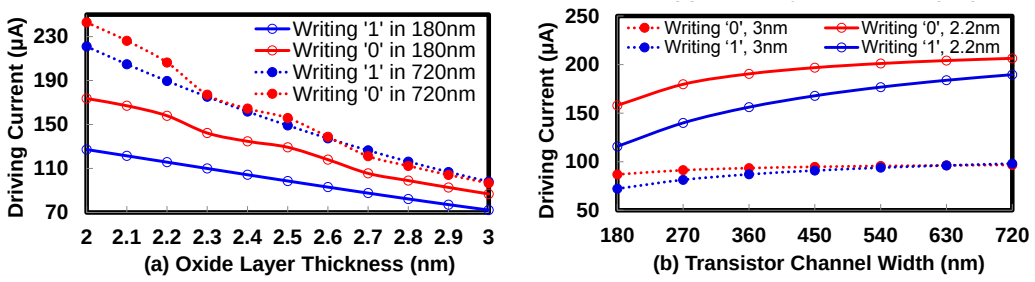

Figure 14: (a) NMOS driving ability varies with oxide layer thickness. (b) NMOS driving ability varies with transistor channel width.

Follows oxide layer thickness increase, the increased MTJ resistance causes the driving ability degradation of the NMOS transistor. Fig. 14(a) and (b) respectively show the changes of the driving ability of the NMOS transistor in the STT-RAM cell when the transistor size and oxide layer thickness vary. When the oxide layer thickness raises from 2nm to 3nm, the driving ability of the 180nm NMOS transistor reduces from $127.2\mu A$ to only $72.2\mu A$. The driving ability degradation ratio becomes severer for a large size NMOS transistor (i.e., 720nm). Fig. 14(b) shows that when the oxide layer is thick (i.e., 3nm), the driving ability of the NMOS transistor quickly saturates when the transistor size increases: since the MTJ resistance is much larger than that of the NMOS transistor, the benefit of increasing the transistor size is offset by the degraded (V_{ds}). Moreover, the NMOS transistor driving abilities at two switching directions merges together when the transistor size increases. The above results show that raising the MTJ resistance may not be a good choice when the NMOS transistor size is large.

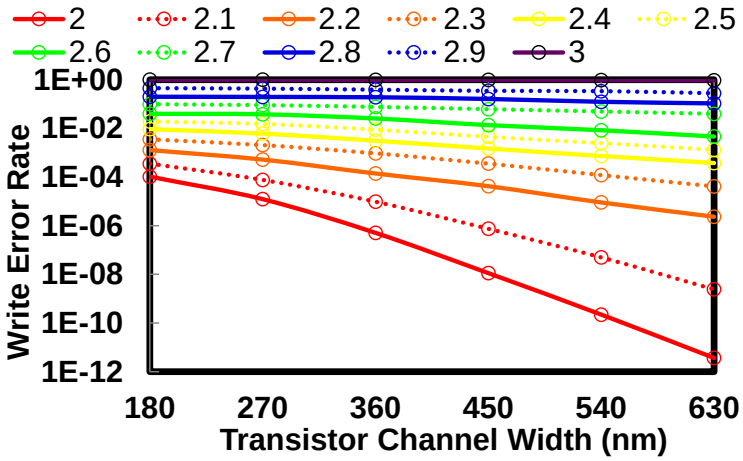

Figure 15: Write error rate under different oxide layer thicknesses.

Fig. 15 shows the writing error rates of the STT-RAM cell at different oxide layer thicknesses and transistor sizes. To have a fair comparison, we only changes one parameter each time, thus in here we fixed the writing pulse width, which means the writing time is the same in each situation. When write pulse width is fixed, increasing the MTJ resistance significantly increases the write error rate of the STT-RAM cell. An extreme case is when oxide layer thickness is 3nm, the write error rate is close to 1! In STT-RAM design, the selection of proper oxide layer thickness depends on not only the corresponding read and write error rates but also the frequencies of read and write.

3.3.2 Word-line Override Designs

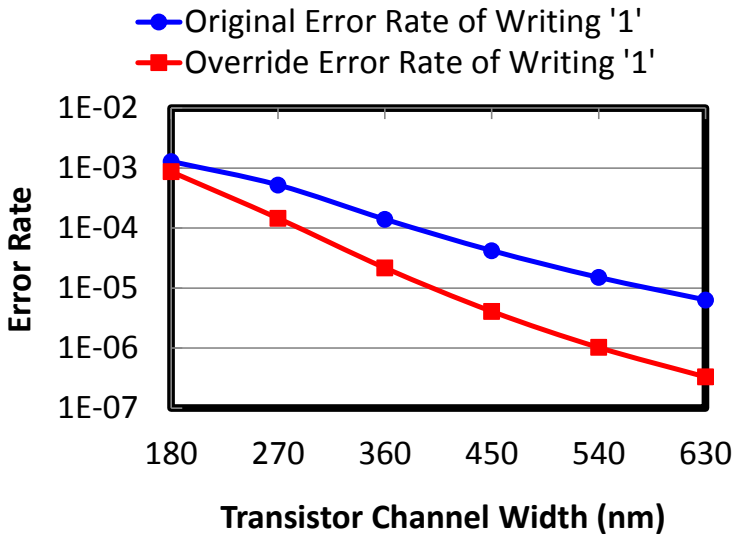

Figure 16: Comparison between original design and override design in writing '1'.

A popular approach to improve write reliability of STT-RAM is 'word-line override', which boosts the word-line voltage to a slightly higher voltage to compensate the loss of V_{gs} during writing '1' [33]. We conducted Monte-Carlo simulations to evaluate the effectiveness of word-line override scheme at different transistor sizes. The word-line voltage is boosted to 1.1V from the normal 1V. Fig 16 shows the write error rate reduction when the NMOS transistor size increases for both conventional design and word-line override design. For simplicity, only the results of the limiting switching direction '0→1' are presented. For the same transistor size, word-line override greatly reduce the write error rate at all the simulated transistor sizes.

3.4 STT-RAM CELL DESIGN OPTIMIZATION FLOW

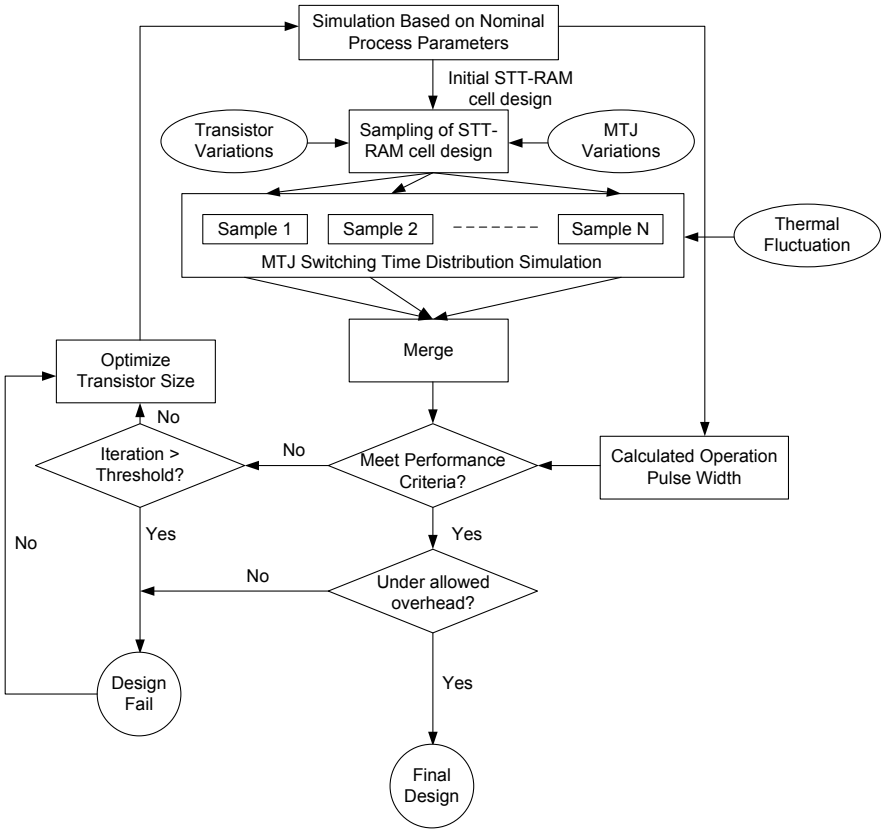

Figure 17: Precess Variation Aware STT-RAM Design Flow.

Fig 17 illustrates our proposed STT-RAM cell design optimization flow to minimize the operation errors. After the device parameters are given, the NMOS transistor size is calculated accordingly based on the designed (nominal) values of both MTJ and CMOS parameters. Meanwhile, a reasonable operation pulse width will be calculated, which is often required to align with the performance requirement. In the second step, the device parameter samples, including both the geometry and the material parameters, are generated based on the process variations of both NMOS transistor and MTJ. These samples are sent to Monte-Carlo-based SPICE simulations to collect the samples of the write current through the MTJs. The third step takes into account the thermal fluctuation effects and the fluctuation of magnetic anisotropy under the given operation pulse width to calculate the distribution of the MTJ switching time and the write errors. Based on the requirements of write

performance and write error rate, we should be able to find the optimal design points for both the NMOS transistor and the MTJ. If the result leads to a design failure, then the word-line override design may be applied. Similar design flow can be applied to the read error rate optimization or the overall STT-RAM error rate optimization.

4.0 MULTI-LEVEL CELL OPERATION ANALYSIS

The multi-level cell (MLC) capability can be implemented by realizing four or more resistance levels in MTJ designs. At least two proposals of MLC MTJ structures have emerged [13, 19] so far, including parallel MLC MTJs and series MLC MTJs. In parallel MLC MTJs, the four resistance states – '00', '01', '10', and '11', are uniquely defined by the four combinations of the magnetic directions of the two magnetic domains in the free layer. The first and the second digit of the two-bit data refer to the resistance state of the hard domain and the soft domain [5]. In series MLC MTJs, the four resistance states are uniquely defined by the combinations of the relative magnetization of the two SLC MTJs. The minimal device size of a parallel MLC MTJ and the small SLC MTJ in a series MLC MTJ can be as the same as that of the normal SLC MTJ, which is defined by the required aspect ratio and the lithography limit.

4.1 VARIABILITY SOURCES IN MLC STT-RAM DESIGNS

The performance and reliability of MLC STT-RAM cells are seriously affected by mainly two types of variabilities, including a) the process variations of MOS and MTJ devices and b) the thermal fluctuations in MTJ switching process.

4.1.1 Process Variations in MLC

The major sources of MTJ device variations mainly include: 1) MTJ shape variations, i.e., the surface area variation; 2) MgO layer thickness variations; and 3) normally distributed localized fluctuation of magnetic anisotropy: $K = M_s \cdot H_k$. Here M_s is saturation magnetization. H_k is the effective anisotropy field including magneto crystalline anisotropy and shape anisotropy. These factors lead to the deviations of MTJ resistance and the required switching current from the nominal values.

The MTJ device variations affect the reliability of the two types of MLC MTJs in the different ways: In parallel MLC MTJs, the two parts of the MTJ with different magnetic domains (For simplicity, we also call them "two magnetic domains" in the rest of this paper) share the same free layer, reference layer and MgO layer. In such a small geometry size, we can assume the MgO layer thickness and the RA (resistance-area) of these two parts are fully correlated. Other parameters, such as the MTJ surface areas, the magnetic anisotropy and the required switching current density can be very different for these two parts because they are determined by the magnetic domain partitioning. In series MLC MTJs, however, all these parameters of two SLC MTJs are close to each other and only spatially correlated.

We note that the MOS device variations also impacts the robustness of MLC STT-RAM designs by causing the magnitude variations of the read and the write currents of the MTJ. In our reliability analysis of MLC STT-RAM, the parametric variability of MOS devices is represented by the variations of the current source output.

4.1.2 Thermal Fluctuations

The thermal fluctuations results in the randomness of the MTJ switching time. As we described in Section 2.3, in general, the impact of thermal fluctuations can be modeled by a normalized thermal induced random field. MTJ switching time becomes a distribution under the impact of thermal fluctuations. A write failure occurs when the MTJ switching time is longer than the write pulse width. The impact of thermal fluctuations is an accumulative effects and determined by

the length of the MTJ switching time. The reduction of switching current does not only prolong the MTJ switching time but also increases the ratio between the standard deviation and the mean value of the switching time [8], indicating a larger impact of thermal fluctuations. Hence, in MLC STT-RAM designs, the impacts of thermal fluctuations could be stronger than that in the SLC STT-RAM designs when the MTJ switching current density is lower than that of the SLC MTJ (e.g., during the soft-domain flipping in parallel MLC MTJs).

4.2 READABILITY ANALYSIS OF MLC MTJS

4.2.1 Nominal Analysis of the Readability of MLC MTJs

We assume that the resistances of the hard domain and the soft domain in a parallel MLC MTJ are R_1 and R_2, respectively. The corresponding the high and the low resistance states of the two domains are R_{1H}, R_{1L}, R_{2H}, and R_{2L}, respectively. The TMR ratio of each domain is defined as: $\frac{R_{iH}-R_{iL}}{R_{iL}}, (i = 1, 2)$. As aforementioned in Section 4.1.1, the two magnetic domains share the same magnetic structure and MgO layer within a small proximity. Thus, we can safely assume the RAs and the $TMRs$ of the two domains are the same, or $RA_{1j} = RA_{2j}, (j = HorL)$ and $\frac{R_{1H}}{R_{1L}} = \frac{R_{2H}}{R_{2L}}$. For the existing in-plane MTJ technology, the typical TMR ratio is $1 \sim 1.2$ [13]. Because the size of the hard domain is larger than that of the soft domain, we have $R_{1H} < R_{2H}$ and $R_{1L} < R_{2L}$. In the simulations in our work, we assume the surface area of the parallel MLC MTJ is a $45nm \times 90nm$ ellipse, which is the minimum shape that satisfies the shape anisotropy requirement [11, 28] and is allowed by the lithography constraint of 45nm CMOS fabrications process.

Sense margin is one of the major concerns in MLC STT-RAM designs because the resistance state distinction of the MTJ is partitioned into multiple levels. Read errors happen when the distributions of the two adjacent resistance states (i.e., 00 vs. 01, 01 vs. 10, and 10 vs. 11) overlap with each other, or the distinction between the two resistance states is smaller than the sense amplifier resolution. The reading error rate can be reduced by maximizing the distinctions between every two adjacent states. Without considering the process variations, the goal of the nominal design method of MLC STT-RAM cell is to maximize the distinctions between the designed values of every two adjacent resistance states.

In the real implementation of parallel MLC MTJs, $R_{00} = R_{1L}||R_{2L}$ and $R_{11} = R_{1H}||R_{2H}$ are fixed by the MTJ designs. The changes of R_{01} and R_{10} are not independent and determined by the partitioning of the free layer. If we assume the surface area of the parallel MLC MTJ is A and the surface area of the hard domain is A_1, we have:

$$R_{1L}\cdot A_1 = R_{2L}\cdot(A - A_1) = R_{00}\cdot A, R_{1H}\cdot A_1 = R_{2H}\cdot(A - A_1) = R_{11}\cdot A. \tag{4.1}$$

Here $A_1 > A/2$. The distinctions between every two adjacent resistance states can be calculated as:

$$D_{00-01} = R_{01} - R_{00} = \frac{TMR\cdot RA}{A}\cdot\frac{A-A_1}{A+A_1\cdot TMR} \tag{4.2}$$

$$D_{01-10} = R_{10} - R_{01} = \frac{[TMR\cdot(TMR+1)\cdot RA](2A_1-A)}{(A+TMR\cdot A_1)[TMR\cdot(A-A_1)+A]} \tag{4.3}$$

$$\begin{aligned} D_{10-11} &= R_{11} - R_{10} \\ &= \frac{TMR\cdot(TMR+1)\cdot RA}{A}\cdot\frac{A-A_1}{TMR\cdot(A-A_1)+A} \end{aligned} \tag{4.4}$$

We calculated the derivatives of D_{00-01}, D_{01-10}, and D_{10-11} with respect to A_1 and have: $\frac{dD_{00-01}}{dA_1} < 0$, $\frac{dD_{10-11}}{dA_1} < 0$, and $\frac{dD_{01-10}}{dA_1} > 0$ when $A_1 \in [A/2, A]$. In other words, D_{00-01} and D_{10-11} monotonically decrease when A_1 increases from $A/2$ to A and D_{01-10} monotonically increases in the same range. Also, since $A - A_1 < A_1$ and $TMR \geq 1$, D_{10-11} is always larger than D_{00-01} based on Eq. (4.2) and (4.4). Therefore, the optimal design of parallel MLC MTJs happens when $D_{00-01} = D_{01-10}$ or:

$$(TMR + 1)(\frac{R_{2L}}{R_{1L}})^2 - \frac{R_{2L}}{R_{1L}} = 2(TMR + 1) \tag{4.5}$$

Here $R_{1L}||R_{2L} = R_{00}$.

In a series MLC MTJ, the optimal MTJ design happens when $D_{00-01} = D_{01-10} = D_{10-11}$, or:

$$R_{1L} = \frac{1}{2}R_{2L} \tag{4.6}$$

Here R_{2L} is usually the low resistance state of the SLC MTJ with the minimum surface area (say, A). The optimal design parameters of a typical parallel MLC MTJ and a typical series MLC MTJ are: $RA = 20\Omega\mu A$, $TMR = 1.2$, The limitation sizes is 45nm×90nm.

4.2.2 Statistical Analysis of the Readability of MLC MTJs

All the optimizations in Section 4.2.1 are based on the nominal values of the device parameters of MLC MTJs. In this section, we will analyze the impacts of process variations on the readability of MLC STT-RAM cells.

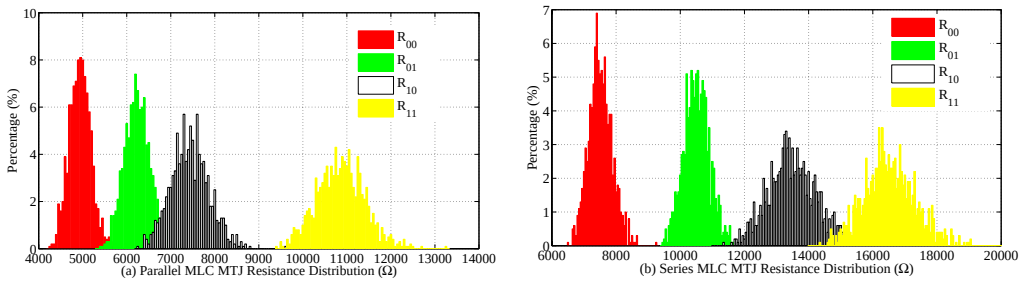

Figure 18: Four state resistance distributions of (a) Parallel MLC MTJ and (b) Series MLC MTJ, optimized by nominal design method.

Fig. 18(a) and Fig. 18(b) shows the distributions of the four resistance states in a parallel MLC MTJ and a series MLC MTJ, respectively. Both MTJs are optimized by using the nominal optimization method presented in Section 4.2.1. The standard deviations (1σ) of RA and TMR are 7% and 9%, respectively, based on the measurement data in [13]. In the nominal optimized parallel MLC MTJ, $\frac{R_1}{R_2} = 1.66$. In the nominal optimized series MLC MTJ, the surface area of the larger MTJ is 64nm×127nm, which corresponds to a low resistance state of $R_{2L} = 2500\Omega$. After the process variations are taken into account, the distributions of the resistance states overlap with each other, resulting in the read errors of the MLC MTJs. Because of the different deviations of every resistance state, the original nominal optimization that maximizes the distinctions between the nominal values of the adjacent resistance states is no longer able to guarantee the minimal overlaps between the adjacent resistance state distributions. A statistical optimization method is required for the minimization of the read error rate of MLC STT-RAM cells.

4.2.2.1 Optimization of Parallel MLC MTJs In our design, we assume the size of the parallel MLC MTJs is the same as the minimum size of the SLC MTJ or 45nm×90nm. The resistances of the two magnetic domains can be adjusted by changing the partition of the free layer. The surface

areas of the whole MTJ follows Gaussian distributions and the surface areas of the two magnetic domains follow a joint Gaussian distribution. To sense the four resistance states in a four-level parallel MLC MTJ, three reference resistances, i.e., R_I, R_{II}, R_{III}, are needed. The read error rates of reading R_{00}, R_{01}, R_{10} and R_{11} can be respectively expressed as:

$$\begin{aligned} P_{e00} &= P(R_{00} > R_I) \\ P_{e01} &= P(R_{01} < R_I) + P(R_{01} > R_{II}) \\ P_{e10} &= P(R_{10} < R_{II}) + P(R_{10} > R_{III}) \\ P_{e11} &= P(R_{11} < R_{III}) \end{aligned} \tag{4.7}$$

We note that the impacts of the read error rates of each resistance states are not accumulative in MLC STT-RAM designs: For a MLC STT-RAM cell, the highest read error rate is the maximum one of all resistance states, or, $P_e = Max(P_{e00}, P_{e01}, P_{e10}, P_{e11})$. To minimize the $P_{ei}, i = 00, 01, 10, 11$, the R_I, R_{II}, ideally, R_{III} must be selected at the cross point of the two adjacent distributions. In memory designs, P_e can be used to determine the required error tolerance capability. The read errors due to the MTJ resistance variations can be corrected or tolerated in the design practices by using error correction code (ECC) and design redundancy etc.

In Fig. 18(a), the overlaps of the resistance state distributions of the parallel MLC MTJ generate the read error rates of $P_{e00} = 0.73\%$, $P_{e01} = 6.44\%$, $P_{e10} = 6.05\%$ and $P_{e11} = 0.018\%$. High read error rates happen at R_{00} and R_{01}, which are incurred by the large overlaps between these two resistance states. Fig. 19(a) depicts read error rate under the different ratios of the nominal resistances of the two magnetic domains (R_2/R_1). P_{e11} is always lower than P_{e00} due to the bigger distinction between R_{10} and R_{11} compared to the one between R_{00} and R_{01}. Following the increase of R_2/R_1 from 1.6, both P_{e00} and P_{e11} increase, indicating the reduced distinction from the adjacent resistance states. However, the increase of R_2/R_1 decreases the P_{e01} and P_{e10} by raising the distinction between R_{01} and R_{10}. When $R_2/R_1 = 2.2$, the parallel MLC MTJ achieves its lowest maximum read error rate as $P_{e00} = 3.31\%$, $P_{e01} = 2.97\%$, $P_{e10} = 0.73\%$ and $P_{e11} = 0.23\%$. The change of the optimal R_2/R_1 ratios in the nominal and statistical optimizations comes from the correlation between the standard deviation and the nominal values of the MTJ resistance state: the higher resistance is, the larger standard deviation of the resistance will be [30].

4.2.2.2 Optimization of Series MLC MTJs

In series MLC MTJ, the serially connected SLC MTJs are fabricated separately. The parameters of these two MTJs are partially correlated due to the spatial correlations. The two resistance states of the small SLC MTJ with the minimum size are $R_{2L} = 5000\Omega$ and $R_{2H} = 11000\Omega$, respectively. The distinctions between two adjacent resistance states can be adjusted by changing the surface area of the large SLC MTJ.

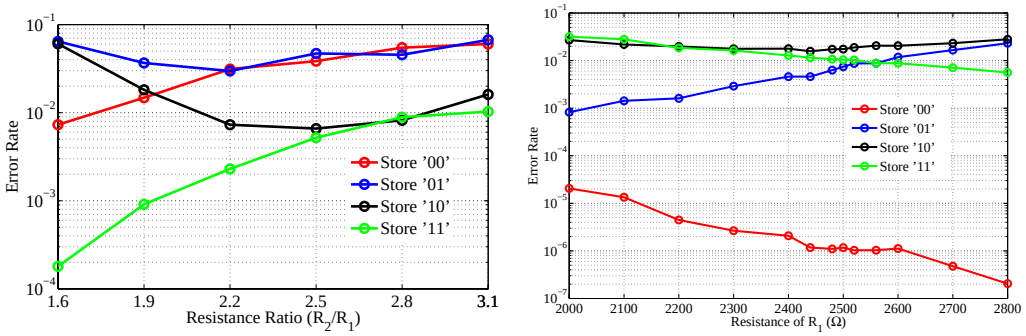

Figure 19: (a) Error Rate vs. R_2/R_1 Ratio Sweep, (b)Error Rate vs. Resistance of Hard Domain Sweep.

Fig. 19(b) shows the read error rates of the four resistance states of the series MLC MTJ when the size of the large SLC MTJ changes. The variation of the large SLC MTJ size is represented by its low resistance state(R_{1L}). The lowest maximum read error rate happens when $R_{1L} = 2440\Omega$, or the MTJ size is 64.5nm×129nm. It is very close to the result of the nominal optimization method – $R_{1L} = 2500\Omega$, or the MTJ size of 64nm×127nm. The corresponding read error rates of each resistance states are $P_{e00} = 0.000118\%$, $P_{e01} = 0.46\%$, $P_{e10} = 1.57\%$ and $P_{e11} = 1.15\%$. Compare to parallel MLC MTJs, series MLC MTJs demonstrated significantly lower read error rate under the same fabrication conditions. Although the read error rate has not achieved the commercial requirement yet, these results are still very encouraging.

4.3 WRITABILITY ANALYSIS OF MLC MTJS

In SLC MTJ designs, increasing the switching current density can effectively reduce the MTJ switching time and improve the write error rate of the SLC STT-RAM cell. In MLC MTJ designs, however, increasing the switching current when programming the MTJ to an intermediate resistance state may overwrite the MTJ to the next resistance level. The thermal fluctuations further complicate the situations of MLC MTJ programming by incurring the additional variability of MTJ switching time. In this section, we will discuss the impacts of these variations and the multi-level programming mechanisms on the writability of the MLC MTJs.

4.3.1 Write Mechanism of MLC STT-RAM Cells

The write operation of a MLC STT-RAM cell is much more complex than that of a SLC STT-RAM cells – Both the polarizations and the amplitude of the switching current must be carefully tuned according to the current and the target resistance states, it need a different directions as the single level STT-RAM do, the amplitudes of it should also be differential for 2 bit writing.

The write scheme of parallel MLC MTJs has been discussed in [6]; In general, the soft domain can be switched by a small current (density) while the hard domain must be switched by a relatively large current (density). It means that the soft domain can be switched alone but the hard domain switching is always associated with the soft domain switching *if the original magnetization directions of the two domains are the same*. Hence, some resistance state transitions require two switching steps. For example, when a parallel MLC MTJs switches from R_{00} to R_{10}, a large current is applied first to switch the MTJ from R_{00} to R_{11}. Then a small current is applied to complete the transition from R_{11} to R_{10}.

For easy analysis, we assume that the bits of a MLC MTJ from '00' to '11' follow the resistance value from low to high. As summarized in [5], the transitions of the MTJ resistance states can be classified into three types:

1. Soft transition (ST), which switches only the soft domain in a parallel MLC MTJ or the small SLC MTJ in a series MLC MTJ;
2. Hard transition (HT), which switches the both domains in a parallel MLC MTJ or both SLC MTJs in a series MLC MTJ to the same magnetization direction;
3. Two-step transition (TT), which utilizes two steps to switch the MLC MTJ to the target resistance states, i.e., one HT followed by one ST.

4.3.2 Impacts of Thermal Fluctuations

We define the threshold switching current (density) as the minimal current (density) required to switching a MTJ within a switching time. The relationship between the magnetization switching time (t_w) and the nominal value of the threshold switching current density (J_C) can be divided in three working regions [25]. When $t_w < 10ns$, the reduction of t_w requires the dramatic increase of the J_C. Also, due to the asymmetry of MTJ switching, the threshold switching current density of writing '1' is usually larger than that of writing '0' [39].

The thermal fluctuation demonstrates different impacts on the MTJ switching performance in the different working regions: For a low switching current density or a $T_w > 10ns$, the thermal fluctuation is dominated by the thermal component of internal energy; the MTJ switching time follows a Poisson distribution. For a high switching current density or a $T_w < 3ns$, the thermal fluctuation is dominated by the thermally active initial angle of procession; the MTJ switching time follows a Gaussian distribution [8]. The distribution of the MTJ switching time in the middle of these two regions follows a combination of the two distributions. In the write operations of MLC STT-RAM, the two parts of the MLC MTJs, i.e., the two magnetic domains in the parallel MLC MTJ or the two SLC MTJs in the series MLC MTJ, may experience different switching current densities, thermal fluctuations and even different threshold current densities (mainly exist in the parallel MLC MTJs). The MTJ switching could ends up with multiple possible resistance states with different probabilities, as we shall show in following sections.

4.3.3 Write Operations of Parallel MLC MTJs

During the write operations of parallel MLC MTJs, the voltage (V) applied to the two terminals of the two magnetic domains are the same. For each domains, the switching current density has:

$$J_i = \frac{V}{R_i \cdot A_i} = \frac{V}{\frac{RA_i}{A_i} \cdot A_i} = \frac{V}{RA_i}, i = 1, 2. \tag{4.8}$$

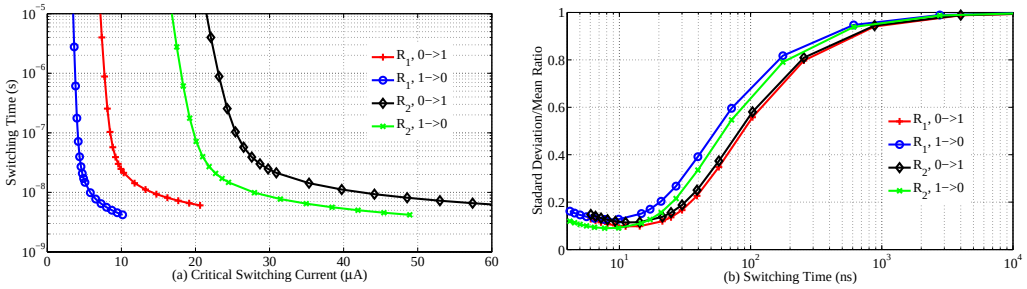

Figure 20: Switching properties of the two domains for a parallel MLC MTJ. (a) switching time vs. switching current. (b) switching time standard deviation vs. switching current.

It shows that after V is fixed, the switching current density through each domain is uniquely determined by the RA of the domain. Here $RA_i = RA_L$ or $RA_L \cdot (TMR + 1)$ for the low- or the high-resistance state, respectively. RA_L is the RA of the low resistance state. As we discussed in Section 4.1.1, the two magnetic domains of a parallel MLC MTJ have the exactly same RA when they are in the same resistance state. In such a case, the two magnetic domains have the the same current density. However, if the two domains are in the opposite resistance states, the current densities of them will be different.

Fig. 20(a) shows our simulation results of the relationships between the T_w and J_C for the two domains in a typical parallel MTJ. The MTJ parameters are scaled from the measured data of a 90×180nm elliptical MTJ device in [19]. Two domains demonstrate different J_C even under the same T_w due to the different shape anisotropy's etc. The write asymmetry is also observed in the result, i.e., the J_C of '0'→'1' transition of the magnetic domain is always higher than that of '1'→'0' transition for the same T_w. The relative deviations of the T_w of the two magnetic domains at the whole working region are shown in Fig. 20(b).

During the write operations of parallel MLC STT-RAM cells, the write current must be applied to switch only the domain(s) that need(s) to be flipped. However, the variability in the magnetization switching of the two domains can introduce write errors. Different from the SLC MTJ where the write error is only incurred by incomplete switching, the writing errors of the parallel MLC MTJ come from either the incomplete switching of the target domains (incomplete write) or overwriting the other domain to an undesired resistance state (overwrite). In a HT transition, only incomplete writes will happen because the write operations require either both domains flip together or only the hard domain flips if the soft domain has already been in the target resistance state. In such a case, increasing the switching current can effectively improve the switching performance of both domains and suppress the write error rate. In a ST transition, the situation can be divided into two scenarios: 1) If the destination resistance state is boundary state, i.e., R_{00} and R_{11}, then only incomplete write failures are possible; 2) If the destination resistance state is intermediate state, i.e., R_{01} and R_{10}, then both incomplete write and overwrite failures may occur. An appropriate switching current must be selected to achieve a low combined writing error rate. We denote the transitions in 2) as "dependent" transitions and the transitions in 1) and HT transitions as "independent" transitions.

Monte-Carlo simulations are conducted to evaluate the write error rates of the dependent transitions, i.e., 00 $\rightarrow$ 01 or 11 $\rightarrow$ 10, as shown in Fig. 21. Here we assume the MTJ switching current is supplied by an adjustable on-chip current source, whose output magnitude has an intrinsic standard deviation of 2% of the nominal value [15]. For a 10ns write pulse width, the optimal switching current for the transitions of '00'$\rightarrow$'01' and '11'$\rightarrow$'10' are 46.5μA and 49.9μA, respectively. Fig. 21 also shows the changes of incomplete and overwrite errors over the whole simulated range. When the switching current decreases from the optimal value, the incomplete writes start to dominate the write errors; When the switching current increases from the optimal value, the overwrite errors of the hard domain start to dominate the write errors. Nonetheless, the error rates of the two dependent transitions are still high (8.2%), indicating a large overlap area between the threshold switching current distributions of the hard domain and the soft domain.

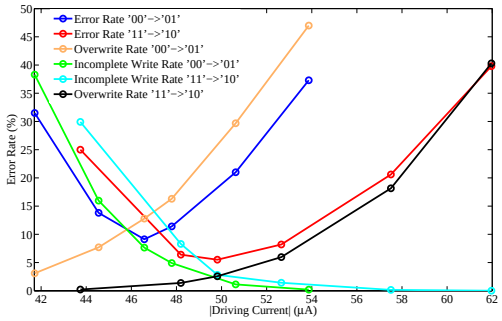

Figure 21: Writing error rate in parallel MLC STT-RAM cell at T_w = 10ns. Notes: The total error rate is not necessarily equal to the sum of incomplete error and overwrite error, which are the errors overwriting the hard domain or incurring the incomplete soft domain flipping, respectively.

Fig. 22(a) shows the write error rates of the dependent transitions of the parallel MLC MTJ at different switching currents when T_w = 3ns, 10ns, and 100ns, respectively. The lowest write error rate is achieved at T_w = 3ns. It is because that when T_w reduces, the required MTJ switching current increases. The impact of the thermal fluctuations on the MTJ switching is suppressed and the distributions of the T_w are compressed. This fact indicates that the parallel MLC MTJ better work at a fast working region to minimize the write error rate.

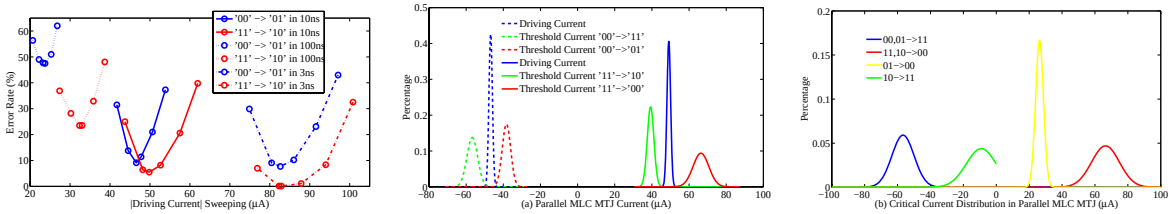

Figure 22: (a)Writing error rate in a parallel MLC STT-RAM cell at different T_w, Threshold current distributions of resistance state trasitions for the parallel MLC MTJ.(b) Dependent transitions. (c) Independent transitions.

We can also map the uncertainties in the switching time of the parallel MLC MTJ under the fixed switching current into the distributions of the required switching currents for fixed switching time. Fig. 22(b) shows the distributions of the threshold switching current of the dependent transitions for the parallel MLC MTJ at a 10ns write pulse width. The distributions of the MTJ write

current supplied by the on-chip current source are also depicted. Take the transition of '00'→'11' as an example, a write current is selected between the threshold current distributions of the transitions of '00'→'01' and '00'→'11'. The two types of write errors, including incomplete write and overwrite, are represented by the overlap between the distributions of the write current and the threshold switching current of '00'→'01' and the overlap between the distributions of the write current and the threshold switching current '00'→'11', respectively. Fig. 22(c) shows the distributions of the threshold switching current of the independent transitions for the parallel MLC MTJ at a 10ns write pulse width. Since only the target magnetic domain will flip during the independent transitions, a sufficiently large write current can be always applied to suppress the incomplete write errors without incurring any overwrite errors.

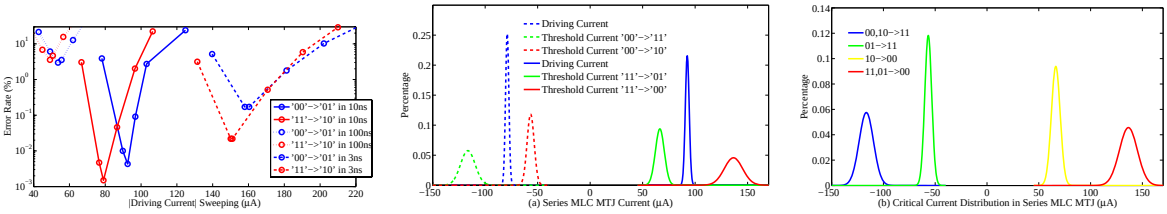

Figure 23: (a)Writing error rate in a series MLC STT-RAM cell at different T_w, Threshold current distributions of resistance state transitions for the series MLC MTJ.(b) Dependent transitions. (c) Independent transitions.

Similar to the distributions of the MTJ switching time, the distributions of the threshold switching current of the parallel MLC MTJ are also dependent on the working regions of the MTJ. After the distributions of the switching current of the resistance state transitions are obtained, the optimal write current can be derived as Fig. 22(a).

4.3.4 Write Operations of Series MLC MTJs

In a series MLC MTJ, the magnitudes of the currents passing through the two SLC MTJs are the same. However, the applied current densities on the two SLC MTJs are different and determined by the different surface areas of them. In Section 4.2.2.2, the analysis on the read reliability of the series MLC MTJs shows that the optimal surface area ratio between the two MLC MTJs is around

2, or 45nm×90nm and 64.5nm×129nm at 45nm technology node. In our simulations, we also assume the two SLC MTJs maintain the same aspect ratios and were fabricated under the same conditions. Thus, they have the same switching properties, i.e., the same relationships between threshold switching current density and the switching time. Again, the switching current density on each SLC MTJ is controlled by the on-chip write current source.

Fig. 23(a) shows the write error rates of the dependent transitions of the series MLC MTJ under different switching currents for a 10ns write pulse width. The optimal switching current for the transitions of '00'→'10' and '11'→'01' are 79.0μA and 92.5μA, respectively. Compared to parallel MLC MTJs, the write error rates of the dependent transitions are significantly reduced: the minimum write error rates of the transitions of '00'→'10' and '11'→'01' are only 0.0015% and 0.0043%, respectively. The improvement of the write reliability is because of the larger distinction between the threshold switching current distributions of the dependent transition and the adjacent resistance state transition, as shown in Fig. 23(b). For comparison purpose, the results of the independent resistance state transitions are shown in Fig. 23(c).

Fig. 23(a) also shows the write error rates of the dependent transitions of the serial MLC MTJ at different switching currents when T_w = 3ns and 100ns, respectively. Similar dependency of the write error rate on the MTJ working region is observed. Interestingly, the minimum write error rate occurs when T_w = 10ns, since the standard deviation/mean ratio reaches its minimum value (see Fig. 23(a)). Compared to parallel MLC MTJs, series MLC MTJs demonstrate much higher write reliability at the same technology node,, while requiring slightly larger switching current and higher write energy consumption.

5.0 DIFFERENTIAL SENSING SCHEME TO IMPROVE THE READ PERFORMANCE OF STT-RAM

5.1 MOTIVATION

Previous conventional wisdom for STT-RAM is that writes are slower and require more power than their conventional SRAM counterparts. Several architectural solutions such as hybrid caches with fast and slow writing memory components [35, 18], various methods or preempting, avoiding, and bypassing writes [41, 10, 24], and leveraging the asymmetry of writing different logic values [24] have been proposed to mitigate the write performance problem. However, due to scaling effects, performance and reliability of STT-RAM reads, not writes will become the ultimate bottleneck at technologies of 45nm and below. Read performance, the dominant operation in caches [1], suffers from increased sense amplifier delays for detecting increasingly small sense margins and higher read error rates. In contrast, due to reduced energy barriers at smaller technology nodes, writes will become faster at lower energy, although this leads to higher susceptibility to read disturbance (inadvertent writes from applying a read current).

5.2 ADAMS TECHNOLOGY

By examining the pros and cons of the existing STT-RAM cell structures, we are able to propose *ADAMS – Asymmetric Differential STT-RAM Cell Structure* which can substantially promote the robustness and performance of STT-RAM designs. In this section, we will illustrate the cell structure of ADAMS and discuss its read and write operations.

5.2.1 Regular Differential Sensing Scheme (RDAMS)

During read operations of an 1T1J cell, a sensing current is injected into the cell while the generated voltage on the bit-line (BL) is compared to a reference level. The maximum sense margin is only $\frac{1}{2} \cdot (R_{high} - R_{low})$. Here R_{high} and R_{low} denote the high- and the low-resistance state of the MTJ, respectively. To further improve the readability of STT-RAM cells, differential sensing scheme may be applied, as shown in Fig. 24(a). A complete differential STT-RAM cell includes two separate 1T1J cells, which can be referred to 'positive' cell (P-cell) and 'negative'cell (N-cell), respectively. The resistance states of these two cells are always opposite, say, the one in the P-cell is high and the one in the N-cell is low for storing '1'. We refer to this design as regular differential STT-RAM cell structure (RDAMS). During the read operation, the sensing currents with the same magnitude are injected into both P-cell and N-cell and the generated voltages on each bit-line will be compared. The corresponding maximum sense margin is $(R_{high} - R_{low})$, which is doubled from the one of 1T1J cell. Both the read latency and the device variation tolerance of the STT-RAM cell are improved. Obviously, the capacity of RDAMS is only half of the one of 1T1J cell.

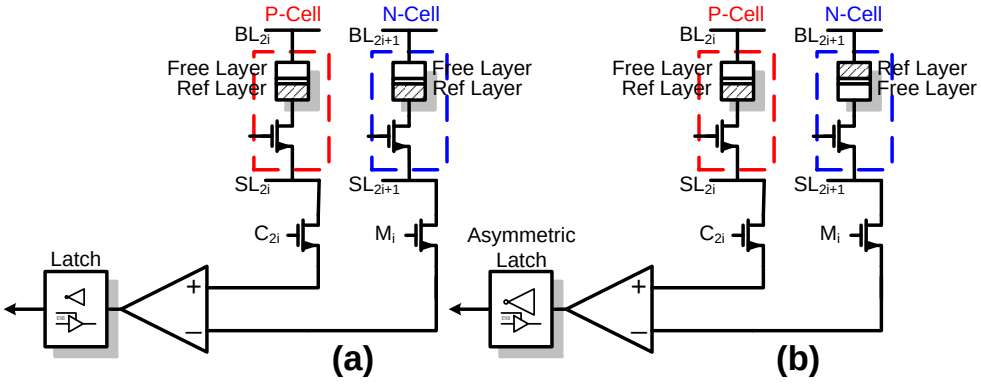

Figure 24: Structure of (a) RDAMS. (b) ADAMS.

However, RDAMS aggravates the read disturbance issue: Between the P-cell and the N-cell, there is always one has the chance to be flipped by the sensing current regardless the value of the data stored in the RDAMS cell. Also, compared to 1T1J cell, the write error rate of the RDAMS cell is doubled as both MTJs must be successfully programmed in one correct write operation. Note that the write performance of a RDAMS cell is limited by the longest write latency between the P-cell and the N-cell which always switch at the opposite directions.

5.2.2 Asymmetric Differential Cell Structure (ADAMS)

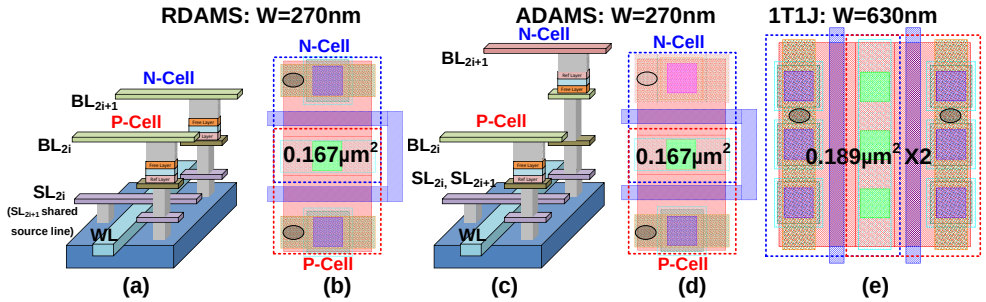

Figure 25: (a) 3D view of RDAMS. (b) Layout of RDAMS. (c) 3D view of ADAMS. (d) Layout of ADAMS.(e) layout of 1T1J.

Fig. 24(b) shows the schematic of an ADAMS cell. The MTJ in the P-cell is reversely connected to the NMOS transistor. In the implementation of ADAMS, the MTJ in the N-cell can be prepared at different layer from the one in the P-cell. Fig. 25(a)–(d) shows the 3D-views and layouts of RDAMS and ADAMS cells at 45nm technology. The width of the NMOS transistors is set to 270nm. For comparison purpose, we also include the layout of an 1T1J cell where the NMOS transistor channel width is 630nm, as shown in Fig. 25(e). In all designs, the channel lengths of the NMOS transistors keep minimum (45nm). The resistance states of the P-cell and N-cell in an ADAMS cell are also always opposite, maintaining the same sense margin as that of an RDAMS cell. However, ADAMS has some interesting characteristics which are different from RDAMS.

5.2.3 Read and Write Robustness of ADAMS

5.2.3.1 Read robustness Different from RDAMS where the read disturbance could happen when sensing the data of any values, ADAMS limits the occurrence of the read disturbance only when '1' is sensed: Assuming the sensing current is applied from BL to SL during the read operation, reading '0' in ADAMS is read-disturbance-free as the P-cell stores '0' and the N-cell stores '1' ($\bar{0}\bar{1}$). Here we use $\bar{x}\bar{y}$, x, $y = 0$ or 1 to denote the ADAMS state where x and y are the stored bit of the P-cell and the N-cell, respectively. When the ADAMS state is $\bar{1}\bar{0}$ ('1'), read disturbance could happen on both P-cell and N-cell. However, the probability that read disturbance simulta-

neously happens at both cells is usually very low. Then the final states of an ADAMS cell after read disturbance occurs are most likely $\bar{0}\bar{0}$ and $\bar{1}\bar{1}$, neither of which is a valid state during normal operations. Thus, if we encounter an invalid ADAMS state (i.e., $\bar{0}\bar{0}$ and $\bar{1}\bar{1}$), we may assume the original ADAMS cell state is $\bar{1}\bar{0}$ ('1').

5.2.3.2 Write robustness In the write operation of an ADAMS cell, the possibility that both P-cell and N-cell are unsuccessfully programmed is also very low. When the write error happens in only one cell, the final state of the ADAMS cell will stop at $\bar{0}\bar{0}$ or $\bar{1}\bar{1}$. In such a case, we may not be able to directly figure out the original and target state of the write operation of the ADAMS cell because the incomplete write can happen in either P-cell or N-cell. However, as we shall show in Section 5.3.1.3, if we assume the invalid final states $\bar{0}\bar{0}$ and $\bar{1}\bar{1}$ always considered as the target state of $\bar{1}\bar{0}$ (writing '1'), the write performance and reliability of the ADAMS cell can be substantially improved.

5.2.4 Asymmetric SenAmp and Latch Design

A sense amplifier (SenAmp) is used in RDAMS or ADAMS to compare the resistance difference between the P-cell and the N-cell. However, if the two cells store the same value, i.e., $\bar{0}\bar{0}$ and $\bar{1}\bar{1}$, the SenAmp may not be able to output a stable result due to the small signal difference at its inputs. As we discussed in Section 5.2.3.1, if ADAMS can output the same results for the invalid ADAMS states $\bar{0}\bar{0}$ and $\bar{1}\bar{1}$ as that for $\bar{1}\bar{0}$, then the majority of the read disturbance errors can be hidden. To realize this function, we propose the following Asymmetric SenAmp and latch designs:

5.2.4.1 Asymmetric SenAmp As shown in Fig. 26(a), we carefully increase the sizes of the PMOS transistors PMA and PMB in the SenAmp. The enhanced driving abilities of PMA and PMB will pull up the OUT signal at the beginning of the sensing process. If the ADAMS cell is in a valid state, i.e., $\bar{0}\bar{1}$ or $\bar{1}\bar{0}$, the *Out* signal will quickly reach Ground or Vdd, respectively; If the ADAMS cell is in an invalid state, i.e., $\bar{0}\bar{0}$ or $\bar{1}\bar{1}$, the *Out* signal will gradually approach Ground or Vdd depending on the relative small voltage level difference at the inputs. In this case, however, the jump-up of the *Out* signal at the beginning will delay its decay to Ground. Since the decay of

the *Out* signal of sensing $\bar{0}\bar{0}$ or $\bar{1}\bar{1}$ is normally slower than that of sensing $\bar{0}\bar{1}$, we may be able to differentiate these two cases by carefully choosing the cutoff point.

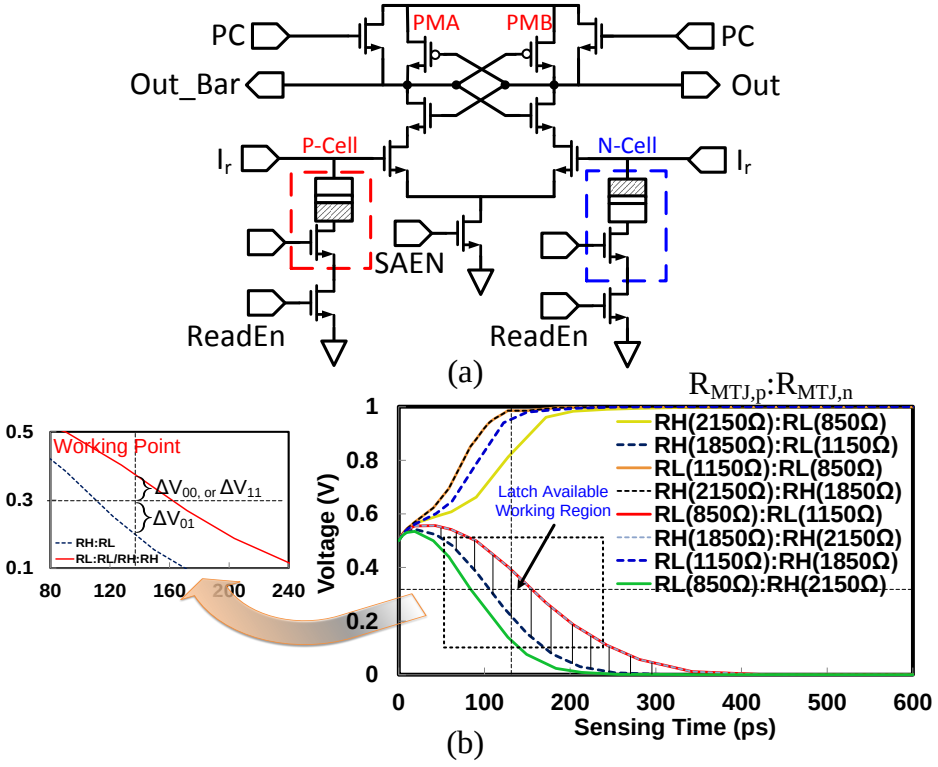

Figure 26: (a) Asymmetric sense amplifier (SenAmp) design. (b) Simulation results of SenAmp Out signal at different corner cases.

Fig. 26(b) illustrates the SPICE simulation results of the *Out* signal of our asymmetric SenAmp design. We use $R_{mtj,p}$ and $R_{mtj,n}$ to denote the resistance of the MTJs in the P-cell and N-cell, respectively. We also assume the nominal values of R_{high} and R_{low} are 1000Ω and 2000Ω, respectively, and their standard deviations are both 5%. The SenAmp is designed with PTM 45nm technology [3]. We simulated the *Out* signals at $\pm 3\sigma$ corners of all possible ADAMS states. The simulation results show that at the valid ADAMS states like $\bar{1}\bar{0}$ and $\bar{0}\bar{1}$, the *Out* signal always quickly reaches Vdd or Ground, respectively, at all corners. At the invalid ADAMS states like $\bar{0}\bar{0}$ and $\bar{1}\bar{1}$, the *Out* signal ends up with Vdd if $R_{mtj,p} > R_{mtj,n}$. When $R_{mtj,p} < R_{mtj,n}$, the *Out* signal slowly decay to ground. However, the difference between the worst-corner *Out* signal in such a case (i.e., $R_{mtj,p} - R_{mtj,n} = 300\Omega$) and that of the state $\bar{0}\bar{1}$ (i.e., $R_{mtj,p} = 1150\Omega$ and $R_{mtj,n} = 1850\Omega$) becomes the possible working region for the output latch to differentiate $\bar{0}\bar{0}/\bar{1}\bar{1}$ and $\bar{0}\bar{1}$.

5.2.4.2 Asymmetric Latch Fig. 27(a) shows the schematic of our asymmetric latch design for ADAMS. The forward inverter has a small size PMOS transistor (PM0) and a large size NMOS transistor (NM0) while the feedback tristate inverter has large size PMOS transistors (PM1 and PM2) and small size NMOS transistors (NM1 and NM2). The unbalanced driving ability between NMOS and PMOS transistors creates a working point for latching '0' and '1' below $\frac{Vdd}{2}$.

Fig. 27(b) shows the simulated worst-case results of the asymmetric latch at the ADAMS states of $\bar{0}\bar{0}$ and $\bar{1}\bar{0}$. The working point of the asymmetric latch is designed to 0.3V. The output of the asymmetric SenAmp is captured at 135ps. As shown in the microscope view in Fig. 26(b), at this time the *Out* signal of the SenAmp of the ADAMS state $\bar{0}\bar{1}$ is below 0.3V while that of the invalid states $\bar{0}\bar{0}$ and $\bar{1}\bar{1}$ are above 0.3V. The generated sense margins, i.e., ΔV_{00}, ΔV_{11} and ΔV_{01}, ensure that the states $\bar{0}\bar{0}$ and $\bar{1}\bar{1}$ are detected as '1' while the state $\bar{0}\bar{1}$ is detected as '0'.

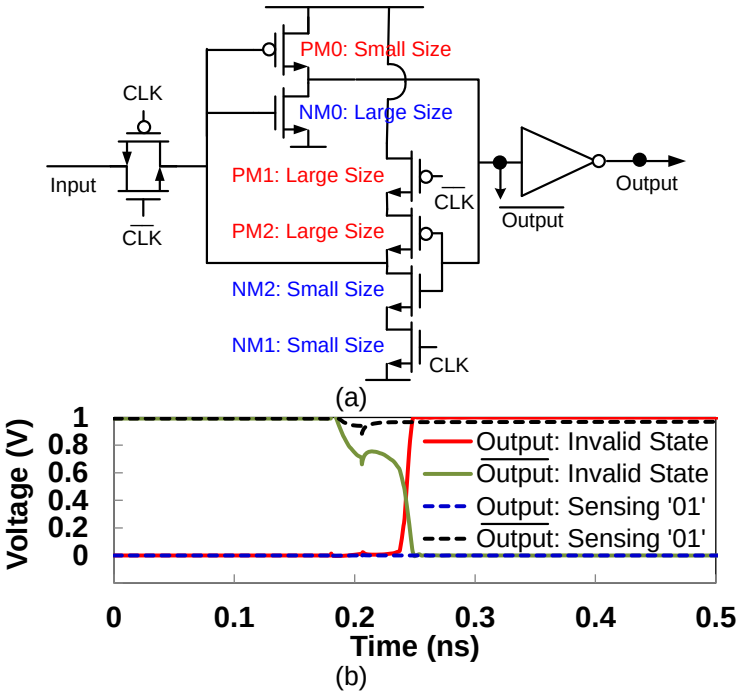

Figure 27: (a) Circuit of Asymmetric Latch. (b) Asymmetric Latch Output Results.

5.2.5 Reconfigurable Scheme STT-RAM

Another advantage of ADAMS is that it can be dynamically reconfigured into two independently functional 1T1J cells in case that the memory capacity is critical. As shown in Fig. 28, the operation of an ADAMS cell can be switched between two modes: high-reliable (HR) mode and high-capacity (HC) mode. A multiplexer is used to select the reference signal of an 1T1J cell from either external or its complimentary 1T1J cell (where the MTJ is reversely connected) depending on the operation mode. The performance, reliability and capacity of the STT-RAM can be flexibly adjusted by switching between the HR and HC modes.

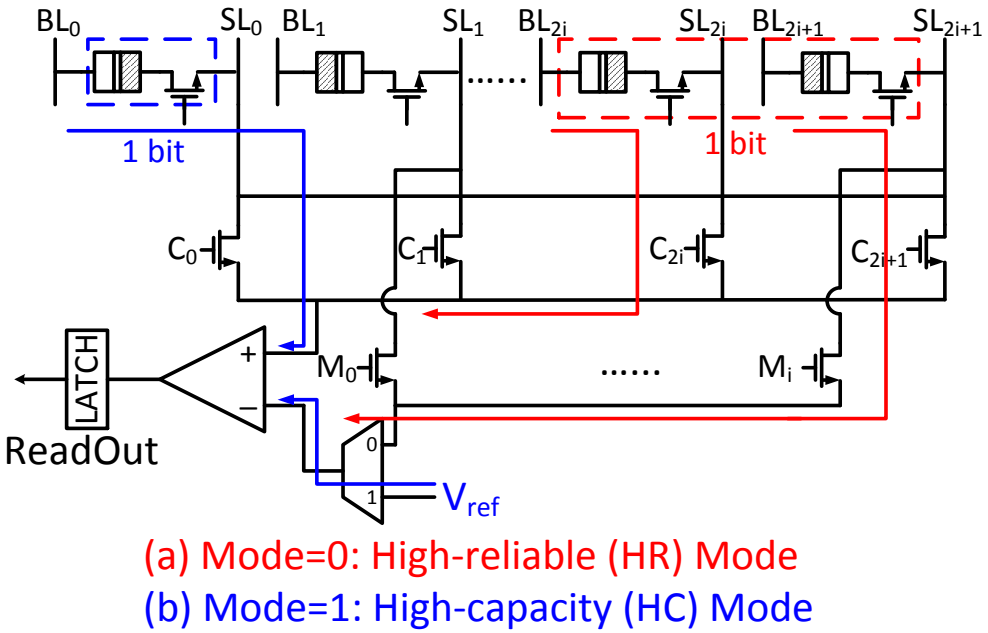

Figure 28: Reconfigurability of ADAMS. Mode = 0: High-reliable (HR) mode; Mode = 1: High-capacity (HC) mode.

5.3 ADAMS DESIGN OPTIMIZATION AND ANALYSIS

5.3.1 Write Operation Analysis

5.3.1.1 Asymmetric Write Analysis Fig. 29 shows the relationship between the MTJ switching current and switching time, including both $1 \rightarrow 0$ and $0 \rightarrow 1$ switching's. The data comes from a $45nm \times 90nm$ elliptical MTJ device model, which have been calibrated with the measurement of a real fabricated device from a leading magnetic recording company. Following the decrease in MTJ switching time, the difference between the nominal values of the required MTJ switching current at two switching directions becomes more and more significant. Fig. 29(b) shows the SDMR (Standard Deviation and Mean Ratio) of different MTJ switching times. In general, the MTJ switching time at $0 \rightarrow 1$ switching suffers from a larger variation than $1 \rightarrow 0$ switching.

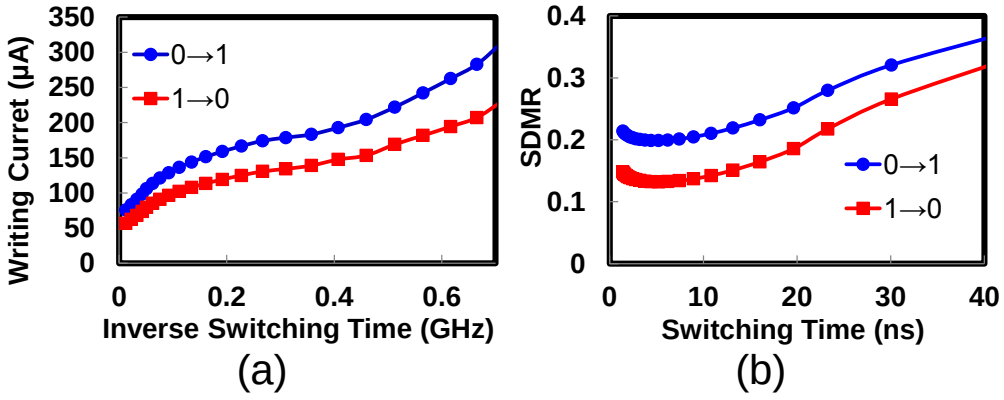

Figure 29: (a) Switching current vs. Inverse of switching Time. (b) Switching time mean vs Standard deviation and mean ratio (SDMR).

As aforementioned in Section 3.1, the MTJ switching current during the write operations is determined by the bias conditions of the NMOS transistor as well as the process variations of the NMOS transistor and the MTJ. We conduct SPICE Monte-Carlo simulations to obtain the MTJ switching current and its distribution at different NMOS transistor sizes and bias conditions. The device parameters adopted in the simulations are summarized in TABLE 4.

Fig. 30 shows the simulation results of the MTJ switching current in P-cell and N-cell at different NMOS transistor sizes and switching directions. The reliability of different cell structures is

limited by the different switching directions e.g., $0 \rightarrow 1$ in P-cell and $1 \rightarrow 0$ in N-cell, respectively. Also, the limiting switching direction always suffers from a larger SDMR than the other switching direction.

Table 4: Summary of Device Parameters

Device	Parameters	Mean	Std. Dev.
Transistor	Channel Length L	45nm	$5\%{\cdot}F$[1]
	Channel Width W	design dependent	$5\%{\cdot}F$
	Threshold Voltage V_{th}	0.466V	30mV
MTJ	Low Resistance R_l	1000Ω	$5\%{\cdot}mean$
	High Resistance R_h	2000Ω	

[1] $F = 45nm$.

5.3.1.2 Definition of Write Error Rate Fig. 31 shows the data storage states of 1T1J, RDAMS and ADAMS and the transitions between different states. The blue circles denote the states storing '0' while the red circles denote the states storing '1'. The black circles denote the prohibited states during the operation and directly correspond to an error. A successful write is defined as the transition between a blue circle and a red circle, and shown as a solid line; A unsuccessful write is defined as the transition between two states marked with the same color, or ending with a prohibited state. The occurrence of an unsuccessful write indicates an write error.

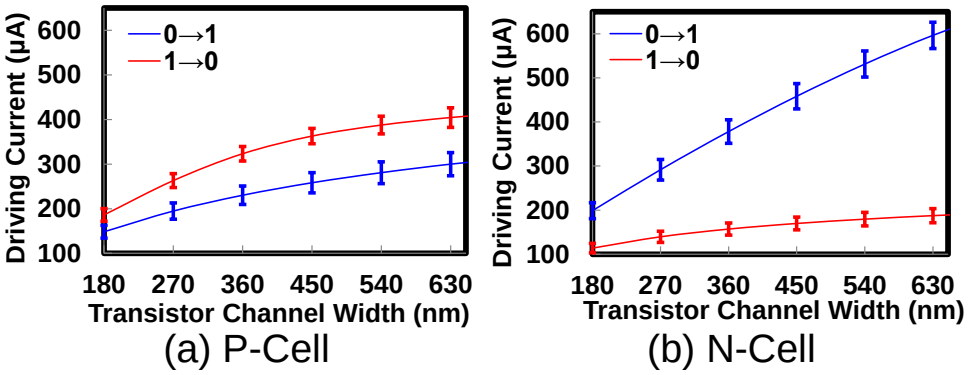

Figure 30: MTJ switching current vs. NMOS transistor size. (a) P-cell. (b) C-cell.

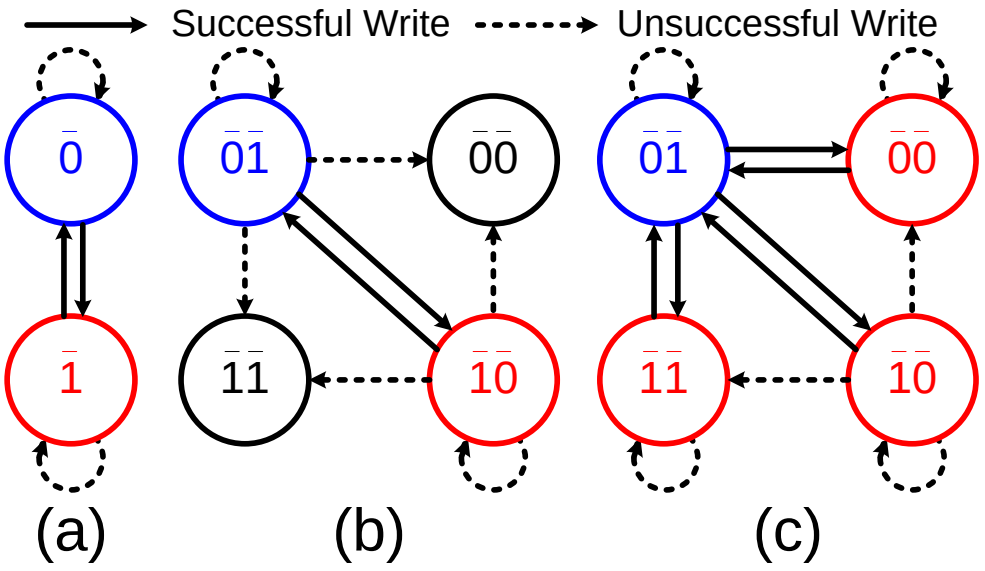

Figure 31: STT-RAM writing state. (a) 1T1J. (b) RDAMS. (c) ADAMS

In an 1T1J cell, the write error rate P_{WF} at different switching directions can be defined as the probability that the MTJ switching time τ is longer than the write pulse width T_W, or:

$$\begin{aligned} P_{WF,0\to1} &= P(\tau_{0\to1} > T_{W0}) \\ P_{WF,1\to0} &= P(\tau_{1\to0} > T_{W1}) \end{aligned} \tag{5.1}$$

Similarly, the write error rates of P-cell and N-cell at different switching directions can be summarized as:

$$\begin{aligned} P^i_{WF,0\to1} &= P(\tau^i_{0\to1} > T^i_{W0}) \\ P^i_{WF,1\to0} &= P(\tau^i_{1\to0} > T^i_{W1}), i = p \text{ or } n. \end{aligned} \tag{5.2}$$

Here the superscripts 'p' and 'n' denote the parameters for P-cell and N-cell, respectively. Fig. 32(a) and (b) shows the write error rates of the P-cell and N-cell at different switching directions when the NMOS transistor size changes. When the NMOS transistor is small, the write error rates at different switching directions are close. However, when the NMOS transistor size grows, the gap between the write error rates at different switching directions quickly increases. Nonetheless, the limiting switching directions of each cell structure, i.e., $0 \to 1$ in P-cell and $1 \to 0$ in N-cell, suffer from much higher write error rate than other switching directions.

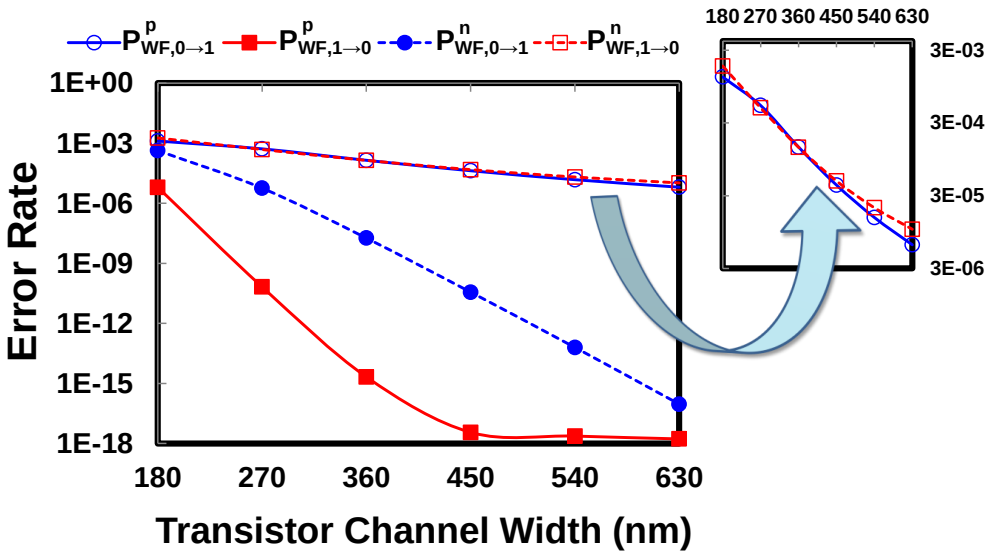

Figure 32: Write error rate at 10ns write pulse width.

The switching probabilities of the transitions from state $\bar{x}_s\bar{y}_s$ to state $\bar{x}_e\bar{y}_e$ in a RDAMS cell and an ADAMS cell can be respectively represented by: $P^R_{\bar{x}_s\bar{y}_s\to\bar{x}_e\bar{y}_e}$, and $P^A_{\bar{x}_s\bar{y}_s\to\bar{x}_e\bar{y}_e}$. Here superscript 'R' and 'A' denote the parameters belonging to RDAMS and ADAMS, respectively. Hence, the total write error rate of a RDAMS cell can be calculated by:

$$\begin{aligned} P^R_{WF} &= \alpha\beta\times P^R_{10\to00}\bigcup P^R_{10\to11} \\ &\quad +(1-\alpha)(1-\beta)\times P^R_{01\to00}\bigcup P^R_{01\to11}. \end{aligned} \tag{5.3}$$

Here α is the probability of the memory cell storing '1' ($\bar{1}\bar{0}$), β is the probability that the memory cell will be programmed to '0' ($\bar{0}\bar{1}$).

5.3.1.3 Write Optimization of ADAMS As discussed in Section 5.2.2, the states $\bar{0}\bar{0}$, $\bar{1}\bar{1}$ and $\bar{1}\bar{0}$ are all detected as '1' in ADAMS cell designs. Therefore, a correct output can be still read out even only one of the P-cell and the N-cell is successfully programmed during writing '1' ($\bar{0}\bar{1}\to\bar{1}\bar{0}$). The rite error rate of an ADAMS cell can be expressed as:

$$\begin{aligned} P^A_{WF} &= \alpha\beta\times P^A_{10\to00}\bigcup P^A_{10\to11} \\ &\quad +(1-\alpha)(1-\beta)\times P^A_{01\to00}\bigcap P^A_{01\to11}. \end{aligned} \tag{5.4}$$

Comparing Eq. (5.4) to Eq. (5.3) we found that, the write error rate contributed by writing '1', which is the second item at the right side of the equations, is significantly reduced in ADAMS.

Fig. 33 shows the write error rates when writing '0' and '1' for a RDAMS cell and an ADAMS cell, which are respectively denoted by $P^{u}_{WF,1\to0}$ and $P^{u}_{WF,0\to1}$, ($u = R$ or A), at different write pulse widths. As the NMOS transistor size increases, all the write error rates reduce though $P^{u}_{WF,1\to0}$ decrease faster than $P^{u}_{WF,0\to1}$. The highest write error rate is $P^{R}_{WF,0\to1}$, which dominates the write error rate of the RDAMS cell. When the write pulse width is long, (i.e, 10ns and 8ns, as shown in Fig. 33(a) and (b),) $P^{A}_{WF,1\to0}$ and $P^{A}_{WF,0\to1}$ in turn dominate the write error rate of the ADAMS cell when the transistor size is small and large, respectively. When shortening the write pulse width (e.g, 3ns), however, $P^{A}_{WF,0\to1}$ dominates the write error rate over the whole transistor size range, as shown in Fig. 33(d). It indicates a higher sensitivity of the error rate of writing '1' to the write pulse width.

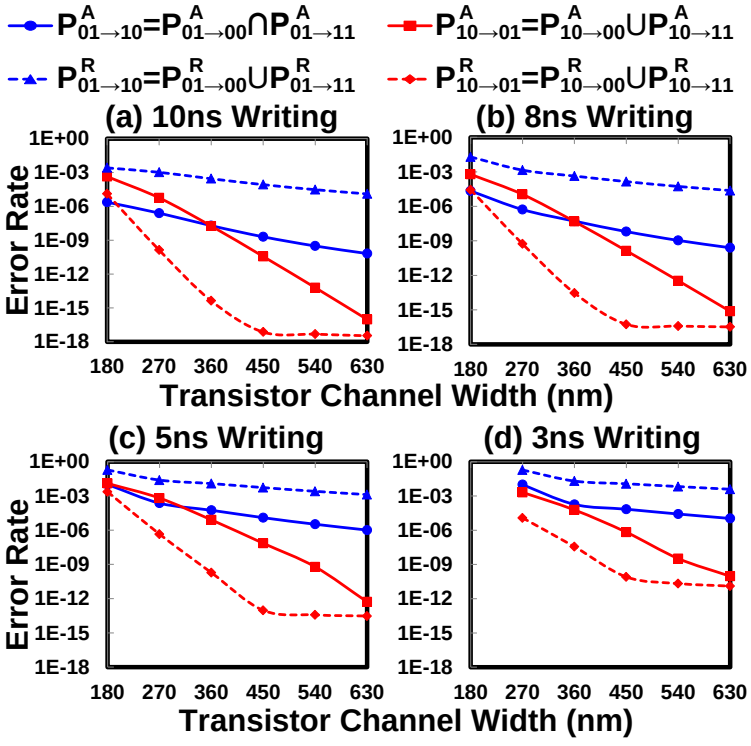

Figure 33: Write error rates of the RDAMS and ADAMS cells when the write pulse width is set to (a) 10ns; (b) 8ns; (c) 5ns; and (d) 3ns.

Comparing Fig. 33(a) with Fig. 32, we found that $P^{p}_{WF,0\to1}$ at the transistor channel width of 630nm (6.34×10^{-6}) is 12× higher than $P^{A}_{WF,1\to0}$ at the transistor channel width of 270nm (5.26×10^{-7}). As shown in Fig 25(d) and (e), the layout areas of the corresponding 1T1J cell and ADAMS cell are $0.189\mu m^2$ and $0.167\mu m^2$, respectively. Note that $P^{p}_{WF,0\to1}$ and $P^{A}_{WF,1\to0}$ dominate the write error rate of the 1T1J and ADAMS cells, respectively. Therefore, it means that ADAMS does not necessarily occupy a larger cell area than 1T1J cell under a certain reliability requirement.

5.3.2 Read Operation Analysis

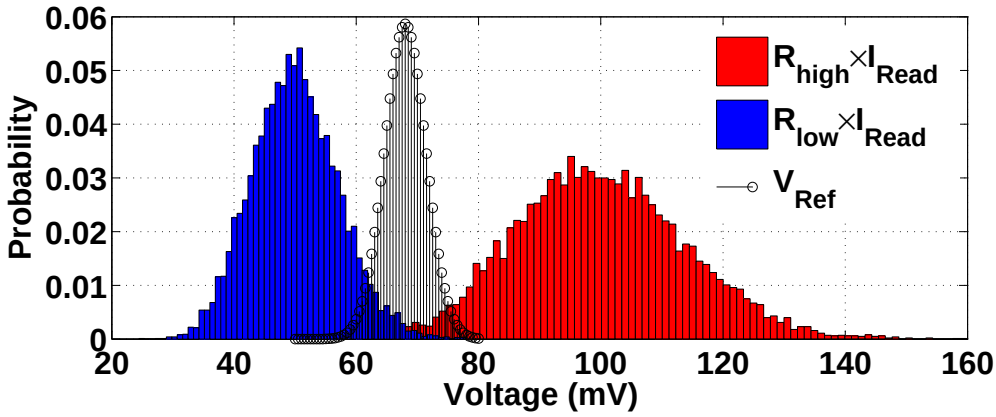

Figure 34: Example of BL voltages distribution of a 1T1J cell.

5.3.2.1 Read Reliability Analysis The two resources of read errors are read disturbance and sensing error. Sensing errors happen if the resistance state of the MTJ is erroneously detected by the SenAmp under the influences of the NMOS transistor and MTJ resistance variations. Fig. 34 shows the Monte-Carlo simulation results on the distributions of the voltages generated on the BL of a 1T1J cell when the MTJ is in the low- and high-resistance states. The simulation parameters are depicted in Table 4. In the 1T1J cell, the reference voltage is generated by a reference cell, which also suffers from device variations. Although some robust devices with small process variations, e.g., resistors, can be used to implement the reference cell, the overlap between the distribution of the reference voltage and the BL voltage still generate a considerable sensing error rate. Usually the reference cell of the 1T1J cell is carefully designed so as to achieve the equal sensing error rates at both '0' and '1'. RDAMS and ADAMS can dramatically reduce the sensing error

rate by directly comparing the resistance states of two complimentary MTJs. The corresponding sensing error rate, which is indicated by the overlaps between the distributions of two BL voltages, is significantly less than that of the 1T1J cell. Note that RDAMS and ADAMS have the same sensing error rate as they all compare the BL voltages generated from the complimentary 1T1J cells. Our simulations show that at an sensing current of $66\mu A$, the sensing error rates of the 1T1J cell and the RDAMS/ADAMS cell are 4×10^{-4} and 5.37×10^{-6}, respectively. Here the conventional SenAmp and our asymmetric SenAmp/latch designs are used in the sensing of the 1T1J cell and the CDAMS/ADAMS cell, respectively.

Because the sensing error is generated by the device variations, it can be reduced by leveraging design redundancy and discarding the memory cells with large device variations. Also, as we shall show later, ADAMS has lower read disturbance error rate than other cell structures under the same sensing current. Hence, the sensing current magnitude in an ADAMS cell may be increased to suppress the sensing error rate.

The MTJ switching probability P_{SW} can be modeled as:

$$P_{SW} = 1 - exp\{-\tau_p/\tau_0 exp[-E/k_BT(1 - I_c/I_{c0})]\}. \tag{5.5}$$

Here I_{c0} and τ_0 are the MTJ threshold switching current and switching time at 0K. I_c is the current applied on the MTJ. τ_p is the pulse width of the applied current. Eq.(5.5) implies that the read disturbance could happen under any sensing current magnitude and pulse width as long as the original resistance state of the MTJ is different from the possibly flipped one.

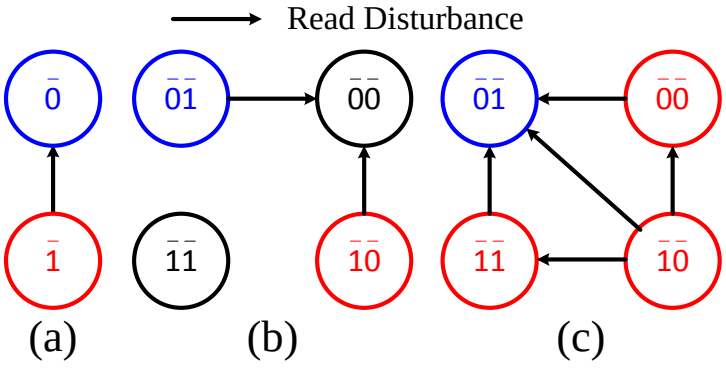

Figure 35: STT-RAM reading state. (a) 1T1J. (b) RDAMS. (c) ADAMS

Fig. 35(a)-(c) shows the state transitions of different STT-RAM cell structures at all possible read disturbances. In 1T1J cell, the read disturbance can happen only when sensing '1' and may flip the state of the MTJ to '0'. In RDAMS, the read disturbance can happen when sensing two non-prohibited states $\bar{0}\bar{1}$ and $\bar{1}\bar{0}$, and may flip them to $\bar{0}\bar{0}$. The read error rates of both 1TJ cell and RDAMS when the stored state is fixed (i.e., '1' for 1T1J and $\bar{0}\bar{1}/\bar{1}\bar{0}$ for RDAMS, respectively) can be calculated by:

$$P^{1}_{dierr} = P^{R}_{dierr} = P^{p}_{dis}. \tag{5.6}$$

In ADAMS, the read disturbance can happen when sensing any four states. Since states $\bar{1}\bar{1}$ and $\bar{0}\bar{0}$ can be read out as '1', a read disturbance will result in a read error in only the following two situations: 1) The state $\bar{1}\bar{0}$ is stored in the ADAMS cell, but read disturbances occur in both P-cell and N-cell in the read operation; 2) Due to the unsuccessful write or the read disturbance happened before, the state stored in the ADAMS cell is either $\bar{0}\bar{0}$ or $\bar{1}\bar{1}$. A read disturbance happens in either P-cell or N-cell during the read operation and flips the state of the ADAMS cell to $\bar{0}\bar{1}$. As a consequence, the read error rate of an cell induced by the read disturbance can be calculated by:

$$P^{A}_{diserr} = S_{10}P^{p}_{dis}P^{n}_{dis} + S_{11}P^{p}_{dis} + S_{00}P^{n}_{dis}. \tag{5.7}$$

In Eq.(5.6) and (5.7), P^{p}_{dis} and P^{n}_{dis} are the read disturbance probability in the P-cell and N-cell, respectively, during the read operations. S_{10}, S_{11} and S_{00} are the probabilities of the ADAMS cell storing $\bar{1}\bar{0}$, $\bar{1}\bar{1}$, and $\bar{0}\bar{1}$, respectively, which are determined by the historical operations of the ADAMS cell.

We measure the read reliability of the ADAMS cell by assuming the cell state starts with $\bar{0}\bar{1}$ and then is written into $\bar{1}\bar{0}$. The corresponding read disturbance error rate of the ADAMS cell can be derived from Eq.(5.7) as:

$$\begin{aligned} P^{A}_{diserr} &= (1 - P^{A}_{01\to11} - P^{A}_{01\to00})P^{p}_{dis}P^{n}_{dis} \\ &+P^{A}_{01\to11}P^{p}_{dis} + P^{A}_{01\to00}P^{n}_{dis}. \end{aligned} \tag{5.8}$$

Here $P^{A}_{01\to00}$ and $P^{A}_{01\to11}$ are the probabilities that the state of the ADAMS cell is wrongly programmed from $\bar{0}\bar{1}$ to $\bar{0}\bar{0}$ and $\bar{1}\bar{1}$, respectively. The corresponding definitions can be found in Section 5.3.1.2.

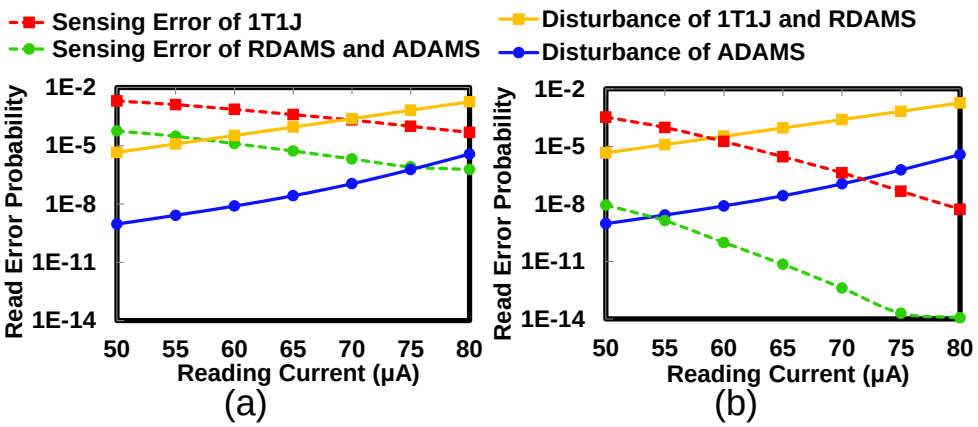

Figure 36: Sensing errors and disturbance errors of different cell structures. (a) Without redundancy. (b) With 3% redundancy.

Fig. 36(a) shows the Monte-Carlo simulation results of both sensing error rate and read disturbance error rate of all three cell structures at different read current. The NMOS transistor channel width is set to 630nm in the 1T1J cell and 270nm in the RDAMS and ADAMS cells. The sensing current pulse width is set to 1ns. As expected, the RDAMS cell and the ADAMS cell have the same sensing error rate, which is significantly lower than that of the 1T1J cell. Similarly, the read disturbance error rate of the RDAMS cell is the same as that of the 1T1J cell as illustrated by Eq.(5.6). The ADAMS cell, however, achieves much lower read disturbance errors than the 1T1J cell and the RDAMS cell by tolerating the invalid states $\bar{0}\bar{0}$ and $\bar{1}\bar{1}$. Each cell structure has different optimal working point, which refers to the sensing current magnitude ensuring the equal sensing error rate and read disturbance error rate. Among all three cell structures, ADAMS offers the lowest combined read error rate at its optimal working point,i.e., 1.4×10^{-6} at the sensing current of $76\mu A$.

Fig. 36(b) shows the sensing and read disturbance error rate of three structures after a 2-bit redundancy is applied to every 64-bit memory bits (3% area overhead). The sensing error rates of both ADAMS cell and CDAMS cell are reduced by more than 3 orders of magnitude. It shows the effectiveness of redundant designs on reducing the sensing errors. However, design redundancy does not help to reduce the intermittent read disturbance errors. As a result, the read disturbance errors dominate the read errors of the ADAMS cell as well as the RDAMS cell. Nonetheless, the combined read error rates of both RDAMS and ADAMS cells are reduced substantially.

5.3.2.2 Read Latency Analysis In an ADAMS cell, the modified working point of the asymmetric SenAmp and latch designs may prolong the sensing latency w.r.t. the conventional design under the same sense margin. As shown in Fig. 26(b), the worse-case read latency of ADAMS is bounded by the sensing of states $\bar{0}\bar{0}/\bar{1}\bar{1}$ and $\bar{0}\bar{1}$. After the SenAmp design is fixed, the total read latency is also affected by the process variations as well as the data capturing time of the latch. To reduce the total read latency, the data must be captured as early as possible once the *Out* signal of the SenAmp corresponding to the state $\bar{0}\bar{1}$ crosses the working point of the latch. In all cell structures, we refer to the SenAmp latency as the time period from the SenAmp starts to function until the *Out* signal reaches the 0.1Vdd when sensing '0' ($\bar{0}\bar{1}$).

Fig. 37(a) shows the Monte-Carlo results of the distributions of the SenAmp latency of different cell structures. At the same sensing current of $60\mu A$, which achieves the lowest read error rate of the 1T1J cell and the RDAMS cell in Fig. 36(b), both the ADAMS cell and the RDAMS cell demonstrate a better sensing latency distribution compared to the 1T1J cell due to the enhanced sense margin. The RDAMS cell has a SenAmp latency slightly shorter than the ADAMS cell though it suffers from a much higher combined read error rate. We can increase the sensing current in the ADAMS cell from $60\mu A$ to $70\mu A$ to significantly improve the SenAmp latency while still maintaining a combined read error rate 463.2× and 302.2× lower than that of the 1T1J cell and the RDAMS cell, respectively, at a sensing current of $60\mu A$, as shown in Fig. 36(b).

In the ADAMS cell, successfully sensing the invalid states $\bar{0}\bar{0}$ and $\bar{1}\bar{1}$ requires the timing coordinations between the SenAmp and the latch. Hence, we simulate the SenAmp and latch latencies of in the ADAMS cell at 3σ corner when the sensing current sweeps, as shown in Fig. 37(b). Following the increase in sensing current, the SenAmp latency decreases while the latch latency grows as the working point of the latch deviates farther from $\frac{Vdd}{2}$. The optimal working point happens when the sensing current equals $65\mu A$, leading to a total read latency of 266.7ps. In the RDAMS cell and the 1T1J cell, the latch latency is only about 20 ps. Nonetheless, the 3σ total read latency of the ADAMS cell (266.7ps) is shorter than that of the 1T1J cell (477.6ps) and the RDAMS cell (321.8ps) by 44.2% and 17.1%, respectively. The corresponding combined read error rate of the 1T1J cell, the RDAMS cell and the ADAMS cell are 5.14×10^{-5}, 3.35×10^{-5}, and 2.66×10^{-8}, respectively.

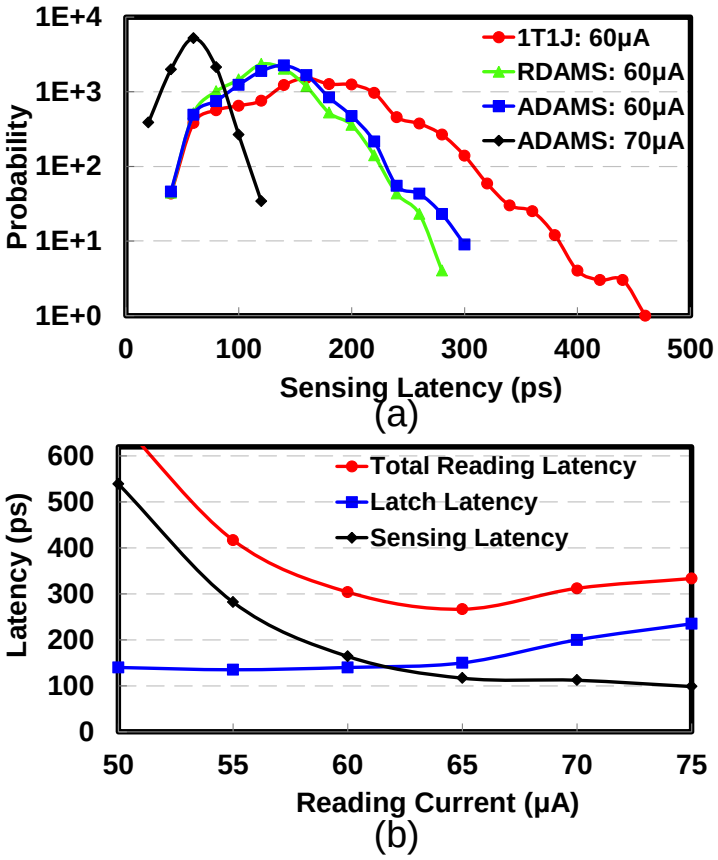

Figure 37: (a) Latency distribution of SenAmps. (b) SenAmp latency, latch latency and total read latency of the ADAMS cell.

6.0 OTHER PROPOSED STT-RAM IMPROVEMENT WORKS

In Chapter 5, we present a purely design improvement of STT-RAM. In this Chapter, we will introduce two improvement, based on a novel alternative operation scheme, and a new structure of the MTJ device.

6.1 BASIC CONCEPT OF FA-STT

The read operations of STT-RAM require a sufficient distinction (sense margin) between the MTJ resistance states and the reference signal. However, the variations of the MTJ resistance can significantly degrade the sense margin or even cause a false detection of the resistance state. Also, process variations and thermal fluctuations introduce a distribution of STT-RAM write speed. A sufficient margin, for example, a write pulse width longer than the nominal value, must be reserved to cover the distribution.

In this work, we propose a *field-assisted STT-RAM design* (FA-STT) to enhance the read and write reliability of STT-RAM simultaneously. Figure 38(a) illustrates the FA-STT design by using a row of memory cells that share the same word-line control. An extra metal wire is placed above the memory row. Applying a current through the metal wire will generate an external magnetic field orthogonal to the magnetization orientation of the MTJ reference layer. As a result, the magnetization of the MTJ free layer is deviated from the original orientation that is parallel or anti-parallel to that of the reference layer, as shown in Figure 38(b).

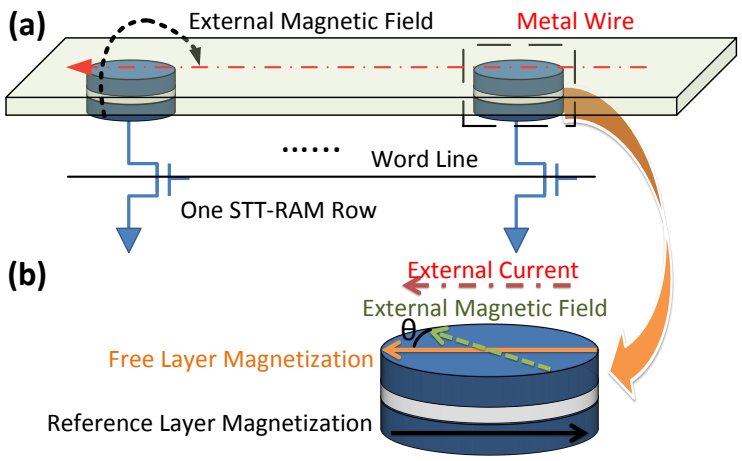

Figure 38: (a) 3D view of FA-STT scheme. (b) MTJ intermediate resistance state generation.

FA-STT leverages this phenomenon to assist the read and write operations:

Read operations: MTJ resistance is determined by the relative angle between the magnetization of two ferromagnetic layers θ. The angular dependence of the magneto-resistance in an in-plane MTJ can be described as [32]:

$$R(\theta) = R(0) + \Delta R \frac{1 - \cos\theta}{2 + \lambda(1 + \cos\theta)}, \tag{6.1}$$

where λ is a fitting parameter. The deviation of the magnetization of the free layer from parallel ($\theta = 0$) or anti-parallel ($\theta = 180$) position generates an intermediate resistance state between R_H and R_L of the MTJ. The relative resistance change between the intermediate state and the initial state of the MTJ can be used to determine the data stored in the STT-RAM cell.

Write operations: The external magnetic field introduce another spin torque component that can accelerate the magnetization switching of the MTJ free layer in write operations.

6.2 FA-STT READ SCHEME

6.2.1 Self-reference Sensing Scheme in FA-STT

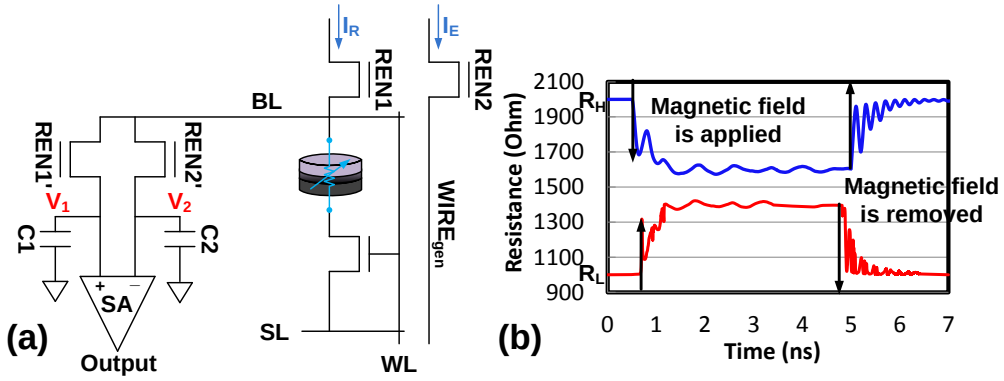

Figure 39: (a) Self-reference circuit design. (b) MTJ resistance during read operation.

Because the intermediate resistance state of the MTJ generated by the external magnetic field is in the middle of the the low- and high-resistance states of the MTJ, we can conduct a two-step sensing scheme to detect the data stored in the MTJ by comparing the relative change between the intermediate and the original resistance states of the MTJ. The conceptual design of FA-STT read circuit is illustrated in Fig. 39(a) and the procedure of the corresponding self-reference sensing scheme can be summarized as follows:

1. **First read**: A read current I_R is applied on the STT-RAM cell to generate a BL voltage V_1, which is stored in a capacitor C_1. $V_1 = V_{1L}$ or V_{1H} when the MTJ is at the low- or high-resistance state, respectively;
2. **Intermediate state generation**: The transistor R_{en2} that is connected to the metal wire W_{gen} is turned on. The external magnetic field is generated by the current passing through $WIRE_{gen}$. As the generated magnetic field is orthogonal to the magnetization orientation of the free layer of the MTJ, it will force the magnetization orientation of the free layer to deviate from the original position, putting the MTJ into the intermediate state;

3. **Second read**: The same read current I_R is applied on the BL again and generates another BL voltage V_2. $V_2 = V_{2L}$ or V_{2H} if the initial state of the MTJ is low- and high-resistance, respectively. V_2 could be also stored in capacitor C_2. Since the Intermediate state is between the low- and high-resistance state, we have: $V_{2H} < V_{1H}$ and $V_{2L} > V_{1L}$;
4. **Sensing**: The data will be readout by comparing the voltages on two capacitors, i.e., '0' ($V_2 > V_1$) or '1' ($V_1 > V_2$).
5. **Remove magnetic field**: The external magnetic field must be removed once the sensing step completes. The magnetization orientation of the MTJ will go back to its original position.

Fig. 39(b) shows an example of the MTJ resistance change during our proposed self-reference sensing scheme. When the magnetic field is applied, the resistance decreases from the high-resistance state and gradually reaches a stable resistance lower than R_{1H}. After the sensing step completes, the applied magnetic field is removed and the MTJ resistance will go back to the original value.

Table 5: Design Parameters

Parameter	Mean	1σ deviation
RA ($\Omega\mu m^2$)	8.1	7%
Surface Area (nm^2)	45×90	5%×*technode*
Oxide Thickness (nm)	2.2	2%
TMR ratio	1	5%
High Resistance (R_H)(Ω)	2000	design dependent
Low Resistance (R_L)(Ω)	1000	design dependent
Reading Current(μA)	20	design depdendent
Transistor Size (nm^2)	45×180	5%×*technode*

6.2.2 Read Operation Analysis

6.2.2.1 Read disturbance A major error in STT-RAM read operations is read disturbance, which denotes that the read current may flip the resistance of the MTJ under the impact of thermal fluctuations. In FA-STT sensing scheme, the probability of MTJ state flipping could be aggravated by the externally applied magnetic field.

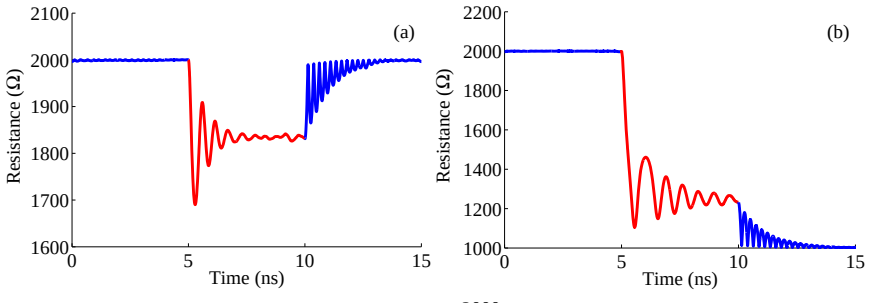

Figure 40: (a) Intermediate state generation. (b) Read disturbance of intermediate state.

We simulated the dynamic MTJ resistance change during FA-STT self-reference sensing process. Table 5 depicts the statistic information of the parameters adopted in our simulations [4]. The R_H and R_L are set at 2000Ω and 1000Ω, respectively. To avoid a large disturbance from the reading current, a relatively small current ($20\mu A$) is selected.

As shown in Fig. 40(a), after applying external magnetic field, the MTJ resistance (and the magnetization orientation of the free layer) experiences an oscillation before it reaches a stable state. A large oscillation momentum will increase the possibility of flipping the resistance state of the MTJ under the impact of the applied read current and thermal fluctuations, that is, the angle between the biased magnetization orientation of the free layer and the original position of the magnetization orientation permanently crosses 90°. Fig. 40(b) shows the case that the applied magnetic field is so large that when the read current is applied, the MTJ flips to the low resistance state.

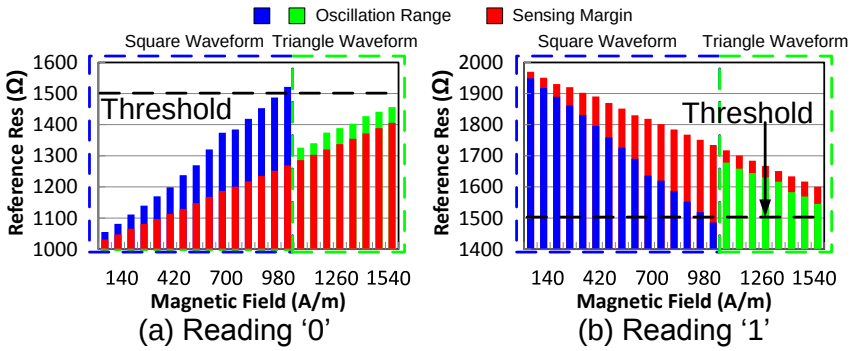

Figure 41: (a) MTJ resistance changes in reading '0'. (b) MTJ resistance changes in reading '1'.

Fig. 41(a) and (b) depict the simulation results of the MTJ resistance change under different external magnetic field magnitudes in reading '0' and '1', respectively. The magnitude of magnetic field sweeps within the range that the MTJ state will not be flipped even considering the worst-case thermal fluctuations. Here we assume the control transistor R_{en2} in Fig. 39(a) is turned on sharply by a step signal. The difference between the stable intermediate state of the MTJ resistance and the original resistance state (i.e., 1000Ω in Fig. 41(a) and 2000Ω in Fig. 41(b), respectively) reflects the sense margin under specific magnetic field magnitude. However, the sense margins in both cases are severely limited (< 200Ω) by the high read disturbance rate incurred by the large momentum of MTJ resistance oscillation.

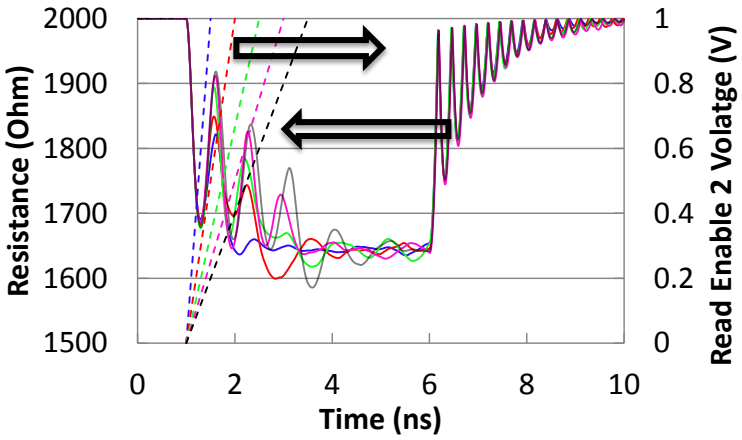

Figure 42: MTJ resistance change under different magnetic field applying speed.

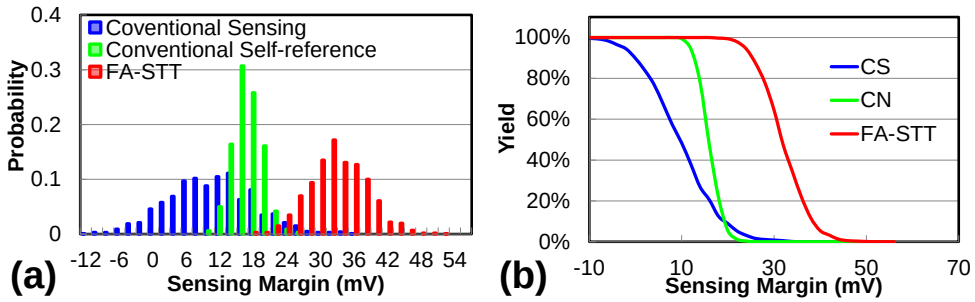

Figure 43: (a) Sensing margin distributions. (b) Memory yields under different sensing margins.

To minimize the oscillation momentum generated in FA-STT sensing, we propose to slowly turn on the transistor R_{en2} with a gradually increased control signal, as shown in Fig. 42. By extending the slope of the R_{en2} control signal to 3ns, the sense margin of FA-STT sensing scheme can be safely raised to 350Ω. Note that sharpening the slope of R_{en2} control signal may shorten the convergence time of the MTJ resistance oscillation and improve the read performance but it also increases the read disturbance rate by raising the oscillation momentum of the MTJ resistance.

6.2.2.2 Sensing margin To evaluate the impact of read error rates in different sensing schemes on memory array yield, Monte-Carlo simulations are conducted to obtain the sense margin distribution of three sensing schemes – FA-STT sensing (FA-STT), conventional nondestructive self-reference sensing (CN) [29], and conventional STT sensing (CS), which directly compares the MTJ resistance with a reference of $(R_L + R_H)/2$. An 64*64 (4Kb) STT-RAM array is simulated while every sense amplifier is shared by eight columns. Read current as $20\mu A$ is adopted in all three sensing schemes to ensure a negligible read disturbance rate.

The sense margin distributions of different sensing schemes are shown in Fig. 43(a). Negative sense margins appear in the distribution of CS sensing as the R_L (R_H) of some MTJs are higher (lower) than the reference value, resulting in false detections of the STT-RAM cell data. CN and FA-STT sensing schemes, however, always produce positive sense margin for all STT-RAM cells because of the nature of self-referencing. Although FA-STT sensing has a wider sense margin distribution than CN sensing, it still offer better read reliability due to the significantly improved sense margin.

Fig. 43(b) shows the memory yields of different sensing scheme under different minimum sense margin requirements. CS sensing has the lowest memory yield among all sensing schemes. Both CN and FA-STT sensing schemes demonstrate a high yield when the required sense margin is small. The yield of CN sensing, however, drops quickly when sense margin requirement raises beyond 10mV. As a comparison, FA-STT can tolerate a minimum sense margin requirement of more than 20mV, which is doubled from the one of CN sensing scheme, for a memory yield of 99.99%.

6.3 FA-STT WRITE SCHEME

6.3.1 Field-assisted MTJ Switching

As aforementioned in Section 6.1, the external magnetic field introduced in FA-STT design can also accelerate the MTJ switching during the write operation of STT-RAM cells. Figure 45(a), (b) and (c) show the magnetization motion of the MTJ free layer when a standard STT-RAM cell switches from '1' to '0', a FA-STT cell switches from '1' to '0' and '0' to '1', respectively. Here the external magnetic filed is applied on the FA-STT cell during the write operations. By comparing these three figures, it can be easily observed that the external magnetic field accelerates the convergence of the magnetization oscillation and speeds up the MTJ resistance switching:

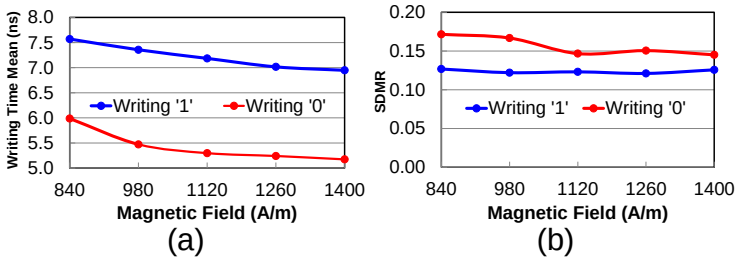

Figure 44: (a) The mean of MTJ switching time vs. the magnetic field. (b) The SDMR of MTJ switching time vs. the magnetic field.

In our simulation, all '1'→'0' switching's start at coordinate $(x, y, z) = (0, 0, 1)$. The '1'→'0' switching of the standard STT-RAM cell ends at $(0, 0, -1)$. The '1'→'0' switching of the FA-STT cell, however, ends at $(0, 0.3, -0.95)$ under the influence of the applied external magnetic field. The magnetization orientation of the MTJ free layer in the FA-STT cell goes back to $(0, 0, -1)$ only when the external magnetic field is removed after the write operation completes. A similar scenario happens in the '0'→'1' switching too.

The external magnetic field accelerates the MTJ switching by turning the magnetization orientation of the free layer toward 90° relevant to its initial position, no matter if it is initially parallel or anti-parallel to the magnetization orientation of the reference layer. However, after the magnetization orientation of the free layer crosses over 90°, the external magnetic field starts to hinder the stabilization of the new MTJ resistance state. Hence, applying the external magnetic field throughout the entire write operation might not be necessary. Based on the MTJ switching theory, after the magnetization orientation of the free layer crosses over 90°, a small amount of switching current is sufficient to retain the switching momentum and complete the switching. Thus, the external magnetic field may be removed earlier than the write current pulse to improve the write performance and save the write energy.

6.3.2 Write Performance Evaluation

Figure 44(a) shows the mean of the MTJ switching time under different magnetic field magnitudes. As the magnetic field increases, the MTJ switching time decreases first and then becomes saturated. The variations of the switching time is measured by *the standard deviation over mean ration* (SDMR), which is shown in Figure 44(b). In general, the variation of the MTJ switching time in writing '1' keeps constant while that in writing '0' decreases slightly as the magnetic field increases. Also, writing '1' has a smaller SDMR than writing '0', mainly because writing '0' has a smaller nominal value of MTJ switching time. Considering both write performance and its variation, we choose $1.26 \times 10^3 A/m$ as the optimal magnitude of the external magnetic field in the following simulations.

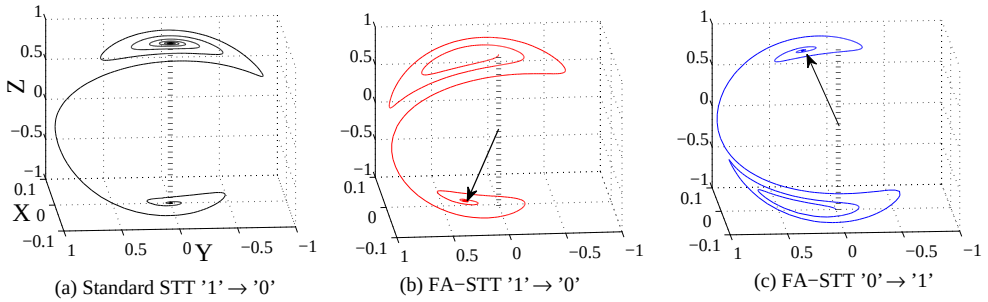

Figure 45: The motion behavior of MTJ free layer magnetization: (a) the standard STT-RAM $1 \rightarrow 0$; (b) FA-STT $1 \rightarrow 0$; and (c) FA-STT $0 \rightarrow 1$.

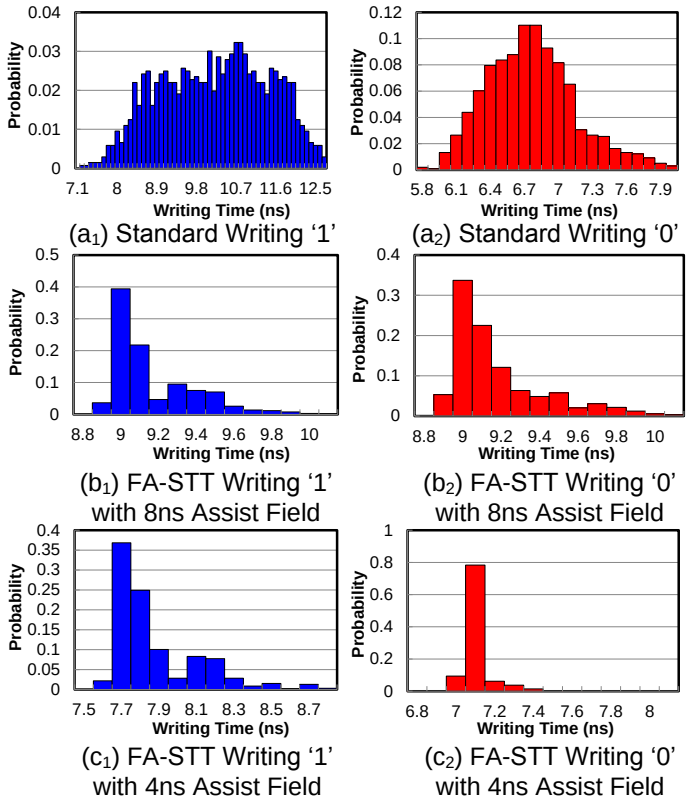

Figure 46: The write time distributions.

Figure 46 shows the distributions of STT-RAM write time obtained from Monte-Carlo simulations. We assume that the select transistor in the STT-RAM cell has a dimension of W/L =

$180nm/45nm$ and include both process variations and thermal randomness. Three designs were compared, including: (a) the standard STT-RAM, (b) the FA-STT with $8ns$ of magnetic field, and (c) the FA-STT with $4ns$ of magnetic field. The write time is defined as the time period for the free layer completely switches its magnetization to the parallel or anti-parallel state. In the figure, the distributions of writing '0' and writing '1' are given separately.

Compared with the standard STT-RAM design, the magnetic field in FA-STT dramatically improves the write speed as well as reduces the variations in write time. Furthermore, the asymmetric writes in the standard STT-RAM design (i.e., writing '1' is much harder and requires longer time than writing '0') is relaxed in FA-STT. For example, as shown in Figure 46(b_1) and (b_2), writing '1' and '0' in FA-STT with $8ns$ assisting field have the similar write time and the corresponding distributions. Reducing the assisting field to $4ns$ makes writing '1' and '0' in FA-STT a little unbalanced, as shown in Figure 46(c_1) and (c_2). This is because the duration of the magnetic field occupies smaller portion of the total write time, resulting less contribution to the MTJ switching. Nonetheless, the small difference between the results of 4ns and 8ns magnetic field indicates that 4ns is sufficient for MTJ switching assistance.

Table 6: Comparison of write error rates under 10ns write period.

	Writing '1'	Writing '0'
Standard STT-RAM	0.42	2.05×10^{-5}
FA-STT with 8ns Assist Field	3.68×10^{-4}	9.60×10^{-5}
FA-STT with 4ns Assist Field	2.29×10^{-9}	5.45×10^{-14}

6.3.3 Write Error Rate

Table 6 compares the write error rates of the above three STT-RAM designs, assuming a fixed $10ns$ write period and a NMOS select transistor of $W/L = 180nm/45nm$. In the standard STT-RAM, the errors of writing '1' dominates the write errors, i.e., a 42% error rate that is unaffordable in real design [34]. Raising the transistor size and/or prolonging the write period become necessary to ensure a reliable write with an acceptable error rate. Compared with the standard STT-RAM,

FA-STT with $8ns$ magnetic field reduces the error rate by three-orders-of-magnitude in writing '1'. The writing '0' error rate slightly increases because the assist field lasts too long. Decreasing the magnetic field to $4ns$ dramatically reduces the error rates in writing '1' and writing '0' down to 2.29×10^{-9} and even lower. Relaxing the write error requirement can further improve the write speed or reduce the STT-RAM cell area.

6.4 LAYOUT DESIGN CONSIDERATION

In FA-STT design, a metal wire is placed above the STT-RAM cells to generate the external magnetic field. The amplitude of the generated external magnetic field can be calculated by Biot-Savart law as [16]:

$$dH = \frac{1}{4\pi}\frac{Idl\times\vec{r_0}}{|r|^2}. \tag{6.2}$$

Here $\vec{r_0}$ and r is the unit vector and the distance between the metal wire and the MTJ. dl is a vector of which the magnitude is the length of the differential element of the wire. H and I is the generated magnetic field and the applied current, respectively.

To minimize the required magnitude of the current, the metal wire should be placed close to the MTJ. Fig. 47(a) and (b) show two options of the wire placement by assuming the MTJ is fabricated between metal 1 and metal 2: (1) the metal wire is placed at metal 1 between the source and the drain of the transistor; and (2) the metal wire is placed at metal 3 on top of the MTJ. Based on Eq. (6.2) The magnitude of the current required to generate a magnetic field of $1.26\times10^3 A/m$ is $782\mu A$ in Fig. 47(a) and 2.6mA in Fig. 47(b), respectively, assuming 10% variation tolerance.

However, according to the design rule of 45nm technology, there is not enough space to place a sufficiently wide metal wire for the required current magnitude with a $W/L = 180nm/45nm$ select transistor in option (1). Hence, option (2) is chosen in our FA-STT design and the corresponding layout is shown in Fig. 47(c). According to wire width requirements of ITRS [14], we are able to fit a sufficiently wide metal wire into this layout structure to carry a current of 2.6mA. This structure, however, requires at least 4 metal layers in the STT-RAM array area by reserving one metal layer solely for the wires generating the magnetic field. Note that although option (2) is selected in our

FA-STT design, option (1) may be still utilized for read/write energy reduction if a wide transistor size is adopted in STT-RAM cell designs, e.g., a multi-level cell structure.

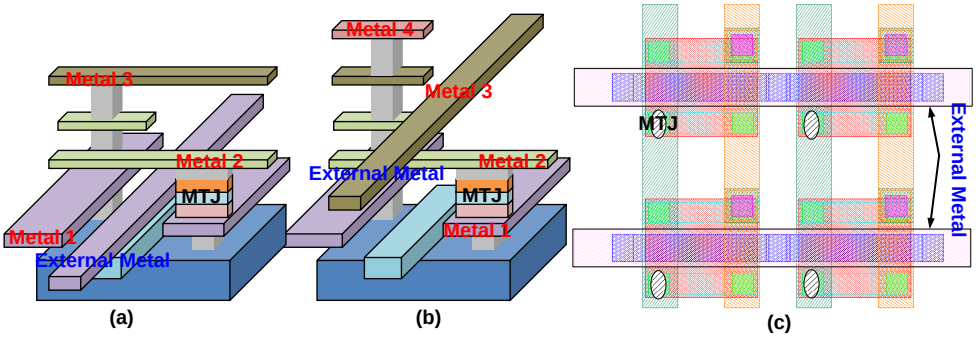

Figure 47: 3D View of External Metal Placing.

6.5 GSHE SPIN LOGIC STRUCTURE

6.5.1 Basic Logic Functions

Since the switching threshold of GSHE MTJ can be changed by manufacturing without using other material, it is possible to build device with various threshold. This property makes such a device possible to achieve basic logic functions, (such as 'AND', 'OR', 'NAND', and 'NOR'). Fig. 48 illustrates the circuit design of basic two inputs logic gates and corresponding truth table. The logical operation performed by each of these GSHE MTJ elements is determined by appropriate connecting direction of input nodes A and B, as well as selecting of input nodes (n) and switching threshold (m). The input currents of each device are determined by the output resistance states of upper level devices (d_1 and d_2). With different resistance states (either R_H or R_L) of two input devices, the current I_1+I_2 will present approximately in three region: $R_H, R_H(0,0)$, $R_H, R_L(0,1)$, and $R_L, R_L(1,1)$ under the same supply voltage. Devices will be switched once the current is larger than the threshold. If the threshold is near R_H, R_L, the device will perform as an 'OR' gate. Otherwise, if the threshold is around R_L, R_L, the device will then perform as an 'AND' gate. 'NAND' and

'NOR' gates can be achieved by an opposite connection of 'AND' and 'OR' device. Truth tables of these four types of logic gates are shown in fig. 48 (a).

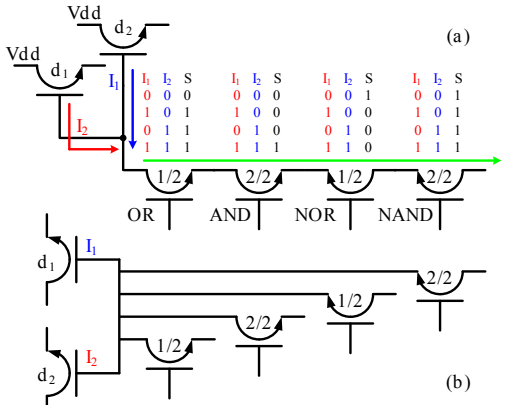

Figure 48: Examples of Basic Logic Functions. (a) Serial Connection, (b) Parallel Connection.

For more detail, although the resistance of GSHE strip is much smaller than that of MTJ, it still needs to be considered when calculating the required threshold. We assume that MTJ resistance is $R_L = \frac{1}{2}R_H = R$. Then the relationship between two resistance states and writing current $I_{XY} = I_1 + I_2$ can be as follow:

$$\begin{aligned} I_{00} &= \frac{4V}{4R+(4N+1)R_S} \\ I_{01} = I_{10} &= \frac{V(3R+R_S)}{(2R+\frac{R_S}{2})(R+\frac{R_S}{2})+NR_S(3R+R_S)} \\ I_{11} &= \frac{2V}{R+(2N+1)R_S} \end{aligned} \tag{6.3}$$

Where R_S is the equivalent resistance of GSHE strip, and N is the number of elements that connected on the output path. As aforementioned, the switching threshold for 'AND' gate and 'OR' gate should be placed between I_{11}, I_{10}, and I_{10}, I_{00}, respectively. In order to tolerate more process variation effect, the margin between the charging and threshold current should be both maximized. Therefore, the switching current threshold of 'AND' gate (I_{AND}) and 'OR' gate (I_{OR}), could be:

$$\begin{aligned} I_{OR} &= \frac{V[10R^2+(12N+6.5)RR_S+(4N+1)R_S^2]}{[4R+(4N+1)R_S][(2R+\frac{R_S}{2})(R+\frac{R_S}{2})+NR_S(3R+R_S)]} \\ I_{AND} &= \frac{V[3.5R^2+(3N+6.5)RR_S+(N+0.75)R_S^2]}{[R+(2N+1)R_S][(2R+\frac{R_S}{2})(R+\frac{R_S}{2})+NR_S(3R+R_S)]} \end{aligned} \tag{6.4}$$

We assume that $R = 2.5k\Omega$, $R_S = 100\Omega$, and the fan-out number N is 4, thus, the switching threshold ratio between I_{AND} and I_{OR} is approximately: $\frac{I_{OR}}{I_{AND}} \approx 0.7745$. The ratio could be changed

with a even large number of N, since the threshold is fixed after fabrication, it can not be changed based on different fan-out, thus, the number of fan-out is limited. However, the fan-out number is still much larger than the input elements number, with proper design, it is possible to achieve any combination logic based on the basic logic gates.

To further reduce the affect from fan-out device, a parallel structure is designed as shown in fig. 48 (b). As load resistance of writing devices is reduced by parallel structure, the current is mainly determined by MTJ resistance. However, such a structure will distribute charging current through each strip, the current will largely reduced depends on how many output elements there is. Thus to reduce the dynamic power consumption, the serial structure will mostly be adopted, the parallel structure will be a better candidate when a large enough power supply is provided.

6.5.2 GSHE Logic Operation Scheme

In GSHE logic device, performing a logic function determined by what resistance state the input device stored, it requires that these input devices should be stable with its data. The devices in a path cannot be all written at the same time. To write one device, the data stored in both its upper and under level devices cannot be changed at the same time. It makes that, in a complex circuit structure, the devices cannot be written all together. Another issue is that each device could be both input device and writing target element, the supply voltage should be able to apply on each of these device in the circuit. To achieve such a writing scheme, a multi-step writing should be applied in GSHE logic device writing. On the other hand, since writing of these device depends on the direction of charging current, once a device has been written, it cannot perform the same function again. Once the device has been switching, it cannot be switched back under the same direction writing. Thus, a preset step is required before each operation, an opposite current is applied to ensure that every device is at the initial state before it can perform a correct function.

Therefore, an unique multi-stage structure is designed to achieve GSHE logic functions. In the scheme, input devices and target devices are assigned into different stage to avoid writing conflict. As an example, a three-step structure is shown in fig. 49 (a). Three control lines ($\Phi 1$, $\Phi 2$, and $\Phi 3$) are applied to separate these three stages. Since GSHE device has three terminals, for each device, its input is connected with one or two outputs of upper stage devices depends on its corresponding

function, the output is connected with inputs of several under stage devices, and its third terminal is connected with the control line of its own stage. Writing/control signal of such a circuit is shown as fig. 49 (b). For each writing stage, there are two steps: preset step and set step. In the preset step of stage M, control line $\Phi(M-1)$ is connected to supply power $2V_{dd}$, control line ΦM is grounded, and the rest lines are all floated. Therefore, the charging current will only apply through GSHE strip of the target devices and the power supply can be large enough to write all these devices back to their initial states. On the other hand, switching control line ΦM to V_{dd} and grounding line $\Phi(M-1)$ will perform a writing in set step. During the writing step, the floating lines can isolate none used devices from input and target device, so that the performed functions will not be disturbed.

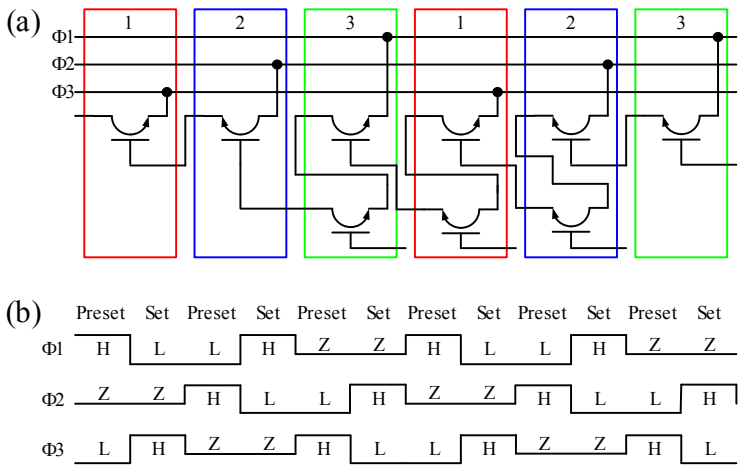

Figure 49: (a) Circuit of Three-stage Operation Scheme, (b) Control Signal Diagram.

The proposed control line writing scheme is a sequenced writing. Under such a scheme, the timing performance of all these devices are all controlled by the control lines, no extra clocks are necessary in the design. At the same time, by leveraging the non-volatility property of GSHE MTJ, data can stored in these device, thus, the requirement of latches or flip-flops can also be reduced. Besides that, since the writing is divided to several steps, the same output data can be readout from the output in each step, when the logic devices haven't catch a new input. The accuracy of a circuit can be verified by comparing these output data, which largely increase the reliability of GSHE combination logic.

6.6 DIODE-GSHE STRUCTURE

6.6.1 Sneak Path Issues

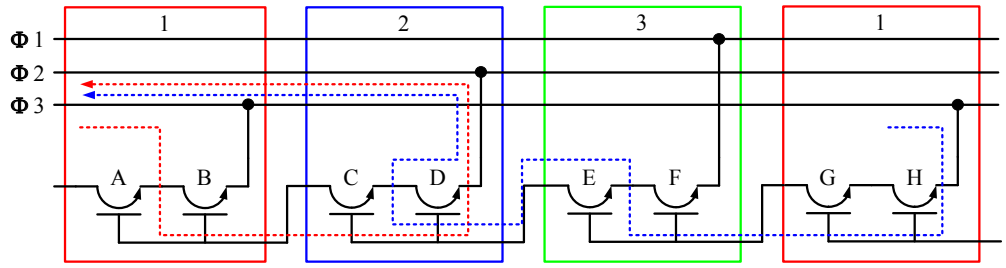

Figure 50: An example of a real case where current sneaks through undesired paths.

Same as many other resistive device, sneak path could always be an issue that high-resistance cells being ”short-circuited” by paths of devices in low-resistance state. Fig. 50 shows an example of a real case of current sneaks in GSHE logic. The current flows through some sneak paths (blue line) beside the desired one (red line). These paths contains uncontrolled parallel resistance, with various data stores in device A to H, the charging current of the desired path will be heavily impacted. The added resistance of sneak paths significantly narrows the operation current margin. To reduced the affects of sneak paths, a much larger number of writing stages is required to operation GSHE logic. As a matter of fact, to avoid overwriting these undesired devices, at least 7 stages are required in the writing scheme. More stages will leads to more non-used stages during each operation stage, and its throughput will also be largely reduced with these unoccupied stages.

6.6.2 Proposed Diode-GSHE Structure

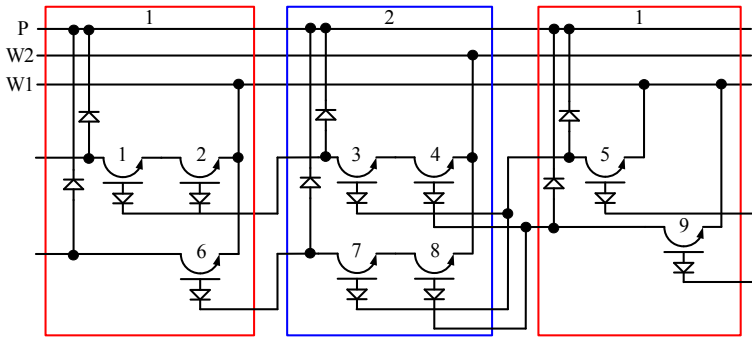

Figure 51: Proposed Diode-GSHE Structure.

The sneak path issue can be eliminated by the proposed diode-GSHE structure. By using a non-linear device as a function of diode [31, 12], the current flow would be limited only within desired direction. Fig. 51 shows our proposed design. With applying non-linear diode on the connection of each blocks, the current can only go through the direction from input device to target device, and undesired sneak current will be largely reduced. Since the writing stage can be blocked by the non-linear devices, no more then two stages are necessary in the desire. With a diode-GSHE structure, there are only two operation steps: when the first stage is used as inputs, the second stage will be programmed; on contrary, writing the first stage is based on the data stored in the second stage. During the whole process, devices in the circuit are always occupied. With the same logic structure, the throughput of Diode-GSHE structure can be further improved.

Table 7: Control Signal of Diode-GSHE Structure

Control Lines	P	W_1	W_2
Preset1	0	V_{low}	Z
Set1	Z	0	V_{high}
Preset2	0	Z	V_{low}
Set2	Z	V_{high}	0

Since only two stages are used, two control lines are required. However, in Diode-GSHE structure, there are an extra control line which is used for preset step. Thus, in Diode-GSHE, the preset will go through a preset line, instead of going back through the writing path. Apparently, there are two mainly advantages as the preset control line is designed. First, since the preset path only go through a single GSHE strip, the supply voltage for preset can be very small to provide a large enough switching current. With a much lower preset voltage, power consumption of the whole system can also be reduced. More important, if the preset current doesn't go through the input device, the input device will not be disturbed by preset control. The preset line design, will also improve the reliability of the whole system. To achieve a two step programming, the input signal will be designed as shown in TABLE 7.

By leveraging non-linear devices, and extra preset line, such a scheme is able to limit the direction of each current flow, so that the sneak path in both preset stage and programming stage will be reduced. Meanwhile, both power consumption and programming throughput can be improved by such a structure.

6.7 CASE STUDY

6.7.1 Full Adder Design

As an example, a full adder has been build based on Diode-GSHE logic structure. Fig 52 shows the structure of 1 bit full adder. Since every GSHE strip comes with its intrinsic resistance, it is not easy to control the current flow compare with relatively precise switching threshold of each device. Thus, in Diode-GSHE design, the fan-out is limited in 1~3. Another rule is that input devices can't be shared with multiple devices, they also can not be connect to a same output if they are in different stages. To follow this rule, a buffer has been introduced in the design, a buffer is an one input node GSHE device, that will pass the upper stage data to the under stages. Although it will cost some power and more design area, it makes a combination logic easier to design.

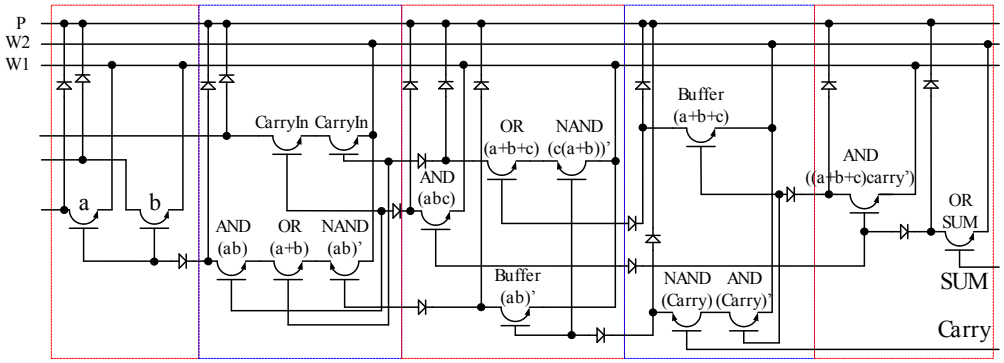

Figure 52: Example of Diode-GSHE Based Full Adder.

Leveraging the property of non-volatility and its self-sequential control, an N-bits adder can be achieve by one single-bit full adder that has a carry-out connected with carry-in (as shown in fig. 53). Since the higher bit shall always wait the carry-out signal from lower bit, such a structure doesn't need to sacrifice the operation latency. It is possible to design an N-bit adder by only one single-bit full adder without any extra overheads. Thus, the adder can largely reduce the design area and power consumption while maintain almost same throughput.

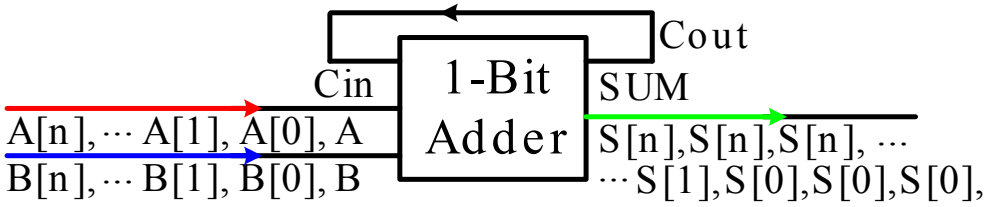

Figure 53: N-bit Adder Structure basd on 1-bit Adder.

On the other hand, as shown in fig. 52, there are three circles from lower bit carry-in input to the outputs, thus, for each bits operation, there will be three same results provided to the output, these two more results that calculated by the same progress, can be used to verify the correction of first result. This scheme can largely increase the GSHE logic reliability which is one of the biggest issue in resistive devices.

6.7.2 Experimental Results

A verilog-A GSHE MTJ model was created for our proposed design. The single bit full adder has been built with such a verilog-A model. With the same function, CMOS based full adder has also been simulated with PTM 22nm, 32nm, and 45nm technology model [3]. All simulations were conducted under Cadence Spectre Analog environment.

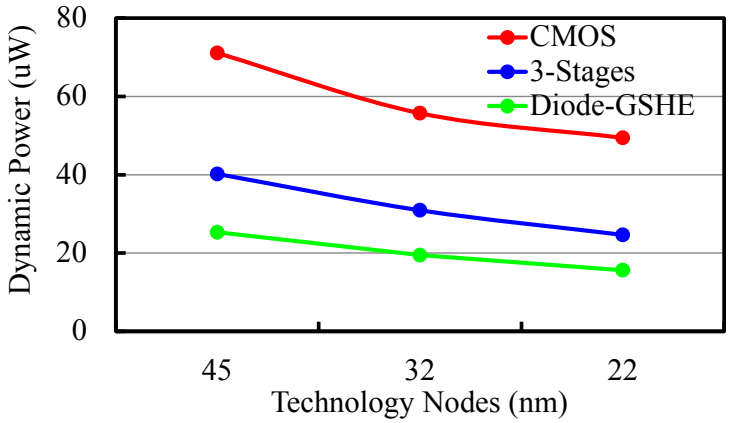

Figure 54: Dynamic Power Consumption Under 22nm, 34nm, and 45nm tech nodes.

The summary of GSHE MTJ performance has been provide in TABLE 8 [23]:

Table 8: Summary of GSHE MTJ Parameters

Parameter	Values
Critical Current (I_0)	$50\mu A$
Switching Latency (t_0)	$5ns$
High Output Resistance (R_{high})	5000Ω
Low Output Resistance (R_{low})	2500Ω
GSHE Strip Resistance (R_S)	100Ω
Surface Area (S_A)	$110 \times 65nm^2$

In order to estimate the advantages of the present circuit, the comparison of evaluated performance has been given between CMOS based and Diode-GSHE based full adder. Fig. 54 compares the simulation results of total power consumptions of 16 bits full adder at 500MHz based on 1)

3-stage GSHE-logic structure, 2) Diode-GSHE logic structure, 3) conventional CMOS structure. GSHE based configurations reduce total power by 2.0×, and 3.16× over CMOS, respectively, when maintains the same bit size. Compare with 3-stage structure, diode-GSHE has extra power consumption in diode using, however, it saves even more power with a much lower preset current. As a result, Diode-GSHE logic has a 36.6% lower power dissipation then 3-stage GSHE structure. No to mentioned, to maintain a high enough reliability, 6 or 7 stages may used in multi-stage GSHE logic, which will comes with a even higher power consumption. When scales down the technology nodes from 45nm to 22nm, power consumption of all three designs will decrease. For both GSHE based logic structure, the power reduction is proportional to the technology nodes. However, the reduction of CMOS tech is slower than GSHE logics. It proves that GSHE could be a better candidate for technology scaling.

Table 9: Comparison of Full Adders between CMOS Circuit and Proposed Diode-GSHE Circuit.

	CMOS	Diode-GSHE
Dynmaic Power	49.4μW	15.6μW
Write Time	1ns/bit	10ns/bit
write Energy	2pJ/bit[21]	20pJ/bit
Static Power	1.5nW	0.3nW
Area (Device Counts)	42MOSs	14 GSHE MTJ + 20 Diodes

Besides dynamic power consumption, TABLE 9 summarizes the comparison between CMOS circuit and our proposed structure circuit, under 22nm tech nodes: except the reduction of dynamic power, the static power is also largely reduced. The leakage power in our proposed structure circuit comes from the control line, GSHE device doesn't have ability to float the line, thus, there is one pass-gate applied on each control line to switch the control line from supply voltage, ground, and floating state. Even though, the usage of CMOS transistor is much less than conventional CMOS based circuits, thus, its static power consumption will be much smaller.

The proposed non-volatile logic circuits make it possible not only to eliminate the power consumption, but also to reduce the chip area. With a full adder, a CMOS based full adder with steam bit structure (which has a latch connect with the carry-out output) will cost around 42 MOSs, while GSHE logic requires only 14 GSHE MTJ + 20 diodes, even by using with conventional diodes in

this structure, the total area of GSHE based logic is still much smaller than CMOS circuits. Write time is one of the most important disadvantage in GSHE logic, it also dominates the write energy when updating stored data. GSHE logic has already had a larger update progress comparing with conventional spin-logic utilizing STT-RAM. We can expect that it is possible to further improved the operation speed.

7.0 CONCLUSION

It has been four decades since the discovery of tunneling magnetoresistance effect by Julliere. Since then, the improving technologies and new discoveries have pushed spin-transfer torque memory (STT-RAM) to become one of the leading candidate for future non-volatile memory technology. As we mentioned in this thesis, STT-RAM has huge benefits like non-volatility, high operation speed, and high integration density. However, the benefits can be greater if not for the conflicting design requirements that STT-RAM needs to overcome to meet read, write and reliability design targets. In this thesis, we systematically analysed these requirements, and discuss the advantage and disadvantage of both single level cell and multi level cell STT-RAM, and follow with several improvement design to overcome the disadvantages. With all the researches, we may prove that STT-RAM can fulfill its potential as the truly universal next-generation memory technology.

BIBLIOGRAPHY

[1] B. Amrutur and M. Horowitz, "Speed and power scaling of sram's," *IEEE Journal of Solid-State Circuits*, vol. 35, no. 2, feb 2000.

[2] L. Berger, "Emission of spin waves by a magnetic multilayer traversed by a current," *Phys. Rev. B*, vol. 54, pp. 9353 –9358, Oct 1996. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevB.54.9353

[3] Y. Cao and et. al., "New paradigm of predictive mosfet and interconnect modeling for early circuit design," in *IEEE Custom Integrated Ckt. Conf.*, 2000, pp. 201–204, http://www-device.eecs.berkeley.edu/ ptm.

[4] Y. Chen and et.al., "A nondestructive self-reference scheme for spin-transfer torque random access memory (stt-ram)," in *Design, Automation Test in Europe*, 2010, pp. 148–153.

[5] Y. Chen, X. Wang, W. Zhu, H. Li, Z. Sun, G. Sun, and Y. Xie, "Access scheme of multi-level cell spin-transfer torque random access memory and its optimization," in *53rd IEEE International Midwest Symposium on Circuits and Systems*, Aug. 2010, pp. 1109 –1112.

[6] Y. Chen, W.-F. Wong, H. Li, and C.-K. Koh, "Processor caches built using multi-level spin-transfer torque ram cells," in *International Symposium on Low Power Electronics and Design 2011*, Aug. 2011, pp. 73 –78.

[7] K. C. Chun, H. Zhao, J. Harms, T.-H. Kim, J. ping Wang, and C. Kim, "A scaling roadmap and performance evaluation of in-plane and perpendicular mtj based stt-mrams for high-density cache memory," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 2, pp. 598–610, Feb 2013.

[8] Z. Diao, Z. Li, S. Wang, Y. Ding, A. Panchula, E. Chen, L. Wang, and Y. Huai, "Spin-transfer torque switching in magnetic tunnel junctions and sspin-transfer torque random access memory," *Journal of Physics: Condensed Matter*, vol. 19, p. 165209, 2007.

[9] T. Gilbert, "A Lagrangian Formulation of the Gyromagnetic Equation of the Magnetization Field," *Phys.Tev.*, vol. 100, no. 1243, 1955.

[10] X. Guo, E. Ipek, and T. Soyata, "Resistive computation: avoiding the power wall with low-leakage, STT-MRAM based computing," in *Proc. of ISCA*, 2010.

[11] Y. Huai, "Spin-transfer torque mram (stt-mram): Challenges and prospects," *AAPPS Bulletin*, vol. 18, no. 6, pp. 33–40, 2008.

[12] C.-H. Huang, J.-S. Huang, S.-M. Lin, W.-Y. Chang, J.-H. He, and Y.-L. Chueh, "Zno1-x nanorod arrays/zno thin film bilayer structure: From homojunction diode and high performance memristor to complementary 1d1r application," *ACS Nano Letters*, 2012.

[13] T. Ishigaki, T. Kawahara, R. Takemura, K. Ono, K. Ito, H. Matsuoka, and H. Ohno, ""A Multi-level-cell Spin-transfer Torque Memory with Series-stacked Magnetotunnel Junctions"," in *Symposium on VLSI Technology*, Jun. 2010, pp. 47 –48.

[14] "The international technology roadmap for semiconductors," http://www.itrs.net, 2008.

[15] M.-Y. Kim, H. Lee, and C. Kim, "Pvt variation tolerant current source with on-chip digital self-calibration," *IEEE Transactions on Very Large Scale Integration Systems*, vol. 20, no. 4, pp. 737 –741, Apr. 2012.

[16] H.-B. Lee and et.al., "Efficient magnetic field calculation method for pancake coil using biot-savart law," in *12th Biennial IEEE Conference on Electromagnetic Field Computation*, 2006, pp. 193–193.

[17] J. Li, C. Augustine, S. Salahuddin, and K. Roy, "Modeling of Failure Probability and Statistical Design of Spin-Torque Transfer Magnetic Random Access Memory (STT MRAM) Array for Yield Enhancement," in *the 45th Design Automation Conference*, june 2008, pp. 278 –283.

[18] Y. Li, Y. Chen, and A. K. Jones, "A software approach for combating asymmetries of non-volatile memories," in *Proc. of ISLPED*, 2012.

[19] X. Lou, Z. Gao, D. V. Dimitrov, and M. X. Tang, "Demonstration of multilevel cell spin transfer switching in mgo magnetic tunnel junctions," *Applied Physics Letters*, vol. 93, no. 24, p. 242502, 2008. [Online]. Available: http://link.aip.org/link/?APL/93/242502/1

[20] J. Mathon and A. Umerski, "Theory of Tunneling Magnetoresistance in a Disordered Fe/ Mg O/ Fe (001) Junction," *Physical Review B*, vol. 74, no. 14, p. 140404, 2006.

[21] S. Matsunaga, J. Hayakawa, S. Ikeda, K. Miura, H. Hasegawa, T. Endoh, H. Ohno, and T. Hanyu, "Fabrication of a nonvolatile full adder based on logic-in-memory architecture using magnetic tunnel junctions," *Applied Physics Express*, vol. 1, no. 9, p. 091301, 2008.

[22] A. Nigam, C. Smullen, V. Mohan, E. Chen, S. Gurumurthi, and M. Stan, "Delivering on the promise of universal memory for spin-transfer torque ram (stt-ram)," in *International Symposium on Low Power Electronics and Design*, Aug. 2011, pp. 121 –126.

[23] C.-F. Pai, L. Liu, Y. Li, H. Tseng, D. Ralph, and R. Buhrman, "Spin transfer torque devices utilizing the giant spin hall effect of tungsten," *Applied Physics Letters*, vol. 101, no. 12, pp. 122 404–122 404–4, Sep 2012.

[24] M. Qureshi, M. Franceschini, A. Jagmohan, and L. Lastras, "PreSET: Improving read-write performance of phase change memories by exploiting asymmetry in write times," in *Proc. of ISCA*, 2012.

[25] A. Raychowdhury, D. Somasekhar, T. Karnik, and V. De, "Design space and scalability exploration of 1t-1stt mtj memory arrays in the presence of variability and disturbances," in *IEEE International Conference on Electron Devices Meeting*, Dec. 2009, pp. 1–4.

[26] S. R. Sarangi, B. Greskamp, R. Teodorescu, J. Nakano, A. Tiwari, and J. Torrellas, "Varius: A model of process variation and resulting timing errors for microarchitects," *IEEE Transactions on Semiconductor Manufacturing*, vol. 21, no. 1, pp. 3–13, Feb 2008.

[27] G. Sun, X. Dong, Y. Xie, J. Li, and Y. Chen, "A novel architecture of the 3d stacked mram l2 cache for cmps," in *the 15th International Symposium on High-Performance Computer Architecture*. IEEE, 2009, pp. 239–249.

[28] J. Z. Sun, "Spin-current interaction with a monodomain magnetic body: A model study," *Phys. Rev. B*, vol. 62, pp. 570–578, Jul 2000. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevB.62.570

[29] Z. Sun and et.al., "Voltage driven nondestructive self-reference sensing scheme of spin-transfer torque memory," *Transactions on VLSI Systems*, vol. 20, no. 11, pp. 2020–2030, 2012.

[30] Z. Sun, H. Li, Y. Chen, and X. Wang, "Variation tolerant sensing scheme of spin-transfer torque memory for yield improvement," in *IEEE/ACM International Conference on Computer-Aided Design*, Nov. 2010, pp. 432–437.

[31] A. Tulapurkar, Y. Suzuki, A. Fukushima, H. Kubota, H. Maehara, K. Tsunekawa, D. Djayaprawira, N. Watanabe, and S. Yuasa, "Spin-torque diode effect in magnetic tunnel junctions," in *Nature*, vol. 438, Nov 2005, pp. 339–342.

[32] S. Urazhdin and et.al., "Noncollinear spin transport in magnetic multilayers," *Phys.Rev.B*, vol. 71, no. 10, p. 100401, Mar. 2005.

[33] X. Wang, Y. Zheng, H. Xi, and D. Dimitrov, "Thermal fluctuation effects on spin torque induced switching: Mean and variations," *Journal of Applied Physics*, vol. 103, no. 3, pp. 034 507–034 507–4, Feb. 2008.

[34] W. Wen, Y. Zhang, Y. Chen, Y. Wang, and Y. Xie, "Ps3-ram: A fast portable and scalable statistical stt-ram reliability analysis method," in *49th Design Automation Conference*, June 2012, pp. 1187–1192.

[35] X. Wu, J. Li, L. Zhang, E. Speight, R. Rajamony, and Y. Xie, "Hybrid cache architecture with disparate memory technologies," in *Proc. of ISCA*, 2009.

[36] W. Xu, H. Sun, Y. Chen, and T. Zhang, "Design of last-level on-chip cache using spin-torque transfer ram (stt ram)," in *IEEE Trans. on VLSI System*. IEEE, 2011, pp. 483–493.

[37] Y. Ye, F. Liu, S. Nassif, and Y. Cao, "Statistical modeling and simulation of threshold variation under dopant fluctuations and line-edge roughness," in *the 45th Design Automation Conference*, june 2008, pp. 900 –905.

[38] S. Yuasa, T. Nagahama, A. Fukushima, Y. Suzuki, and K. Ando, "Giant room-temperature magnetoresistance in single-crystal fe/mgo/fe magnetic tunnel junctions," *Nature materials*, vol. 3, no. 12, pp. 868–871, 2004.

[39] Y. Zhang, Y. Li, A.K.Jones, X. Wang, and Y. Chen, "Asymmetry of mtj switching and its implication to the stt-ram designs," *Design Automation and Test in Europe*, Mar. 2012.

[40] W. Zhao, L. Torres, L. V. Cargnini, R. M. Brum, Y. Zhang, Y. Guillemenet, G. Sassatelli, Y. Lakys, J.-O. Klein, D. Etiemble, *et al.*, "High performance soc design using magnetic logic and memory," in *VLSI-SoC: Advanced Research for Systems on Chip*. Springer, 2012, pp. 10–33.

[41] P. Zhou, B. Zhao, J. Yang, and Y. Zhang, "Energy reduction for STT-RAM using early write termination," in *Proc of ICCAD*, 2009.