DEVELOPING VARIATION AWARE SIMULATION TOOLS, MODELS, AND DESIGNS FOR STT-RAM

by

Enes Eken

B.S. in Electrical and Electronics Engineering

Selcuk University, 2009

Master of Science in Electrical Engineering

University of Pittsburgh, 2014

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

Enes Eken

It was defended on May 8, 2017

and approved by

Yiran Chen, Ph.D., Associate Professor

Department of Electrical and Computer Engineering

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

Department of Electrical and Computer Engineering

Ervin Sejdic, Ph.D., Associate Professor

Department of Electrical and Computer Engineering

William E. Stanchina, Ph.D., Professor

Department of Electrical and Computer Engineering

Bo Zheng, Ph.D., Assistant Professor

Industrial Engineering Department

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

Department of Electrical and Computer Engineering

Copyright © by Enes Eken
2017

DEVELOPING VARIATION AWARE SIMULATION TOOLS, MODELS, AND DESIGNS FOR STT-RAM

Enes Eken, PhD

University of Pittsburgh, 2017

Technology scaling imposes many challenges on design and manufacturing of conventional memories, such as high leakage and reliability issues of SRAM, DRAM, and NAND flash. Extensive research has been performed to develop new memory technologies that can overcome these challenges, including phase-change memory (PCM), resistive memory (ReRAM), spin-transfer torque random access memory (STT-RAM), etc. Among all these technologies, STT-RAM is particularly identified as a potential replacement of DRAM in future main memory application because of its many attractive features like zero standby power, excellent CMOS-compatibility, etc.

However, like all other memory technologies, STT-RAM has some problems, such as long switching time and large programming energy, being lack of a variation aware simulation tool which are waiting to be solved. In order to solve long switching time and large programming energy problems, Spin-Hall Effect (SHE) assisted STT-RAM structure (SHE-RAM) has been recently invented. In this work, I propose two possible SHE-RAM designs from the aspects of two different write access operations, namely, High Density SHE-RAM and Disturbance Free SHE-RAM, respectively.

In addition to the SHE-RAM designs, I will also propose a simulation tool for STT-RAMs. As an early-stage modeling tool, NVSim has been widely adopted for simulations of emerging non-volatile memory technologies in computer architecture research, including STT-RAM, ReRAM, PCM, etc. I will introduce a new member of NVSim family – NVSim-VX^s, which enables statistical simulation of STT-RAM for write performance, errors, and energy consumption. In this work, I also developed a dynamic macromagnetic model for biaxial MTJ for MLC-STT circuit designs.

Besides simulating the relations between the switching current and the switching time of each MTJ resistance state, this model is also capable to capture the switching transience that can be used to calculate the write error rate of the MLC-STT cell. Write performance and energy consumption of the MLC-STT cell can also be derived and optimized based on the model for different design configurations. Finally, this model allows designers to perform a comprehensive reliability analysis of the MLC-STT cell by taking into account the device parametric variations and the ambient temperature during write operations.

TABLE OF CONTENTS

PREFACE xii

1.0 INTRODUCTION 1

2.0 NVSIM-VX^{S}: AN IMPROVED NVSIM FOR VARIATION AWARE STT-RAM SIMULATION 2

2.1 Introduction 2

2.2 Preliminary 4

2.2.1 Basics of STT-RAM and NVSim 4

2.2.2 Basics of NVSim 5

2.3 NVSim-VX^{s} Framework 5

2.3.1 Temperature-Aware Statistical STT-RAM Switching Time Model 6

2.3.2 STT-RAM Statistical Energy Modeling 13

2.3.3 STT-RAM Write Error Rate 16

2.3.4 Block Level Extension for NVSim-VX^{s} 19

2.3.5 Block Level Energy Consumption 19

2.3.6 Block Level Write Error Rate 21

2.4 Conclusions 21

3.0 SPIN-HALL ASSISTED STT-RAM DESIGN AND DISCUSSION 23

3.1 Introduction 23

3.2 Basics of SHE-RAM 24

3.3 SHE-RAM DESIGNS 26

3.3.1 High Density SHE-RAM 27

3.3.2 Disturbance Free SHE-RAM 31

3.4 Conclusion 35
4.0 MODELING OF BIAXIAL MAGNETIC TUNNELING JUNCTION FOR MULTILEVEL CELL STT-RAM REALIZATION AND RELIABILITY ANALYSIS . 37
4.1 Introduction 37
4.2 Preliminary 39
4.2.1 Basic Operations of STT-RAM Cell 39
4.2.2 Basics of Uniaxial and Biaxial Anisotropies 39
4.3 Modeling of Biaxial MTJ 42
4.3.1 Model Description Of Biaxial MTJ 42
4.3.2 Basic Functions of Biaxial MTJ Model 47
4.3.3 Some Discussions 49
4.4 Model Validation 51
4.5 Reliability Analysis 57
4.5.1 Switching Time and Energy Consumption of Biaxial MTJs 57
4.5.2 Write Errors of Biaxial MTJs 62
4.6 Conclusion 64
BIBLIOGRAPHY 65

LIST OF TABLES

2.1 Device and Circuit Simulation parameters 6
3.1 Summary of device parameters 27
4.1 Device and circuit simulation parameters 47
4.2 Possible switching current and switching time for each state 48
4.3 Device level simulation parameters 56

LIST OF FIGURES

2.1 STT-RAM basics. (a) In-plane MTJ. (b) Perpendicular MTJ. (c) 1T1J cell structure. 4

2.2 Framework of the NVSim-VX^s. 7

2.3 Detailed flow for circuit-level and cell-level simulations in NVSim-VX^s. 8

2.4 Simulated switching time distribution v.s. Log-normal distribution ($45nm$ technology node, $90nm$ transistor width, 4 different temperatures): a) $300K$ b) $325K$ c) $350K$ d) $375K$. 9

2.5 The standard deviation of switching time distributions for 7 transistor widths at $45nm$ technology node. 10

2.6 The mean of switching time distributions for 7 transistor widths at $45nm$ technology node. 11

2.7 The simulated results of our model and Monte-Carlo simulations for key parameters of σ and μ: (a) $m_\sigma(w)$ v.s. I; (b) $\sigma_0(w)$ v.s. I; (c) $m_\mu(w)$ v.s. I; (d) $\mu_0(w)$ v.s.I. . . 12

2.8 Overview of the statistical energy consumption characterization flow. 13

2.9 The write energy consumption distribution comparison of our model v.s. Monte-Carlo at $45nm$ technology node, $90nm$ transistor width, $60ns$ write pulse width. . . 14

2.10 The mean of the energy consumption at $10ns$ write pulse width for 7 different transistor widths at $45nm$ technology node. 16

2.11 The mean of energy consumption at different write pulse widths for $w = 2.5L$ transistor width ($L = 45nm$). 17

2.12 Simulated write error rate v.s. temperature from our model and Monte-Carlo for $22nm$ perpendicular MTJ and $45nm$ in-plane MTJ ($10ns$ write pulse width, '0'→'1' switching). 17

2.13 Simulated write error rate v.s. temperature from our model and Monte-Carlo at different write time (transistor width $w = 2.5L$,'0'→'1' switching). 18

2.14 The mean and standard deviation of block-level energy consumption for 3 different flipping combinations at $10ns$ write time and $w = 3L$ transistor width, $L = 45nm$. . 20

2.15 The block level write error rate for 3 different flipping combinations at $10ns$ write time and $w = 3L$ transistor width, $L = 45nm$. 21

3.1 (a) A MTJ whose reference and free layer are in anti-parallel. The exerted torque is zero as the two magnetization vectors are in parallel. (b) Free layer magnetization vector makes an angle under the effect of the SHE current, and STT becomes available to switch the magnetization of the free layer since it is not in parallel to that of the reference layer any more. 25

3.2 Illustration of High Density SHE-RAM. SHE current is controlled by the SHE control transistor and MTJs are accessed through the connected word line transistors. During write operations, the data is still written by applying appropriate bias on the source line and the bit line. 28

3.3 (a) Free layer's magnetic moment under the effect of SHE and STT currents. The oscillations starts from the deviated alignment of the free layer. (b) Free layer's magnetic moment under the effect of only STT current. Incubation delay takes place. 29

3.4 Simulated MTJ switching time under different transistor widths. Applying SHE current will significantly improve the MTJ switching performance, especially when the transistor width is small (or the STT current is small). 30

3.5 MTJ switching time of a High Density SHE-RAM cell under different SHE current amplitudes for a 0.7ns pulse width. 32

3.6 Disturbance Free SHE-RAM Design. SHE current is shared by the cells on the entire word line and controlled by $\mathrm{W_{x-SHE}}$ transistors. 33

3.7 Switching time comparison for High Density SHE-RAM Design and Disturbance Free SHE-RAM Design for different transistor widths. 34

3.8 Switching time probability distribution for Disturbance Free SHE-RAM Design and for conventional Perpendicular MTJ for $90nm$ and $720nm$ transistor width. . . 35

4.1 STT-RAM basics. a) Uniaxial MTJ. b) Biaxial MTJ. c) 1T1J STT-RAM cell structure. 40

4.2 Energy of the biaxial anisotropy and the changing of the logic states at the energy minima points. 41
4.3 Illustration of FL's magnetization vector in spherical coordinate system. 43
4.4 Change of m_f during the MTJ switchings. a) '0' to '1' switching transience. b) '0' to '1' switching time. c) '0' to '2' switching transience. d) '0' to '2' switching time. e) '0' to '3' switching transience. f) '0' to '3' switching time. 46
4.5 '0' to '1' switchings at two different switching currents. 50
4.6 '0' to '2' and '3' switchings at two different current values. 51
4.7 Changing of biaxial MTJ resistance value during the switching processes. 52
4.8 Four distinct positions of free layer's magnetization vector at $0°$, $90°$, $180°$, and $270°$ for '0', '1', '3', and '2', respectively. 54
4.9 Switching motion of biaxial MTJ under the effect of the applied magnetic field. . . 55
4.10 Model validation against to reference [22] for resistance values and the applied magnetic field. 56
4.11 Biaxial MTJ switching time distribution for '0' to other states at $300K$. 58
4.12 Biaxial MTJ energy consumption distribution during the switchings from '0' to other states at $300K$. 60
4.13 Energy consumption mean and standard deviation values of biaxial MTJ's each switching directions. 61
4.14 Uniaxial MTJ and biaxial MTJ write error rates at different temperatures with $60ns$ write time. 63

PREFACE

I would like to express my appreciation to my advisor, Yiran Chen, for his excellent guidance and patience throughout my graduate education. Without his constructive advices, I would have never completed this work. It was him who taught me not only how to conduct research, but also how to learn by myself, which, I think, one of the most valuable skills I gained throughout this journey.

I would like to take this opportunity to convey my gratitude to the members of Evolutionary Intelligent (EI) Lab at Swanson School of Engineering, for their generous help.

I would like to express my thankfulness to Professor Zhi-Hong Mao, Professor Ervin Sejdic, Professor William E. Stanchina, and Professor Bo Zeng for being my committee members.

I would like to send my deepest appreciation to my mother and father, Hafize and Necmettin, and my brother Muhammed Yaser for their praying for me.

Last, but definitely not the least, I would like to thank my beloved wife, Serife Eken who waited me to come from the school after the PhD proposal exam, to go to hospital for delivering our first baby Beyyine. I hope, one day, she can understand under which conditions she came to this world.

Finally, I would like to thank Turkish Ministry of Education for supporting and funding my scholarship throughout my graduate research and giving me a chance to reach my goals.

1.0 INTRODUCTION

Spin-transfer torque random access memory (STT-RAM) recently received significant attentions for its promising characteristics in cache and memory applications. As an early-stage modeling tool, NVSim has been widely adopted for simulations of emerging nonvolatile memory technologies in computer architecture research, including STT-RAM, ReRAM, PCM, etc. In this work, I propose a new member of NVSim family – NVSim-VX^s, which enables statistical simulation of STT-RAM for write performance, errors, and energy consumption. This enhanced model takes into account the impacts of parametric variabilities of CMOS and MTJ devices and the chip operating temperature. It is also calibrated with Monte-Carlo Simulations based on macro-magnetic and SPICE models, covering five technology nodes between 22nm and 90nm. NVSim-VX^s strongly supports the fast-growing needs of STT-RAM research on reliability analysis and enhancement, announcing the next important stage of NVSim development.

Long switching time and large programming energy of Magnetic Tunneling Junction (MTJ) continue being major challenges in STT-RAM designs. In order to overcome this problem, a Spin-Hall Effect (SHE) assisted STT-RAM structure (SHE-RAM) has been recently invented. In addition to NVSim-VX^s, I will also propose two possible SHE-RAM designs from the aspects of two different write access operations, namely, High Density SHE-RAM and Disturbance Free SHE-RAM, respectively. In High Density SHE-RAM, SHE current is shared by the entire bit line. Such a structure removes the SHE control transistor from each SHE-RAM cell and hence, substantially reduces the memory cell area. In Disturbance Free SHE-RAM, one memory cell contains two transistors to remove the disturbance to the unselected bits and eliminate the possible erroneous flipping of the bits.

2.0 NVSIM-VX^{S}: AN IMPROVED NVSIM FOR VARIATION AWARE STT-RAM SIMULATION

2.1 INTRODUCTION

"Post-silicon" devices have received increasing attentions in solid-state device and circuit society due to the concerns on continuous scaling of conventional CMOS technology. The high leakage power and significantly degraded reliability of mainstream memory technologies [30] inspired the popular research on emerging memory technologies: *spin-transfer torque random access memory* (STT-RAM), *resistive memory* (ReRAM), *phase-change memory* (PCM) [5], etc. In particular, STT-RAM demonstrates many characteristics that are of importance to on-chip cache and memory applications, such as high integration density, zero standby power, nanosecond access time, and excellent CMOS-compatibility [16].

It is known that write error is the major reliability issue in STT-RAM operations. Compared with conventional memory technologies, simulating a STT-RAM cell is very challenging because it requires understandings of both CMOS and magnetic devices. In [4], Chen *et al.* proposed the first combined magnetic and SPICE simulation framework to evaluate the write performance and energy of STT-RAM cells by considering the interaction between transistor and magnetic tunneling junction (MTJ) devices. Besides parametric variabilities that exist in conventional memory cells, thermal-induced switching randomness also significantly affects write operations of STT-RAM cells. Performing statistical analysis on the write reliability, hence, requires very costly and entangled Monte-Carlo simulations on both types of devices.

Block-level STT-RAM models have been also developed to fulfill the need in architectural analysis. Arcaro *et al.* integrated a STT-RAM model into CACTI [1] – a tool was originally used for conventional memory modeling and design [19].

Wu *et al.* presented an architecture-level simulation framework of the advanced perpendicular STT-RAM in [29]. Dong *et al.* released the most widely used STT-RAM block-level model, namely, NVSim [6], which can support the design parameter extraction of not only STT-RAM but also ReRAM and PCM. However, none of the above models are able to simulate the impacts of CMOS or MTJ variations and consequently, the write errors of STT-RAM.

In this work, we introduce a new member of NVSim family – NVSim-VX^s, which enables statistical simulation of STT-RAM for write performance, errors, and energy consumption. Besides the parametric variabilities of both CMOS and MTJ devices, this enhanced model also takes into account chip operating temperature, which significantly affects the write reliability of STT-RAM. As the first systematic model to simulate the entangled relationships between different design parameters and metrics of STT-RAM at block-level, the major novelty we introduce to STT-RAM research can be summarized as follows:

- We derive statistical approximations of STT-RAM design metrics, e.g., switching time and energy consumption, and generate the corresponding compact models for fast statistical analysis;
- We implement the STT-RAM model based on the switching pattern of input bits, which have been proved as the major factor affecting the statistical STT-RAM design metrics;
- We develop the models of both perpendicular and in-plane STT-RAMs to support the scaling of STT-RAM technologies.

The first release of NVSim-VX^s supports 5 technology nodes {22, 32, 45, 65, 90}nm, and driving NMOS transistor size between 2-5$\times$ the minimum feature size. It also supports operating temperature between $300K$ to $375K$. The model has been thoroughly calibrated with the Monte-Carlo simulations based on macro-magnetic and SPICE models to ensure the accuracy.

The rest of this proposal is organized as follows: Section 4.2 presents the basics of STT-RAM and NVSim; Section 2.3 introduces the statistical compact models that are developed for NVSim-VX^s; Section 2.4 concludes our work.

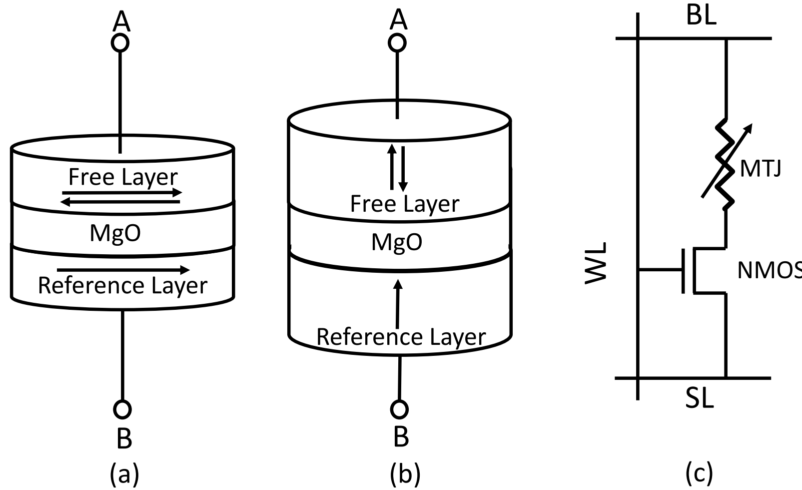

Figure 2.1: STT-RAM basics. (a) In-plane MTJ. (b) Perpendicular MTJ. (c) 1T1J cell structure.

2.2 PRELIMINARY

2.2.1 Basics of STT-RAM and NVSim

In a STT-RAM cell, the data is stored as the resistance of a MTJ device, as shown in Fig. 2.1. The MTJ resistance is determined by the magnetization directions of the two ferromagnetic layers, i.e., in parallel (low-resistance) or anti-parallel (high-resistance). The magnetization of one of the ferromagnetic layers (reference layer) is fixed while that of the other ferromagnetic layer (free layer) can be switched by applying a write current with proper polarization. The magnetization of the ferromagnetic layers can be either in parallel or perpendicular to the surface of the MTJ, namely, in-plane or perpendicular MTJ, as shown in Fig. 2.1(a) and (b), respectively. Fig. 2.1(c) shows a popular "1T1J" STT-RAM cell design where a NMOS transistor supplies the write and read current to the MTJ. The switching process of the MTJ is greatly affected by the amplitude of the current passing through it, which varies with the parametric variations of the MTJ and transistor. It is also affected by the thermal-induced fluctuations of magnetization precession. As pointed out by the prior art [32], this asymmetric structure results in a very unreliable '0(low)$\rightarrow$'1(high)' switching, i.e., with much longer switching time and wider distribution than the other switching direction.

The magnetization switching of the MTJ under a write current can be modeled by *Landau-Lifshitz-Gilbert* (LLG) equation as [26]:

$$\frac{dm_f}{dt} + \alpha(m_f \times \frac{dm_f}{dt}) = \frac{1}{2}\gamma H_k \Sigma_{i=1}^{4}(\frac{\Gamma_i}{l_t K}). \tag{2.1}$$

Here, M_s is free layer magnetization saturation and m_f represents the unit vector of the free layer. α is the Gilbert damping constant; γ is the gyromagnetic ratio; H_k is the Stoner-Wohlfarth switching field; l_t is the free layer thickness. The four torque (Γ) terms represent the factors that affecting the m_f dynamics – the uniaxial anisotropy, the easy-plane anisotropy, the Langevin random thermal field, and the spin torque term from the applied current. The parameters adopted in our macro-magnetic simulations are summarized in TABLE 4.1 [10].

2.2.2 Basics of NVSim

NVSim is a widely used open source simulation framework for circuit-level modeling of emerging nonvolatile memories like STT-RAM, ReRAM, PCM, etc [6]. It is developed to enable early-stage design space exploration before the memory is designed or fabricated. NVSim can extract the memory design metrics, i.e., read/write latency, read/write energy consumption, area, etc. under the given design constraints, or optimize the design parameters. NVSim also allows the users with only device knowledge to obtain block-level design specs via a user-friendly interface.

To simulate an STT-RAM design in NVSim, users are expected to specify the write current value and switching time for both SET and RESET operations. However, obtaining correct values of these parameters requires running macro-magnetic models, which is not supported in the current version of NVSim based on pure circuit-level simulation. Moreover, the current version of NVSim supports neither the statistical analysis of STT-RAM, e.g., the variations of write performance (errors) and energy consumptio, nor the impact of operating temperature.

2.3 NVSIM-VX^{S} FRAMEWORK

Fig. 2.2 presents the framework of our proposed NVSim-VX^{s}, which includes three important new features that are not supported by the existing deterministic STT-RAM simulators: the temperature-

Table 2.1: Device and Circuit Simulation parameters

	Parameter	Symbol	Value	Unit
Device Level	Mag. Saturation	M_s	230	kA/m
	Uniaxial Anisotropy	H_k	200	Oe
	Gilbert Constant	α	0.01	
	Free Later Thickness	l_t	1	nm
Circuit Level	Parameter	Mean	Std. Dev	
	Channel length	$L = 22{\sim}90nm$	$\sigma_L = 0.05L$	
	Channel width	$W = 2L{\sim}5L$	$\sigma_W = 0.05L$	
	Free layer volume	$V = L \times 2L \times 1nm^3$	$\sigma_V = 0.05V$	
	Resistance low	$R_L = 1000\Omega$	$\sigma_{R_L} = 0.05R_L$	
	Resistance high	$R_H = 2000\Omega$	$\sigma_{R_H} = 0.05R_H$	

aware statistical switching time model, the statistical energy consumption model, and the write error rate model. Compared to the current version of NVSIM, NVSim-VX^s possesses a more flexible and user-friendly interface to facilitate its probabilistic design philosophy, i.e. allowing users to set circuit and architecture parameters as the inputs and obtain cell/block level statistical design metrics from the outputs. Furthermore, the important switching pattern (i.e. number of '0'→'1' or '1'→'0' flipping's in write) dependent energy and reliability analysis is also enabled at block level.

2.3.1 Temperature-Aware Statistical STT-RAM Switching Time Model

The MTJ switching time variation is mainly generated from the following two torque terms: the Langevin random field and the spin torque, as suggested by Eq. (4.3). In specific, the randomness sources of the Langevin random field are the variations of MTJ surface area and the thickness of free layer while the spin torque is generated by the driving current, which is affected by process variations of both NMOS transistor and MTJ device [33].

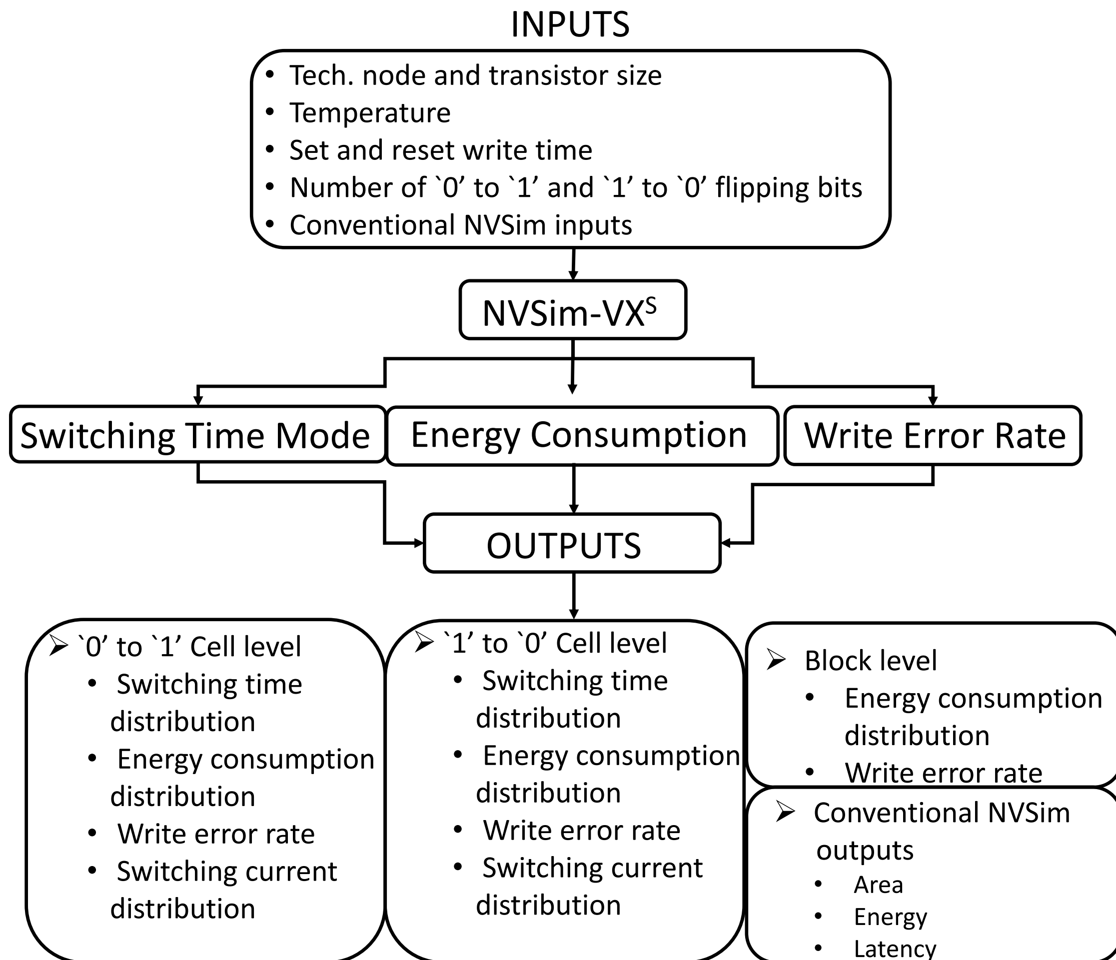

Figure 2.2: Framework of the NVSim-VX^s.

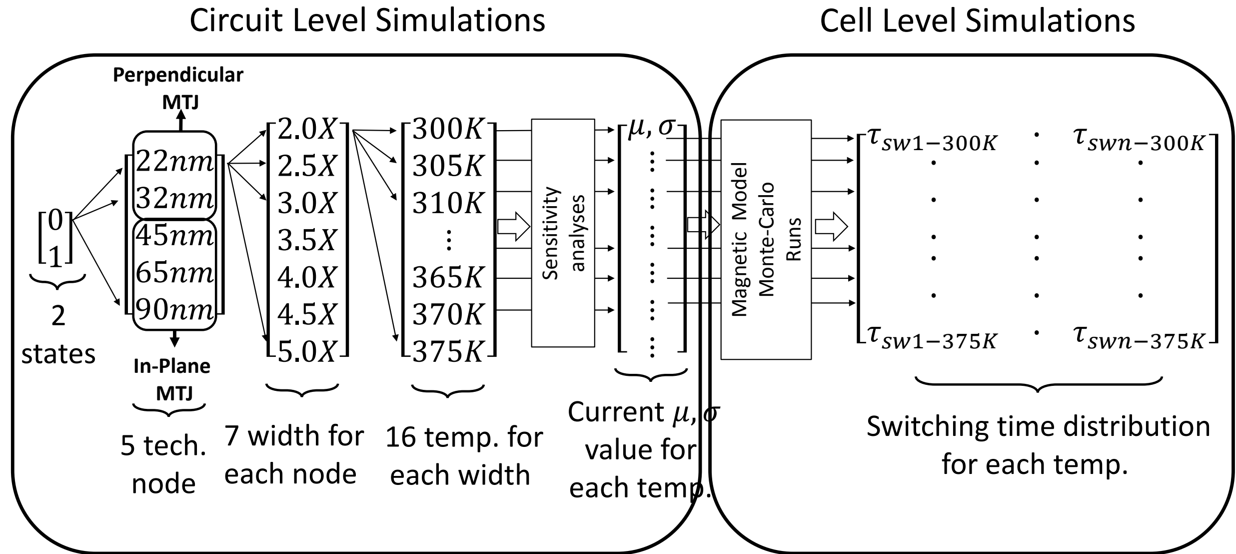

Figure 2.3: Detailed flow for circuit-level and cell-level simulations in NVSim-VX^s.

Note that the above two terms are also significantly affected by the fluctuation of operating temperature.To minimize the costly hybrid CMOS-Magnetic simulations required to capture all the parametric variabilities, our temperature-aware statistical STT-RAM switching time model is derived and simplified from extensive LLG model-based Monte-Carlo simulations, which cover 2 MTJ resistance switching directions, 5 technology nodes, 7 different transistor widths, and 16 different temperature points, as illustrated in Fig. 2.3. In the first step, sensitivity analysis is conducted at different temperatures to characterize the driving current distributions [27]. Variations of the transistor channel length, the transistor width, and the MTJ resistance are also taken into considerations. Simultaneously considering all variability parameters can reduce the computation complexity from $O(N^k)$ to $O(N)$, where K is the number of variability parameters and N is the number of samples for each parameter. In the second step, we integrate both driving current distributions and the Langevin random field into LLG equation under different temperatures to obtain the temperature-aware switching time distributions. Finally, a fast and compact timing model that directly links the switching time variation (i.e., mean and standard deviation) to temperature and driving current can be achieved. Based on the device parameters and simulation setup summarized in TABLE 4.1.

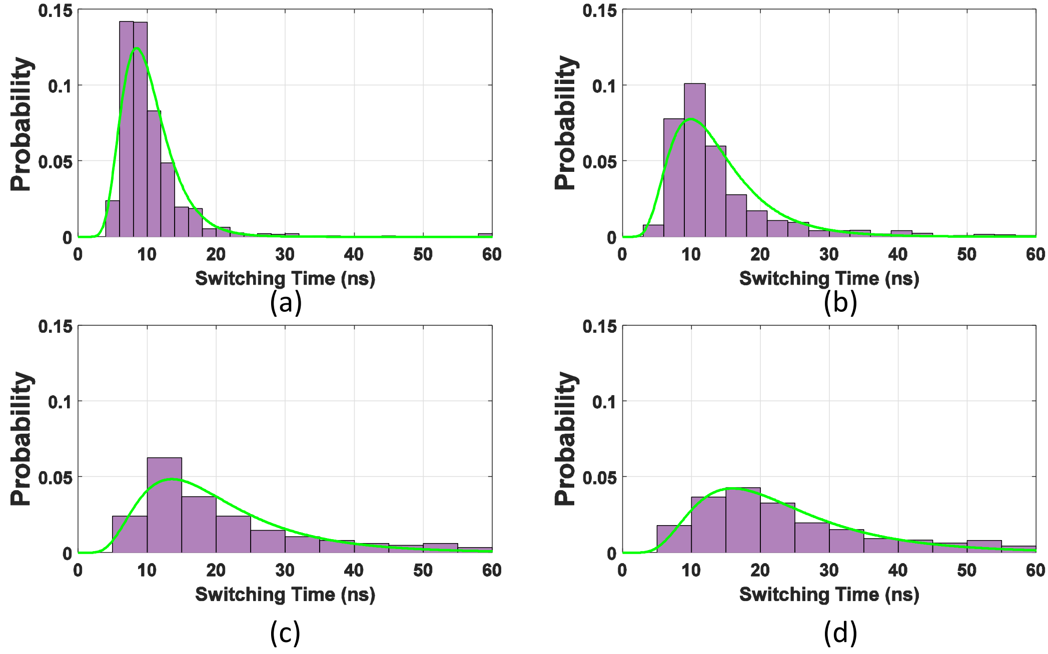

Figure 2.4: Simulated switching time distribution v.s. Log-normal distribution ($45nm$ technology node, $90nm$ transistor width, 4 different temperatures): a) $300K$ b) $325K$ c) $350K$ d) $375K$.

We performed Monte-Carlo simulations to obtain switching time distributions of '0'→'1' switching at 4 different temperatures for an STT-RAM cell with a $w = 2L$ NMOS transistor width at technology node $L = 45nm$, as depicted in Fig. 2.4. All the simulated switching time results are in excellent agreement with the Log-normal distributions at the concerned temperatures. As the temperature increases from $300K$ to $375K$, the distribution of the MTJ switching time becomes broader, indicating the increased impact of temperature on MTJ switching and hence the STT-RAM cell write reliability. As we shall show next, the corresponding mean (μ) and standard deviation (σ) of the Log-normal MTJ switching time distribution can be directly linked with the driving current and temperature using our model. Our further investigation suggests a linear approximation of the relationship between the μ/σ and temperature. This linear approximation of the temperature dependency of μ and σ of the MTJ switching time can be expressed by:

$$\begin{aligned} \sigma(w) &= m_\sigma(w) * T_n + \sigma_0(w), \\ \mu(w) &= m_\mu(w) * T_n + \mu_0(w). \end{aligned} \tag{2.2}$$

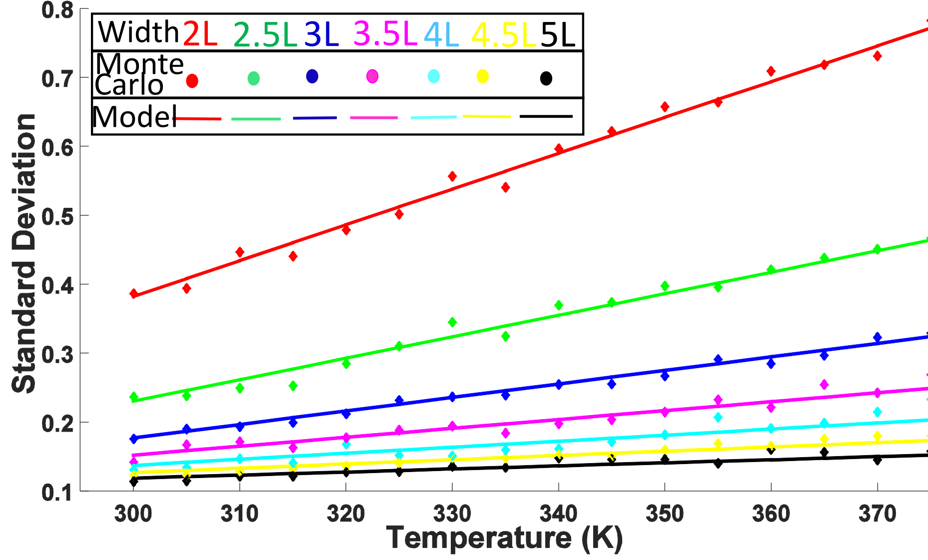

Figure 2.5: The standard deviation of switching time distributions for 7 transistor widths at $45nm$ technology node.

Here $m_\sigma(w)$ and $m_\mu(w)$ are the coefficient representing the temperature dependency of μ and σ at transistor width w. T_n is the normalized temperature. $\sigma_0(w)$ and $\mu_0(w)$ are the initial values of $\sigma(w)$ and $\mu(w)$, respectively, at $T_n = 0$.

Fig. 2.5 and 2.6 show the results of the linear approximation of σ and μ for 7 different transistor widths ($2L \sim 5L$, $L = 45nm$) at different temperatures ($300K \sim 375K$), respectively. For comparison purpose, the results of Monte-Carlo simulations are also presented. It can be observed that our linear model provides very accurate approximation of the Monte-Carlo simulation results at the whole covered ranges of transistor widths and temperatures. As transistor width increases, both the temperature dependency and the initial values of σ and μ monotonically decreases, implying a less sensitivity to the temperature change and improved thermal robustness.

For a specific transistor width w, the temperature dependency and the initial values of σ and μ – ($m_\sigma(w)$, $m_\mu(w)$) and ($\sigma_0(w)$, $\mu_0(w)$), are the functions of the MTJ driving current I as:

$$m_\sigma(w) = a_{m\sigma} * e^{b_{m\sigma} I},$$
$$m_\mu(w) = a_{m\mu} * e^{b_{m\mu} I}. \quad (2.3)$$

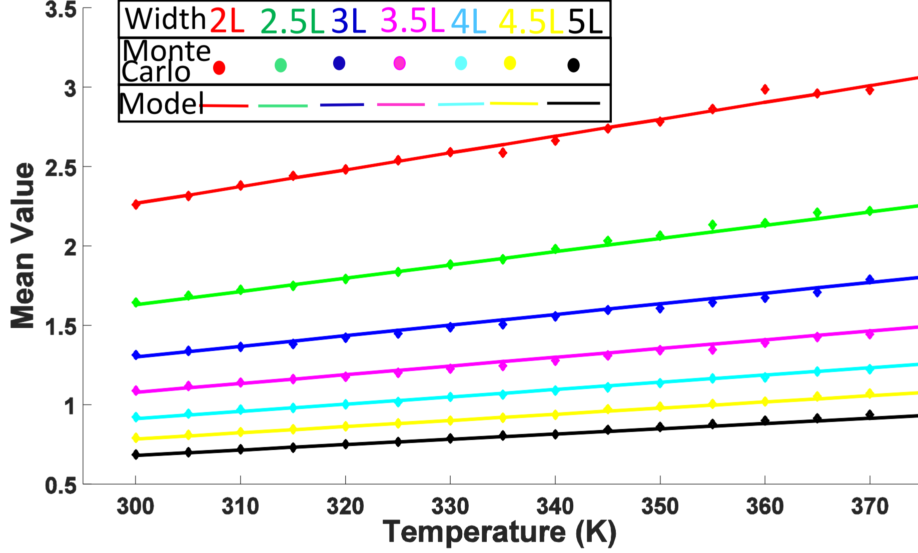

Figure 2.6: The mean of switching time distributions for 7 transistor widths at $45nm$ technology node.

and

$$\sigma_0(w) = a_{c\sigma} * e^{b_{c\sigma}I} + c_{c\sigma} * e^{d_{c\sigma}I},$$
$$\mu_0(w) = a_{c\mu} * e^{b_{c\mu}I} + c_{c\mu} * e^{d_{c\mu}I}. \tag{2.4}$$

Here, the driving current I is determined by the NMOS transistor width w at different technology node. a_i, b_i, c_j and d_j are technology-dependent fitting parameters where $i = m\sigma, m\mu, c\sigma, c\mu$, $j = c\sigma, c\mu$. Fig. 2.7 depicts the simulated relationship between $m_\sigma(w)$, $m_\mu(w)$, $\sigma_0(w)$, $\mu_0(w)$ v.s. I based on our model in Eq. (2.3) and (2.4). The results includes the data at 7 different transistor widths (i.e., $2L \sim 5L$, $L = 45nm$). To validate our model, the Monte-Carlo simulation results are also included. The results show that our model matches the Monte-Carlo simulations very well in all the simulated cases.

By substituting Eq. (2.3) and (2.4) into Eq. (2.2), the MTJ switching time distributions can be expressed by:

$$\sigma(w) = (a_{m\sigma} * e^{b_{m\sigma}I}) * T_n + (a_{c\sigma} * e^{b_{c\sigma}I} + c_{c\sigma} * e^{d_{c\sigma}I}),$$
$$\mu(w) = (a_{m\mu} * e^{b_{m\mu}I}) * T_n + (a_{c\mu} * e^{b_{c\mu}I} + c_{c\mu} * e^{d_{c\mu}I}). \tag{2.5}$$

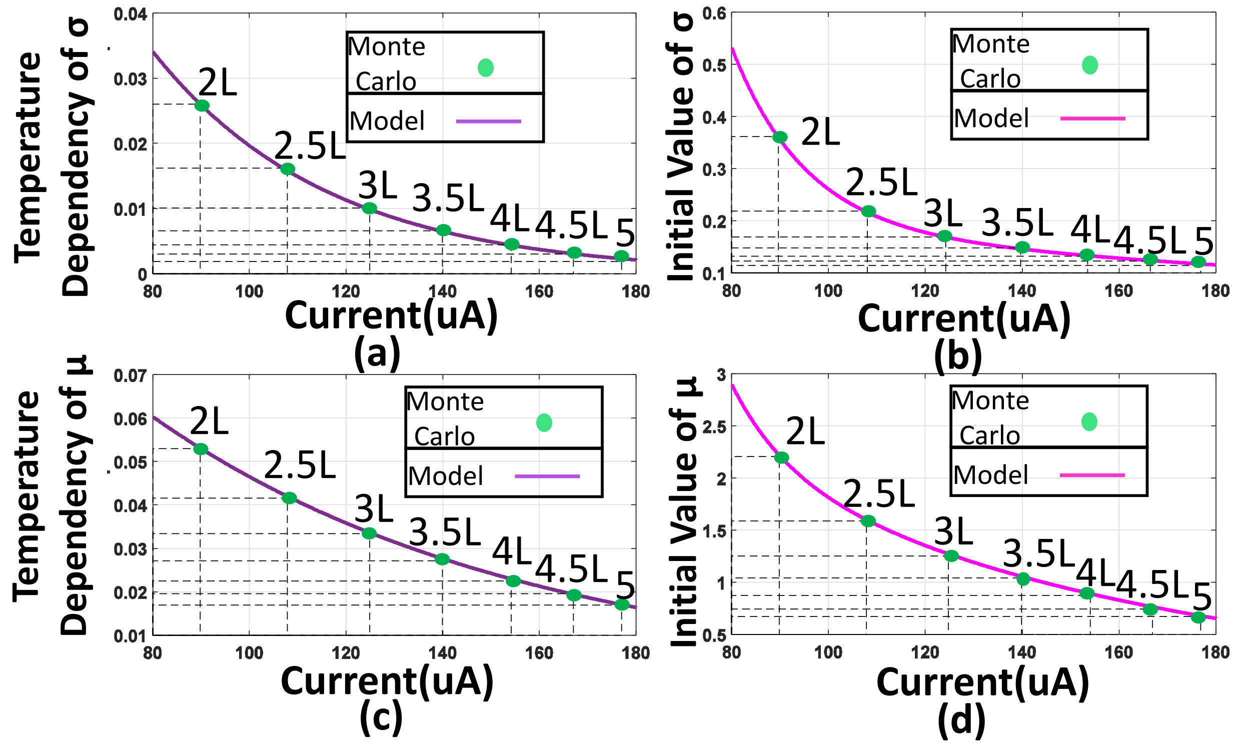

Figure 2.7: The simulated results of our model and Monte-Carlo simulations for key parameters of σ and μ: (a) $m_\sigma(w)$ v.s. I; (b) $\sigma_0(w)$ v.s. I; (c) $m_\mu(w)$ v.s. I; (d) $\mu_0(w)$ v.s.I.

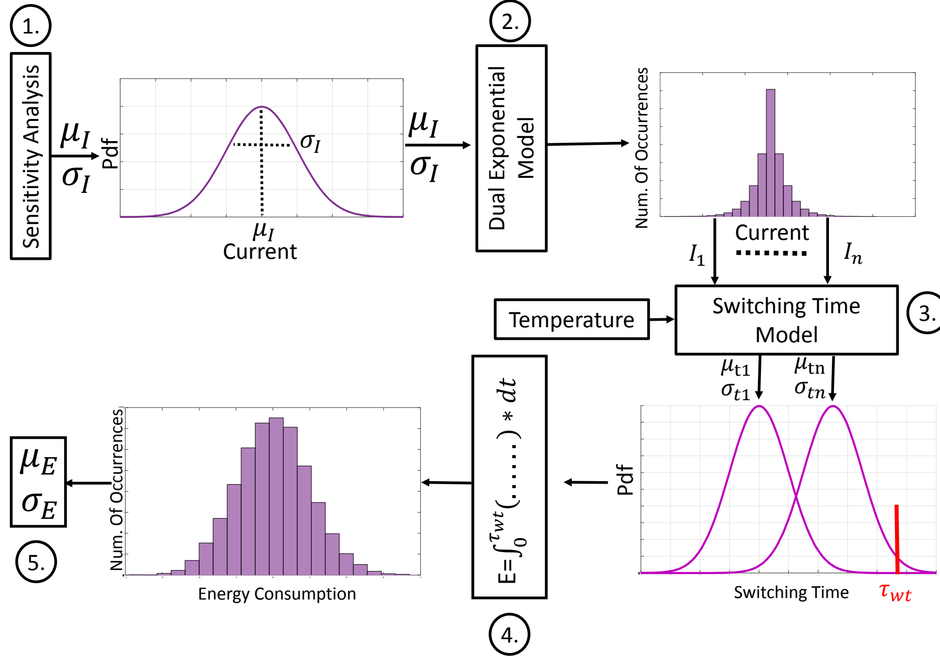

Figure 2.8: Overview of the statistical energy consumption characterization flow.

Although the above illustrated examples are based on 45nm technology, our developed temperature aware statistical STT-RAM switching time model is capable to capture the switching time variaions for different transistor sizes ($2L \sim 5L$), different temperatures ($300K \sim 375K$) at different technology nodes ($22 \sim 90nm$), showing its adaptivity and scalability in advanced technology nodes.

2.3.2 STT-RAM Statistical Energy Modeling

In current version of NVSim, the write energy of STT-RAM is deterministically modeled without considering any fluctuations in write operations. The cell-level write energy is directly extracted from the given SET/RESET current, applied voltage, and write time. However, as previously discussed, the switching time of the MTJ in each STT-RAM cell varies with the parametric variations of the MTJ and the NMOS transistor, and is influenced by thermal fluctuations. The MTJ resistance states, which affect the write current through the device, also follows some distributions. Hence, in NVSim-VX^s, we characterize the statistical STT-RAM write energy consumption.

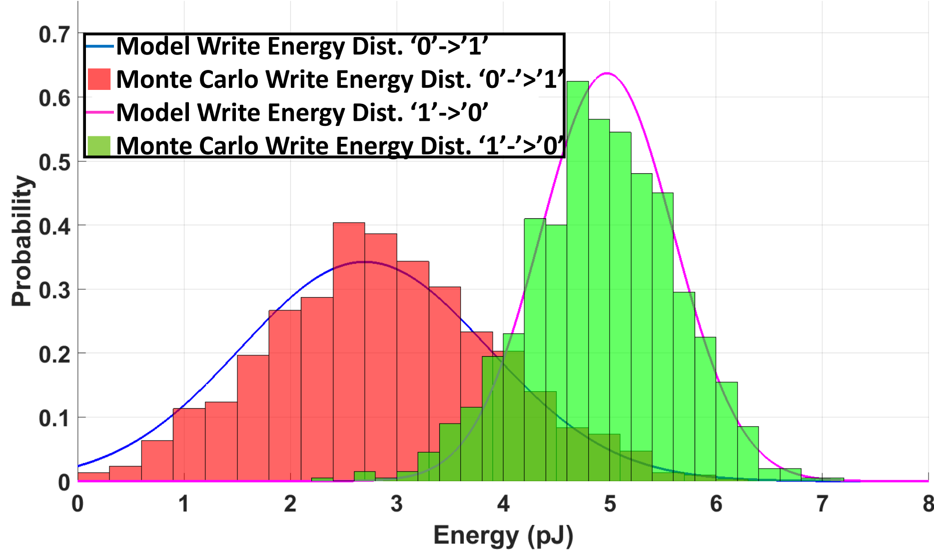

Figure 2.9: The write energy consumption distribution comparison of our model v.s. Monte-Carlo at $45nm$ technology node, $90nm$ transistor width, $60ns$ write pulse width.

The energy consumption of an STT-RAM cell during a write operation, i.e., '0'→'1' switching, can be calculated using Joule's first law as:

$$E = (I_H * \tau_{sw} + I_L * (\tau_{wt} - \tau_{sw})) * V. \tag{2.6}$$

Here I_H is the initial high driving current at low resistance (R_L) state and I_L is the low post-switching current after the resistance state switches to high resistance (R_H). τ_{sw} is the actual MTJ switching time and τ_{wt} is the writing period (write time) for which the programming voltage (V) is applied. Note that we ignore the oscillation of the driving current generated by the magnetic precession. Also, both I_H and τ_{sw} are correlated, and subjected to the variations from CMOS/MTJ device and thermal fluctuations.

Fig. 2.8 depicts the overview of our proposed STT-RAM statistical write energy model, including the following five steps:

1. **Derive current information**: Obtain the driving current statistical information by conducting sensitivity analysis, as is discussed in statistical STT-RAM switching time model;
2. **Generate current sample**: Generate the driving current samples over the dual exponential current distribution and the statistical information [27];

3. **Obtain switching time distribution**: Send the driving current samples and the temperature to the temperature-aware statistical switching time model developed in Section 2.3.1 and generate different switching time distributions for each sample;
4. **Calculate statistical energy**: Calculate the energy by doing integral over user-specified write time at each driving current sample and switching time distribution pair as below:

$$E_i = \int_0^{\tau_{wt}} (I_{Hi} * t + I_L * (\tau_{wt} - t)) * V * f_i(t) * dt. \tag{2.7}$$

   Here I_{Hi} denotes the i_{th} sample of the I_H obtained from dual exponential model, $f_i(t)$ is the probability density function of corresponding switching time distribution of the current sample. I_L is the low post-switching current.
5. **Dump energy distribution**: Calculate the mean (μ_E) and standard deviation (σ_E) of the write energy consumption as:

$$\mu_E = \frac{\sum_{i=1}^{n} E_i * f_i}{\sum_{i=1}^{n} f_i} \quad and \quad \sigma_E = \sqrt{\frac{\sum_{i=1}^{n} (E_i - \mu_E)^2 * f_i}{\sum_{i=1}^{n} f_i}}. \tag{2.8}$$

   Here f_i is the number of occurrences of energy value E_i for the i current sample.

Our simulations show that the write energy consumption roughly follows a Gaussian distribution whose mean and standard deviation can be obtained from step 5. Fig. 2.9 shows the write energy distributions of both MTJ switching directions obtained by our model and Monte-Carlo simulations at $\tau_{wt} = 60ns$. The temperature is 350K and the transistor width $W = 2L$, $L = 45nm$. The results show that our model approximates the Monte-Carlo simulations very closely. Fig 2.10 compares the mean value of the write energy consumption of STT-RAM cell designs with various transistor widths ($2L \sim 5L$, $L = 45nm$) under different temperatures at '1'→'0' switching. Again, our model can always provide the results very close to that of the Monte-Carlo simulations with the simulated transistor sizes. As the temperature increases, energy consumption reduces almost linearly at large transistor widths (i.e. $3.5L \sim 5L$) because of the narrow distribution of the MTJ switching time (τ_{sw}).

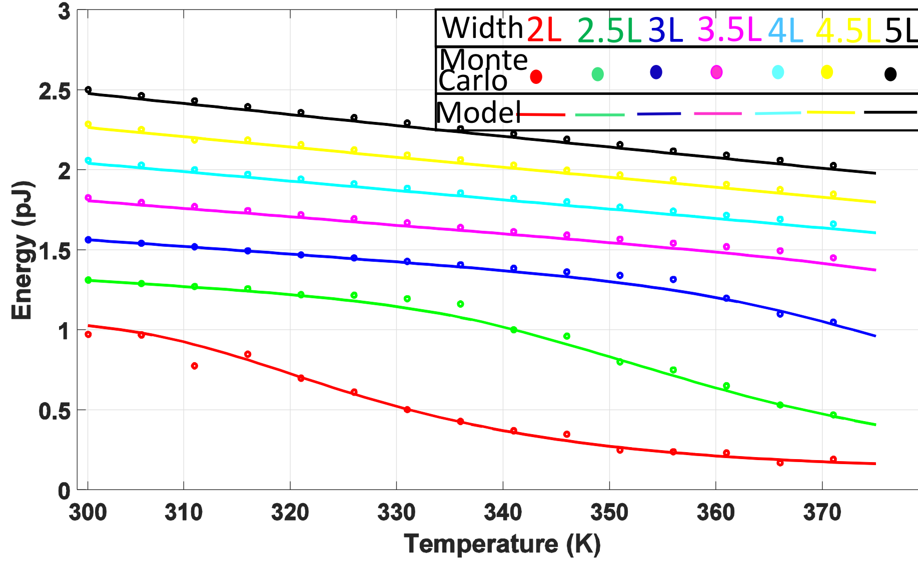

Figure 2.10: The mean of the energy consumption at $10ns$ write pulse width for 7 different transistor widths at $45nm$ technology node.

On the contrary, the changing rate of the energy consumption with temperature becomes non-linear when the transistor width is small, indicating a high sensitivity to temperature change. It again proves that a large access transistor can help reducing the thermal-induced performance variations of STT-RAM. Fig. 2.11 shows the energy consumption of $W = 2.5L$ at three different τ_{wt}. Reducing the τ_{wt} slightly degrades the linearity of the temperature dependency of the energy consumption.

2.3.3 STT-RAM Write Error Rate

An STT-RAM write failure happens if the MTJ switching cannot complete within the applied write pulse width (or the write time τ_{wt}). Following technology scaling, write reliability emerges as one of the main challenges in STT-RAM designs. An accurate and fast prediction of STT-RAM write error rate become essential but cannot be obtained by performing conventional deterministic circuit simulations.

Traditionally, calibrating the write error rate of an STT-RAM cell requires two runs of Monte-Carlo simulations and one sample/distribution processing: Firstly, circuit-level simulations are conducted to get the STT-RAM switching current distribution by considering all parametric vari-

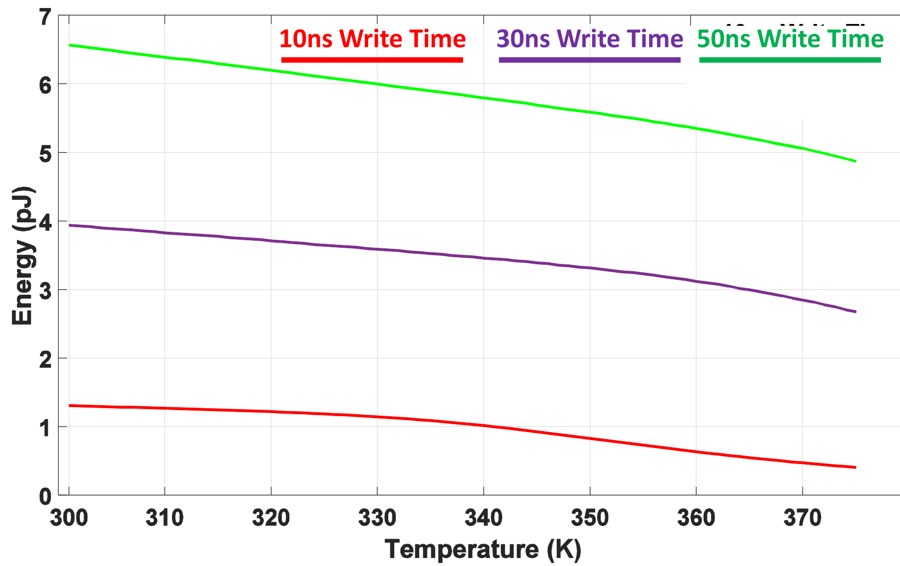

Figure 2.11: The mean of energy consumption at different write pulse widths for $w = 2.5L$ transistor width ($L = 45nm$).

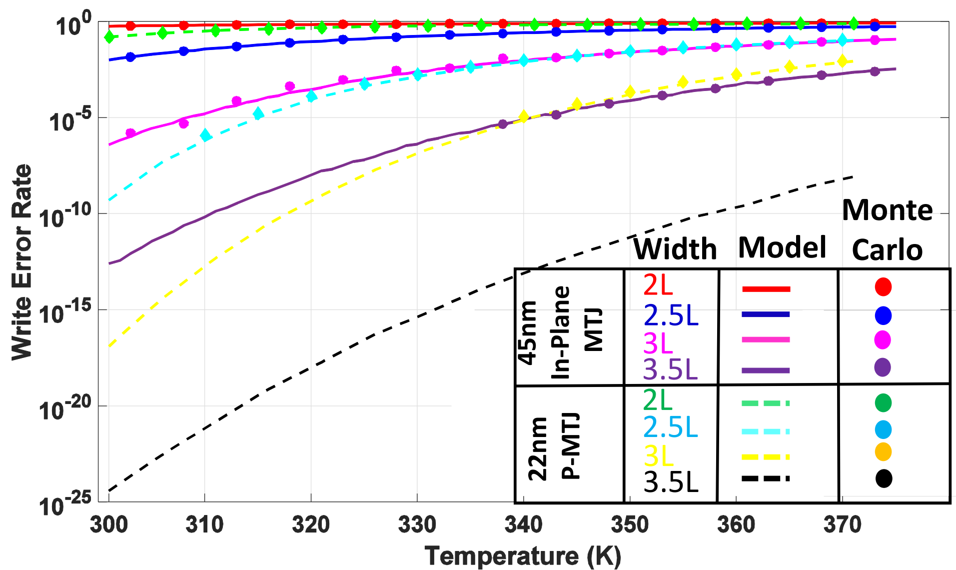

Figure 2.12: Simulated write error rate v.s. temperature from our model and Monte-Carlo for $22nm$ perpendicular MTJ and $45nm$ in-plane MTJ ($10ns$ write pulse width, '0'→'1' switching).

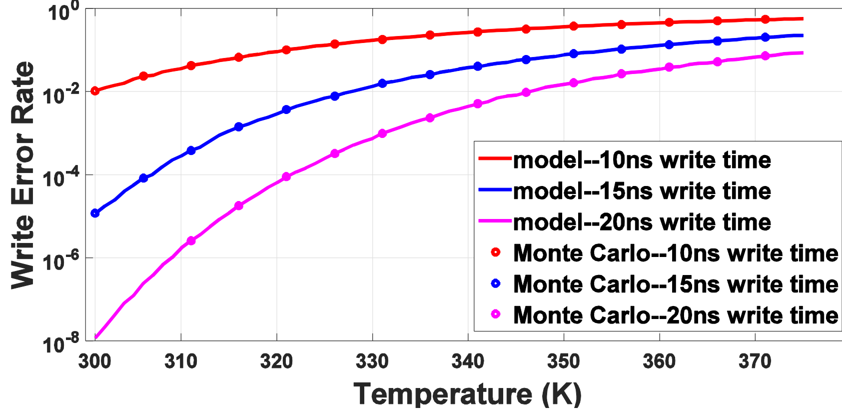

Figure 2.13: Simulated write error rate v.s. temperature from our model and Monte-Carlo at different write time (transistor width $w = 2.5L$,'0'→'1' switching).

abilities; Secondly, the current distribution is sent to a macro-magnetic model and the second-round Monte-Carlo simulations are performed to obtain the STT-RAM switching time distribution; Finally, the generated switching time samples or distribution must be compared with the given write time to calculate the write failure.

In NVSim-VX^s, the samples of the STT-RAM switching time can be obtained from the embeded temperature-aware statistical switching timing model without running the two costly Monte-Carlo simulations. However, the normal write error rate of a STT-RAM cell is so low that a large number of samples of the STT-RAM switching time still need to be generated to calculate the error rate. In our work, we introduce the mixture importance sampling technique in NVSim-VX^s to reduce the write error rate calculation cost as [13, 14]:

$$E_{p(x)}[\theta] = E_{g(x)}[\theta * \frac{p(x)}{g(x)}]. \tag{2.9}$$

Here $p(x)$ denotes the switching time probability density function and $g(x)$ is the distorted sampling function defined as;

$$g_\lambda(x) = \lambda_1 p(x) + \lambda_2 U(x) + (1 - \lambda_1 - \lambda_2) p(x - \mu_s), \tag{2.10}$$

where $0 \leq \lambda_1 + \lambda_1 < 1$. $U(x)$ is the uniform pdf; μ_s is the shifted center and is chosen experimentally.

Fig. 2.12 compares the results of write error rate of an STT-RAM obtained from NVSim-VX^s and Monte-Carlo simulations, respectively, for a $22nm$ perpendicular MTJ and an $45nm$ in-plane MTJ at '0'→'1' switching. Our model can achieve good accuracy at the 4 simulated transistor widths and precisely describe the changing trend of the write error rate with the temperature. Interestingly, Fig. 2.12 shows that the write error rate of the $22nm$ perpendicular MTJ always outperforms the one of $45nm$ in-plane MTJ at the similar relative transistor sizes ($2L \sim 3.5L$) and the same temperature. This result validates the conclusion that perpendicular STT-RAM is more promising at scaled technology node, i.e., below $45nm$.

Fig. 2.13 shows the simulated write error rate over different temperatures for the $45nm$ in-plane MTJ with different write pulse widths. Increasing the write time can greatly reduce the write error rate at low temperatures; However, limited improvement is observed at higher temperatures.

2.3.4 Block Level Extension for NVSim-VX^s

In NVSim, memory is usually organized as three different hierarchies: bank, mat and subarray. Bank is at the top of the hierarchy and it contains multiple mats that can be operated simultaneously. A mat is further composed of multiple subarrays, which are the elementary structure of the NVSim.

To make the NVSim-VX^s suitable for architecture-level simulation, we extended our cell level model to block level. NVSim-VX^s is capable to calculate the block level write energy or error rate more precisely by taking the switching pattern into consideration. To the best of our knowledge, this is the first time that such an important feature is integrated into nonvolatile memory simulators.

2.3.5 Block Level Energy Consumption

Write energy consumption of an STT-RAM cell is distinctive at two switching directions. The block-level energy estimation will be more accurate if the users can provide the switching patterns of the accessed block by considering the difference between the incoming data and the stored data. There are four possibilities of the bit switching: '0'→'1', '1'→'0', '0'→'0', and '1'→'1'. For a switching of 'i'→'j', $i, j = 0$ or 1, the mean and the standard deviation of total energy consumption spent on 'i'→'j' switching's in a memory write can be expressed as:

$$\mu_{eT,ij} = N_{F,ij} * \mu_{e,ij} \quad and \quad \sigma_{eT,ij} = N_{F,ij} * \sigma_{e,ij}. \tag{2.11}$$

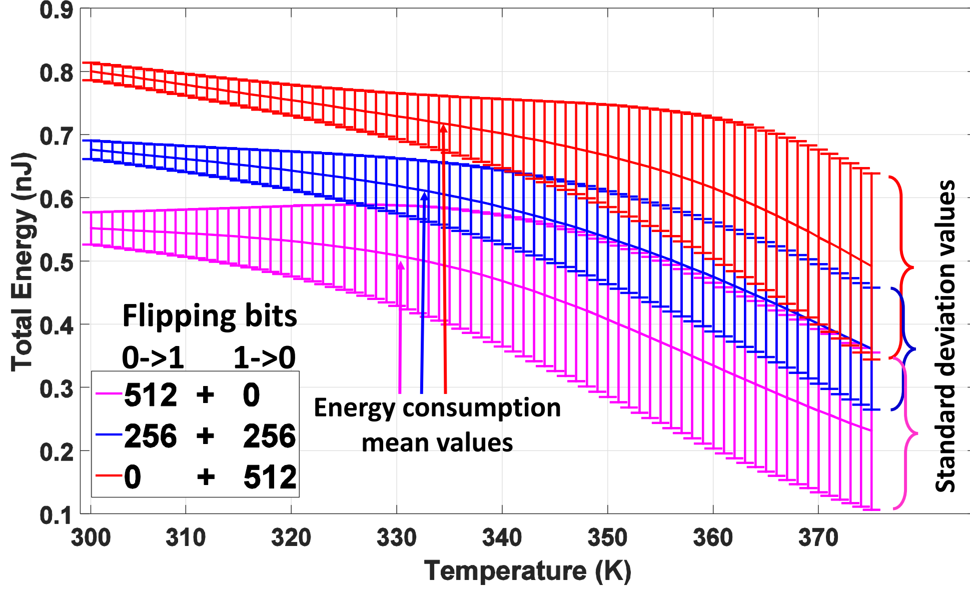

Figure 2.14: The mean and standard deviation of block-level energy consumption for 3 different flipping combinations at $10ns$ write time and $w = 3L$ transistor width, $L = 45nm$.

Here, $\mu_{e,ij}$ and $\sigma_{e,ij}$ are the mean and the standard deviation of the write energy of the STT-RAM cell, respectively, at 'i'→'j' switching.

The distribution of the write energy consumption can be then described by:

$$\mu_{eT} = \sum_{ij} \mu_{eT,ij} \quad and \quad \sigma_{eT} = \sqrt{\sum_{ij} \sigma_{eT,ij}{}^2}. \tag{2.12}$$

When $i = j$, the stored data is actually overwritten by the same value. The write energy can be zeroed by applying a "read-before-write" technique to eliminate this redundant operation. "Read-before-write" is the default mode of NVSim-VX^s and the energy of one read operation is automatically included in the write energy calculation.

Fig. 2.14 shows an example of NVSim-VX^s results of the nominal write energy consumptions and their standard deviations of a 512-bit block with three different switching patterns.

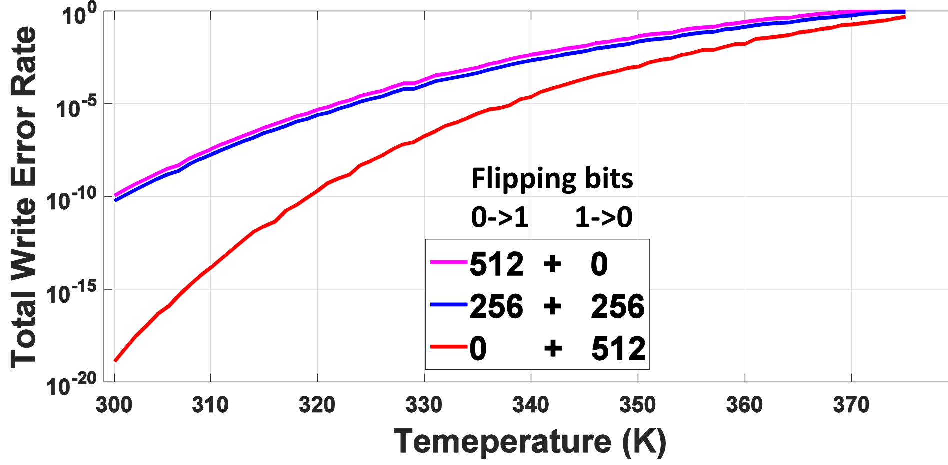

Figure 2.15: The block level write error rate for 3 different flipping combinations at $10ns$ write time and $w = 3L$ transistor width, $L = 45nm$.

2.3.6 Block Level Write Error Rate

Due to the asymmetry of the write error rates at both switching directions of STT-RAM cells, the block level write error rate is also highly related to the switching patterns of the array. The array-level write error rate can be easily extracted from the cell level results for a given array size as:

$$P = 1 - \prod_{i=0}^{1} (1 - W_{ER_i})^{N_{F_i}} \tag{2.13}$$

Here, we assume the location information of bit switchings is known during programming. N_{F_i} and W_{ER_i} denote the number of flipping bits and the bit write error rate, respectively for the switching of 'i'→'$\bar{i}$'. Fig. 2.15 shows an application example of NVSim-VX[s] in simulating the write error rate of a 512-bit block with different switching patterns.

2.4 CONCLUSIONS

The existing deterministic nonvolatile memory simulators are incapable to capture the probabilistic behaviors of STT-RAM incurred by process variations and thermal fluctuations. In this work, we introduce NVSim-VX[s], an enhanced version of NVSim, to facility the need of statistical STT-

RAM modeling for write performance, errors and energy consumption. The example simulation results show that increasing temperature dramatically degrades the STT-RAM's performance and reliability. Moreover, NVSim-VX^s simulations quantitatively validate the advantages of perpendicular STT-RAM over in-plane STT-RAM, showing its promising capability to predict the scaling trend of STT-RAM. We plan to release NVSim-VX^s as the new member of NVMSim family in the near future and continue developing the statistical versions of NVSim for other types of emerging memories with more enhanced features.

3.0 SPIN-HALL ASSISTED STT-RAM DESIGN AND DISCUSSION

3.1 INTRODUCTION

Conventional memory technologies like SRAM, DRAM and Flash memory have been widely utilized in modern computer systems. However, as technology node continues to scale down, these electrical charge-based memory technologies suffer from high leakage power and large process variations that cause severe reliability issues. In order to overcome these problems, many new memory technologies, including Spin-Transfer Torque Random Access Memory (STT-RAM), have been studied. Although STT-RAM features many attractive characteristics like non-volatility, low standby power, and high cell density [20, 31, 36], it also has many drawbacks such as long programming latency and high programming energy etc.

Spin-Hall Effect (SHE) assisted STT-RAM (or SHE-RAM) was recently proposed to solve the challenges in conventional STT-RAM designs [24]. By eliminating incubation delay, programming time and/or energy of SHE-RAM cells can be substantially reduced, compared to conventional STT-RAM.

Several access schemes of SHE-RAM were also discussed in [12]. However, these schemes require either very sharp writing pulse or external magnetic field which introduces additional fabrication process/cost. In another design in [15], a single bit is represented by two MTJs and four transistors. This design reduces storage density and also increases access power consumption. Nonetheless, a design that can maximize the benefits of SHE effects is still highly desired.

In this work, we proposed two SHE-RAM designs aiming at different applications. The first one is named as “High Density SHE-RAM”, which targets off-chip memory application requiring high cell density. The high cell density of High Density SHE-RAM is ensured by deploying a source line shared by all memory bits to supply the SHE current.

Only one transistor is needed to control the whole shared source line. The second one is named as "Disturbance Free SHE-RAM", which targets applications where reliability is the major concern. The potential disturbance to the unselected bits in the High Density SHE-RAM is eliminated in the Disturbance Free SHE-RAM by inserting isolating transistor between the cells and sharing the SHE current among the bits on the same word line.

The remain of this proposal is organized as follows: Section 3.2 presents the basics of SHE-RAM and the model of spin-hall assist; Section 3.3 introduces the designs of High Density SHE-RAM and Disturbance Free SHE-RAM; Section 3.4 concludes our work.

3.2 BASICS OF SHE-RAM

In conventional STT-RAM, data is stored as the resistance state of a *Magnetic Tunneling Junction* (MTJ) device, which consists of two ferromagnetic layers, namely, *reference layer* and *free layer*, and a tunneling oxide layer, as shown in Fig. 3.1(a). The relative magnetization orientations of these two ferromagnetic layers determine the resistance of the MTJ. That is being said, when their magnetization orientations are in parallel (anti-parallel), the MTJ is in its low (high) resistance state. When a current is injected to the MTJ, the current is spin polarized after passing through the RL, and exerts a torque to the FL and change its orientation.

At the beginning of the switching process of the MTJ, the angle between the magnetization vectors of the free layer and the reference layer will be either $0°$ or $180°$. In both cases, the torque exerted by the spin polarized current will be zero because the cross product of two vectors with the same direction is zero. In conventional STT-RAM design, this initial angle may be disturbed by thermal fluctuations [24]. The time needed to disturb the free layer magnetization from the "perfect" alignment with the reference layer, called "incubation delay", may be up to several nanoseconds.

Different from conventional STT-RAM, SHE-RAM contains an electrode underneath the perpendicular MTJs. Here the magnetization orientation of the two ferromagnetic layers of the MTJ are along the axis $\hat{\mathbf{z}}$. When an assist SHE current pulse is applied on this electrode, an in-plane polarized current whose polarization direction is along the axis $\hat{\mathbf{y}}$ is injected into the MTJ.

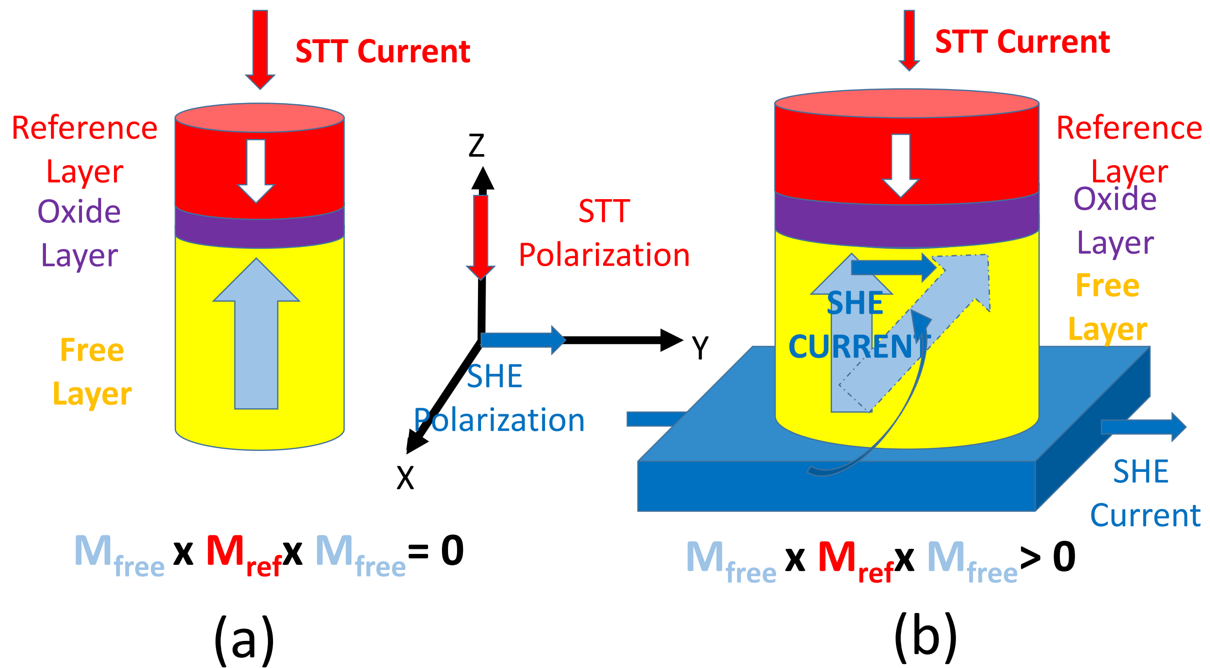

Figure 3.1: (a) A MTJ whose reference and free layer are in anti-parallel. The exerted torque is zero as the two magnetization vectors are in parallel. (b) Free layer magnetization vector makes an angle under the effect of the SHE current, and STT becomes available to switch the magnetization of the free layer since it is not in parallel to that of the reference layer any more.

The SHE current pulls the free layers magnetic vector from $0°$ or $180°$ to an intermediate angle, as shown in Fig. 3.1(b). Since the magnetization orientations of the free layer and the reference layer are no longer in parallel, the torque exerted by the spin polarized current will be larger than zero. Hence, the SHE current is able to assist the spin polarized current to switch the MTJ much faster. The incubation delay is eliminated and hence, both the switching time and the switching energy consumption are reduced compared to the conventional STT-RAM.

The magnetization dynamics of the MTJ free layer in SHE-RAM design can be modeled by solving the Landau-Lifshitz-Gilber (LLG) equation [21] with SHE current modification [24] as:

$$\begin{aligned}
&\tfrac{\partial \mathbf{M}}{\partial t} = -\gamma\mu_0 \left(\mathbf{M} \times \mathbf{H}_{\text{eff}}\right) + \tfrac{\alpha}{M_s}\left(\mathbf{M} \times \tfrac{\partial \mathbf{M}}{\partial t}\right) + \\
&\tfrac{c_{\text{SHE}}}{M_s^2}\left(\mathbf{M} \times \hat{\sigma}_{\text{SHE}} \times \mathbf{M}\right) + \tfrac{c_{\text{STT}}}{M_s^2}\left(\mathbf{M} \times \hat{\mathbf{m}}_{\text{ref}} \times \mathbf{M}\right) + \\
&\tfrac{\beta_{\text{MTJ}}}{M_{\text{s}}}\left(\mathbf{M} \times \hat{\mathbf{m}}_{\text{ref}}\right),
\end{aligned} \tag{3.1}$$

where $\mathbf{M}$ is the magnetization vector of the free layer, $\mathbf{H}_{\text{eff}}$ is the effective magnetic field vector, α is Gilbert damping constant and *Ms* is magnetization saturation, c_{SHE} is spin-Hall torque coefficient, c_{STT} is spin-transfer torque coefficient and β_{MTJ}=0.25c_{STT} as observed experimentally [24], γ is electron gyro magnetic ratio, μ_o is permeability.

3.3 SHE-RAM DESIGNS

In conventional STT-RAM, the MTJ switching can be accelerated by three means: 1) increasing the program current; 2) relaxing the MTJ non-volatility (e.g., by reducing the volume of free layer); and 3) applying an external magnetic field [9]. However, increasing the program current and applying an external magnetic field incur a large power consumption while non-volatility relaxation degrades the retention time of the STT-RAM cell [30]. As a comparison, spin-hall effect offers a very affordable option for programming performance improvement of the MTJ. Based on spin-hall effect, we propose two SHE-RAM designs, namely, High Density SHE-RAM and Disturbance Free SHE-RAM, aiming different applications. The device parameters used in the relevant analysis are summarized in Table 3.1.

Table 3.1: Summary of device parameters

Device	Parameters	Value	Std. Dev.
Tran-sistor	Channel length L	$45nm$	$2.25nm$
	Channel width W	$90 - 720nm$	$2.25nm$
	Threshold voltage V_{th}	$0.466V$	$30mV$
MTJ	MTJ Volueme	$45 \times 90 \times 1nm^3$	5%
	High and low resistance	$2000/1000\Omega$	
	Magnetization saturation	$800emu/cc$	
	Uniaxial anisotropy Hk	$3400Oe$	
	Gilbert damping constant α	0.01	

3.3.1 High Density SHE-RAM

High Density SHE-RAM is designed for capacity demanding applications such as off-chip memory. A High Density SHE-RAM cell contains only one transistor and one MTJ, as shown in Fig. 3.2. A SHE metal wire and a SHE control transistor are shared by all the SHE memory cells connected to the same source line. The programming current to the MTJ of each memory cell is supplied by the transistor connected to the MTJ. During the write operation of a memory cell, both the SHE control transistor and the corresponding cell transistor are turned on simultaneously and the polarization of the programming current is determined by the biases applied on the bit line and the source line.

As a perpendicular MTJ is applied, write operations do not have any requirement on the direction of the SHE current. For example, writing '1' and '0' almost equally benefit from a SHE current flowing from SHE control transistor to source line and vice versa becayse SHE currents with both polarizations can disturb the initial alignment of the magnetization direction of the free layer. Similar to the conventional STT-RAM, when the MTJ programming current flows from the bit (source) line to the source (bit) line, the magnetization orientation of the free layer will switch to the same as (opposite to) that of the reference layer, indicating logic '0' ('1'). The influence of the SHE current on the dynamics of the magnetization vector of the free layer is virtually depicted

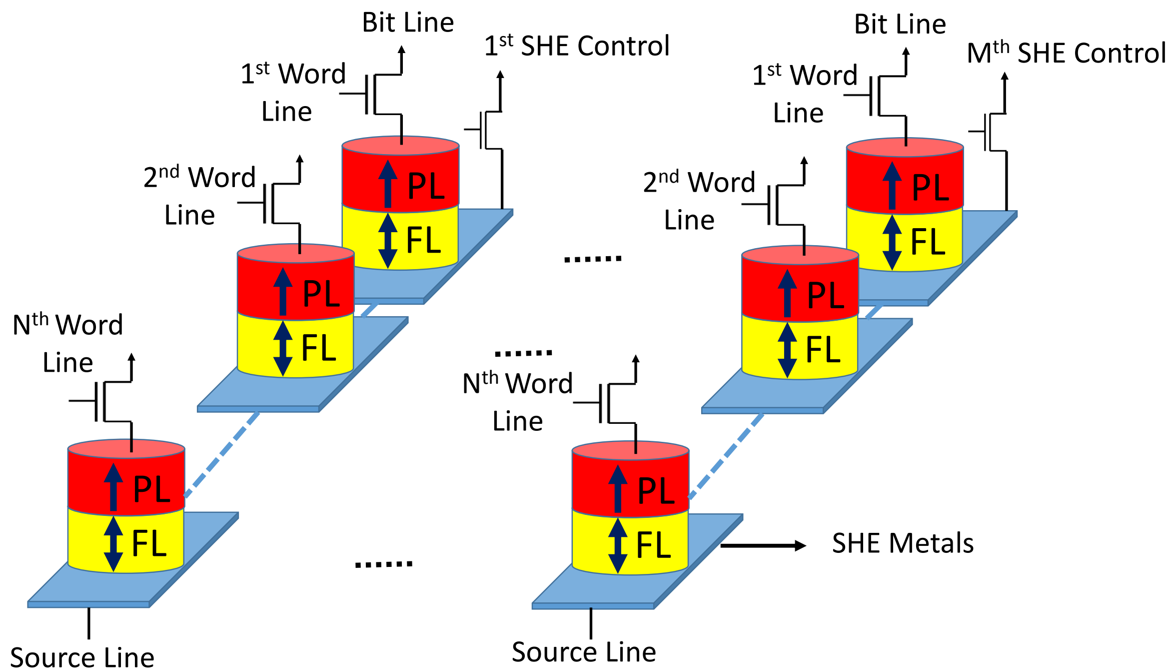

Figure 3.2: Illustration of High Density SHE-RAM. SHE current is controlled by the SHE control transistor and MTJs are accessed through the connected word line transistors. During write operations, the data is still written by applying appropriate bias on the source line and the bit line.

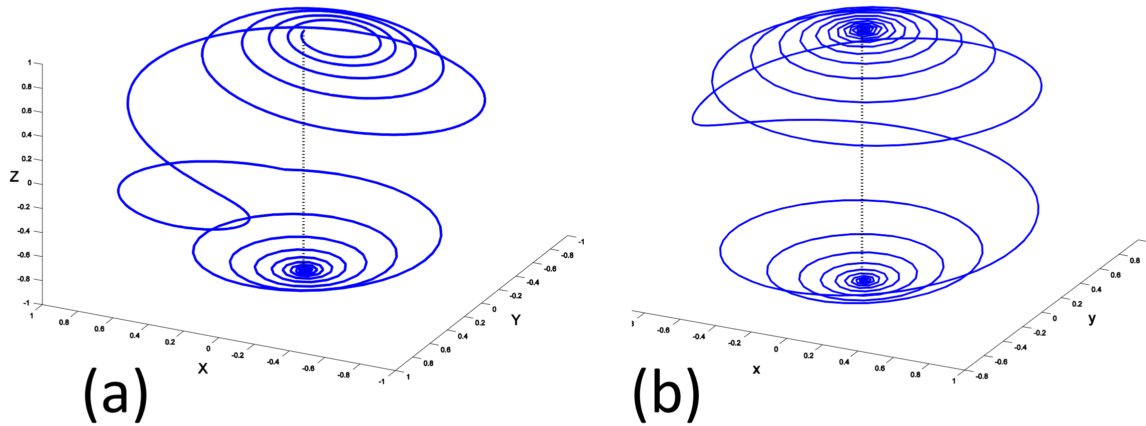

Figure 3.3: (a) Free layer's magnetic moment under the effect of SHE and STT currents. The oscillations starts from the deviated alignment of the free layer. (b) Free layer's magnetic moment under the effect of only STT current. Incubation delay takes place.

in Fig. 3.3(a). When the SHE current is applied, the magnetization vector of the free layer takes almost no time to deviate from the initial position. After that, the magnetization vector of the free layer quickly switches to the target state under the impact of the STT current. As a comparison, if only STT current is applied, significant oscillations occur when the MTJ deviates from the initial position, as shown in Fig. 3.3(b).

Note that the SHE current is required only at the beginning of the write operation to disturb the initial alignment of the free layer. Hence, for energy saving purpose, the SHE current can be supplied for only a very short time, i.e., sub-nanosecond, rather than the whole writing process. Fig. 3.4 compares the switching time of the MTJ when different SHE current pulse width is applied under different sizes of the cell transistor in a High Density SHE-RAM cell. The device and circuit level parameters are summarized in TABLE 3.1. The SHE current is fixed at 20μA. In general, the longer the SHE current pulse width is, the shorter the MTJ switching time will be: when the transistor width equals 90nm, raising the SHE current pulse width from 0.5ns to 0.7ns will reduce the MTJ switching time by half. However, as the transistor width increases, the difference

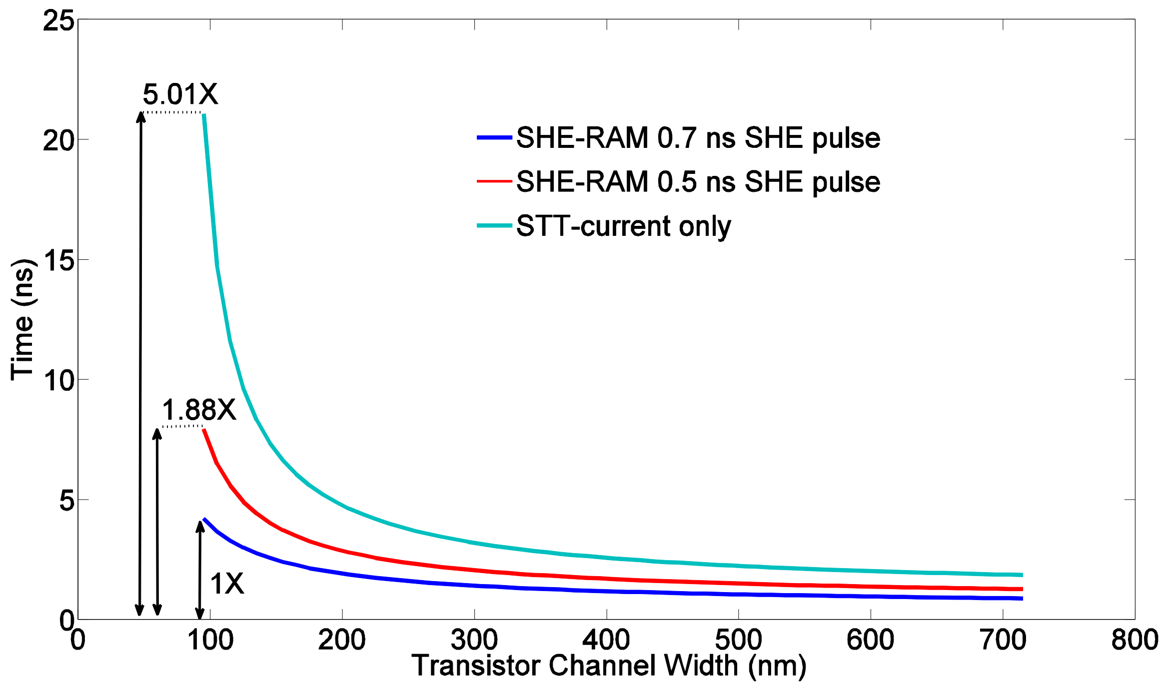

Figure 3.4: Simulated MTJ switching time under different transistor widths. Applying SHE current will significantly improve the MTJ switching performance, especially when the transistor width is small (or the STT current is small).

between the switching times of the MTJ at different SHE current pulse widths shrinks, implying the relatively increased impact of the STT current. For comparison purpose, we also simulated the switching time of the MTJ when only STT current is applied. When the transistor width is small, say, 90nm, applying a 0.7ns SHE current pulse can reduce the MTJ switching time by almost 5× compared to the case where only the STT current is applied. Again, the MTJ switching time reduction incurred by the SHE current becomes less significant when the transistor width increases. Fig. 3.5 shows that increasing the amplitude of the SHE current will slightly improve the MTJ switching performance. But this effect is very limited.

The read operation of the High Density SHE-RAM is similar to the conventional STT-RAM: The word line is asserted and a read current I_r is injected. Depending on the data stored on the MTJ device (or its resistance state), a high or a low voltage will be generated on the bit line. The value of the stored data can be read out by comparing the bit line voltage with a predefined reference voltage.

We note that the SHE current shared by the memory cells connected to the same source line may cause the disturbance to the unselected cells during write operations when thermal fluctuation is taken into account. Such a disturbance can be further aggravated by the process variations of the MTJ, which result in the variability of the MTJ geometry size, the MTJ critical switching current, and the non-volatility. For example, a 20μA SHE current will result in a disturbance rate of 0.072% for the unselected MTJ at 300K. In order to prevent disturbing the unselected bits, we proposed Disturbance Free SHE-RAM.

3.3.2 Disturbance Free SHE-RAM

Instead of sharing the SHE current among the cells on the same source line in High Density SHE-RAM, in Disturbance Free SHE-RAM, the SHE current is shared among the cells on the same word line, as illustrated in Fig. 3.6: One Disturbance Free SHE-RAM cell includes two transistors. One of the transistors connects the MTJ and the word line (e.g., $\mathrm{W_{0-MTJ0}}$) and the other one is inserted between the segments of the SHE line (e.g., $\mathrm{W_{0-SHE0}}$)). During write operations, the SHE current only passes through the cells that are selected by signal $\mathrm{W_{x-SHE}}$ (x = 0, 1, ... as the word line number). No disturbances to the unselected cells are introduced. The write operations can be performed as follows: First, SHE line (e.g., $\mathrm{W_{0-SHE}}$) is activated for a duration of subnanosecond

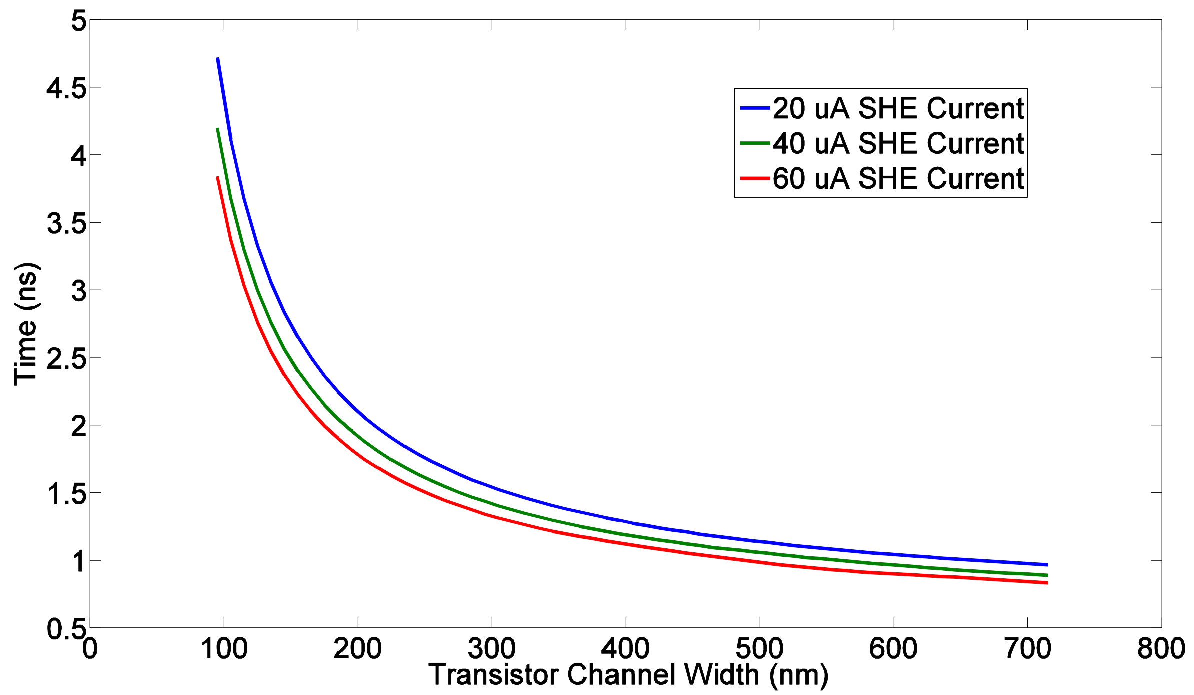

Figure 3.5: MTJ switching time of a High Density SHE-RAM cell under different SHE current amplitudes for a 0.7ns pulse width.

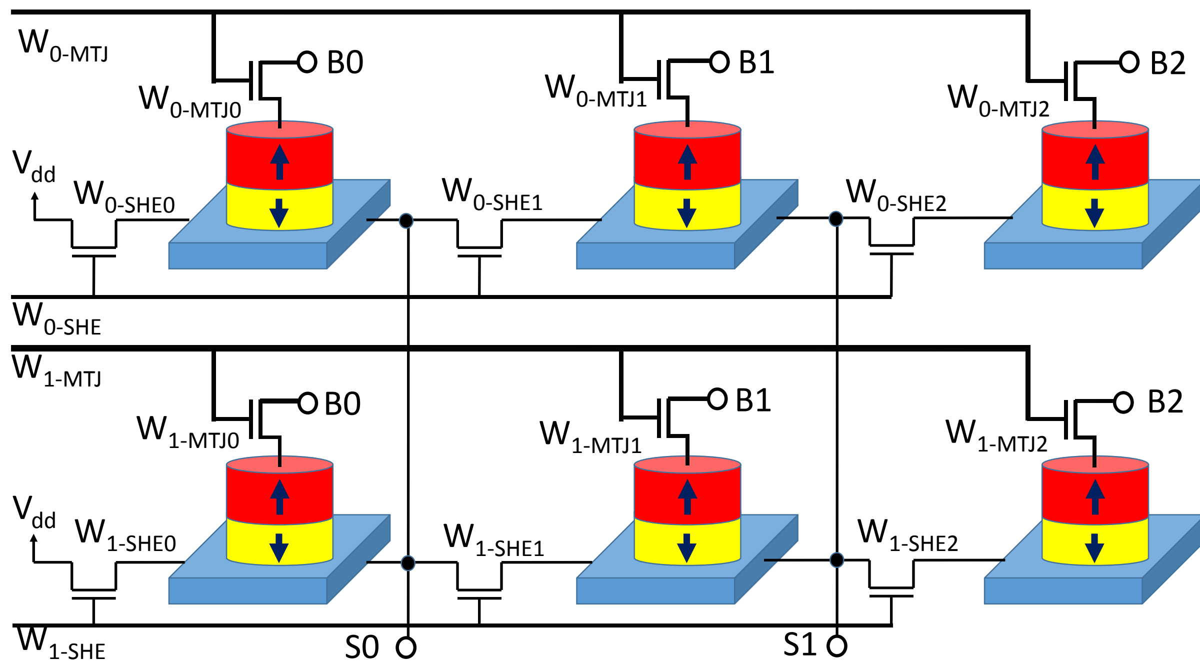

Figure 3.6: Disturbance Free SHE-RAM Design. SHE current is shared by the cells on the entire word line and controlled by $\mathrm{W_{x-SHE}}$ transistors.

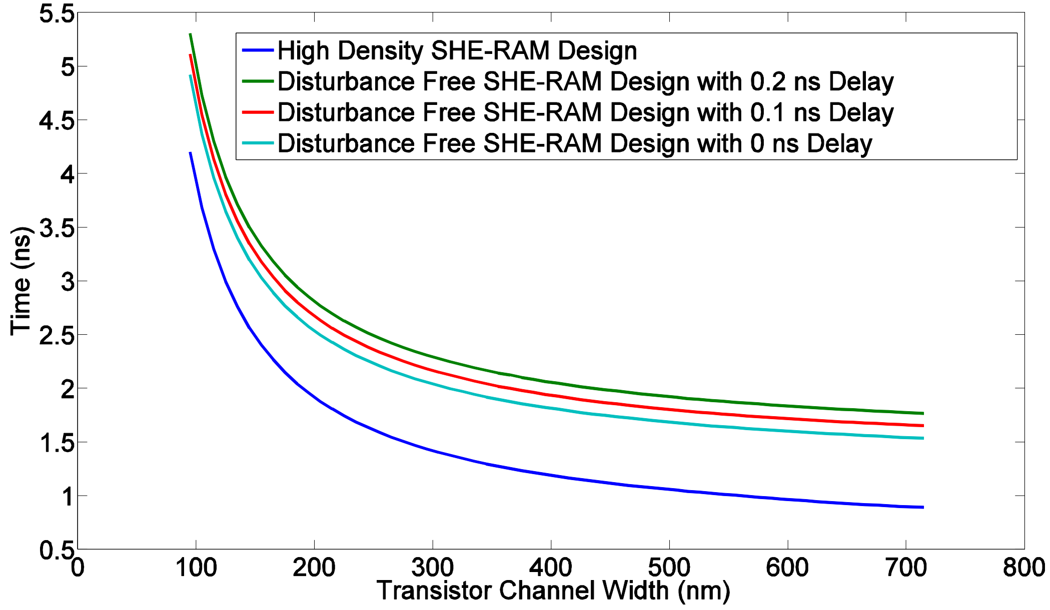

Figure 3.7: Switching time comparison for High Density SHE-RAM Design and Disturbance Free SHE-RAM Design for different transistor widths.

to allow the SHE current to flow underneath all the cells along the entire SHE line and disturb their free layer magnetization alignment. After the SHE line is deactivated, MTJ line (W_{0-MTJ}) is turned on to select the cells along the whole MTJ line and the data are programmed into the cells by applying appropriate biases on the bit line and the source line. Since the SHE line has been deactivated before the MTJ line is activated, each cell can be written independently without causing any inference between them. However, the interval between turning on MTJ line and turning off SHE line may adversely affect the SHE effect, as depicted in Fig. 3.7. Keeping the interval short (but no-zero) is critical for improving the write performance of the Disturbance Free SHE-RAM.

We note that compared to High Density SHE-RAM, Disturbance Free SHE-RAM generally has a longer write operation because the SHE and STT currents are applied at different times, as also shown in Fig. 3.7. After the SHE line is deactivated, the SHE effect immediately starts to decay. As depicted in Fig. 3.7, when the interval between turning off the SHE current and turning on the STT current is increasing, the MTJ switching time increases. Nonetheless, Disturbance Free SHE-RAM still demonstrates significantly enhanced write performance compared to conventional STT-RAM.

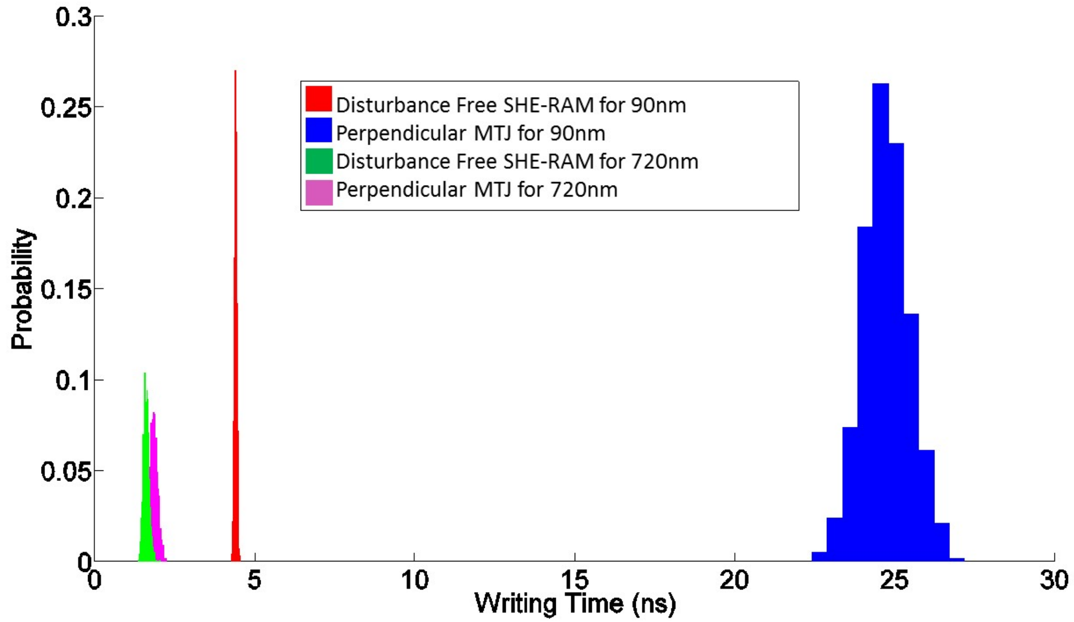

Figure 3.8: Switching time probability distribution for Disturbance Free SHE-RAM Design and for conventional Perpendicular MTJ for $90nm$ and $720nm$ transistor width.

Fig. 3.8 further shows that Disturbance Free SHE-RAM offers not only a faster MTJ switching time but also a tighter distribution of the MTJ switching time when both process variations and thermal fluctuations are considered. This fact holds true over a very wide transistor width range, say, from 90nm to 720nm.

The read operation of Disturbance Free SHE-RAM is also similar to conventional STT-RAM except that transistors $\mathrm{W}_{\mathrm{x-SHE}}$ must be turned off in order to operate on each cell independently.

The estimated memory cell areas of High Density SHE-RAM and Disturbance Free SHE-RAM are 0.0243μm^2 and 0.0567μm^2, respectively, at 45nm technology node.

3.4 CONCLUSION

In this work, we proposed two designs of SHE-RAM, namely, High Density SHE-RAM and Disturbance Free SHE-RAM, for capacity-sensitive and reliable applications, respectively. The introduction of SHE current reduces the amplitude of the required switching current to the MTJ.

This leads to small cell area of High Density SHE-RAM. The disturbance to the unselected cells in High Density SHE-RAM can be then eliminated by the shared word line design in Disturbance Free SHE-RAM though extra cell area overhead is needed.

4.0 MODELING OF BIAXIAL MAGNETIC TUNNELING JUNCTION FOR MULTILEVEL CELL STT-RAM REALIZATION AND RELIABILITY ANALYSIS

4.1 INTRODUCTION

Technology scaling imposes many challenges on design and manufacturing of conventional memories, such as high leakage and reliability issues of SRAM, DRAM, and NAND flash. Extensive research has been performed to develop new memory technologies that can overcome these challenges, including *phase-change memory* (PCM), *resistive memory* (ReRAM), *spin-transfer torque random access memory* (STT-RAM), etc. Among all these technologies, STT-RAM is particularly identified as a potential replacement of DRAM in future main memory application because of its many attractive features like zero standby power, nanosecond access time, high integration density, excellent CMOS-compatibility, etc [34].

In a conventional STT-RAM cell, the data is stored as one of the two resistance states of a magnetic tunneling junction (MTJ) device. The resistance of the MTJ can be switched between the two states by applying a programming current with proper polarization, which is supplied by a select transistor. An important design tradeoff of STT-RAM technology is to balance between the access performance and the storage density [3]: Increasing the programming current can reduce the MTJ switching time while also increasing the select transistor size [28], and consequently, increasing the STT-RAM cell area [30].

Multi-level cell (MLC) technology was recently introduced to STT-RAM designs to improve the storage density, i.e., MLC-STT. MLC-STT allows storing more than 1 bit in a memory cell and hence, needs to realize at least 4 resistance states in a single STT-RAM cell [2]. Popular implementations of MLC-STT include connecting two MTJs with different sizes in series (or in parallel) to construct four different resistance states [35]. In these designs, one MTJ serves as a

soft domain that can be switched by a smaller current while the other MTJ serves as a hard domain that can be switched by only a larger current. Since the soft domain is also flipped when switching the hard domain, a two-step programming procedure is needed and consequently, leading to a long write operation time.

Very recently, a new MTJ structure named *biaxial MTJ* is proposed to enable one-step programming in MLC-STT design [25]. Different from the conventional uniaxial MTJ technology, biaxial MTJ technology can implement four stable resistance states on one MTJ device. Although biaxial MTJ has started to gain increasing attentions in STT-RAM design society [22] , [7], lacking a capable model that can describe the static and dynamic behaviors of biaxial MTJ greatly hinders the study of biaxial MTJ based MLC-STT design. Ideally, such a model shall also be able to simulate the influences of process variations and thermal fluctuations, which are two major factors affecting the reliability of STT-RAM cells [27].

In this work, we developed, to the best knowledge of the authors, the first dynamic macro-magnetic model for biaxial MTJ for MLC-STT circuit designs. Besides simulating the relations between the switching current and the switching time of each MTJ resistance state, our model is also capable to capture the switching transience that can be used to calculate the write error rate of the MLC-STT cell. Write performance and energy consumption of the MLC-STT cell can also be derived and optimized based on our model for different design configurations. Finally, our model allows designers to perform a comprehensive reliability analysis of the MLC-STT cell by taking into account the device parametric variations and the ambient temperature during write operations.

The rest of this paper is organized as follows: Section 2 presents the basics of STT-RAM and uniaxial/biaxial anisotropy; Section 3 introduces our developed biaxial MTJ dynamic switching model; Section 4 provides the validation of our model; Section 5 analyzes the impacts of process variations and thermal fluctuations on the write operation of biaxial MTJ, including both performance and energy consumption; Section 6 concludes our work.

4.2 PRELIMINARY

4.2.1 Basic Operations of STT-RAM Cell

In a STT-RAM cell, data is stored as a resistance value in a MTJ, where a thin insulating layer is stacked between two ferromagnetic layers as shown in Fig. 4.1(a) and (b). One of the ferromagnetic layers has a fixed magnetization direction and is referred to as reference layer (RL) while the other layer has a magnetization direction that can be changed by applying a switching current and is referred to as free layer (FL). The resistance of the MTJ and thereby the stored logical data are determined based on the relative orientations of the FL and the RL: when the two orientations are in parallel (anti-parallel), the MTJ is in its low (high) resistance state.

Fig. 4.1(c) shows the popular one-transistor-one-MTJ (1T1J) STT-RAM cell design where a NMOS transistor supplies the write and the read current to the MTJ. For an uniaxial MTJ, the value being written is determined by the direction of the write current, i.e., from bitline (BL) to sourceline (SL) for ‘0’ or from sourceline (SL) to bitline (BL) for ‘1’. For a biaxial MTJ, the value (out of four possibilities) being written is determined by the combination of the direction and the amplitude of the write current, as we shall show later.

4.2.2 Basics of Uniaxial and Biaxial Anisotropies

Energy function of the uniaxial crystalline anisotropy for a uniaxial MTJ, which is shown in Fig. 4.1(a), can be defined as [18]:

$$E_u = K_u \sin^2(\theta). \tag{4.1}$$

Here, K_u is the uniaxial anisotropy constant and θ is the angle between the FL’s magnetization vector and the easy axis. In Fig. 4.1, we choose easy axis as y-axis. Uniaxial crystalline anisotropy has two minimum energy points along the y-axis at $0°$ and $180°$, respectively. These two minima points correspond to the low (at $0°$) and the high (at $180°$) resistance states of the MTJ, or ‘0’ and ‘1’, respectively. Fig. 4.1(b) shows a biaxial MTJ where the energy function for the mixture of uniaxial and biaxial crystalline anisotropy can be calculated by [18]:

$$E_b = K_u \sin^2(\theta) + \frac{1}{4} K_1 \sin^2(2\theta). \tag{4.2}$$

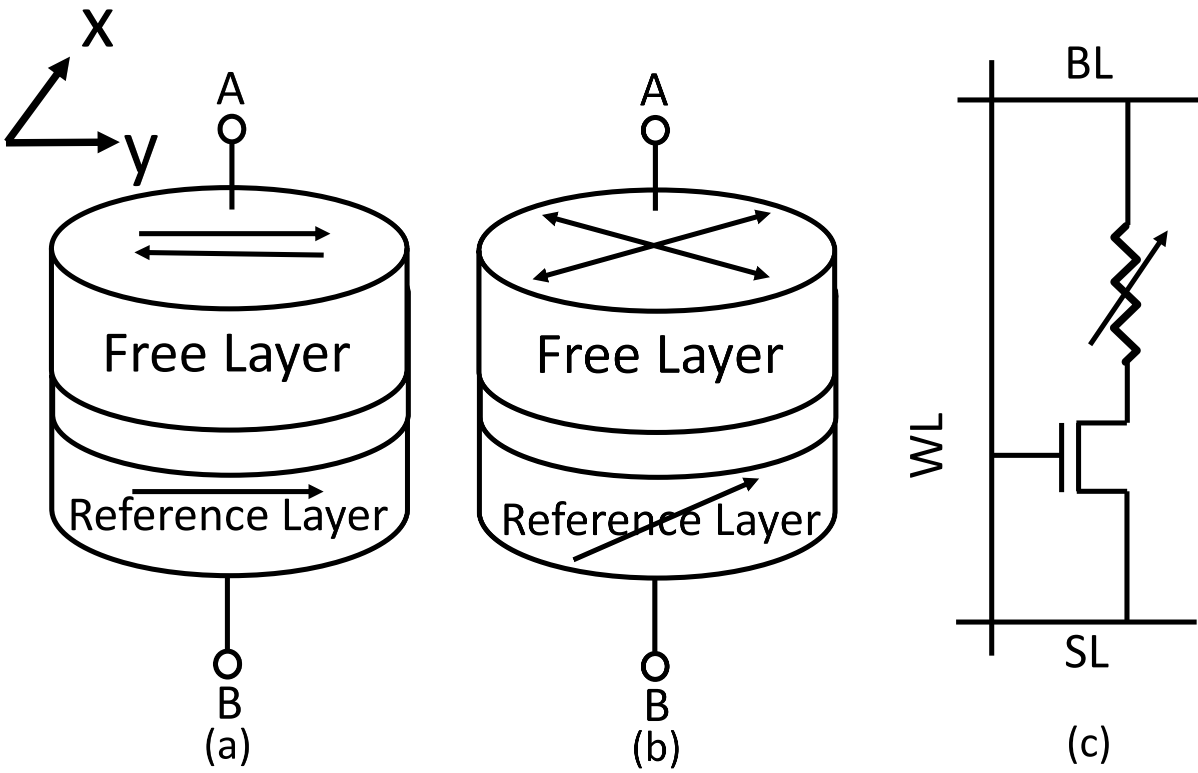

Figure 4.1: STT-RAM basics. a) Uniaxial MTJ. b) Biaxial MTJ. c) 1T1J STT-RAM cell structure.

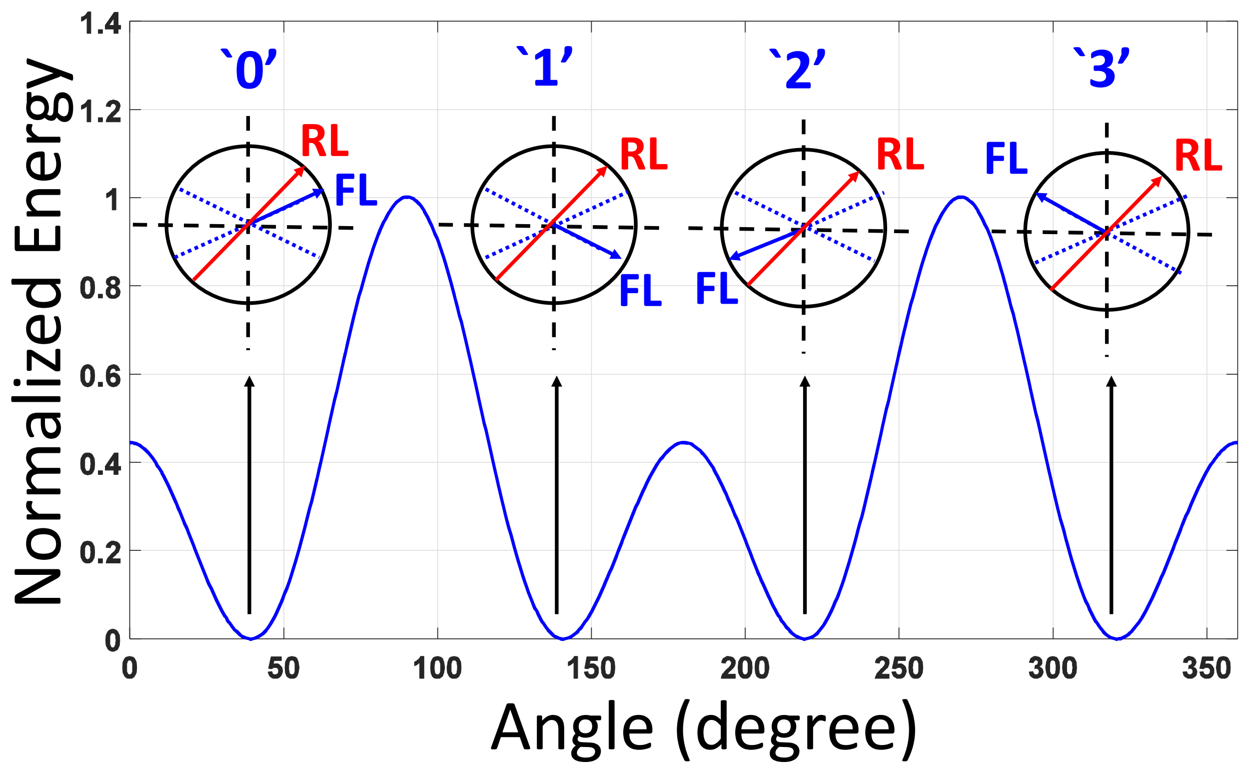

Figure 4.2: Energy of the biaxial anisotropy and the changing of the logic states at the energy minima points.

Here K_1 is the biaxial anisotropy constant. Different from uniaxial, biaxial anisotropy energy has four minimum energy points corresponding to four stable resistance states of the biaxial MTJ, as illustrated in Fig. 4.2. The ratio between K_u and K_1 determines where these energy minima are, or the stable position of the direction of the FL. In order to maximize the resistance differences between different resistance states, in this work, we choose $K_u/K_1 = 2.75/4.75$ [25]. The direction of the RL is also properly tilted to enable sufficiently large distinctions between adjacent resistance states, as shown in Fig. 4.2 and Fig. 4.1(b).

The dynamic magnetic response of a MTJ's FL can be modeled by the Landau-Lifshitz-Gilbert (LLG) equation as [26]:

$$\frac{dm_f}{dt} + \alpha(m_f \times \frac{dm_f}{dt}) = \frac{1}{2}\gamma H \Sigma_{i=1}^{4}(\frac{\Gamma_i}{l_t K}), \tag{4.3}$$

where m_f is the unit vector of the FL's magnetization vector, α is the Gilbert damping coefficient, γ is the gyromagnetic ratio, t is the switching time, H is the anisotropy field, and l_t is the thickness of the free layer. m_f is under the influence of four torque terms (Γ) including biaxial anisotropy (Γ_1), easy-plane anisotropy (Γ_2), Langevin random thermal field (Γ_3), and spin torque term (Γ_4) from the applied current. The solution of LLG equation for uniaxial anisotropy has been extensively discussed in many prior-arts [11, 26]. In this work, we will extend the LLG equation to model the biaxial anisotropy torque in a biaxial MTJ.

4.3 MODELING OF BIAXIAL MTJ

In this section, we will give the details on the mathematical development process of the three-dimensional(3D) modeling for biaxial anisotropy torque, which is denoted as Γ_1 in Eq.(4.3).

4.3.1 Model Description Of Biaxial MTJ

As presented in Section 4.2, the value of biaxial anisotropy energy (see Eq. (4.2)) is determined by the θ angle between the FL's magnetization vector and the easy axis (or the y-axis) in Fig.4.1(b).

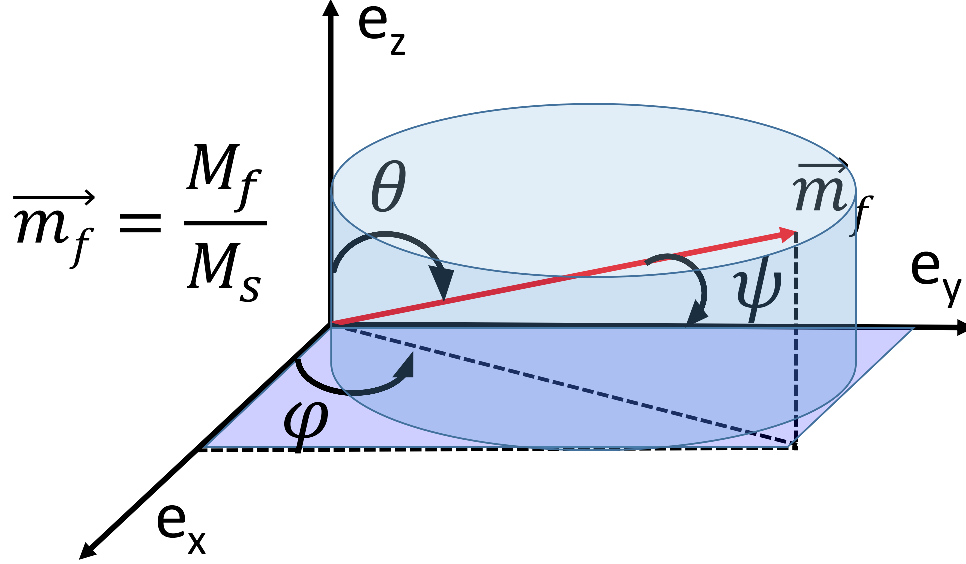

Figure 4.3: Illustration of FL's magnetization vector in spherical coordinate system.

However, this efinition is valid for only a two-dimensional (2D) motion of the FL's magnetization vector and is no longer capable to fully model the 3D motion in a spherical coordinate system. For the simulations where timing the process of magnetic reversal or switching time needs to be precise, a 3D dynamic model will be more appropriate for MTJs [11]. In a 3D spherical coordinate system, as shown in Fig. 4.3, θ or any other angles will not be solely enough to describe biaxial anisotropy energy because the angle between the FL's magnetization vector and the easy axis (e_y) does not depend on only θ but also the angle φ. Hence, ψ is introduced to denote the angle between the FL's magnetization vector (m_f) and the easy axis(e_y) as depicted in Fig. 4.3. Here ψ satisfies

$$\cos\psi = \frac{m_f . e}{\|m_f\| \, . \, \|e\|}, \tag{4.4}$$

where, (.) is dot product. m_f and e are the FL's magnetization vector and the easy axis vector, respectively. m_f and e can be defined by

$$m_f = \frac{M_f}{M_s} = (\sin\theta\cos\varphi\hat{\mathbf{x}}, \sin\theta\sin\varphi\hat{\mathbf{y}}, \cos\theta\hat{\mathbf{z}}), \tag{4.5}$$

and

$$e = (e_x\hat{\mathbf{x}}, e_y\hat{\mathbf{y}}, e_z\hat{\mathbf{z}}), \tag{4.6}$$

respectively, by substituting (4.5) and (4.6) into (4.4), we have

$$\psi = \cos^{-1}(m_f.e) = \cos^{-1}(v). \tag{4.7}$$

Here,

$$v = m_f.e = e_x \sin\theta \cos\varphi + e_y \sin\theta \sin\varphi + e_z \cos\theta. \tag{4.8}$$

After obtaining ψ, the biaxial anisotropy energy function given by Eq. (4.2) can be rewritten to

$$E_b = K_u \sin^2(\psi) + \frac{1}{4} K_1 \sin^2(2\psi). \tag{4.9}$$

In order to obtain the torque exerted by the biaxial anisotropy energy, we need to first calculate the derivatives of the biaxial anisotropy energy function with respect to θ and φ. Since ψ is a function of θ and φ, we can safely apply the chain rule as:

$$\begin{aligned} \frac{\partial E_b}{\partial \theta} &= \frac{\partial E_b}{\partial \psi} \frac{\partial \psi}{\partial \theta} \\ \frac{\partial E_b}{\partial \varphi} &= \frac{\partial E_b}{\partial \psi} \frac{\partial \psi}{\partial \varphi}. \end{aligned} \tag{4.10}$$

The first term on the right side, $\frac{\partial E_b}{\partial \psi}$, is a common item on both equations and can be expressed by:

$$\frac{\partial E_b}{\partial \psi} = K_u \sin(2\psi) + \frac{1}{2} K_1 \sin(4\psi). \tag{4.11}$$

At the next step, in order to get the partial derivatives of ψ with respect to θ and φ on the right side of Eq.(4.10), we have;

$$\begin{aligned} \frac{\partial \psi}{\partial \theta} &= \frac{\partial \psi}{\partial v} \frac{\partial v}{\partial \theta} \\ \frac{\partial \psi}{\partial \varphi} &= \frac{\partial \psi}{\partial v} \frac{\partial v}{\partial \varphi}. \end{aligned} \tag{4.12}$$

Here ψ is a function of v defined in Eq. (4.7). Hence, the first common term on the right side of Eq. (4.12) can be expressed as;

$$\frac{\partial \psi}{\partial v} = \frac{-1}{\sqrt{1 - v^2}}, \tag{4.13}$$

and the last two terms on the right side of Eq. (4.12) can be calculated as:

$$\begin{aligned} \frac{\partial v}{\partial \theta} &= (e_x \cos\theta \cos\varphi + e_y \cos\theta \sin\varphi - e_z \sin\theta) \\ \frac{\partial v}{\partial \varphi} &= (-e_x \sin\theta \sin\varphi + e_y \sin\theta \cos\varphi). \end{aligned} \tag{4.14}$$

By substituting Eq. (4.14) and Eq. (4.13) in Eq. (4.12), we have:

$$\begin{aligned} \frac{\partial \psi}{\partial \theta} &= \frac{-1}{\sqrt{1-v^2}}(e_x \cos\theta \cos\varphi + e_y \cos\theta \sin\varphi - e_z \sin\theta) \\ \frac{\partial \psi}{\partial \varphi} &= \frac{-1}{\sqrt{1-v^2}}(-e_x \sin\theta \sin\varphi + e_y \sin\theta \cos\varphi). \end{aligned} \tag{4.15}$$

Based on Eq. (4.15) and Eq. (4.11), the biaxial anisotropy torque expressed in Eq. (4.10) can be rewritten as:

$$\begin{aligned} \frac{\partial E_b}{\partial \theta} &= (K_u \sin(2\cos^{-1}(v)) + \frac{1}{2}K_1 \sin(4\cos^{-1}(v))) \\ &\quad \frac{-1}{\sqrt{1-v^2}}(e_x \cos\theta \cos\varphi + e_y \cos\theta \sin\varphi - e_z \sin\theta) \\ \\ \frac{\partial E_b}{\partial \varphi} &= (K_u \sin(2\cos^{-1}(v)) + \frac{1}{2}K_1 \sin(4\cos^{-1}(v))) \\ &\quad \frac{-1}{\sqrt{1-v^2}}(-e_x \sin\theta \sin\varphi + e_y \sin\theta \cos\varphi). \end{aligned} \tag{4.16}$$

After obtaining the biaxial anisotropy torque terms as Eq. (4.16), in order to implement them to the LLG equation, biaxial anisotropy effective fields are needed and can be calculated as;

$$\begin{aligned} H_{\theta b} &= -\frac{1}{\mu_0 M_s}\frac{\partial E_b}{\partial \theta} \\ H_{\varphi b} &= -\frac{1}{\mu_0 M_s \sin\theta}\frac{\partial E_b}{\partial \varphi}. \end{aligned} \tag{4.17}$$

We note that the LLG equation given by Eq. (4.3) can be translated into spherical coordinate system by defining the position of the FL's magnetization vector in the 3D space using the angles θ and φ as [11]. Hence, the dynamics of the FL's magnetization can be remodeled as:

$$\begin{aligned} \frac{d\theta}{dt} &= \frac{\gamma_0}{1+\alpha^2}(H_\varphi + \alpha H_\theta) \\ \frac{d\varphi}{dt} &= \frac{\gamma_0}{(1+\alpha^2)\sin\theta}(\alpha H_\varphi - H_\theta). \end{aligned} \tag{4.18}$$

Here H_θ and H_φ are the net effective fields containing biaxial anisotropy, easyplane anisotropy[21]. Langevin random thermal field [17], and STT field [21] for θ and φ components.

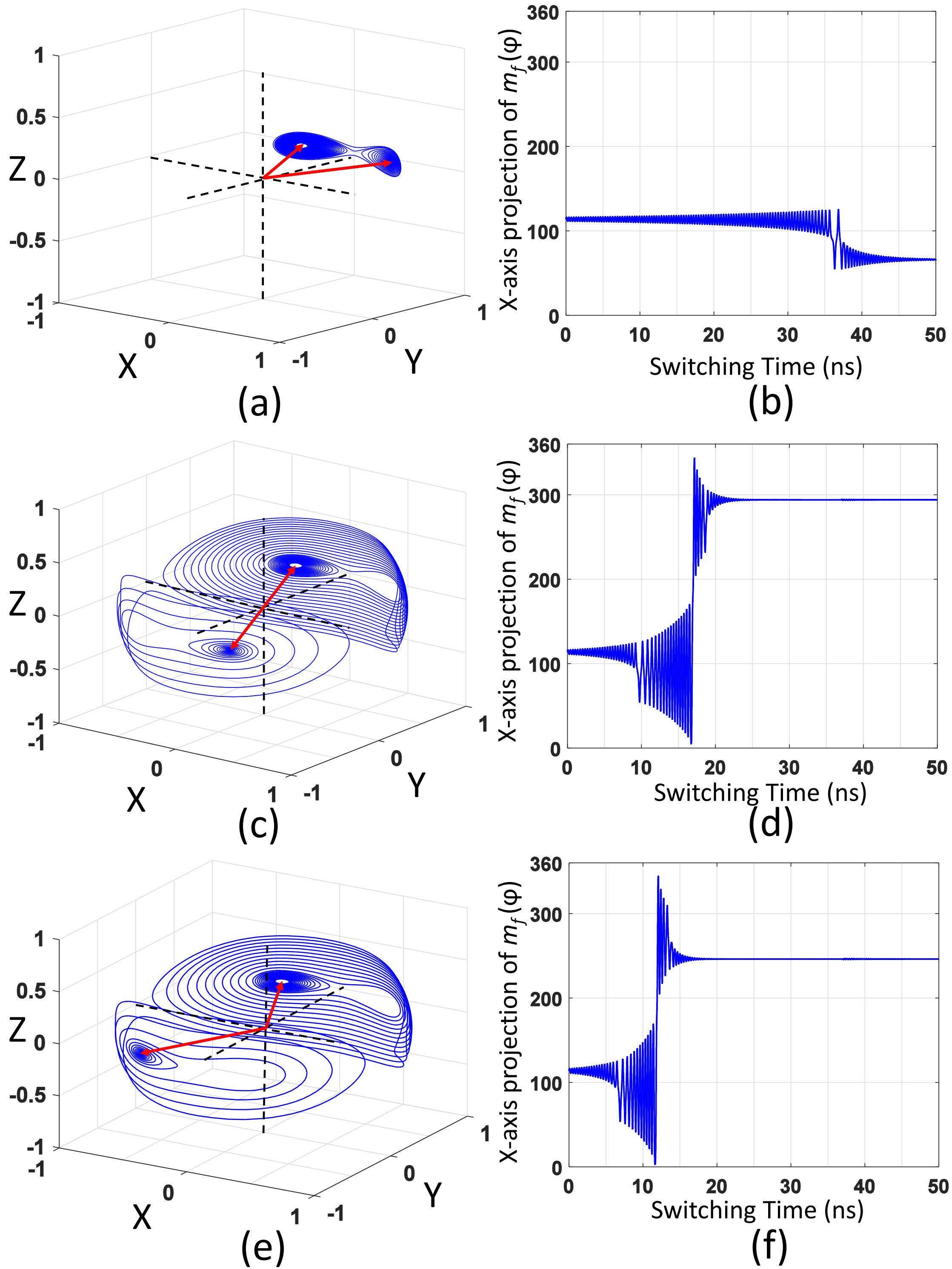

Figure 4.4: Change of m_f during the MTJ switchings. a) '0' to '1' switching transience. b) '0' to '1' switching time. c) '0' to '2' switching transience. d) '0' to '2' switching time. e) '0' to '3' switching transience. f) '0' to '3' switching time.

Table 4.1: Device and circuit simulation parameters

	Parameter	Symbol	Value	Unit
Device Level	Mag. Saturation	M_s	230	kA/m
	Anisotropy Field	H_k	200	Oe
	Gilbert Constant	α	0.01	
	Free Layer Thickness	l_t	1	nm
Circuit Level	Parameter	Mean	Std. Dev	
	Channel length	$L = 65nm$	$\sigma_L = 0.05L$	
	Channel width	$W = 2L$	$\sigma_W = 0.05W$	
	Free layer volume	$V = L \times 2L \times 1nm^3$	$\sigma_V = 0.05V$	
	Resistance low	$R_L = 1000\Omega$	$\sigma_{R_L} = 0.05R_L$	
	Resistance high	$R_H = 3000\Omega$	$\sigma_{R_H} = 0.05R_H$	

4.3.2 Basic Functions of Biaxial MTJ Model

In our work, we assume that a MTJ is a single domain structure and ignore the current generated magnetic field. It is a valid simplification and represents the normal fabrication and operation mode of the MTJ. Table 4.1 summarizes the device and circuit parameters adopted in our work, including the performed experiments. These parameters are consistent with the ones characterized from the fabricated MTJ device and have been validated in some prior-arts [8].

Fig. 4.4 depicts the simulation results of our model about the switching transience and the switching time of the FL's magnetization vector when a programming current is applied. It includes the results when the MTJ switches from '0' to '1', '2', and '3', respectively.

Fig. 4.4(a) shows the 3D motion of the FL's magnetization vector during the switching of '0' to '1'. After applying a switching current (i.e., $76\mu A$), the FL's magnetization vector starts precession from the energy minimum point corresponding to '0' (see Fig. 4.2) and moves to the energy minimum point corresponding to '1' through relaxation spin. Fig. 4.4(b) shows the change of the angle φ between the x-axis and the FL's magnetization vector over time during this '0' to '1'.

Table 4.2: Possible switching current and switching time for each state

From \ To		Logic 0	Logic 1	Logic 2	Logic 3
Logic 0	Current	-	$76\mu A$	$110\mu A$	$130\mu A$
	Time	-	$45ns$	$25ns$	$18ns$
Logic 1	Current	$83\mu A$	-	$110\mu A$	$140\mu A$
	Time	$50ns$	-	$25ns$	$20ns$
Logic 2	Current	$-110\mu A$	$-100\mu A$	-	$-83\mu A$
	Time	$25ns$	$30ns$	-	$55ns$
Logic 3	Current	$-130\mu A$	$-110\mu A$	$-80\mu A$	-
	Time	$20ns$	$30ns$	$50ns$	-

The initial position of the FL's magnetization of '0' is $114°$ with respect to x-axis, and settles down at $66°$ for '1'. Fig. 4.4(c), (e), (d), and (f) show the similar results for the switchings from '0' to '2' and '3', respectively. The corresponding switching currents of the MTJ are $110\mu A$ and $130\mu A$, respectively. The final positions of the FL's magnetization in these two switchings are $294°$ and $246°$, respectively, as shown in 4.4(d) and (f). As aforementioned at Section 4.2, the final positions of the FL's magnetization at different MTJ resistance states are determined by the ratio of K_u/K_1.

An interesting observation from Fig. 4.4 is that, the switching times of writing '1', '2', and '3' keep reducing, i.e., are $45ns$, $25ns$, $18ns$, respectively. This is because that the switching currents in these scenarios are not equal, say, increase from $76\mu A$ to $110\mu A$ and $130\mu A$, respectively.

Besides the results shown in Fig. 4.4, possible switching currents and the corresponding MTJ switching times for other MTJ switchings are summarized in Table 4.2. As we shall show in the next section, the tradeoff between the switching current magnitude and the MTJ switching time in each switching scenario of biaxial MTJs is different from that of uniaxial MTJs.

4.3.3 Some Discussions

The observation that increasing the magnitude of the switching current can always speedup the MTJ resistance switching process of uniaxial MTJs [32] does not always hold any more in the switching of biaxial MTJs between adjacent states. Fig. 4.5 shows the simulated results of a biaxial MTJ switching from '0' to '1' at two different switching currents. Increasing the switching current from $76\mu A$ to $86\mu A$ does not introduce a faster switching time; the increased oscillations actually extends the switching process that settles down at '1'. However, when the biaxial MTJ switches between two states that are not adjacent and takes a long switching time, this observation may still hold. As we can see in Fig. 4.6, increasing the switching current results in a faster switching time when the biaxial MTJ switches from '0' to '2' or '3'. However, it should not be ignored that increasing the switching current may also incur write error, e.g., switching from '0' to '2' may end up with '3' when process variations and thermal fluctuations take place. A more detailed discussion about the write errors of biaxial MTJs will be given in Section 4.5.

As aforementioned in Section 4.2, the magnetization of the RL is not in parallel to the one of the FL at any MTJ resistance state in order to offer sufficient distinction between the states. Then the MTJ resistance at each state can be calculated by [23]:

$$R(\alpha) = R(0) + \Delta R \frac{1-\cos(\alpha)}{2+\lambda(1+\cos(\alpha))}, \tag{4.19}$$

where λ is a fitting parameter, and α is the angle between the magnetization vectors of the RL and the FL at each MTJ resistance state. Since the RL is tilted $135°$ in the x-y plane, we have $\alpha = 135 - \varphi$.

Fig. 4.7 shows the resistance switching transience of different MTJ switchings from '0'. The final resistance states of each logic value are:

logic '0' $\longrightarrow 1.03\ K\Omega$,

logic '1' $\longrightarrow 1.40\ K\Omega$,

logic '2' $\longrightarrow 2.87\ K\Omega$,

logic '3' $\longrightarrow 2.03\ K\Omega$.

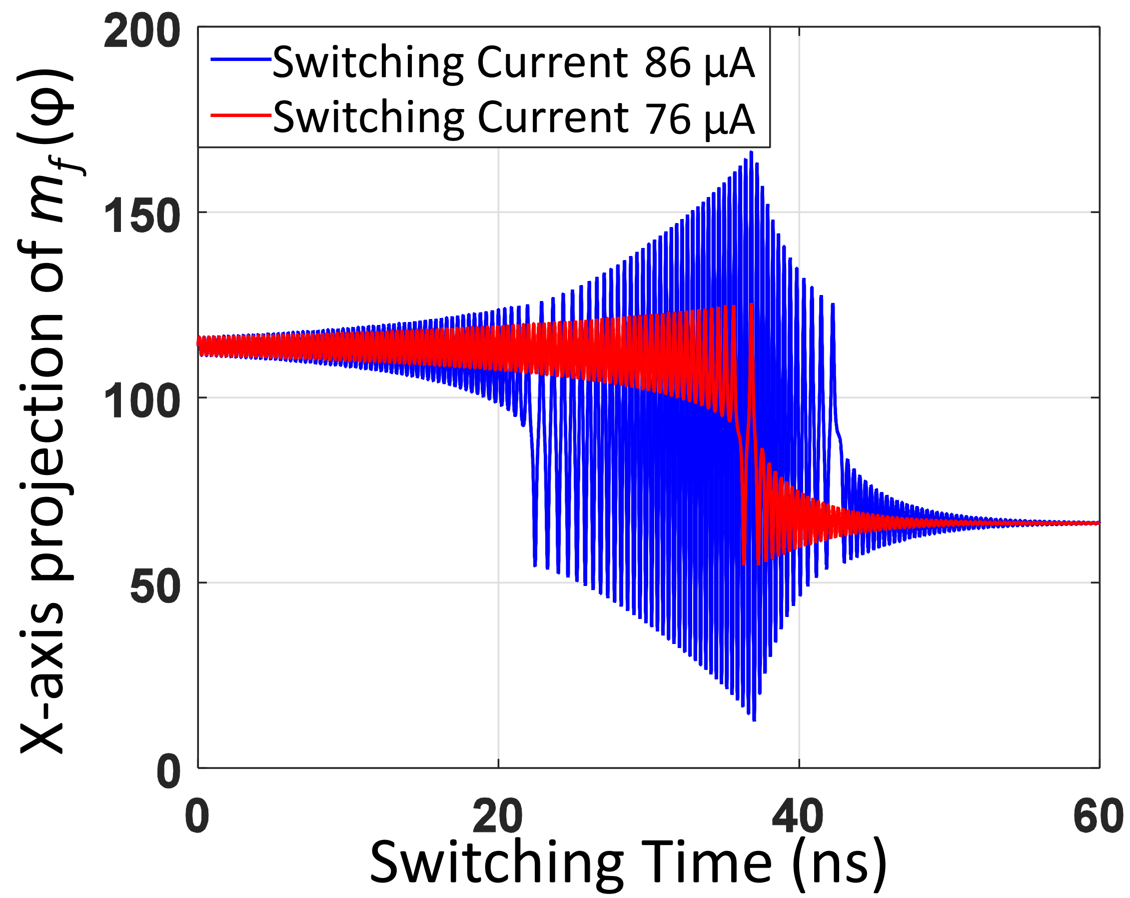

Figure 4.5: ‘0’ to ‘1’ switchings at two different switching currents.

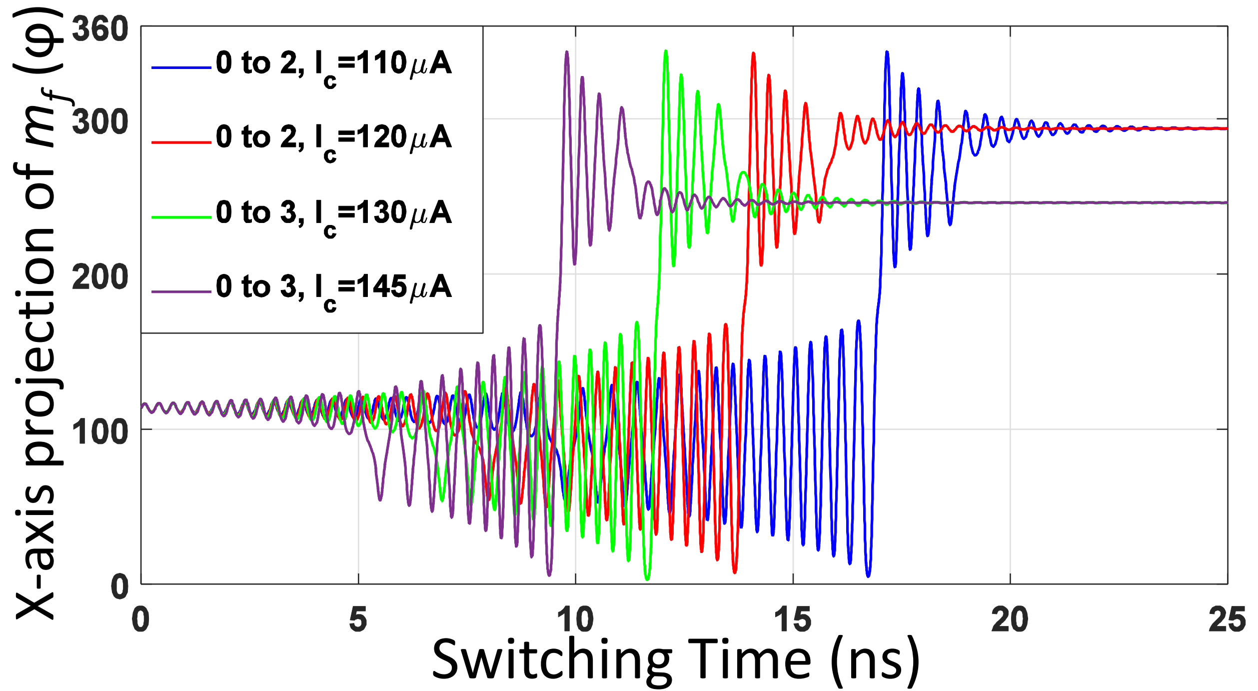

Figure 4.6: '0' to '2' and '3' switchings at two different current values.

4.4 MODEL VALIDATION

We validated our developed biaxial MTJ model against to one of the first four-state MTJ manufactured using epitaxial $Co_{50}Fe_{50}$-MgO-$Co_{50}Fe_{50}$ [22]. To ensure a fair validation, we make the following customization and parameter adjustment of our model to accommodate the parameters of the manufactured device.

1. Epitaxial $Co_{50}Fe_{50}$ electrode has four in-plane magnetization directions toward ϕ_0, $\phi_0 + 90°$, $\phi_0 + 180°$ and $\phi_0 + 270°$ due to the magnetocrystalline anisotropy, where ϕ_0 is the angle between the magnetization vector of the reference layer and the x axis. If ϕ_0 is set to $0°$, two out the four possible magnetization directions, i.e., $90°$ and $270°$, will result the same MTJ resistance. In order to avoid this problem, ϕ_0 is set to be slightly different than $0°$ in [22].

   As aforementioned, angles of the states can be adjusted by the energy constants K_u and K_1, which unfortunately are not disclosed in [22].

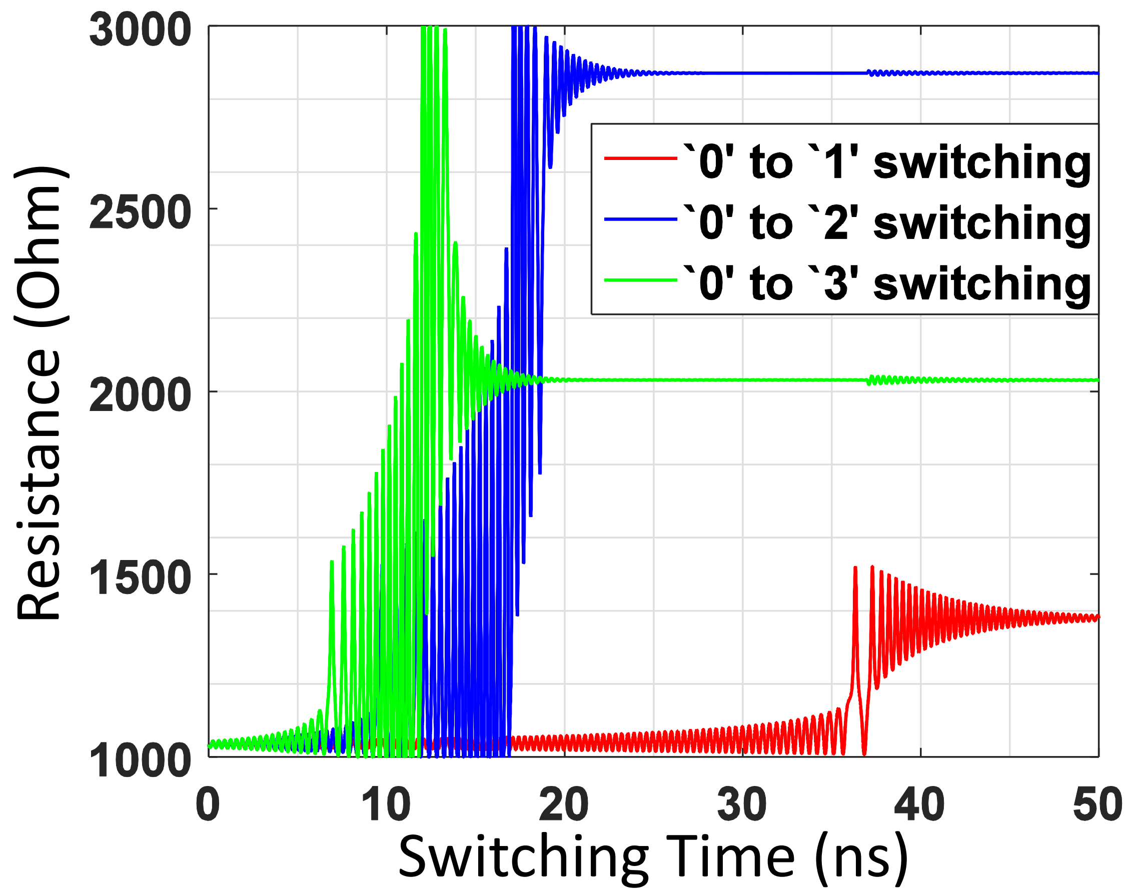

Figure 4.7: Changing of biaxial MTJ resistance value during the switching processes.

In our validation, this angle is well fitted when the K_u/K_1 ratio is $1/5$.

2. It is known that a MTJ can be switched by both magnetic field and spin polarized current. The biaxial MTJ fabricated in [22], however, was tested under magnetic field switching. Note that two (orthogonal) magnetic fields are required to program the biaxial MTJ because the relative magnetization directions of the FL and RL at the stable resistance states are not necessarily in parallel, which is different from uniaxial MTJ. To accommodate this condition, we extended our model by introducing the term representing the applied magnetic field.The energy for the applied magnetic field can be written as:

$$E_m = -\mu_0 V M_S (H_x \sin\theta\cos\varphi + H_y \sin\theta\sin\varphi). \tag{4.20}$$

Where H_x and H_y are the applied magnetic fields at the directions of x and y axes, respectively.The torque terms for the applied magnetic field can obtained as;

$$\begin{aligned}\frac{\partial E_m}{\partial\theta} &= -\mu_0 V M_S (H_x \cos\theta\cos\varphi + H_y \cos\theta\sin\varphi)\\ \frac{\partial E_m}{\partial\varphi} &= -\mu_0 V M_S \sin\theta (H_x \sin\varphi - H_y \cos\varphi).\end{aligned} \tag{4.21}$$

In order to implement these applied field torques to LLG Eq. (4.18), the corresponding effective fields can be expressed as:

$$\begin{aligned}H_{m\theta} &= (H_x \cos\theta\cos\varphi + H_y \cos\theta\sin\varphi)\\ H_{m\varphi} &= (H_y \cos\theta - H_x \sin\theta)\end{aligned} \tag{4.22}$$

and included in the LLG Equation.

We first simulated the final position of the FL of each MTJ resistance state in the absence of external magnetic field, as shown in Fig 4.8. The magnetization direction of the FL may start with any arbitrary position but after some spin relaxations, it will settle down to a stable state corresponding to ‘0’, ‘1’, ‘3’, and ‘2’, respectively.

Fig. 4.9 illustrates the writing process of the biaxial MTJ between four different states using magnetic field. Note that two orthogonal magnetic fields may need to be respectively applied along ±x-axis and ±y-axis in sequence to finish some MTJ programming. For example, to switch the MTJ from ‘0’ to ‘3’, the first magnetic field must be applied at the direction of +y-axis to switch

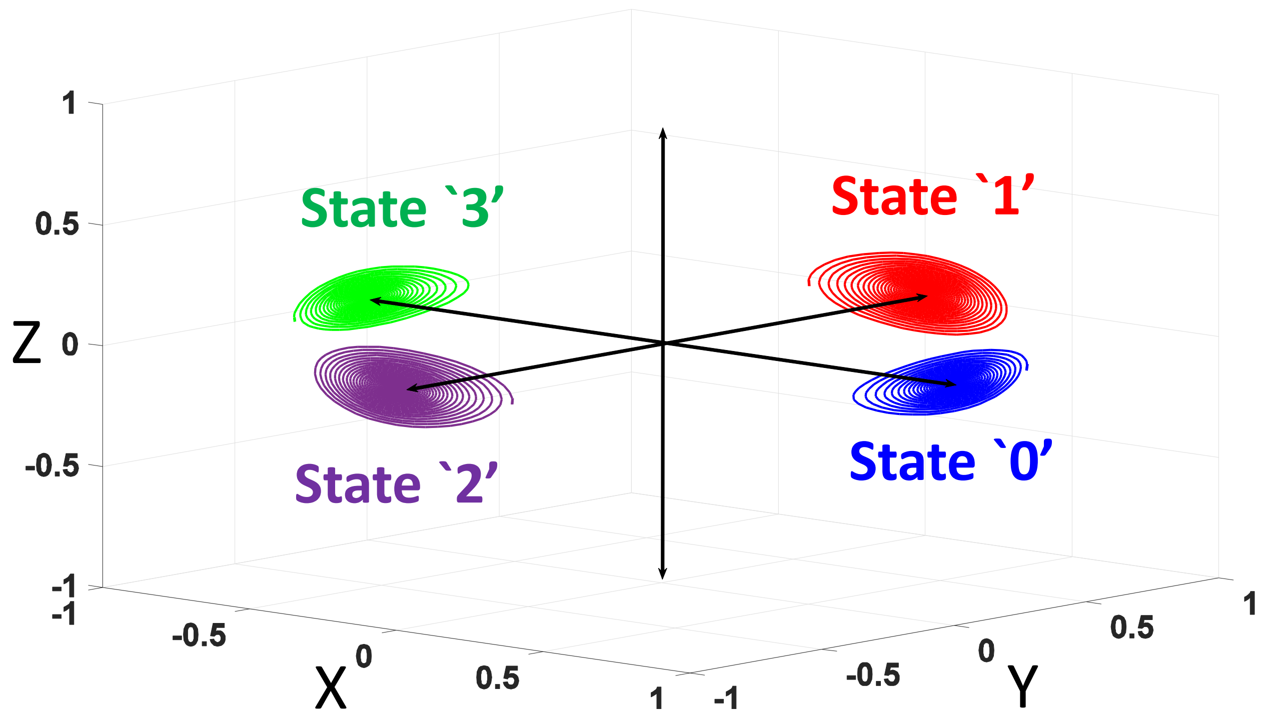

Figure 4.8: Four distinct positions of free layer's magnetization vector at $0°$, $90°$, $180°$, and $270°$ for '0', '1', '3', and '2', respectively.

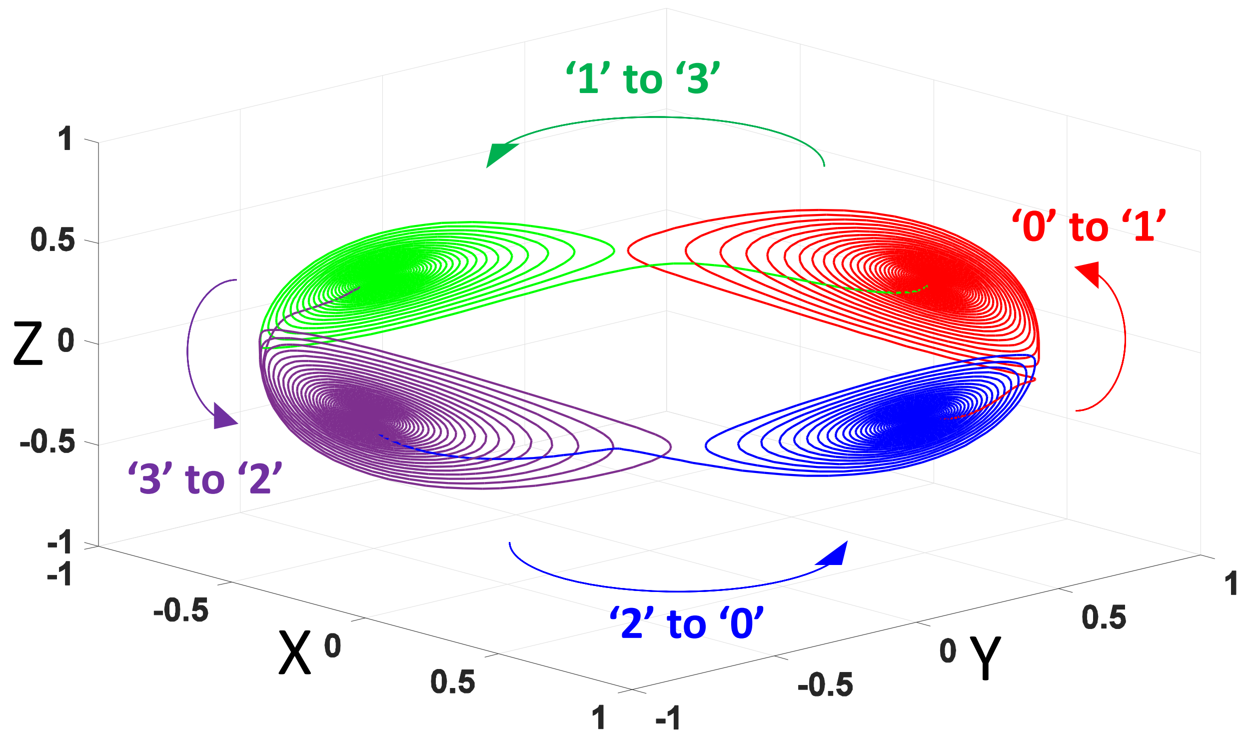

Figure 4.9: Switching motion of biaxial MTJ under the effect of the applied magnetic field.

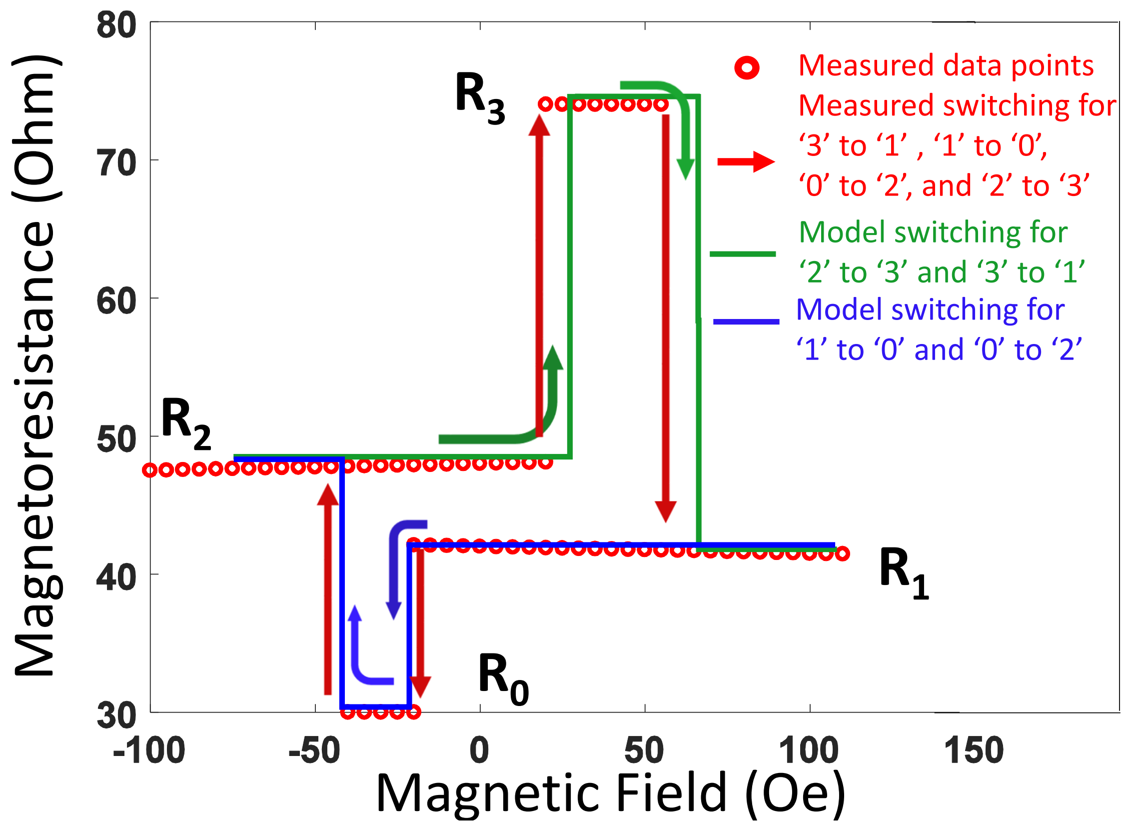

Figure 4.10: Model validation against to reference [22] for resistance values and the applied magnetic field.

Table 4.3: Device level simulation parameters

	Parameter	Symbol	Value	Unit
Device Level	Mag. Saturation	M_s	1050	kA/m
	Anisotropy Field	H_k	800	Oe
	Gilbert Constant	α	0.01	–
	Magneto-resistance Ratio	TMR	145%	–
	Free Layer Thickness	l_t	20	nm
	Free Layer Area	A	10×10	μm^2

the MTJ state to ‘1’ (or -y-axis to switch the MTJ state to ‘2’) and then the second magnetic field must be applied at the direction of -x-axis to switch the MTJ state to ‘3’.

Fig. 4.10 compares the simulated results of different MTJ switchings using our model and the measured data in [22] under the applied magnetic fields. Note that here we have to approximate some devices parameters, e.g., magnetization saturation, anisotropy fields etc., which are not disclosed in [22] to generate our simulation results. These approximated parameters are summarized in Table 4.3. Nonetheless, our model matches the measured data switchings very well.

4.5 RELIABILITY ANALYSIS

All the analysis’ in previous sections are based on the nominal values of the device parameters. However, parametric variabilities (a.k.a. process variations) of MTJ (and CMOS device) and thermal fluctuations under different ambient temperatures greatly influence the MTJ switching process. In this section, we will use our developed model to perform reliability analysis of biaxial MTJs in terms of switching performance, energy consumption, and write errors.

4.5.1 Switching Time and Energy Consumption of Biaxial MTJs

During write operations of a MLC-STT cell, the switching current of the MTJ is determined by the voltage applied on the memory cell. In order to obtain the MTJ switching time distribution, we first conducted 1000 times Monte-Carlo SPICE simulations to collect the MTJ switching current samples at each MTJ resistance state by considering CMOS device parametric variations. In our simulation, we set the operating voltage of the MLC STT-RAM cell to $1.1V$, which generates the nominal switching currents shown in Table 4.2. We then apply the switching current samples to our biaxial macro-magnetic model and run another 1000 times Monte-Carlo simulations at $300K$. Both MTJ device parametric variations and thermal fluctuations are considered in these simulations. All the device parameters adopted in our simulations are summarized in Table 4.1. Here the transistor channel width is set to $130nm$. Fig. 4.11 shows the switching time distributions when the biaxial MTJ switches from ‘0’ to ‘1’, ‘2’, and ‘3’, respectively. The switching energy consumption

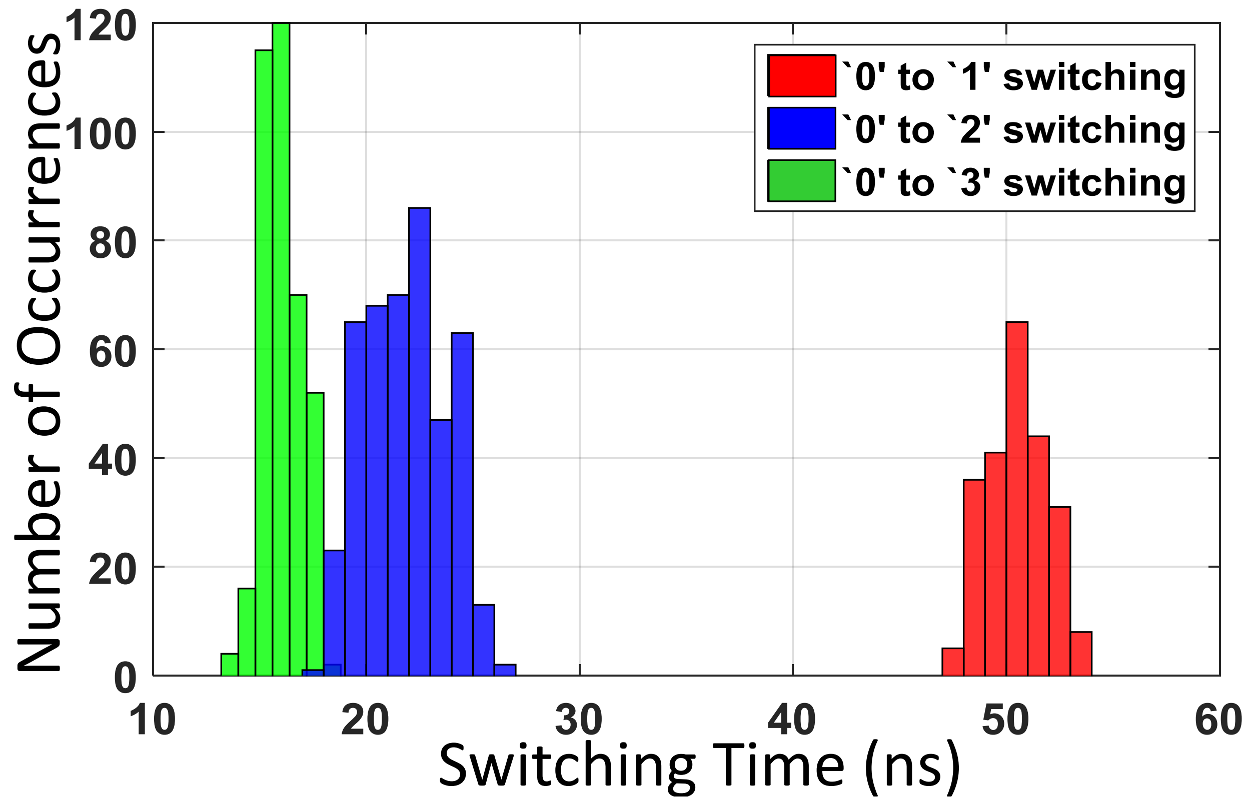

Figure 4.11: Biaxial MTJ switching time distribution for ‘0’ to other states at $300K$.

can be calculated as the integral of the switching current and the supplied voltage (V) over the programming time τ_{wt} as:

$$E_i = \int_{0}^{\tau_{wt}} (I_i * t + I_{pi} * (\tau_{wt} - t)) * V * dt. \tag{4.23}$$

Here I_i is the i_{th} current sample, I_{pi} is the corresponding post-switching current, t is the switching time. We note that oscillations of the FL's magnetization vector during the MTJ switching also cause oscillations of the switching current, which are ignored in our simulations.

We assume a fixed programming time ($\tau_{wt} = 60ns$) for all MTJ switchings (with different programming current amplitudes though). The energy consumption distributions of each MTJ switchings are depicted in Fig. 4.12. As can be seen from the result, the switching from '0' to '1' consumed much less energy than the switchings to other states because of a much lower switching current. However, in real applications, the energy consumption of the switchings from '0' to '1' or '2' can be further reduced when early-termination technique is applied [36], where the switching current is timely removed once the MTJ switching completes. Nonetheless, the energy consumptions of all switchings are at only pJ level.

Our experimental results show that the write energy consumption of biaxial MTJ-based MLC-STT cells can be perfectly modeled by Gaussian distribution. Hence, after receiving the energy consumption samples as above, the mean and the standard deviation of the write energy consumption of the biaxial MTJ can be calculated by [8]:

$$\mu_E = \frac{\sum_{i=1}^{n} E_i * f_i}{\sum_{i=1}^{n} f_i} \quad and \quad \sigma_E = \sqrt{\frac{\sum_{i=1}^{n} (E_i - \mu_E)^2 * f_i}{\sum_{i=1}^{n} f_i}}, \tag{4.24}$$

where f_i is the number of occurrences of energy value E_i. n is the total sample number.

For comparison purpose, we simulated the energy consumption distributions of each MTJ switching. The mean and the standard deviation values of these distributions can be found in Fig. 4.13. Our simulation results show that the highest and the lowest energies are consumed during the switchings of '1' to '3' and '0' to '1', respectively.

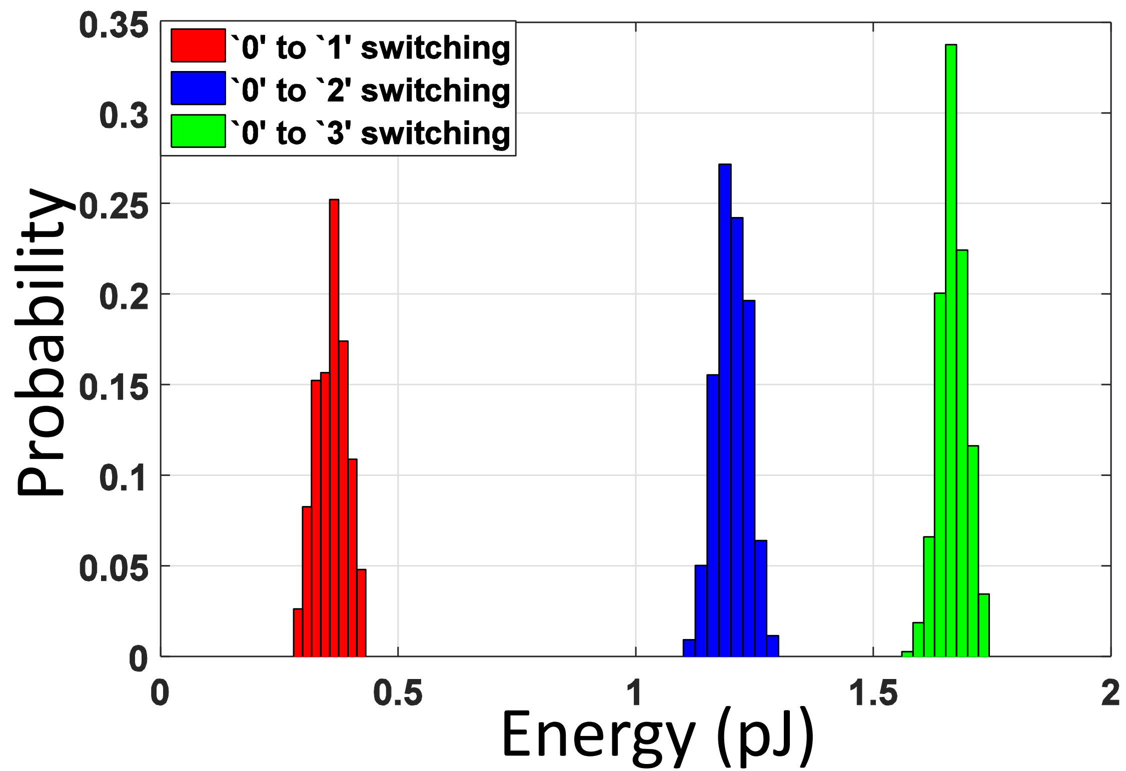

Figure 4.12: Biaxial MTJ energy consumption distribution during the switchings from ‘0’ to other states at $300K$.

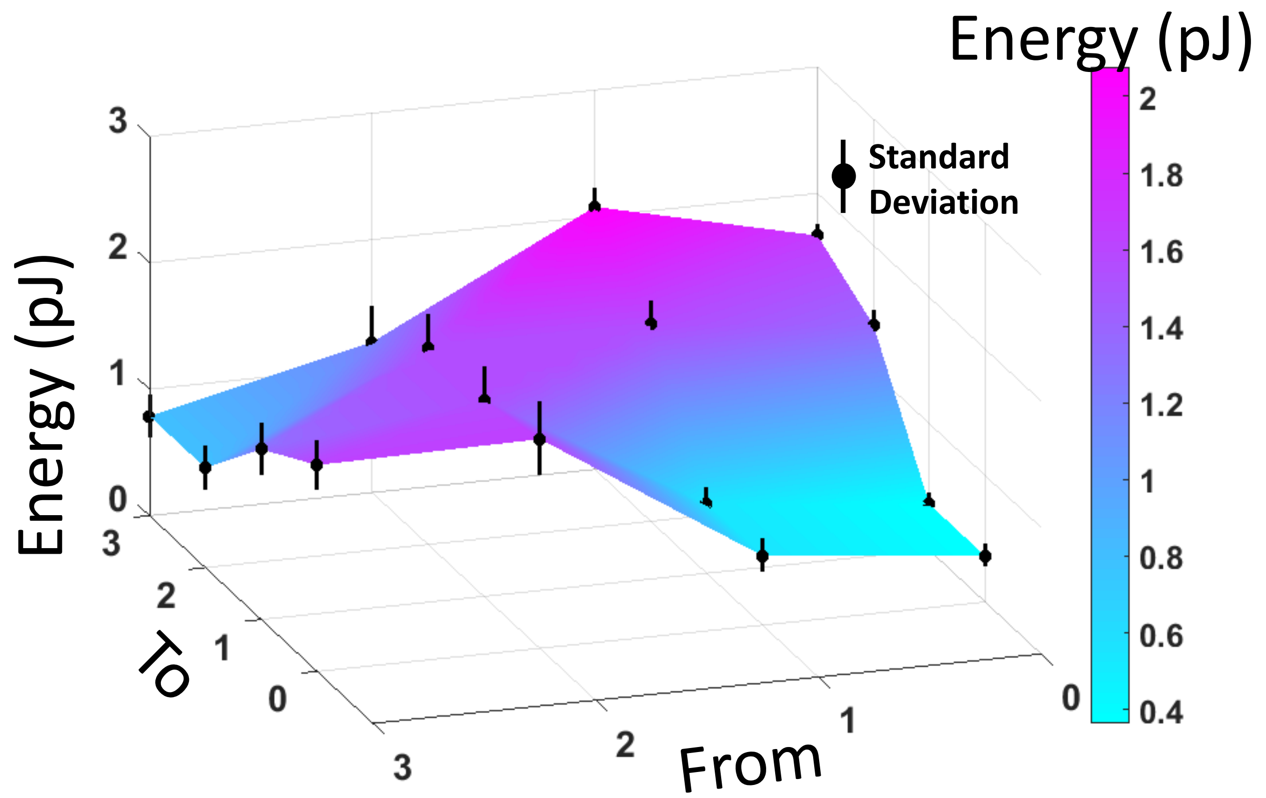

Figure 4.13: Energy consumption mean and standard deviation values of biaxial MTJ's each switching directions.

4.5.2 Write Errors of Biaxial MTJs

Uniaxial MTJs have two energy minimum points, each of which correspond to a resistance state. In order to switch from one state to the other, an energy barrier between these two energy minima must be overcome. If this barrier is not passed over within the write time, a write error of the uniaxial MTJ is induced and reflected as an incomplete write at circuit level. However, the write errors of biaxial MTJs are more complex since there are three energy barriers between four energy minimum points, as shown in Fig. 4.2. Particularly, besides incomplete write, a MLC-STT cell based on biaxial MTJs suffer from another write error type called "overwrite": If the switching current is too large, for example, during the '0' to '1' switching, the biaxial MTJ may cross over multiple energy barriers and settle down to the next two energy minima that correspond to '2' or '3', respectively. Note that overwrite errors happen only when writing an intermediate logic states, i.e., '1' or '2'.

Fig. 4.14 illustrates our simulation results of the write error rates of the biaxial MTJ when switching from '0' to '1', '2', and '3', respectively, with a fixed $60ns$ write time at different temperatures. For comparison purpose, we also plot the write error rate of the uniaxial MTJ with the same size switching from '0' to '1', which is the most erroneous switching direction of the uniaxial MTJ [34]. Instead of running the costly Monte-Carlo simulations, we use an open source tool – NVSim-VX^s [8] that embeds process variations and temperature impacts in our simulations. The MTJ switching time distribution model of NVSim-VX^s is modified with our proposed biaxial MTJ model. Fig. 4.14 shows that the biaxial MTJ switching of '0' to '1' suffers from the highest error rate, which is also significantly higher than that of the uniaxial MTJ. However, the write error rates of other switchings of the biaxial MTJ are close to the one of the uniaxial MTJ. It should be noted that, the '0' to '1' switching is also the one takes the longest time among the simulated biaxial MTJ switchings.

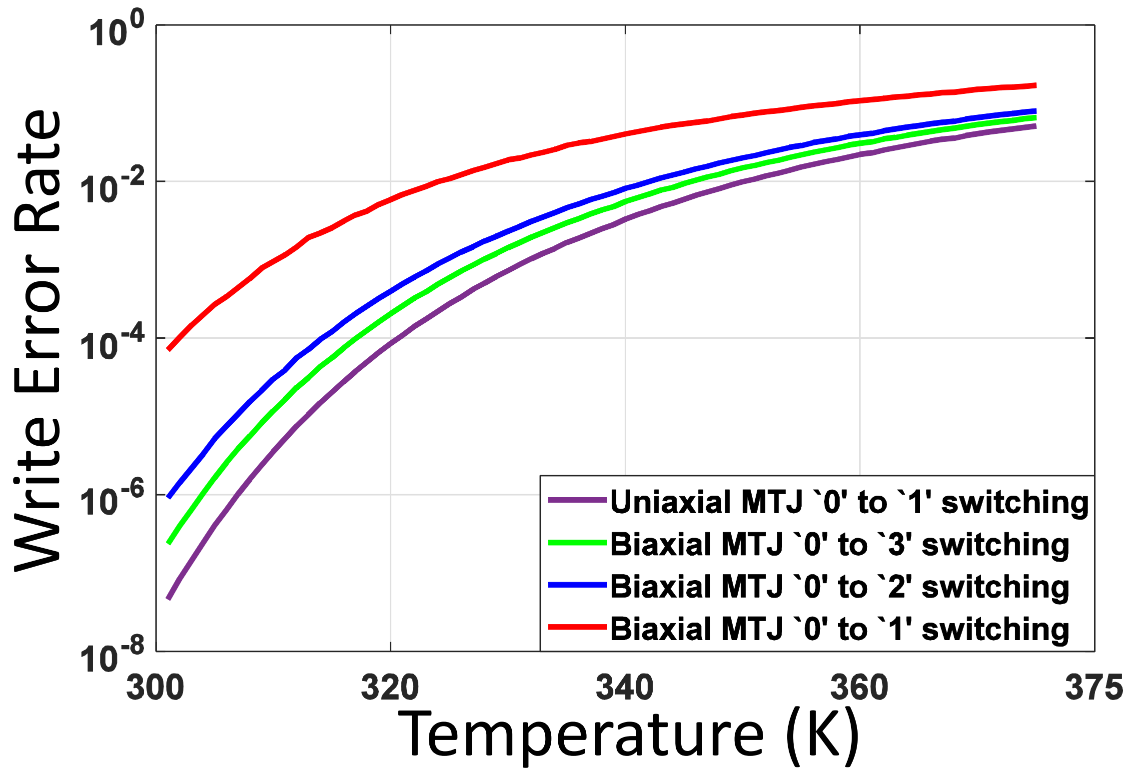

Figure 4.14: Uniaxial MTJ and biaxial MTJ write error rates at different temperatures with $60ns$ write time.

4.6 CONCLUSION

To eliminate the two-step write operations of conventional MLC STT-RAM cells, biaxial MTJ structure is proposed to store more than one bit in one MTJ device. In this work, we developed a dynamic biaxial MTJ model that can capture the switching transience between different resistance states of the biaxial MTJ and validated our model against the data measured from real device. Both process variations and thermal fluctuations can be considered in our model to perform the reliability and energy consumption analysis of the biaxial MTJ. Our results show that the highest energy consumption of the biaxial MTJ happens at the '1' to '3' switching by assuming a fixed write time while the lowest write error rate happens at the '0' to '3' switching. Hence, an adaptive write scheme, e.g., dynamically adjusting the write time of the MTJ to assure the completion of the write like early termination technology [36], may be critical in the MLC STT-RAM design based on the biaxial MTJ for energy reduction and reliability enhancement.

BIBLIOGRAPHY

[1] S. Arcaro, S. Di Carlo, M. Indaco, D. Pala, P. Prinetto, E. Vatajelu, et al. Integration of stt-mram model into cacti simulator. In *International Design & Test Symposium*, pages 67–72. IEEE, 2014.

[2] X. Bi, M. Mao, D. Wang, and H. Li. Unleashing the potential of mlc stt-ram caches. In *Proceedings of the International Conference on Computer-Aided Design*, pages 429–436. IEEE Press, 2013.

[3] X. Bi, Z. Sun, H. Li, and W. Wu. Probabilistic design methodology to improve run-time stability and performance of stt-ram caches. In *Proceedings of the International Conference on Computer-Aided Design*, pages 88–94. ACM, 2012.

[4] Y. Chen, X. Wang, H. Li, H. Xi, Y. Yan, and W. Zhu. Design margin exploration of spin-transfer torque ram (stt-ram) in scaled technologies. *IEEE Transactions on Very Large Scale Integration (VLSI) System*, 18(12):1724–1734, 2010.

[5] Y.-C. Chen, H. Li, W. Zhang, and R. E. Pino. The 3-d stacking bipolar rram for high density. *IEEE Transactions on Nanotechnology*, 11(5):948–956, 2012.

[6] X. Dong, C. Xu, Y. Xie, and N. P. Jouppi. Nvsim: A circuit-level performance, energy, and area model for emerging nonvolatile memory. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 31(7):994–1007, 2012.

[7] N. D'Souza, M. Salehi-Fashami, S. Bandyopadhyay, and J. Atulasimha. Hybrid spintronics-straintronic nanomagnetic logic with two-state elliptical and four-state concave magnetostrictive nanomagnets. In *Device Research Conference (DRC), 2014 72nd Annual*, pages 109–110. IEEE, 2014.

[8] E. Eken, L. Song, I. Bayram, C. Xu, W. Wen, Y. Xie, and Y. Chen. Nvsim-vx s: an improved nvsim for variation aware stt-ram simulation. In *Proceedings of the 53rd Annual Design Automation Conference*, page 70. ACM, 2016.

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

[10] E. Eken, Y. Zhang, W. Wen, R. Joshi, H. Li, and Y. Chen. A novel self-reference technique for stt-ram read and write reliability enhancement. *IEEE Transactions on Magnetics*, 50(11):1–4, 2014.

[11] L. M. Engelbrecht. Modeling spintronics devices in verilog-a for use with industry-standard simulation tools. 2011.

[12] K. Jabeur, L. Buda-Prejbeanu, G. Prenat, and G. Pendina. Study of two writing schemes for a magnetic tunnel junction based on spin orbit torque. *International Journal of Electronics Science and Engineering*, 7(8):501–507, 2013.

[13] J. A. Jess, K. Kalafala, S. R. Naidu, R. H. Otten, and C. Visweswariah. Statistical timing for parametric yield prediction of digital integrated circuits. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 25(11):2376–2392, 2006.

[14] R. Kanj, R. Joshi, and S. Nassif. Mixture importance sampling and its application to the analysis of sram designs in the presence of rare failure events. In *Design Automation Conference*, pages 69–72. ACM, 2006.

[15] Y. Kim, S. H. Choday, and K. Roy. Dstt-mram: Differential spin hall mram for on-chip memories. *arXiv preprint arXiv:1305.4085*, 2013.

[16] M. Motoyoshi, I. Yamamura, W. Ohtsuka, M. Shouji, H. Yamagishi, M. Nakamura, H. Yamada, K. Tai, T. Kikutani, T. Sagara, et al. A study for 0.18μm high-density mram. In *Symposium on VLSI Technology*, pages 22–23, 2004.

[17] 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*, pages 121–126, 2011.

[18] M. Pitke. Biaxial anisotropy for memory applications. *Czechoslovak Journal of Physics B*, 21(4-5):467–469, 1971.

[19] P. Shivakumar and N. P. Jouppi. Cacti 3.0: An integrated cache timing, power, and area model. Technical report, Compaq Computer Corporation, 2001.

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

[21] J. Sun. Spin-current interaction with a monodomain magnetic body: A model study. *Physical Review B*, 62(1):570, 2000.

[22] T. Uemura, T. Marukame, K.-i. Matsuda, and M. Yamamoto. Four-state magnetic random access memory and ternary content addressable memory using cofe-based magnetic tunnel junctions. In *37th International Symposium on Multiple-Valued Logic (ISMVL'07)*, pages 49–49. IEEE, 2007.

[23] S. Urazhdin, R. Loloee, and W. P. Pratt, Jr. Noncollinear spin transport in magnetic multilayers. *Phys.Rev.B*, 71(10):100401, Mar. 2005.

[24] A. van den Brink, S. Cosemans, S. Cornelissen, M. Manfrini, A. Vaysset, W. Van Roy, T. Min, H. Swagten, and B. Koopmans. Spin-hall-assisted magnetic random access memory. *Applied Physics Letters*, 104(1):012403, 2014.

[25] A. Vatankhahghadim and A. Sheikholeslami. A multi-level cell for stt-mram with biaxial magnetic tunnel junction. In *2015 IEEE International Symposium on Multiple-Valued Logic*, pages 158–163. IEEE, 2015.

[26] P. Wang, E. Eken, W. Zhang, R. Joshi, R. Kanj, and Y. Chen. A thermal and process variation aware mtj switching model and its applications in soft error analysis. In *More than Moore Technologies for Next Generation Computer Design*, pages 101–125. Springer, 2015.

[27] 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 *Design Automation Conference*, pages 1191–1196. ACM, 2012.

[28] W. Wen, Y. Zhang, L. Zhang, and Y. Chen. Loadsa: A yield-driven top-down design method for stt-ram array. In *Asia and South Pacific Design Automation Conference*, pages 291–296. IEEE, 2013.

[29] B. Wu, Y. Cheng, Y. Wang, A. Todri-Sanial, G. Sun, L. Torres, and W. Zhao. An architecture-level cache simulation framework supporting advanced pma stt-mram. In *Nanoscale Architectures (NANOARCH), 2015 IEEE/ACM International Symposium on*, pages 7–12. IEEE, 2015.

[30] W. Xu, Y. Chen, X. Wang, and T. Zhang. Improving stt mram storage density through smaller-than-worst-case transistor sizing. In *Design Automation Conference*, pages 87–90. ACM, 2009.

[31] W. Xu, H. Sun, X. Wang, Y. Chen, and T. Zhang. Design of last-level on-chip cache using spin-torque transfer ram (stt ram). *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, 19(3):483–493, 2011.

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

[33] Y. Zhang, X. Wang, and Y. Chen. Stt-ram cell design optimization for persistent and non-persistent error rate reduction: a statistical design view. In *International Conference on Computer-Aided Design*, pages 471–477. IEEE Press, 2011.

[34] Y. Zhang, X. Wang, Y. Li, A. K. Jones, and Y. Chen. Asymmetry of mtj switching and its implication to stt-ram designs. In *Proceedings of the Conference on Design, Automation and Test in Europe*, pages 1313–1318. EDA Consortium, 2012.

[35] Y. Zhang, L. Zhang, W. Wen, G. Sun, and Y. Chen. Multi-level cell stt-ram: Is it realistic or just a dream? In *Proceedings of the International Conference on Computer-Aided Design*, pages 526–532. ACM, 2012.

[36] P. Zhou, B. Zhao, J. Yang, and Y. Zhang. Energy reduction for stt-ram using early write termination. In *Computer-Aided Design-Digest of Technical Papers, 2009. ICCAD 2009. IEEE/ACM International Conference on*, pages 264–268. IEEE, 2009.