Modeling and Design of STT-MRAMs

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Electrical Engineering

by

Richard William Dorrance

2011
The thesis of Richard William Dorrance is approved.

Kang L. Wang

Chih-Kong Ken Yang

Dejan Marković, Committee Chair

University of California, Los Angeles
2011
To my parents
# Table of Contents

1 Introduction ................................................................. 1  
  1.1 Motivation for STT-MRAM .......................................... 1  
  1.2 Thesis Outline ...................................................... 4  

2 Magnetic Tunnel Junctions .............................................. 6  
  2.1 Introduction to Spintronics ........................................ 6  
    2.1.1 History ...................................................... 7  
    2.1.2 Principle of Operation ..................................... 8  
    2.1.3 Other Devices and Applications ............................ 9  
  2.2 The Magnetic Tunnel Junction ..................................... 10  
    2.2.1 Resistance Hysteresis ....................................... 10  
    2.2.2 Critical Switching Current ................................. 12  
    2.2.3 Tunnel Magnetoresistance Temperature Dependency .... 15  
    2.2.4 Bias Voltage Effects ....................................... 15  
    2.2.5 Other Important MTJ Characteristics ..................... 16  

3 Modeling MTJ Characteristics ........................................ 18  
  3.1 Modeling Dynamic Behavior ....................................... 18  
    3.1.1 Magnetization Dynamics .................................... 18  
    3.1.2 Tunnel Magnetoresistance .................................. 21  
  3.2 Model Verification ................................................ 23  
    3.2.1 Comparison to Measured Devices .......................... 23  

iv
### List of Figures

1.1 Comparison of memory technologies. ........................................ 2
1.2 SEM photo of an MTJ. ......................................................... 3
1.3 MTJ in (a) parallel and (b) antiparallel configuration. .............. 4

2.1 Spintronic operation of a spin polarizer. ............................... 8
2.2 Spintronic operation of a spin filter. ................................... 8
2.3 Resistance hysteresis of an MTJ. ......................................... 11
2.4 MTJ switching regimes. ..................................................... 13
2.5 Switching probability vs. pulse duration. ............................... 14

3.1 Sketch of basic MTJ structure. ............................................ 19
3.2 Efficiency factor of spin-polarization vs. $\theta$. ...................... 20
3.3 Normalized magnetization saturation. ................................... 22
3.4 Fitted MTJ parameters. .................................................... 23
3.5 TMR vs. temperature. ....................................................... 24
3.6 TMR vs. bias voltage. ....................................................... 24
3.7 R-H hysteresis. .............................................................. 25
3.8 Process flow for evaluating Verilog-A model. .......................... 26
3.9 Resistance vs. time. ......................................................... 27
3.10 Measured MTJ devices. ..................................................... 28
3.11 MTJ free layer dimensions. .............................................. 29

4.1 1T-1MTJ memory cell architectures. .................................... 31
List of Tables

3.1 Measured device statistics. ........................................ 28

5.1 $J_C(P \rightarrow AP)$ for an RA of $5 \, \Omega \cdot \mu m^2$ ........................................ 47
5.2 $J_C(P \rightarrow AP)$ for an RA of $10 \, \Omega \cdot \mu m^2$ ........................................ 47
5.3 $J_C(P \rightarrow AP)$ for an RA of $15 \, \Omega \cdot \mu m^2$ ........................................ 47

6.1 Time to read $R_P$ (90nm) ........................................ 52
6.2 Time to read $R_{AP}$ (90nm) ........................................ 52
ACKNOWLEDGMENTS

First, I would like to thank my advisor, Professor Dejan Marković, without whom this thesis could never have been written. I am sincerely grateful for the help and support he has given me through the course of this work. I would also like to acknowledge Professors Chih-Kong Ken Yang and Kang L. Wang and Dr. Pedram Khalili. Professor Yang’s knowledge and experience in designing memories was instrumental to this project. Without the help and insight of Professor Wang and Dr. Khalili I could never have comprehended the physics of spintronics half as well as I do now.

This work could not have progressed without the infrastructure and support provided by research group. I would like to whole heartedly thank Henry Chen, Kevin Dwan, and Yuta Toriyama for their help in the editing of this manuscript. I would also like to thank the rest of the STT-RAM circuits design team. Without their hard work and dedication, this project could never have gotten off the ground. I would also like to thank graduate students Juan G. Alzate, Pramey Upadhyaya, and Mark Lewis from the STT-RAM MTJ design team. They provided an extensive number of device characterizations and simulations for the development of my MTJ macro-model.

Most of all, I would like to thank my parents, Gary Dorrance and Karen Lawrence, for all the love and support they have provided over the years. Not the least of which included the editing of early drafts of this thesis. Last, but certainly not least, I would like to thank the mighty Tyrannosaurus Rex, king of the dinosaurs. May he live on in the hopes and dreams of a man still a boy at heart.
Abstract of the Thesis

Modeling and Design of STT-MRAMs

by

Richard William Dorrance

Master of Science in Electrical Engineering
University of California, Los Angeles, 2011
Professor Dejan Marković, Chair

Spin-Torque Transfer Magnetoresistive Random Access Memory (STT-MRAM) is an emerging memory technology with the potential to become a true universal memory: the density of DRAM, the speed of SRAM, and the non-volatility of Flash. STT-MRAM uses a Magnetic Tunnel Junction (MTJ) device as a non-volatile magnetic memory storage element and the recently discovered spin-torque phenomenon to switch magnetic states. In this work, the fundamental quantum mechanical nature of the MTJ is explored to develop a highly accurate physics-based model of its spintronic operation. Innovative design-space analysis techniques are introduced to investigate existing and proposed STT-MRAM architectures. Three test chips were fabricated using these new design methodologies at 90nm, 65nm, and 45nm technology nodes. Each chip has a memory density more than two times greater and a read/write performance more than 10 times greater when compared to published state-of-the-art STT-MRAMs. Theoretical and observed scaling trends show flash-like densities, with SRAM-equivalent access times, while using 10 times less energy in more advanced technology nodes (below 32nm).
CHAPTER 1

Introduction

STT-MRAM\(^1\) is an emerging memory technology that exploits the recently discovered phenomena of spin-torque transfer (STT) in MTJs. This chapter provides a brief motivation for STT-MRAM, as well as outlines the rest of the thesis.

1.1 Motivation for STT-MRAM

Currently, three types of memory exist, with each technology doing a single thing very well: Static RAM (SRAM), Dynamic RAM (DRAM), and Flash memory. SRAM has excellent read and write speeds, but has a very large cell size (requiring 6 or more transistors per cell). The speed of SRAM makes it ideally suited for embedded applications, particularly cache memory, where performance is more important than memory density. SRAM is volatile, but requires very little active power for data retention. DRAM is able to provide much better memory density through its use of a single transistor with a storage capacitor. However, charge tends to leak off of the capacitor, requiring a power hungry refresh cycle every few milliseconds. DRAM is typically used as the main system memory in a computer, where memory density and performance are more important than

\(^1\)In literature, Spin-Torque Transfer Random Access Memory (STT-RAM) and Spin Random Access Memory (SPRAM) are used interchangeably with STT-MRAM. However, STT-MRAM is more common and is, therefore, used exclusively in this thesis.
power consumption. Flash memory technologies are very attractive for mobile applications where non-volatility and very high densities are required. While Flash does have reasonably fast read access times, write speeds are very slow and endurance rates are very low (< 100,000 cycles). To optimize for power, performance, and cost, a typical system must integrate all three types of memory. STT-MRAM promises to be a “universal memory”, combining all of the advantages of SRAM, DRAM, and Flash. Such a memory would eliminate the need multiple application specific memories, improving system performance and reliability, while also lowering cost and power consumption.

Magnetoresistive Random Access Memory (MRAM) is a technology that has existed in one form or another since the late 1970s [1]. MRAM is based on the concept of using the direction of magnetization to store binary information, while exploiting magnetoresistive properties for data retrieval. The mid-1990s saw a

<table>
<thead>
<tr>
<th></th>
<th>SRAM</th>
<th>DRAM</th>
<th>Flash (NOR)</th>
<th>Flash (NAND)</th>
<th>FeRAM</th>
<th>MRAM</th>
<th>PRAM</th>
<th>RRAM</th>
<th>STT-MRAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-volatile</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Cell Size (F^2)</td>
<td>50-120</td>
<td>6-10</td>
<td>10</td>
<td>5</td>
<td>15-34</td>
<td>16-40</td>
<td>6-12</td>
<td>6-10</td>
<td>6-20</td>
</tr>
<tr>
<td>Read Time [ns]</td>
<td>1-100</td>
<td>30</td>
<td>10</td>
<td>50</td>
<td>20-80</td>
<td>3-20</td>
<td>20-50</td>
<td>10-50</td>
<td>2-20</td>
</tr>
<tr>
<td>Write/Erase Time [ns]</td>
<td>1-100</td>
<td>15</td>
<td>1μs/1ms</td>
<td>1ms/0.1ms</td>
<td>50/50</td>
<td>3-20</td>
<td>50/120</td>
<td>10-50</td>
<td>2-20</td>
</tr>
<tr>
<td>Endurance</td>
<td>(10^{18})</td>
<td>(10^{16})</td>
<td>(10^3)</td>
<td>(10^3)</td>
<td>(10^{18})</td>
<td>(&gt;10^{18})</td>
<td>(10^8)</td>
<td>(10^8)</td>
<td>(&gt;10^{18})</td>
</tr>
<tr>
<td>Write Power</td>
<td>Low</td>
<td>Low</td>
<td>Very High</td>
<td>Very High</td>
<td>Low</td>
<td>High</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Other Power Consumption</td>
<td>Leakage</td>
<td>Refresh</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>High Voltage Required</td>
<td>No</td>
<td>3V</td>
<td>6-8V</td>
<td>16-20V</td>
<td>2-3V</td>
<td>3V</td>
<td>1.5-3V</td>
<td>1.5-3V</td>
<td>&lt;1.5V</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Existing Products</th>
<th>Prototypes</th>
</tr>
</thead>
</table>

Figure 1.1: Comparison of memory technologies (source: Wolf et al. [1]).
resurgence of interest in MRAM technologies with the discovery of room temperature tunneling magnetoresistance (TMR) in magnetic tunnel junctions (MTJs). Fig. 1.2 shows an SEM photo of a single MTJ nanopillar. The spintronic operation of the MTJ is discussed later in this work. For now, it is sufficient to understand the MTJ as a pair of ferromagnets separated by a thin insulating layer. Two possible magnetic states arise, the parallel combination of the two layers (Fig. 1.3(a)) and the antiparallel combination (Fig. 1.3(b)). The parallel configuration leads to a low resistive state ($R_P$), while the antiparallel configuration leads to a high resistive state ($R_{AP}$).

The spin-torque transfer effect was first theoretically predicted and demonstrated by J. C. Slonczewski in 1996 [2] and has formed the basis of next generation MRAMs. STT-MRAM can scale well below 65nm, while reducing writing currents by more than a hundredfold [1]. Before STT, writing currents increased exponentially with MRAM scaling, causing electromigration and power concerns that prevented scaling below 90nm. The nonvolatile nature, low power, high performance, and memory density of STT-MRAM make it an excellent candidate for
the first commercially available universal memory. However, the lack of an accurate, compact macro-model, incorporating temperature and bias voltage effects, is the largest obstacle to the design of high performance STT-MRAMs. Without such a design tool, it is impossible to verify timing and yield or predict device behavior with scaling. Another big challenge is the integration of MTJs with CMOS. The process flow, while in principle fully compatible with CMOS, adds extra design and layout constraints. An integrated MTJ also behaves slightly differently, further stressing the need for an accurate macro-model.

1.2 Thesis Outline

Chapter 2 begins with a brief introduction to the field of spintronics, highlighting the general principles of operation and potential applications. The second half of Chapter 2 is dedicated entirely to the operation and device characteristics of the magnetic tunnel junction. Chapter 3 introduces a compact MTJ macro-model capable of accurately capturing and modeling the quantum mechanical behavior of MTJs. STT-MRAM memory cell and subarray architectures are outlined in
Chapter 4, with design-space analysis techniques introduced in Chapter 5. The analysis techniques introduced in Chapter 5 are used in Chapter 6 in the design of three separate STT-MRAM memory chips at 90nm, 65nm, and 45nm technology nodes. Finally, Chapter 7 presents ongoing and future work, and concludes the thesis.
CHAPTER 2

Magnetic Tunnel Junctions

The focus of this chapter is to introduce the MTJ device, as well as the field of spintronics. The first section provides a brief background on spintronics—its history and fundamental physical operation. Alternative devices (e.g. spin FETs, MBTs, and spin LEDs) and applications are discussed before defining the characteristics and unique properties of the MTJ device.

2.1 Introduction to Spintronics

Spintronics, the amalgamation of the words “spin” and “electronics,” involves the active control and manipulation of electron spin in solid-state electronics [3]. In traditional electronic devices, information processing works on the principle of control over the flow of charge through a semiconductor material. Large scale, non-volatile memories (e.g., hard disk drives or HDDs) exploit ferromagnetism to store information by forcing the spin alignment of many electrons [4]. Spintronics, as a whole, aims to merge information processing and storage through the use of spin-polarized currents [1].
2.1.1 History

Early work into spintronics began in the mid-1930s with the discovery of unusual resistance behavior in ferromagnetic materials at extremely low temperatures [3]. Electron tunneling measurements played a key role in early experimental work, with several key experiments in the early 1970s demonstrating the viability of spin filters (discussed later). In 1975, Jullierè [5] formulated his now-famous conductance model describing the change of conductance between the parallel and antiparallel states of an MTJ. However, it wasn’t until the mid-to-late 1980s that the room temperature magnetoresistive effects were discovered. Anisotropic magnetoresistive (AMR) layers were first used to construct AMR-MRAM to replace bulky and heavy plated-wire radiation-hard memories [1]. AMR was quickly replaced by the discovery of giant magnetoresistance (GMR) in 1988 [6]. Since the discovery of GMR, electron spin has formed the basis of almost all electronic information storage [7].

In the early 1990s, MTJ materials with higher TMRs (on the order of 20% at room temperature) were discovered [1]. Since then, MTJ structures (using MgO insulating barriers) with TMRs on the order of 1000% have been demonstrated at room temperature [8]. Within ten years of its discovery, spintronics has grown into a billion dollar industry, with commercial sales exceeding $3 billion in 2005 [6]. Despite these successes, spin injection from ferromagnetic layers into semiconductors remains a significant bottleneck in semiconductor-based spintronics. Recently, much emphasis has been placed in trying to induce ferromagnetism in a semiconductors to produce dilute magnetic semiconductors (DMS) [7]. DMS has the potential to improve the Curie temperature and magnetic bandgap of future spintronic devices [9].
2.1.2 Principle of Operation

Electron spin is a “pseudovector” with a fixed magnitude but a variable direction (spin polarization). The spin polarization of an electron can be made bistable by placing it in a magnetic field. In the presence of a magnetic field, only spin polarizations parallel or antiparallel to the field are possible [1]. This property introduces the concept of a spin polarizer (Fig. 2.1). A thin ferromagnetic layer can act as a spin polarizer. When a spin unpolarized current passes through the
ferromagnetic layer, it tends to become spin-polarized in the direction of magnetization [6]. Another key aspect to spintronics is the concept of a spin filter (Fig. 2.2). A spin filter will only pass a current if the two are polarized in the same direction. If the current and filter are completely antiparallel, no current is passed. Ferromagnetic films also display the properties of a spin filter [6]. A “spin valve” can be constructed by using a spin polarizer in conjunction with a spin filter [4]. By controlling the angle of magnetization between the polarizer and the filter, a magnetically controlled spin valve can be formed. The spin valve effect is exploited in MRAMs to use MTJs as the memory storage element [1].

2.1.3 Other Devices and Applications

Several kinds of “spin transistors” exist, including the spin field-effect transistor (spin FET), the magnetic bipolar transistor (MBT), and hot-electron spin transistors [3, 7]. Structurally similar to a MOSFET, a spin FET sandwiches the conducting channel between two ferromagnetic layers. When the ferromagnets are aligned in the parallel configuration, the spin FET behaves like a normal MOSFET. However, when configured in the antiparallel alignment, transistor will be shut off [10]. Spin FETs can be easily integrated into existing CMOS circuitry and provide much larger ON/OFF current ratios [3]. MBTs are essentially BJTs with the addition of a ferromagnetic spin injector attached to the emitter. In an MBT, the gain factor $\beta$ heavily depends upon nonequilibrium spin polarization and is called magnetoamplification [11]. MBTs can be used to generate almost 100% spin-coherent currents that can be very long lived [3,11].

Another potential application of spintronics is optics, specifically, through the use of spin light emitting diodes (spin LEDs) and spin selective Kerr rotators [7]. In a spin LED, the polarization of the light emitted is modulated through the
application of an external magnetic field \[12\]. Variable polarized LEDs promise
to provide more energy efficient displays and significantly higher signal-to-noise
ratio (SNR) in optical communications \[7\]. A Kerr rotator takes advantage of the
magneto-optic Kerr effect (MOKE), the unique optically-reflective properties of
magnetic materials, to manipulate the polarity of reflected light. Traditionally,
Kerr rotators have many applications in the microscopic imaging of magnetic
domains, magnetic media, and terahertz lasers \[13\]. A spin selective rotator,
with the application of a bias voltage, can be made to reflect incident light either
with or without a large Kerr rotation angle \[7\].

2.2 The Magnetic Tunnel Junction

This section is intended to describe the major device characteristics observed
in MTJs. The science responsible for each effect, as well as their importance to
the MTJ model, is discussed.

2.2.1 Resistance Hysteresis

The large resistance hysteresis present in MTJs makes them very well-suited
as a non-volatile memory element. The source of this hysteresis is very nicely
explained by the spin-valve structure of an MTJ \[3\]. As mentioned before in
Fig. 3.1, an MTJ consists of a thin insulating layer sandwiched between two fer-
romagnetic layers. The electromagnetic dynamics of the system allows for only
two possible states: parallel or antiparallel \[6\]. The two ferromagnetic layers are
magnetized in the same direction while in the parallel state and in the oppo-
site directions while in the antiparallel state. When a current flows through the
MTJ, one ferromagnetic layer acts as a spin polarizer and the other as a spin
Figure 2.3: Resistance hysteresis of an MTJ. Switching from $P \rightarrow AP$ (blue arrows) and $AP \rightarrow P$ (red arrows).

filter. In the parallel state, since the two ferromagnetic layers are aligned, the current is passed undisturbed, creating a low resistive state ($R_P$). However, in the antiparallel state, the spin filter will block the antiparallel current generated by the polarizing layer, creating a high resistive state ($R_{AP}$). Tunnel magnetoresistance (TMR) is a metric for determining the efficiency of spin-valve operation in an MTJ [14]. TMR is defined as:

$$TMR = \frac{R_{AP} - R_P}{R_P}. \quad (2.1)$$

Fig. 2.3 shows a sample resistance hysteresis of an MTJ by sweeping the bias voltage. $R_P$ and $R_{AP}$ are clearly evident, along with several other characteristics to be discussed: critical switching currents, switching asymmetry, and the bias
 voltage dependance of TMR.

### 2.2.2 Critical Switching Current

#### 2.2.2.1 Asymmetric Switching Currents

It should be noted that the critical switching currents are asymmetric, with $I_C(\text{P} \rightarrow \text{AP}) > I_C(\text{AP} \rightarrow \text{P})$ [15]. This effect was predicted by Slonczewski [2] with his discovery of the spin-torque transfer phenomena. This asymmetry is proportional to and increases linearly with TMR [16]. The simplest explanation of this behavior is that the antiparallel configuration is a lower energy state than the parallel case [3], making it is easier to switch to the antiparallel state than the parallel state. Several techniques exist to minimize the asymmetry. Lee et al. [17] were able to tune the magnetostatic offset field (using an external magnetic field) with exceptional results, reducing the asymmetric current ratio from 1.51 to 1.04. Yao et al. [18] were able to reduce the offset from 1.50 to 1.28 with the introduction of a nanocurrent-channel layer to the MTJ stack.

#### 2.2.2.2 Switching Regimes

In MTJs, two types of magnetic switching occur due to spin-torque transfer: precessional and thermally activated switching [19, 20]. Precessional switching occurs on a nanosecond time scale, while thermally activated switching occurs at much larger time scales [14]. The transition between these two switching regions lies between 1 and 10ns, which is depicted in Fig. 2.4. The dynamics of precessional switching are well described by the Landau-Lifshitz-Gilbert equation (LLGE) [21, 22], given by:

$$\frac{\partial \vec{m}}{\partial t} = -\gamma M_S \vec{m} \times \left( \vec{h}_{\text{eff}} - \alpha \frac{\partial \vec{m}}{\partial t} \right).$$  \hspace{1cm} (2.2)
Equation 2.2, with the addition of Slonczewski’s spin-torque transfer term [2], will be discussed in much more detail in Chapter 3.

Switching occurs on much longer time scale when the current though the MTJ is less than the critical switching current [19]. In the thermally activated regime, the switching current is a function of pulse duration $\tau$:

$$ I_C = I_{C0} \left[ 1 - \frac{\ln(\tau/\tau_0)}{\Delta} \right], $$

where $\Delta$ is the thermal stability of the MTJ, $\tau_0$ is the natural time constant, and $I_{C0}$ is the critical switching current [23].
2.2.2.3 Probabilistic Switching

Due to thermal agitation, the initial angle between the magnetizations of the fixed and free layers are in constant flux [25]. Combined with other finite temperature effects, this leads to a time-varying critical switching current [26]. This effect is very well modeled as a single critical switching current with a probabilistic distribution [27]. Fig. 2.5 shows one such measurement of the probabilistic distribution for a $135 \times 65nm^2$ CoFeB/MgO/CoFeB device. For memory applications, devices exhibiting very sharp transitions are highly desirable [28].
2.2.3 Tunnel Magnetoresistance Temperature Dependency

The sensitivity of TMR to temperature is well documented in literature [5, 29–32]. The effect at zero bias voltage is very well described by the Jullière conductance model [29]. The Jullière model decomposes the conductance of the MTJ into two parts: (i) $G_T$, the conductance due to direct elastic tunneling, and (ii) $G_{SI}$, the conductance due to imperfections in the insulating layer (assumed to be unpolarized). The total conductance ($G$), as a function of the angle $\theta$ is given by:

$$G(\theta) = G_T \{1 + P_1P_2 \cos(\theta)\} + G_{SI},$$

where $P_1$ and $P_2$ are the factors of spin-polarization for the two ferromagnetic layers, and $\theta = 0^\circ$ for parallel and $\theta = 180^\circ$ for anti-parallel magnetization. The temperature dependence of spin-polarization has been extensively studied and shown to be:

$$P(T) = P_0 \left(1 - \alpha_{sp}T^{3/2}\right).$$

It should be noted that variations in $G_T$ due to temperature are almost negligible, whereas $G_{SI} \propto T^{4/3}$ has been confirmed both theoretically and experimentally [33].

2.2.4 Bias Voltage Effects

The Jullière conductance model is not perfect, being only able predict TMR at zero bias voltage [30]. Fig. 2.3 illustrates the effect of the so called “zero bias anomaly” in an MTJ structure [34]. The source of the bias voltage dependence of TMR is still not very well understood [35]. However, it is suspected that elastic currents play a role at low voltages [36] and redistribution of the density of states at higher voltages [35]. At higher voltages, Simmons’ formula can be used to
model the density of states to predict degradation of TMR to a bias voltage [31].

2.2.5 Other Important MTJ Characteristics

2.2.5.1 Self Induced Heating

Due to small device sizes and large write currents, the power density of a write operation in an MTJ can be very high. These high power densities can lead to localized heating or self induced heating in MTJs [37]. Hotspots (weak areas in the insulating barrier) and pinholes (direct contact between the magnetic layers) cause nonuniform current flow through the MTJ [38]. This leads to nonuniform heating across the tunneling barrier, affecting spin-polarization efficiency and causing inelastic electron scattering [38]. Simulations show that consecutive write operations produce a 9-15°C increase in the temperature of the MTJ [37]. Additionally, a large number of writes followed by a read leads to degraded sensing margin. Self induced heating is exploited as the writing mechanism in Thermal Assisted Switching MRAMs (TAS-MRAMs) [39]. However, in STT-MRAMS, lower RAs are generally used to avoid self induced heating [37].

2.2.5.2 Backhopping

Backhopping is a recently discovered phenomenon, whereby increasing the bias voltage beyond the apparent switching threshold causes the MTJ to process back and forth before settling to its original state [40]. This results in a lowered probability of switching at bias voltages beyond the threshold, causing non-monotonicity in the probability switching curves [41]. Backhopping is also much more pronounced in switching from an antiparallel to a parallel state [40,41]. This suggests that backhopping is related to the interlayer exchange coupling be-
between the free and fixed layers. Backhopping is more pronounced on longer time scales, where self induced heating could be lowering the thermal energy barrier and causing hot-electron events [40]. Another explanation is that certain noise processes (discussed in the next section) might be responsible [41].

2.2.5.3 Noise

Many different mechanisms are responsible for noise in MTJs. Among these are thermal noise (Johnson-Nyquist), shot noise (current), flicker noise (1/f), random telegraph noise (RTN), and noise due to charge-trapping in the oxide barrier [28,42–45]. Due to the strong coupling between magnetization and junction resistance in MTJs, noise in the magnetic domain is responsible for random resistance fluctuations [42]. These resistance fluctuations are responsible for 1/f noise as well as RTN [45]. Magnetic impurities inside the tunneling barrier are responsible for charge-trapping [42].

Thermal noise dominates at low bias voltages before quickly being overpowered by shot noise [44]. At room temperatures, shot noise typically dominates for bias voltages greater than 50mV [43]. The thermal noise of an MTJ is given by \( S_V = 4k_B T R_{MTJ} \), where \( k_B \) is Boltzmann’s constant, \( T \) is in Kelvin, and \( R_{MTJ} \) is the resistance of the MTJ [28]. Similarly, shot noise can be expressed as \( S_V = 2eI R_{MTJ}^2 \), where \( e \) is the charge of an electron and \( I \) is the current through the device [28].

Another significant contribution to low-frequency noise is due to domain wall hopping between pinning sites [28,42]. These pinning sites are created by edge roughness, interface defects, bulk defects, and random film anisotropy [42]. The low-frequency noise characteristics of an MTJ can be significantly reduced by improving the smoothness of the ferromagnetic/insulator interface [45].
CHAPTER 3
Modeling MTJ Characteristics

Recent advances in MgO-based MTJs show strong potential for STT-MRAMs [46]. STT-MRAM has the potential to rival the densities of DRAM, the speed of SRAM, and is non-volatile without degrading over time like Flash [47]. The greatest hindrance in the design of STT-MRAM, and other spintronics circuits, is the lack of a compact MTJ model capable of accurately modeling temperature and voltage dependencies. Capturing these dependencies, in a compact model compatible with circuit simulators, is crucial for performing accurate Monte Carlo simulations to place yield and performance bounds on STT-MRAM. This chapter presents such a model implemented in Verilog-A. The model’s simulation results were also compared to a model implemented using the LLG Micromagnetics Simulator [48] and actual device measurements from 135nm by 65nm CoFeB/MgO/CoFeB MTJs.

3.1 Modeling Dynamic Behavior

3.1.1 Magnetization Dynamics

The precessional motion of magnetization (∆M) of the free layer of a MTJ, in the presence of an external magnetic field (∆H_{eff}), can be very accurately modeled by the LLGE, Eq. 2.2 [23]. With the introduction of Slonczewski’s spin-torque
transfer term [2], the normalized LLGE with STT is given by:
\[
\frac{\partial \vec{m}}{\partial t} = -\gamma M_S \vec{m} \times \left( \vec{h}_{\text{eff}} + \frac{J_e}{J_C} b(\theta) (\vec{m} \times \vec{p}) - \alpha \frac{\partial \vec{m}}{\partial t} \right),
\]  
(3.1)

where \( M_S = |\vec{M}| \), \( \gamma \) is the absolute value of the gyromagnetic ratio \((\gamma_e \mu_0)\), \( \vec{m} \) is the unit vector in the direction of \( \vec{M} \), \( \vec{p} \) is the unit vector in the direction of the magnetization of the fixed layer, \( \vec{h}_{\text{eff}} = \vec{H}_{\text{eff}}/M_S \), \( J_e \) is the current density (see Fig. 3.1), \( \theta \) is the angle between \( \vec{m} \) and \( \vec{p} \), and \( \alpha > 0 \) is the material-dependent Gilbert damping constant. The efficiency factor of spin-polarization \((b(\theta)) \), see Fig. 3.2) is defined as:
\[
b(\theta) = \left[ -4 + (1 + P)^3 \frac{\{3 + \cos(\theta)\}}{4P^{3/2}} \right]^{-1},
\]  
(3.2)

where \( P \) is the percentage of electrons polarized in the \( \vec{p} \) direction. The switching current density \((J_C) \) has been modified to include thermally-activated switching [23]. For a constant pulse of duration \( \tau \), \( J_C \) is given by:
\[
J_C = J_{C0} \left[ 1 - \frac{\ln(\tau/\tau_0)}{\Delta} \right],
\]  
(3.3)

where \( \Delta \) is the thermal stability of the MTJ and \( \tau_0 = (\gamma M_S)^{-1} \) is the natural time constant. Furthermore, the characteristic current density \((J_{C0}) \) is defined
Figure 3.2: Magnitude of the efficiency factor of spin-polarization vs. $\theta$ for $P = 0.65$.

as:

$$J_{CO} = \gamma M_S \frac{eM_Sd}{g_e\mu_B},$$  \hspace{1cm} (3.4) $$

where $e$ is the absolute value of electron charge, $d$ is the thickness of the free layer, $g_e$ is the Landé factor for electrons, and $\mu_B$ is the Bohr magneton [49].

3.1.1.1 Effective Magnetic Field

The effective magnetic field ($\vec{H}_{eff}$) is given by:

$$\vec{H}_{eff} = \vec{H}_{ext} + \vec{H}_{dem} + \vec{H}_{an},$$  \hspace{1cm} (3.5) $$

where $\vec{H}_{ext}$ is the external applied magnetic field, $\vec{H}_{dem}$ is the demagnetization field, and $\vec{H}_{an}$ is the magnetocrystalline anisotropy field. The demagnetization field (shape anisotropy) varies with the geometry of the free layer and is modeled as $\vec{H}_{dem} = N\vec{M}$. If the free layer is assumed to be a very flat ellipsoid, the factors of the demagnetization tensor $N$, calculated by Osborn [50], are:

$$N_X = \frac{d}{L} \left(1 - e^2\right)^{1/2} \frac{K - E}{e^2},$$  \hspace{1cm} (3.6) $$
\[ N_Y = \frac{d K - (1 - e^2) E}{L} \frac{E}{e^2 (1 - e^2)^{1/2}}, \] (3.7)

\[ N_Z = 1 - \frac{d}{L} \frac{E}{(1 - e^2)^{1/2}}, \] (3.8)

where \( K \) and \( E \) are the complete elliptic integrals of the first and second kind whose argument is:

\[ e = \left(1 - \frac{W^2}{L^2}\right)^{1/2}. \] (3.9)

### 3.1.1.2 Temperature Dependencies

In the dynamic equations, only the magnetization saturation (\( M_S \)) and the spin-polarization (\( P \)) vary with temperature. For temperatures below the Curie temperature (\( T_C \)), we can use the Weiss theory of ferromagnetism [51] to model:

\[ M_S(T) = M_{S0}(1 - T/T_C)^\beta, \] (3.10)

where \( M_{S0} \) is the magnetization saturation at absolute zero and \( \beta \) is the material-dependent critical exponent (see Fig. 3.3) [52]. Similarly, the temperature dependence of spin-polarization has been extensively studied and shown to be:

\[ P(T) = P_0 \left(1 - \alpha_{sp} T^{3/2}\right) \] (3.11)

where \( P_0 \) is the spin-polarization at absolute zero and \( \alpha_{sp} \) is a material and geometric dependent constant [33].

### 3.1.2 Tunnel Magnetoresistance

Temperature variations in MTJ conductance \( G(\theta) \) are modeled in Shang et al. [33] by modifying the Julliere model. Julliere’s model, Eq. 2.4, is reproduced
Figure 3.3: Normalized plot of magnetization saturation for generic ferromagnetic materials.

here with $P_1 = P_2 = P$:

$$G(\theta) = G_T \{1 + P^2 \cos(\theta)\} + G_{SI}. \quad (3.12)$$

As a reminder, the variation of $G_T$ due to temperature is negligible, whereas $G_{SI} \propto T^{4/3}$. Using $\theta = 0^\circ$ for parallel magnetization and $\theta = 180^\circ$ for anti-parallel, the tunnel magnetoresistance with zero applied bias voltage ($TMR_0$) can be expressed as:

$$TMR_0(T) = \frac{2P_0^2(1 - \alpha_{sp}T^{3/2})^2}{1 - P_0^2(1 - \alpha_{sp}T^{3/2})^2 + \frac{G_{SI}(T)}{G_T(T)}}. \quad (3.13)$$

The Jullière model fails to predict the effects of a bias voltage on TMR [31]. However, this can be rectified with the addition of a simple fitting function:

$$TMR(T, V) = \frac{TMR_0(T)}{1 + \left(\frac{V}{V_0}\right)^2}, \quad (3.14)$$

where $V_0$ is the voltage at which TMR is halved.
<table>
<thead>
<tr>
<th>Geometric Parameters</th>
<th>LLGE Damping</th>
<th>Conductance</th>
</tr>
</thead>
<tbody>
<tr>
<td>W 65 [nm]</td>
<td>$\alpha$ 0.05</td>
<td>$G_T$ 1.07 [mS]</td>
</tr>
<tr>
<td>L 135 [nm]</td>
<td></td>
<td>$G_S$ 0 [mS]</td>
</tr>
<tr>
<td>d 1.8 [nm]</td>
<td>$M_{S0}$ 1100 [emu/cc]</td>
<td></td>
</tr>
<tr>
<td>$t_{ox}$ 0.9 [nm]</td>
<td>$T_C$ 1420 [K]</td>
<td></td>
</tr>
<tr>
<td>Spin Polarization</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$P_0$ 0.725</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\alpha_{sp}$ 2x10^{-5} [K^{-3/2}]</td>
<td>$V_0$ 65 [nm]</td>
<td></td>
</tr>
</tbody>
</table>

**Demagnetization Tensor (Calculated)**

| $N_X$ 0.0113 | $N_Y$ 0.0198 | $N_Z$ 0.9689 |

**Figure 3.4:** Fitted MTJ parameters.

### 3.2 Model Verification

 Implemented in Verilog-A, the compact model is comprised of two electrical terminals, an externally applied field vector, the initial state of magnetization, the demagnetization factors, and 13 device-specific parameters: 4 geometric parameters, 8 material-dependent parameters, and 1 empirically-derived parameter. The model was fitted to a 135nm by 65nm CoFeB/MgO/CoFeB MTJ (see Fig. 3.4). For validation of the model, we compare to detailed micromagnetic simulations, previously published data, and experimental results from fabricated MTJ nanopillars.

#### 3.2.1 Comparison to Measured Devices

The ability of Eqs. 3.12 and 3.13 to accurately model the temperature dependence of $R_P$ and TMR is well established in literature. Shang et al. [33] managed to obtain excellent fitting for $\text{Al}_2\text{O}_3$ based MTJs before deviating at high temperatures due to the crystallization of the amorphous insulating layer. Similarly, Kou et al. [53] and Wiśniowski et al. [55] reported extremely good fitting for MgO-
Figure 3.5: TMR vs. temperature: Verilog-A model (line), reported in [53] (triangles), reported in [54] (squares), and fabricated devices (circles).

Figure 3.6: TMR vs. an applied bias voltage at 300K: Verilog-A model (black line) and fabricated devices (red circles).
and IrMn-based devices respectively. Fig. 3.5 contains a loosely fitted curve of (3.13) for limited empirical data, as well as reported TMR values from literature. An excellent fitting of (2.1) to experimental data was obtained (Fig. 3.6), with an accuracy of ±3%. The steady-state accuracy of (3.1) at modeling switching thresholds for an applied external field is quite good and can be seen in Fig. 3.7.

### 3.2.2 Comparison to Micromagnetic Simulations

It is extremely difficult to accurately measure the switching characteristics of fabricated MTJs in the nanosecond regime. However, micromagnetic simulations are fully capable of accurately predicting their behavior [48]. As such, micromagnetic simulations were used to evaluate the switching behavior of the Verilog-A model in the nanosecond regime at different temperatures (see Fig. 3.8). Fig. 3.9 shows the time evolution of the resistance model ($R(t)$) and the micromagnetic
derived resistance ($R'(t)$) at 300K and 380K (expected operating temperature when integrated with CMOS).

Being based on a macrospin model, $R(t)$ does not account for non-uniformities in the free layer magnetization during switching. Despite this, the pre-switching oscillations and underdamped behavior of $R'(t)$ are still observed to a point in $R(t)$. This effect is captured by the shape anisotropy modeled by the demagnetization tensor (Eqs. 3.6, 3.7, and 3.8). Also, $R(t)$ manages to track the switching delay of $R'(t)$ across a wide range of temperatures and pulse shapes.

### 3.3 Statistical Characterization of MTJ Devices

#### 3.3.1 MTJ Device Variability

While statistical variation of CMOS is generally well understood, similar characteristics for MTJs have not been well documented. This section uses a combination of fundamental equations and measured device characteristics to model the statistical behavior of MTJs. Figure 3.10(a) contains a plot of measured $R_P$
Figure 3.9: Resistance vs. time for an applied $\pm 2\text{V}$ 100MHz square-wave at (a) 300K and (b) 380K. Verilog-A model ($R(t)$) is the black, dashed line and micromagnetic simulations ($R'(t)$) in the red, solid line.
Figure 3.10: Measured (a) $R_{AP}$ vs. $R_P$ and (b) TMR vs. RA for MTJ nanopillars measuring 150 × 45nm$^2$ (X), 130 × 50nm$^2$ (Y), and 170 × 45nm$^2$ (Z)

<table>
<thead>
<tr>
<th></th>
<th>X</th>
<th>Y</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>$TMR$ [%]</td>
<td>105.7</td>
<td>107.3</td>
<td>105.3</td>
</tr>
<tr>
<td>$\sigma_{TMR}$ [%]</td>
<td>4.7</td>
<td>2.7</td>
<td>4.6</td>
</tr>
<tr>
<td>$RA$ [$\Omega \cdot \mu m^2$]</td>
<td>4.88</td>
<td>5.51</td>
<td>5.22</td>
</tr>
<tr>
<td>$\sigma_{RA}$ [$\Omega \cdot \mu m^2$]</td>
<td>0.342</td>
<td>0.297</td>
<td>0.311</td>
</tr>
</tbody>
</table>

Table 3.1: Measured device statistics.

vs. $R_{AP}$ for 105 MTJ nanopillars of varying size and target RAs. Variations in resistance and TMR are due to a combination of lithographic variations in the physical dimensions of the nanopillar and minute variations in the thicknesses of the up to 20 different layers in state-of-the-art MTJ processes [56]. The cumulative effects of these variations on RA and TMR can be easily measured [57], as shown in Fig. 3.10(b) and Table 3.1.
3.3.2 Scaling of MTJ Current and Resistance

It would be useful to understand how the resistance and switching current of an MTJ changes as the device is scaled. A precessional-based switching model, modified to include thermally-activated switching, was used as a starting point. The switching current of an MTJ is given by:

\[ I_C = I_{C0} \left[ 1 - \ln\left(\frac{\tau}{\tau_0}\right) \right], \tag{3.15} \]

with the critical switching current \( I_{C0} \) given by:

\[ I_{C0} = \frac{\alpha 4\pi e}{\eta \hbar} M_S^2 V, \tag{3.16} \]

where \( k_B \) is Boltzmann’s constant, \( T \) is the absolute temperature in Kelvin, \( H_K \) is the out-of-plane uniaxial anisotropy, and \( E \) is the energy of anisotropy [58,59].

For an MTJ with free layer dimensions \( l > w >> d \) [60], as shown in Fig. 3.11, the thermal stability of an MTJ is approximately:

\[ \Delta = \frac{E}{k_B T} = \frac{H_K M_S}{2k_B T} V \approx d \left( \frac{1}{w} - \frac{1}{l} \right) \frac{M_S^2}{k_B T} V. \tag{3.17} \]

Dimensional scaling is performed to maintain a constant \( \Delta \) in order to ensure the long-term non-volatility of the MRAM. If dimensions \( l \) and \( w \) of the MTJ are scaled by a factor \( \lambda \) to manipulate \( I_{C0} \) and \( R_{P/AP} \), then to keep \( \Delta \) constant, \( d \) must scale by \( \lambda^{-1/2} \). This results in \( I_{C0} \propto lwd \rightarrow \lambda^{3/2} \) and \( R_{P/AP} \propto l^{-1}w^{-1} \rightarrow \lambda^{-2} \).
CHAPTER 4

Memory Architectures

Several different types of memory architectures exist for STT-MRAMs. At the cell level, many architectures are tailored to certain MTJ characteristics, more specifically to the ratio of the critical writing currents $I_C(P \rightarrow AP)$ and $I_C(AP \rightarrow P)$. Other cell architectures attempt to exploit the different thresholds between reading and writing current to increase effective memory density. At the array level, several different subarraying techniques are employed to maximize performance and minimize area.

4.1 Cell Architectures

4.1.1 1T-1MTJ

There are two widely used 1T-1MTJ STT-MRAM cell architectures, the “conventional” cell (Fig. 4.1(a)) and the “reverse” cell (Fig. 4.1(b)) [61]. The “conventional” architecture gets its name from the fact that most MTJ are deposited with the fixed layer on the bottom. A smooth deposition surface is required to form a high quality pinning layer capable of generating the fixed layer [62]. The surface roughness introduced by various film deposition steps generally makes depositing the pinning layer on the top of the MTJ stack impractical. This means that it is easier to connect the fixed layer of the MTJ to the access transistor and
the free layer to the bitline.

The “reverse” structure is built exactly as it sounds, with the fixed and free layers connected in the reverse fashion of the “conventional” cell. The “reverse” architecture attempts to match the inherent driving current asymmetry of the access transistor to the asymmetric switching currents of the MTJ [16]. It is widely assumed that the trade-off between these two architectures depends solely upon the ratio of the critical writing currents $I_C(P \rightarrow AP)$ and $I_C(AP \rightarrow P)$. If this ratio is greater than 1, a “reverse-connected” architecture should be used. Otherwise, a “conventional” architecture should be used. However, this is not exactly true as this assumption fails to take into account both the $V_{GS}$ and $V_{DS}$ operating points of the access transistors during the write operation. For state-of-the-art technology nodes with supply voltages below 1V, the cutoff point between these architectures is a writing current ratio closer to 1.5.
4.1.2 Shared

Access device "sharing" is one potential technique for increasing cell density. As shown in Fig. 4.2, one access transistor is connected to multiple MTJ devices, with additional bitlines to support independent access. This also allows the access transistor to be sized up to provide higher write current while maintaining the same overall memory density. However, there are several shortcomings associated with this technique which will be discussed in future sections.

4.1.3 Stacked

MTJ device “stacking” is another potential technique to increase cell density. “Stacking” works by connecting several different types of MTJ in series with one access transistor, shown in Fig. 4.3, in a similar fashion to multi-bit Flash cells.
To ensure functionality, the resistance and critical writing current of each MTJ need to be sufficiently different. Reading and writing to a cell would require multiple cycles, one for each bit.

4.2 Subarraying

4.2.1 1T-1MTJ

As stated before, subarraying is necessary for larger memories. Single, large memory arrays are slow and require additional buffering to drive very long wires. Breaking it up into several smaller subarrays allows the memory to operate faster and share peripheral circuitry. For 1T-1MTJ cell architectures, the number of cells per bitline is limited by the capacitance of the access transistor and the MTJ itself. Generally, a single bitline can support no more than 256 cells.

4.2.2 Shared Architectures

As mentioned earlier, there are several shortcomings to a “shared” MTJ architecture. During the write operation, there are multiple parasitic current paths that siphon current from the device being written to, forcing the access device to be sized up. These parasitic currents also have the potential to flip cells not being
Figure 4.4: Shared architecture with $M$ MTJs per transistor and $N$ wordlines per subarray.
accessed. When reading, these parasitic paths lower the effective TMR that can be observed.

For $M$ MTJs per access device and $N$ wordlines per subarray, the effective TMR ($TMR_{eff}$) can be expressed as:

$$TMR_{eff} = \frac{2(N + M - 1) - NM}{NM + (N - 1)(M - 1) \cdot TMR} \cdot TMR.$$  \hspace{1cm} (4.1)

The worst case TMR degradation can be easily derived by finding the largest possible $R_P$ and the smallest possible $R_{AP}$. If $R_\parallel$ is the parasitic parallel resistance shown in Fig. 4.4, then $R_{P,MAX}$ can be calculated as $R_P \parallel R_\parallel$, where:

$$R_\parallel = \frac{N + M - 1}{(N - 1)(M - 1)} \cdot (1 + TMR) \cdot R_P$$ \hspace{1cm} (4.2)

for the case of all parasitic resistances in the antiparallel state. Similarly, $R_{AP,MIN}$ can be calculated as $R_P \cdot (1 + TMR) \parallel R_\parallel$, where:

$$R_\parallel = \frac{N + M - 1}{(N - 1)(M - 1)} \cdot R_P$$ \hspace{1cm} (4.3)

for the case of all parasitic resistances in the parallel state. For $M > 2$, $TMR_{eff}$ is negative, limiting sharing to only two MTJs per access device.

In order to quantify the impact of sharing on the writing operation, the maximum allowable disturbance current a parasitic device can handle before a significant probability of switching occurs must first be calculated. For $M = 2$, since the ability to read is already limited, the four worst corner cases are shown in Fig. 4.5. For the corner case of Fig. 4.5(a), we require:

$$R_P \cdot I_{WRITE}(P \rightarrow AP) - R_P \cdot I_{READ}(P \rightarrow AP) \leq 2R_P \cdot I_{READ}(P \rightarrow AP).$$  \hspace{1cm} (4.4)

Solving:

$$\frac{I_{READ}(P \rightarrow AP)}{I_{WRITE}(P \rightarrow AP)} \geq \frac{1}{3}.$$ \hspace{1cm} (4.5)
Similarly, for the other three cases, $I_{\text{READ}}/I_{\text{WRITE}} \geq 1/3$. This means that sharing can only be successfully implemented if the MTJ can tolerate a reading current greater than one third of the writing current without a significant probability of flipping.
CHAPTER 5

Design-Space Analysis

In order to design an STT-MRAM with adequate design margin for high yield, one must consider all the implications of MTJ/CMOS integration. The design depends considerably on the underlying transistor technology, since a given CMOS technology constrains the design space due to the overhead and impact of the access transistor in each memory cell. The feasibility and yield of the memory are also heavily dependent upon the variation of the MTJs [63]. Previous work, such as Raychowdhury et al. [64], has failed to address these dependencies and provide the necessary framework for large-scale design. This chapter introduces the concept of the memory cell design space and a sensitivity analysis to optimize yield, power, and density for an STT-MRAM.

5.1 Defining the Design Space

The analysis in this chapter is done for a conventional 1T-1MTJ cell architecture as shown in Fig. 5.1, but can be generalized to any cell architecture. The writing currents for flipping the cell resistance are defined as $I_C(P \rightarrow AP)$ and $I_C(AP \rightarrow P)$. The design space of a single STT-MRAM memory cell can be illustrated using an $R_{AP}$ vs. $R_P$ plot as is shown in Fig. 5.2. The feasibility region is indicated by the shaded region. It contains all points ($R_P$, $R_{AP}$) in the design space so that a memory cell made with such an MTJ is functional. In the
design space, the two lower bounds are set by the read margin of the cell, while the two upper bounds are set by the write margin of the cell.

Lower bound $R_{P, MIN}$ is dependent on the implementation of the sense amplifier, and represents the minimum resistance required for reliable circuit operation. Additionally, $R_{AP, MIN}$ is determined by $TMR_{MIN}$ (Fig. 5.3), the minimum TMR required for the read amplifier to differentiate between $R_P$ and $R_{AP}$. Regardless of the specifics of the implementation, all sense amplifiers can be classified as either a voltage- or current-sensing topology. For a generic current-sensing read circuit, $TMR_{MIN}$ can be expressed as:

$$TMR_{MIN} = \frac{2\Delta I_{ref}/I_{ref}}{1 - \Delta I_{ref}/I_{ref}}.$$  \hspace{1cm} (5.1)

For $I_{ref}$ flowing through the reference resistance $R_{ref}$, $I_{ref} + \Delta I_{ref,1}$ flows through $R_P$ and $I_{ref} - \Delta I_{ref,2}$ through $R_{AP}$. When $\Delta I_{ref,1} = \Delta I_{ref,2} = \Delta I_{ref}$, $TMR_{MIN}$ is minimized. Under this condition, $R_{ref} = 2(R_P \parallel R_{AP})$ and we can express
Figure 5.2: Design space, in a 65nm process, for $W_N = 2.0 \mu m$, $I_C(P \rightarrow AP) = 500 \mu A$, $I_C(AP \rightarrow P) = 375 \mu A$, with an overlay of device X from Table 3.1.

$TMR_{MIN}$ as a function of the normalized fractional sensing current ($\Delta I_{ref}/I_{ref}$). In Eq. 5.1, $\Delta I_{ref}$ must be chosen so that the read amplifier correctly evaluates across all transistor PVT variations. Similarly, $TMR_{MIN}$ can be derived for voltage-sensing topologies. It should be noted that the lower bounds $R_{P,MIN}$ and $R_{AP,MIN}$, while critical to the readability of the cell, are almost completely independent of the MTJs used. The only requirement is that sensing time and current ($I_{ref}$) be small enough so as not to disturb the cell during the read
The upper bounds, $R_{P,MAX}$ and $R_{AP,MAX}$, are the maximum allowable resistances such that the access transistor, in a 1T-1MTJ configuration, is still able to provide the minimum critical writing currents $I_C(P \rightarrow AP)$ and $I_C(AP \rightarrow P)$. These upper bounds are consequently very sensitive to the specific characteristics of the MTJ device used. Transistor-level simulations are used to determine the relationship between $R_{MAX}$, $I_C$, and cell size (transistor width $W_N$) for a technology. Fig. 5.4 shows an example of such a simulation in a 65nm process. Using the conventional configuration from Fig. 5.1, $W_N$ is swept along with $R_{MAX}$. The contours of the simulated current are shown.

Fig. 5.2 shows a specific MTJ cell and its associated statistical variation (the concentric ovals around point $B$) overlaid on the design space. The design-space
margin (DSM) can be defined as the number of $\sigma$’s of MTJ variation before crossing any of the previously defined bounds. Defining DSM in terms of $\sigma$ simplifies feasibility characterization to a single variable and thus allows yield to be quickly calculated. To a first order, $3\sigma$, $4\sigma$, $5\sigma$, and $6\sigma$ of design margin roughly correspond to being able to reliably produce 1kbit, 32kbit, 4Mbit, and 1Gbit memory arrays.

5.2 Sensitivity Analysis and Design Example

Many variables, at both the circuit and device levels, affect the design space. In order to optimize a variable for a target memory specification, it must be determined how such a variable impacts the design space. This section introduces a design-space sensitivity (DSS) as a metric to quantify the behavior of the change
in design space as a function of various design parameters ($V_{DD}$, $\lambda$, $J_C$, RA, TMR, $W_N$, etc.).

5.2.1 Design-Space Sensitivity Analysis

First consider the points $A$, $B$, and $C$ in Fig. 5.2. Points $A$ and $C$ correspond to the corner values of $R_P$ and $R_{AP}$ in the feasible design space. Point $B$ represents the nominal MTJ at the center of the MTJ device distribution. For a positive design margin to exist, point $B$ must fall somewhere between points $A$ and $C$.

A “better” design space can be achieved from altering a design parameter, if a larger distribution of the MTJs (the number of $\sigma$) falls within the feasible region. Note that the improved design space is not simply increasing the area of the feasibility region, since the motion of point $B$ must be considered as well. Recall that point $A$ depends only slightly on the MTJ parameters. Therefore, the improvement (or deterioration) of the design space depends mostly on the change in DSM between points $B$ and $C$ as a function of a particular design variable.

Therefore, the design-space sensitivity to the parameter $X$ is defined as:

$$DSS(X) = \frac{\partial(\frac{R_C-R_B}{\sigma})_{P/AP}}{\partial X},$$

(5.2)

where $R_B$ and $R_C$ are taken as either $R_P$ or $R_{AP}$ at points $B$ and $C$, thus defining the DSS along each dimension of the design space. $\frac{R_C-R_B}{\sigma}$ is the normalized distance between points $B$ and $C$ in the design space along the $R_{P/AP}$ dimension. Intuitively, the $DSS(X)$ describes the instantaneous rate of change in DSM to a particular design parameter $X$. The derivative loses positional information, and so the DSS is used in conjunction with the original plot of the design space to determine the benefit of tuning the design parameter $X$. For both the $R_P$ and
$R_{AP}$ dimensions, if $DSS(X) > 0$, then the DSM is improved by increasing $X$, and if $DSS(X) < 0$, then DSM is improved by decreasing $X$. When the design-space sensitivities for the two dimensions conflict, the size of the design space in each dimension should then be taken into account.

5.2.2 Design Example

In this section, the sensitivity analysis was used to design a 4Mbit STT-MRAM with a $30F^2$ cell size (comparable to eDRAM) in a 65nm technology. Device X from Table 3.1, with $I_C(P \rightarrow AP) = 450\mu A$ and $I_C(AP \rightarrow P) = 300\mu A$, is the nominal MTJ and can be scaled by $\lambda$. Also, approximately $5\sigma$ of design margin is required for reasonable yield.

Fig. 5.5(a) shows the design space for a nominal $V_{DD} = 1.0V$ and $\lambda = 1.0$. The inner red oval is the $3\sigma$ variation of the MTJ, while the dashed, black oval represents the $5\sigma$ variation of the MTJ. Clearly, with nominal $V_{DD}$ and $\lambda$, the memory is not functional. Fig. 5.6 shows that the design space is much more sensitive to $V_{DD}$ than it is to $\lambda$. Therefore, we choose to scale $V_{DD}$ to 1.4V. Fig. 5.5(b) shows the new design space, with the $3\sigma$ bound at the edge of the design boundary.

Scaling $V_{DD}$ alone proves insufficient to meet the $5\sigma$ design margin required, and so $\lambda$ was simultaneously scaled. Fig. 5.6(b) shows that scaling $\lambda$ results in conflicting DSS. The $R_{AP}$ margin improves more by scaling $\lambda$ up, while the $R_P$ margin improves by scaling $\lambda$ down. However, Fig. 5.5(b) indicates that $R_{AP}$ has considerable margin, and we can trade off some of that margin for improved margin in $R_P$. Therefore, we choose to scale $\lambda$ down to 0.7. As we can see in Fig. 5.5(c), the desired $5\sigma$ bound on MTJ variation is essentially enclosed within the design space.
Figure 5.5: Design space, in a 65nm process, for a $30F^2$ cell ($W_N = 0.65\mu m$) for device X from Table 3.1: $I_C(P \rightarrow AP) = 450\mu A$, $I_C(AP \rightarrow P) = 300\mu A$. Inner red oval represents $3\sigma$ of MTJ device variation. Dashed, black oval corresponds to $5\sigma$ of MTJ variation.
(a) Sensitivity to $V_{DD}$ vs. $V_{DD}$.
(b) Sensitivity to $\lambda$ vs. $\lambda$.

Figure 5.6: Design-space sensitivity of parameters (a) $V_{DD}$ and (b) $\lambda$ in a 65nm technology.

5.3 Future Scalability

Scalability is an important feature for the success of a memory technology. Fig. 5.7 shows how scaling of the transistor technology impacts the design margin when using an MTJ that has a current density ($J_C$) of $6 \times 10^6 \text{ A/cm}^2$ for 10ns P to AP switching and RA of $5\Omega \cdot \mu\text{m}^2$. The figure shows that SRAM equivalent sizes ($120F^2$) scale well, with a design margin more than sufficient to construct gigabit memories ($> 6\sigma$). However, as the cell size is decreased, the design margin begins to degrade below 45nm. The design margin practically disappears once an eDRAM-equivalent cell size ($30F^2$) is reached. However, if the MTJ current density scales with technology, also shown in Fig. 5.7, then this trend is reversed. By scaling $J_C$ by as little as 4.5%, a constant design margin can be achieved between each technology node below 45nm.
Figure 5.7: Design margin vs. technology node. For constant $J_c(P \rightarrow AP) = 6 \times 10^6$ A/cm$^2$ and $J_c(P \rightarrow AP)$ scaling by 4.5% each technology generation.

Tables 5.1, 5.2, and 5.3 contain critical switching current densities for Flash-, eDRAM-, and SRAM-equivalent cell sizes for RAs of 5, 10, and 15Ω·µm$^2$, respectively. All table values correspond to 5σ of design margin and sub-10ns switching times. It should be noted that while larger RAs require smaller current densities (in order to meet voltage headroom constraints), they scale much better between successive technology nodes. Also, Flash-like cell sizes (6$F^2$) require current densities below $3 \times 10^6$ A/cm$^2$ in current technologies and less than $2 \times 10^6$ A/cm$^2$ in upcoming 22nm and 16nm technology nodes. Aggressive scaling of MTJ currents will be required to achieve Flash-like densities in future technologies.
### Table 5.1: $J_C(P \rightarrow AP)$ for an RA of $5 \ \Omega \cdot \mu m^2$

<table>
<thead>
<tr>
<th>Equivalent</th>
<th>$J_C$ vs. Technology Node ($10^6 \ A/cm^2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell Size</td>
<td>90nm</td>
</tr>
<tr>
<td>FLASH ($6F^2$)</td>
<td>2.72</td>
</tr>
<tr>
<td>eDRAM ($30F^2$)</td>
<td>4.60</td>
</tr>
<tr>
<td>SRAM ($120F^2$)</td>
<td>6.67</td>
</tr>
</tbody>
</table>

‡Predicted

### Table 5.2: $J_C(P \rightarrow AP)$ for an RA of $10 \ \Omega \cdot \mu m^2$

<table>
<thead>
<tr>
<th>Equivalent</th>
<th>$J_C$ vs. Technology Node ($10^6 \ A/cm^2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell Size</td>
<td>90nm</td>
</tr>
<tr>
<td>FLASH ($6F^2$)</td>
<td>2.22</td>
</tr>
<tr>
<td>eDRAM ($30F^2$)</td>
<td>3.27</td>
</tr>
<tr>
<td>SRAM ($120F^2$)</td>
<td>4.27</td>
</tr>
</tbody>
</table>

‡Predicted

### Table 5.3: $J_C(P \rightarrow AP)$ for an RA of $15 \ \Omega \cdot \mu m^2$

<table>
<thead>
<tr>
<th>Equivalent</th>
<th>$J_C$ vs. Technology Node ($10^6 \ A/cm^2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell Size</td>
<td>90nm</td>
</tr>
<tr>
<td>FLASH ($6F^2$)</td>
<td>1.83</td>
</tr>
<tr>
<td>eDRAM ($30F^2$)</td>
<td>2.56</td>
</tr>
<tr>
<td>SRAM ($120F^2$)</td>
<td>3.17</td>
</tr>
</tbody>
</table>

‡Predicted
CHAPTER 6

Memory Design

In this chapter, the design flow of three test chips implemented in 90nm, 65nm, and 45nm processes is described. Several architectures from Chapter 4 were selected for testing. Each design was subjected to the analysis outlined in Chapter 5 in order to optimize read/write performance, memory density, and energy considerations. Each chip was designed to operate with MTJs fabricated by UCLA’s Western Institute of Nanoelectronics (WIN). MTJ device specifications are detailed in [65], [66] and [67]. However, manufacturing requirements and restrictions made integration of MTJs with CMOS not possible on the test chips. A brief explanation is provided with a discussion of the process flow for MTJ/CMOS integration before the chip design is described.

6.1 MTJ/CMOS Integration

As mentioned before, MTJs are well suited for integration into a commercial CMOS process flow. In this flow, the deposition of the insulating oxide barrier is critical to the performance of the MTJ. If the layer is too thin (< 0.7nm) the MTJ does not exhibit any TMR, due to the formation of pin holes and soft points shorting the barrier. If the layer is too thick (> 2.5nm), then the resistance of the device is too large [68]. The deposition surface also needs to be very smooth, whereas typical Al interconnects (with a ⟨111⟩ texture) are far too
rough. However, the Cu interconnects available in the thin metal layers of modern state-of-the-art fabrication process are ideal for MTJ deposition \cite{69}. MTJs are typically integrated after the thin Cu layer, usually M4 in most processes. Fig. 6.1 shows the side view of a typical 1T-1MTJ with full integration at M4. MTJ pillar dimensions down to 30nm can be accomplished with e-beam lithography, focused ion beam etching, or double patterning \cite{70}.

Unfortunately, it was not possible to integrate the MTJs from \cite{65}, \cite{66} and \cite{67} into the test chips due to fabrication restrictions. There are 45nm wafers ready for MTJ deposition once the fabrication flow is finalized. Meanwhile, the characterizations of test chips are done through simulation or with the addition of dummy resistor arrays on CMOS.
6.2 Test Chips

6.2.1 90nm Bulk CMOS

Fig. 6.2 shows a block diagram of the layout of our 90nm bulk CMOS test chip. The design work was performed for a conventional cell architecture and was intended to test the integration process flow. With a total of 6kbit, the memory had two 1kbit memory arrays, using RVT transistors with a cell size of $55F^2$, and two 2kbit memory arrays, using LVT transistors with a cell size of $30F^2$. For purposes of comparison, an SRAM cell in this technology is approximately $75F^2$. Included in the design are two resistor arrays with values ranging from a few hundred ohms to several kilohms (RSEL<16:0> and RSEL<33:17> in Fig. 6.2), representing a range of TMR from 0% to 1000%. Simulated and measured results for the read performance of $R_P$ (logical 0) are shown in Table 6.1. Similar results for the read performance of $R_{AP}$ (logical 1) are shown in Table 6.2. The drive current of the $50F^2$ cell was also measured to be approximately $300\mu A$. Estimates show that this level of drive current should easily allow for thermally activated switching with write times on the order of 10 to 20 nanoseconds.

6.2.2 65nm Bulk CMOS

In the 65nm process, the memory array was increased to 16kbit and included three cell sizes: $28F^2$, $35F^2$, and $50F^2$. Again, the design work assumed a conventional cell architecture. A “short-pulse” reading scheme was also introduced with a bidirectional write driver (Fig. 6.3) to improve read/write performance. Short-pulse reading works by delivering a large, but very short, current pulse to the MTJ and latching in its value. A dual-wordline voltage boosting scheme (dual-boosting) was implemented that allowed the drive current to be increased
Figure 6.2: Block diagram of the 90nm test chip.
Table 6.1: Time to read $R_P$ (90nm)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>100</td>
<td>3.77</td>
<td>7.20</td>
<td>30</td>
</tr>
<tr>
<td>500</td>
<td>100</td>
<td>3.17</td>
<td>5.20</td>
<td>55</td>
</tr>
<tr>
<td>670</td>
<td>50</td>
<td>7.30</td>
<td>8.50</td>
<td>30</td>
</tr>
<tr>
<td>670</td>
<td>50</td>
<td>5.18</td>
<td>10.2</td>
<td>55</td>
</tr>
<tr>
<td>670</td>
<td>200</td>
<td>3.04</td>
<td>4.20</td>
<td>30</td>
</tr>
<tr>
<td>670</td>
<td>200</td>
<td>2.69</td>
<td>4.80</td>
<td>55</td>
</tr>
</tbody>
</table>

Table 6.2: Time to read $R_{AP}$ (90nm)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>100</td>
<td>2.00</td>
<td>2.10</td>
<td>30 &amp; 55</td>
</tr>
<tr>
<td>670</td>
<td>50</td>
<td>2.60</td>
<td>2.70</td>
<td>30 &amp; 55</td>
</tr>
<tr>
<td>670</td>
<td>200</td>
<td>2.37</td>
<td>2.70</td>
<td>30 &amp; 50</td>
</tr>
</tbody>
</table>

Figure 6.3: Read/Write driver for short-pulse reading.
Figure 6.4: Cadence layout of the 65nm test chip: (1) 8kbit memory array, (2) resistor array, (3) muxes, (4) read/write circuitry, (5) read delay measurement circuitry, (6) configuration scan chain, (7) pulse generator, (8) data scan chain, and (9) decoder.

by 70% in our smallest cell sizes and 130% in our largest cells. The resistor arrays from the 90nm design were kept for CMOS characterization. Fig. 6.4 shows a layout of the 65nm design.

6.2.3 45nm SOI CMOS

The 45nm design moved the bulk CMOS to an SOI process. The memory size was increased to 32kbit and the same memory architecture was kept from our 65nm design (dual-boosting with short-pulse reading). It was possible to decrease the cell sizes to $17F^2$, $25F^2$, and $40F^2$, while still maintaining the same
Figure 6.5: Cadence layout of the 45nm test chip: (1) 8kbit memory array, (2) resistor array, (3) muxes, (4) read/write circuitry, (5) configuration scan chain, (6) pulse generator and read delay measurement circuitry, (7) data scan chain, and (8) decoder.

drive current as the 65nm design. Overall, large improvements to the memory layout and organization were made. Fig. 6.5 shows a layout of the 45nm design.

6.2.4 Design Comparison

A comparison of the 90nm, 65nm and 45nm chips presented in this work is done with state-of-the-art STT-MRAMS from [61], [71], [72], and [73].
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Designer</td>
<td>This Work</td>
<td>This Work</td>
<td>This Work</td>
<td>Qualcomm</td>
<td>NEC</td>
<td>Fujitsu &amp; UT</td>
<td>Toshiba</td>
</tr>
<tr>
<td>Power Supply [V]</td>
<td>1.1/1.4</td>
<td>1.2/1.6</td>
<td>1.2</td>
<td>1.1/1.8</td>
<td>1/1.5</td>
<td>1.2/3.3</td>
<td>1.2</td>
</tr>
<tr>
<td>Architecture</td>
<td>Conventional (Dual-Boosted)</td>
<td>Conventional (Dual-Boosted)</td>
<td>Conventional</td>
<td>Reversed 2T-1MTJ (Boosted)</td>
<td>Conventional</td>
<td>2T-1MTJ</td>
<td></td>
</tr>
<tr>
<td>Memory Size</td>
<td>32kbit</td>
<td>16kbit</td>
<td>6kbit</td>
<td>32Mbit</td>
<td>32Mbit</td>
<td>16kbit</td>
<td>64Mbit</td>
</tr>
<tr>
<td>Process [nm]</td>
<td>45</td>
<td>65</td>
<td>90</td>
<td>45</td>
<td>90</td>
<td>130</td>
<td>65</td>
</tr>
<tr>
<td>Cell Size [F²]</td>
<td>17, 25, 40</td>
<td>28, 35, 50</td>
<td>30, 60</td>
<td>11</td>
<td>70</td>
<td>140</td>
<td>36</td>
</tr>
<tr>
<td>Read Time</td>
<td>1-3ns</td>
<td>3-5ns</td>
<td>3-10ns</td>
<td>&lt;100ns</td>
<td>60ns</td>
<td>8ns</td>
<td>11ns</td>
</tr>
<tr>
<td>Write Time</td>
<td>3-5ns</td>
<td>3-5ns</td>
<td>10-20ns</td>
<td>10ns-1ms</td>
<td>91ns</td>
<td>9-10ns</td>
<td>30ns</td>
</tr>
</tbody>
</table>

Figure 6.6: Design comparison of STT-MRAMs.
CHAPTER 7

Conclusion

The potential for STT-MRAM to become a true universal memory has been advanced by fabricating three test chips that demonstrate increased memory densities and performance when compared to published state-of-the-art STT-MRAMs. The 45nm test chip has a memory density 3.8 times greater on average when compared to published state-of-the-art STT-MRAMs. It has read access times ranging from 5 to 100 times faster, and is capable of delivering over 500μA of drive current to write in less than 5ns. While MTJ/CMOS integration was not accomplished, for reasons outside of our control, the chip represents the next step forward in making STT-MRAM a true universal memory. Analysis of the future scalability of MTJs shows that with a 22nm technology node, STT-MRAM will be able to achieve Flash-like densities, with sub-nanosecond performance, while using ten times less power.

7.1 Summary of Research Contributions

Specific accomplishments of this research are:

- Development of a physics-based MTJ macro-model capable of accurately modeling and predicting device behavior across temperature. The model was implemented in Verilog-A.
• Development of a design-space analysis toolset specifically for STT-MRAMs. These tools allowed an accurate and fair assessment of different memory architectures at the same time in order to maximize design margin, and minimize power and area for a fixed level of performance.

• Design of three STT-MRAM test chips:
  ◦ 90nm bulk CMOS: 6kbit memory array, 2.5x density improvement over SRAM, 2-10ns read, 200\(\mu\)A write current.
  ◦ 65nm bulk CMOS: 16kbit memory array, 4.3x density improvement over SRAM, new high-speed read/write architecture, 400\(\mu\)A write current.
  ◦ 45nm SOI CMOS: 32kbit memory array, 7.1x density improvement over SRAM, average 3.8x density improvement over state-of-the-art STT-MRAMs, 500\(\mu\)A write current.

7.2 Future Work

Additional steps required to demonstrate that STT-MRAM can be developed into a true universal memory are:

• Extend the compact MTJ model to include localized heating effects, the backhopping phenomenon, and noise models.

• Explore stacked memory architecture to improve cell density.

• Continue to refine the DSM and DSS of the design-space analysis tools for integration with circuit simulators.

• Finalize the fabrication flow for MTJ/CMOS integration, including running performance and yield characterizations on an integrated 45nm test chip.
REFERENCES


