Testing STT-MRAM: Manufacturing Defects, Fault Models, and Test Solutions

Wu, L.

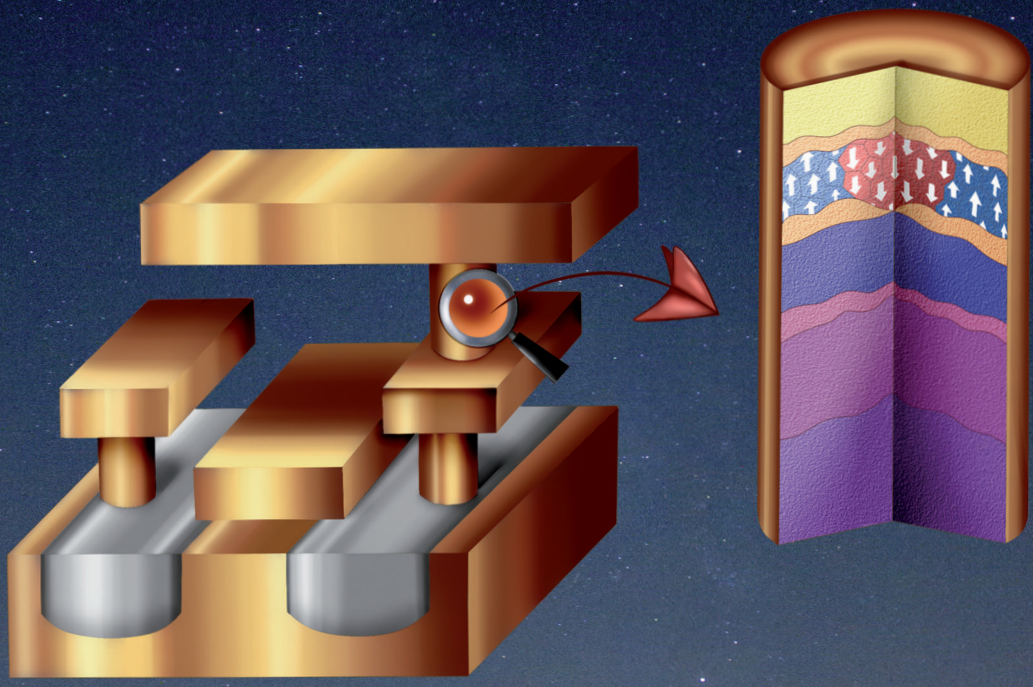
DOI
10.4233/uuid:088a3991-4ea9-48a0-9b92-cc763748868c

Publication date
2021

Document Version
Final published version

Citation (APA)
Wu, L. (2021). *Testing STT-MRAM: Manufacturing Defects, Fault Models, and Test Solutions*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:088a3991-4ea9-48a0-9b92-cc763748868c

Testing STT-MRAM: Manufacturing Defects, Fault Models, and Test Solutions



Lizhou Wu

TESTING STT-MRAM: MANUFACTURING DEFECTS, FAULT MODELS, AND TEST SOLUTIONS

TESTING STT-MRAM: MANUFACTURING DEFECTS, FAULT MODELS, AND TEST SOLUTIONS

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus prof. dr. ir. T.H.J.J. van der Hagen
chair of the Board for Doctorates
to be defended publicly on
Monday 22 February 2021 at 15:00 o'clock

by

Lizhou WU (吴利舟)

Master of Engineering in Computer Science & Technology
National University of Defense Technology, China
born in Quzhou, Zhejiang, China

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus	chairperson
Prof. dr. ir. S. Hamdioui	Delft University of Technology, promotor
Dr. ir. M. Taouil	Delft University of Technology, copromotor
Independent members:	
Prof. dr. K.A.A. Makinwa	Delft University of Technology
Prof. dr. M. Sachdev	University of Waterloo, Canada
Prof. dr. P. Girard	LIRMM Laboratory, France
Dr. S. Rao	IMEC, Belgium
Dr. B. Kruseman	NXP Semiconductors, the Netherlands
Prof. dr. ir. W.A. Serdijn	Delft University of Technology, reserve member

Keywords: memory test, device-aware test, manufacturing test, STT-MRAM, MTJ, manufacturing defect, fault model, robust design, magnetic coupling

Printed by: Ipskamp Printing, the Netherlands

Front & Back: designed by Yu Zhang & Lizhou Wu

Author email: njuwulizhou@163.com

ISBN 978-94-6384-199-3

An electronic version of this dissertation is available at
http://repository.tudelft.nl/.

Dedicated to:
my parents for raising me up and supporting me in education,
my wife for encouraging and accompanying me in this journey.

SUMMARY

As one of the most promising emerging memory technologies, spin-transfer torque magnetic random access memory (STT-MRAM) offers non-volatility, fast access speed, high density, nearly unlimited endurance, radiation immunity, and low-power consumption. Thanks to these advantageous features, STT-MRAM is customizable as both embedded and discrete memory solutions for a variety of applications such as enterprise SSD, AIoT, automotive, and aerospace. Therefore, numerous start-ups (e.g., Everspin) have been founded focusing on STT-MRAM commercialization, and major foundries worldwide (e.g., TSMC, Samsung, and Intel) also invest heavily on it. As STT-MRAM mass production and deployment in industry is around the corner, high-quality yet cost-efficient manufacturing test solutions are needed to ensure the required quality of products being shipped to end customers.

This dissertation mainly focuses on robust design and high-quality test for STT-MRAM. We first investigate the manufacturing process of STT-MRAM and physical defects that may take place in each step based on literature survey and silicon measurements. Special attentions are given to those unique steps and defects related to the fabrication of magnetic tunnel junction (MTJ) devices, which are the data-storing elements in STT-MRAMs. We build a complete STT-MRAM simulation platform, composed of a Python simulation controller and an STT-MRAM circuit design. The former controls and automates all simulation procedures, whereas the latter is a circuit netlist consisting of a 1T-1MTJ memory array and peripheral circuits such as write drivers and sense amplifiers. To enable fast and accurate electrical/magnetic co-simulations of STT-MRAM, we propose a magnetic-field-aware compact model for MTJs with perpendicular magnetic anisotropy; it is optimized and calibrated with comprehensive measurement data of MTJ devices fabricated at imec. This model can be used for robust device/circuit co-design of STT-MRAM under PVT variations and various magnetic configurations including external disturbance fields and internal magnetic coupling effects.

Based on this simulation platform, we explore STT-MRAM testing with the conventional fault modeling and test approach. In this approach, any physical defect irrespective of its physical nature is modeled as a linear resistor (i.e., open, short, or bridge), which is then injected into our STT-MRAM netlist for fault analysis. Test development is also covered based on the fault modeling results. Although it is convincing to model defects in interconnects as linear resistors, this approach has never been validated for defects inside semiconductor devices such as MTJ. Based on comprehensive characterization on fabricated MTJ devices, we demonstrate that modeling an MTJ-internal defect as a linear resistor is inaccurate. This is because linear resistors cannot reflect the defect-induced changes in MTJ's magnetic properties which are as important as electrical ones. Furthermore, we experimentally observed extremely low, intermediate, and extremely high resistances in some defective MTJs; these resistance values are out of the specification of logic '0' and '1'. We also observed that some MTJ faulty behaviors are intermittent

rather than permanent. Hence, the conventional fault modeling and test approach is unable to derive high-quality test solutions for STT-MRAMs.

To address these issues, we propose Device-Aware Test (DAT) approach which goes beyond cell-aware test and specifically targets device-internal defects. DAT consists of three steps: 1) device-aware defect modeling, 2) device-aware fault modeling, and 3) device-aware test development. In the first step, a physical defect is characterized and modeled physically; the impact of the defect on the technology parameters of the defective device is determined. Subsequently, such impact is incorporated into the device's electrical parameters to obtain a parameterized defective device model which can be calibrated by silicon data if available. In the second step, we define a complete fault space using an upgraded fault primitive notation to cover all possible resistive states in STT-MRAMs; a systematic fault analysis is then performed to validate realistic faults within the pre-defined fault space in the presence of the defect. Finally, the obtained faults are used to develop appropriate test solutions; e.g., March tests, DfT designs, and stress tests.

We have applied the proposed DAT to three key types of MTJ-internal defects as case-studies in this thesis. They are pinhole defects, synthetic anti-ferromagnet flip (SAFF) defects, and intermediate (IM) state defects. For each type of MTJ defects, we perform comprehensive characterization on fabricated MTJ devices, and develop a defective MTJ compact model with defect parameters as inputs; the model is also calibrated with the measured silicon data. By applying device-aware fault modeling, accurate and realistic faults are obtained. Comparing the results to those obtained using the conventional approach reveals two observations: 1) The conventional approach leads to wrong fault models which in turn would lead to test escapes and a waste of test time and resources; 2) Our DAT approach results in more accurate fault models which reflect the physical defects, thus ensuring high-quality tests at minimal cost. With the obtained faults using our DAT approach, we propose optimized test solutions for the above-mentioned three types of MTJ-internal defects.

SAMENVATTING

Als een van de meest veelbelovende opkomende geheugentechnologieën biedt spin-transfer torque magnetic random access memory (STT-MRAM) niet-vluchtigheid, hoge lees- en schrijfsnelheid, hoge dichtheid, nagenoeg onbeperkt schrijfuithoudingsvermogen, een hoge robuustheid tegen straling en een laag energieverbruik. Deze eigenschappen zorgen dat STT-MRAM zowel kan worden toegepast als ingebedde en als discrete geheugenoplossing voor een verscheidenheid aan applicaties, zoals enterprise SSD, kunstmatige intelligentie voor het internet der dingen en in de auto-, lucht- en ruimtevaartindustrie. Daarom zijn er talloze startups (bijv. Everspin) opgericht die zich richten op de commercialisering van STT-MRAM, maar ook grote chipproducenten over de gehele wereld (bijv. TSMC, Samsung en Intel) doen grote investeringen in deze technologie. Aangezien de massaproductie en -implementatie van STT-MRAM nabij is, zijn hoogwaardige maar toch kostenefficiënte productietestoplossingen cruciaal om de vereiste kwaliteit van de producten die naar de eindklanten worden verzonden te garanderen.

Dit proefschrift richt zich voornamelijk op het robuust ontwerpen en het ontwikkelen van hoogwaardige tests voor STT-MRAM. We onderzoeken eerst het fabricageproces van STT-MRAM en de fysieke defecten die in elke stap kunnen optreden op basis van literatuuronderzoek en siliciummetingen. Speciale aandacht wordt geschonken aan de unieke stappen en defecten die verband houden met de fabricage van de magnetische tunneljunctie (MTJ), welke het gegevensopslagelement is in STT-MRAM's. We ontwikkelen een compleet STT-MRAM-simulatieplatform, bestaande uit een simulatiecontroller geschreven in Python en een STT-MRAM-circuitontwerp. De eerstgenoemde controleert en automatiseert alle simulatieprocedures, terwijl de laatstgenoemde een circuitbeschrijving is van een 1T-1MTJ-geheugenraster en randcircuits zoals schrijfcircuits en leesversterkers. Om snelle en nauwkeurige elektrische/magnetische co-simulaties van STT-MRAM mogelijk te maken, introduceren we een compact model voor MTJ's met loodrechte magnetische anisotropie, dat rekening houdt met het magnetisch veld. Het model is geoptimaliseerd en gekalibreerd met uitgebreide meetgegevens van MTJ's die bij imec zijn vervaardigd. Dit model kan worden gebruikt om robuuste MTJ's en STT-MRAM-circuits te ontwerpen met inachtneming van proces-, spannings- en temperatuurvariaties alsook verschillende magnetische configuraties, waaronder externe storingsvelden en interne magnetische koppelingseffecten.

Op basis van dit simulatieplatform bestuderen we het testen van STT-MRAM middels de conventionele foutmodellerings en testbenadering. In deze benadering wordt elk fysiek defect, ongeacht zijn fysieke aard, gemodelleerd als een lineaire weerstand (d.w.z. als een open verbinding, een kortsluiting of een overbruggingsverbinding), die vervolgens wordt toegevoegd aan ons STT-MRAM-circuitontwerp voor foutanalyse. Testontwikkeling wordt ook behandeld op basis van de foutmodelleringsresultaten. Hoewel het overtuigend is om defecten in verbindingen te modelleren als lineaire weerstanden, is deze benadering nooit gevalideerd voor defecten in halfgeleidercomponenten zelf, zoals

in MTJ's. We laten op basis van uitgebreide karakterisering van gefabriceerde MTJ's zien dat het modelleren van een intern MTJ-defect als een lineaire weerstand onnauwkeurig is. Dit komt doordat lineaire weerstanden de door defecten veroorzaakte veranderingen in de magnetische eigenschappen van MTJ, die even belangrijk zijn als elektrische, niet kunnen weerspiegelen. Daarnaast hebben we extreem lage en extreem hoge weerstanden alsook weerstanden die tussen de twee gewenste weerstandswaarden liggen experimenteel waargenomen in sommige defecte MTJ's. Deze weerstandswaarden vallen buiten de specificaties van een logische '0' en '1'. We hebben ook vastgesteld dat het foutieve gedrag van sommige defecte MTJ's niet permanent is maar met tussenpozen optreedt. Om deze redenen is het onmogelijk om met de conventionele foutmodelleringsen testbenadering hoogwaardige testoplossingen voor STT-MRAM te genereren.

Om deze problemen aan te pakken, stellen we een Componentbewuste Test (CBT) -benadering voor die verder gaat dan celbewuste tests en specifiek gericht is op interne defecten van de component. De CBT-benadering bestaat uit drie stappen: 1) componentbewuste defectmodellering, 2) componentbewuste foutmodellering en 3) componentbewuste testontwikkeling. In de eerste stap wordt een productiedefect gekarakteriseerd en de consequenties ervan gemodelleerd. Hiermee wordt de uitwerking van het defect op de technologieparameters van de defecte component bepaald. Vervolgens wordt de impact opgenomen in de elektrische parameters van de component om een geparametriseerd model van de defecte component te verkrijgen dat, indien beschikbaar, kan worden gekalibreerd met siliciummetingen. In de tweede stap definiëren we een volledige foutruimte met behulp van een verbeterde notatie van foutprimitieven om alle mogelijke weerstandstoestanden in STT-MRAM's te beschrijven. Vervolgens wordt een systematische foutanalyse uitgevoerd om de foutruimte te valideren in de aanwezigheid van een defect en dus realistische fouten te determineren. Ten slotte worden de gevalideerde fouten gebruikt om geschikte testoplossingen te ontwikkelen. Dit kunnen bijvoorbeeld marcheertests, ontwerp-voor-test-structuren en stresstests zijn.

In dit proefschrift passen wij als casestudy de CBT-benadering toe op drie belangrijke MTJ-defecten. Deze defecten zijn: een minuscule gaten in de MgO-tunnelbarrière, synthetische anti-ferromagnetische omkeringsdefecten en tussenliggende-toestanddefecten. Voor elk MTJ-defect voeren we een uitgebreide karakterisering uit op gefabriceerde MTJ's en ontwikkelen we een compact MTJ-defectmodel met defectparameters als invoer; het model is tevens gekalibreerd aan de hand van de uitgevoerde metingen. Door componentbewuste foutmodellering toe te passen, worden nauwkeurige en realistische fouten gevonden. Wanneer de resultaten van deze aanpak vergeleken worden met die verkregen middels de conventionele benadering, kunnen de volgende twee waarnemingen gemaakt worden. (1) De conventionele aanpak leidt tot verkeerde foutmodellen die op hun beurt zouden leiden tot valsnegatieve testresultaten en verspilling van testtijd en -middelen. (2) Onze CBT-benadering resulteert in nauwkeurigere foutmodellen die de werkelijke productiedefecten beschrijven, waardoor hoge testkwaliteit tegen minimale kosten kan worden gegarandeerd. We leggen voor elk van de drie bovengenoemde interne MTJ-defecten geoptimaliseerde testoplossingen voor die gebaseerd zijn op de fouten verkregen middels onze CBT-benadering.

ACKNOWLEDGEMENTS

The year 2020 represents crisis and hardship to most people, as the novel coronavirus hits every corner in the world. To me, this year has a special meaning apart from COVID-19. After doing research for four years in pursuit of the PhD degree, I am so delighted that I will be soon hitting the finish line. Four years' investment with hard working day and nigh, eventually results in a small book advancing the state of the art a little bit in the knowledge of human being. Pretty cool and worth it! During this tough "marathon", there are obviously a lot of ups and downs. At this moment, when looking back, I feel so grateful to all who appear in my life, especially those who helped me in accomplishing this dissertation which I consider as the best achievement in my life so far.

First of all, I would like to express my deepest gratitude to my supervisory team at TU Delft: Prof.dr.ir. **Said Hamdioui** and Dr.ir. **Mottaqiallah Taouil**. A big thanks to my promotor and daily supervisor Said. Thank you for getting me on board and providing me with an unbelievable platform to carry out research work. Thank you for giving me hard time in writing papers and in preparing presentations. In my first year, you taught me how to do research by quoting an old Chinese proverb: teaching someone how to fish is better than just giving him a fish (授人以鱼不如授人以渔). You taught me how to think critically and independently. This covers the entire cycle in research; it applies when we review someone else's work, when we look for research issues, when we discuss my progress, when we revise papers together, when we polish slides for conference presentations. No matter how busy you are, you always manage to join my progress meetings and correct my paper manuscripts. In addition, You emphasize the importance of collaboration in research, which I benefit from significantly especially when collaborating with imec. Looking back, there are so many vivid moments popping up in my mind. When you were correcting my first ITC paper at TU Delft in 2017, you praised me for doing a good job in writing the introduction section. Frankly speaking, that was such a great encouragement to me after being overwhelmed by depression in my first year. But of course, I was happy to hear that and did not tell you the paper had been corrected by Motta many times already. Earlier this year, we were preparing another paper for ITC submission. On a beautiful Sunday, April 12, you sent me an email, saying that "I have to say that you have really learned how to write! I am very impressed with the quality which shows how much you improved!" Being your student, I am used to your critical style of education, and this came so surprisingly. Thank God, this made me happy for the entire week. After four years' training process with your guidance, I am proud to say that I am an independent "fisherman" now. I would also like to thank Motta who is my co-promotor and daily supervisor as well. Thank you for brainstorming with me and helping me improve my writing skills. Honestly, the first year of my PhD is definitely not a sweet memory. I will not forget the comfort and encouragement you gave me when I was in frustration. Also, I will take the original manuscripts of my first ITC paper with full of your corrections and comments back to China and keep them for the rest of my

life. In recent years, you have your hardware security group and are getting busier and busier, but you still manage to be an active member in our STT-MRAM testing team. I am very lucky to have a supervisor like you who is also my friend and mentor. When I have problems in work or personal issues, you are always there for help. You are a big bro who cares about students' feelings and always clears up a messy situation for us. You also serve as a bridge between students and Said. Said is the bitter flavor in life, and you are the sweet one. Don't say that you are too nice to students, we appreciate your nice trait which is so precious. Thank you, Motta, for adding sugar in my PhD life.

I also want to thank my supervisors: Dr. **Siddharth Rao** and Dr. **Erik Jan Marinissen** from imec, Belgium. Special thanks to Dr. **Gouri Sankar Kar** for continuously supporting this collaboration project. Sid, thank you for being my daily supervisor at imec. The internship at imec opens a new world for me. I can clearly remember how you trained me to use the RRAM characterization tool to characterize MTJ devices. A lot of manual setups before the start of an auto. measurement. I messed up a couple of times. But you never got upset and were always patient to me. Thank you. Later on, we luckily moved to the Hprobe tool for our measurements, which was so much easier to use. I have also learned a lot from you in MTJ physics, device characterization and modeling, data analysis etc. You are dubbed as a magneticist in our team. By the way, you are also the guy who led my way to Python. Before going to imec, I was a MATLAB guy. Because of your persuasion, I changed to Python, which I found so much better and powerful that I had been using it until now. Putting aside work, you are also a very good friend of mine. Without you, I would not have adapted to the working environment at imec very quickly. **Angelo** is a nice, smart Italian guy. We had a lot of fun together including lunch breaks and several times of BBQ with his Italian friends. **Giuseppe** cooked tasty pasta with all ingredients shipped from Italy by his lovely Mom. I would also like to thank everyone in the MRAM device team: **Kevin**, **Woojin**, **Jackson**, and **Simon**, for their help and forming a friendly and relaxing atmosphere. I cherish the memory of beer gatherings at Café Belge. Belgium beers are amazing, although I can only name a few brands. I would like to thank the other supervisor: Erik Jan. I feel very honored to work with you, frankly. The first time we met is still fresh and clear in my mind as if it happened yesterday. On June 23, 2017, I delivered a presentation to you about STT-MRAM technology, design, and test. I was a bit nervous despite the fact that I had prepared this presentation for a week (Said told me there was a big guy visiting us from imec). In the end, this meeting went well and we had some Q&As during the presentation in a relaxing vibe. Since then, you became an indispensable member in our STT-MRAM testing team. There are countless moments of you in my mind that I would never forget. For example, When I moved to Leuven, you invited me along with some other students to your home for dinner. That was so sweet and unforgettable. When I gave my first ITC talk at ITC'18 in Phoenix, Prof. Mehdi asked me a tricky question, which I did not response to well. But luckily, you defended me in a very concise and elegant way. In the summer of 2019, when most people went for vacation, we had a lot of discussions on the magnetic coupling work via emails and confcalls. This work ended up winning the best paper award at DATE'20. I will also remember all the long emails we exchanged for discussing our work and papers. Writing these emails were never easy, neither reading them. Thanks for investing so much time on me. All in all, many thanks to both of you: Sid and Erik Jan. The winter in Leuven is

as cold and windy as in Delft. But because of you two, I feel warm inside.

I would like to thank all my previous and current office mates: **Daniël**, **Innocent**, **Peyman**, **Moritz**, and **Guilherme**. Daniël, no matter how my office and colleagues change, you are always the one who sits beside me. When I need to read some Dutch documents or have troubles in my personal life, I always turn to you for help. Thanks for dealing with all this stuff for me over the past years. I will never forget the amazing conference trips we had in Phoenix and Baden-Baden. Innocent, thanks for nice conversations with you in the office. I also appreciate your invitations to join church prayers. Peyman, thanks for your lucky coin when I was in depression. Moritz, thanks a lot for translating both my thesis summary and propositions with many annoying updates. Don't forget to shout to the audience at least three times, device-aware test, when you give your conference presentations. Guilherme, thanks for running the simulations for our ETS paper. I am impressed by your enthusiasm in our office discussions, which are always very fruitful obviously. BTW, guys, our FP notation and naming scheme are the best innovation in the world. Sharing the same office with you gives me a lot of good memories: complaining about work and life, survival run crossing TU campus, chatting in a bar, etc. I wish all of you a bright future.

I would like to extend my thanks to all colleagues at the QCE department. Thanks to Prof. **Koen Bertels** for your efforts in creating a nice working environment and supporting all different kinds of social events such as football, Karting, bowling, barbecue, borrel, and Xmas party etc. **Lei**, **Jintao**, thank you for sharing your experience with me in living in the Netherlands and working with Said. Thanks to **Anh** and **Guilherme** for organizing QCE colloquia, which broaden my knowledge beyond my PhD topics. Many thanks to my Chinese fellow PhD friends: **Shanshan**, **Jian**, **Xiang**, **Lingling**, **Yande**, **Wanghe**, and **Baozhou**. Although we have different supervisors, we have a lot of things in common. With you guys, my four-year PhD life in Delft becomes colorful. **Mahdi**, **Muath**, **Abdulqader**, thank you for being my colleagues. I have learned many things from you especially about different cultures. We had an amazing week in Beijing during the 2019 Sino-Dutch Summer School, together with **Abid**. Thanks to **Cezar**, **Abdullah**, **Haji**, **Troya**, **Mark**, for having daily lunch and interesting conversations with jokes and laughs. These are the things that I miss so much in the days of working from home. I also want to acknowledge **Lidwina**, **Joyce**, **Laura**, **Trisha**, and **Paul** for taking care of management, paperwork, and other secretary-related tasks. Thanks to **Erik** for fixing computer problems and maintaining websites, servers, software etc.

I would like to mention our QCE indoor football game that takes place every week. I enjoyed it so much during my entire PhD. It not only serves as a social event for all colleagues in our department, but also provides a good chance to do some exercise to refresh myself. Thanks to all who have participated in this event. Special thanks to **Imran** for organizing it and sending us a reminder email every Wednesday. This role shifted to my dear colleague **Daniël** in the recent two years, thank you as well. I cannot stop laughing when recalling the moments of **Leon** shouting with a Megaphone "Guys, football, time for football" in the corridor. Said, Motta, Lei, Jintao, Daniël, Innocent, Luca, Peyman, Mohammod etc., it was so much fun playing football with you guys. I will keep my QCE football shirt as a memento forever.

I also want to express my gratitude to some of my Chinese PhD fellows at TU Delft,

friends, and professors. **Bowen**, it is my luck to be your friend. We shared some amazing moments living in DUWO studios at Roland Holstlaan in the first year. I have to say you are an excellent cook. **Shuaiqiang**, **Yande**, thanks for being my house mates at Arthur Schendelplein. **Xiaohui**, thank you for organizing some interesting activities. Special thanks to **Zhan**. We are both Erik Jan's students. This special network connects us and builds our friendship. I love talking to you about research and any other matters. Thanks to **Yachao** and **Luge**. We built our friendship many many years ago back in China. After moving to Europe to do our PhDs, we always keep in touch and share joys and sorrows in work and life. I enjoy all the trips with you in Europe. **Xindi**, thanks for handling all stuff for me in China. May our friendship last forever! Many thanks to Prof. **Liu Fang** and **Xiao Nong** for supporting me in doing a PhD overseas and many fruitful discussions related to my research.

Last but most importantly, I would like to express my deepest thanks to my family. My wife, **Zhang Yu**, is the most beautiful girl on earth (in my heart). Years ago in China, you did not even know how to cook. But you are apparently a Michelin 3-star chef now. The amazing food you cook gives me endless energy for work. Thank you for always staying with me no matter where I go, from China to Netherlands, then to Belgium, and back to Netherlands. The tulips at Keukenhof, the windmills at Kinderdijk, the world cup cheers at Leuven, and the church bells at Delft, all have witnessed our love. Thanks to my parents-in-law and brother-in-law for understanding and supporting me in the past few years. My parents, there are no words that I can express my gratitude to you. Your love and support give me courage to face any challenges in my life. I would also like to thank my sister, brother-in-law, and big uncle for contributing to a lovely and warm big family. Once in a while, we have a family video call on weekends and share interesting daily stories to each other. This really helps to get my mind off work and cheer me up sometimes. Many thanks to all of you!

Lizhou Wu
Delft, October, 2020

CONTENTS

1

INTRODUCTION

Spin-transfer torque magnetic random access memory (STT-MRAM) is considered as one of the most promising non-volatile memory technologies. After more than 40 years' research and development, its mass production is around the corner as numerous foundries and start-ups worldwide swarm into its commercialization. Like any semiconductor product, effective yet cost-efficient test solutions are of great importance to ensure high-quality STT-MRAM products being shipped to end customers. The main subject of this dissertation is to investigate STT-MRAM-specific manufacturing defects, accurately model them to derive realistic fault models, and eventually develop high-quality test solutions for STT-MRAMs. This chapter serves as a brief introduction to this dissertation. We start with highlighting the role of VLSI test, its importance, and basic concepts. Second, we introduce emerging non-volatile memory technologies covering three main classes: PCM, RRAM, and MRAM. Their working principles are briefly reviewed and their performance is compared to each other as well as to existing charge-based memories: SRAM, DRAM, and flash. Their development status, potential applications, and positioning in the present memory hierarchy is also discussed, with an emphasis on STT-MRAM. Third, we present the state of the art in both conventional memory testing and STT-MRAM testing. Fourth, we explain the research topics explored over the course of this PhD project. Fifth, we present the main contributions of this dissertation advancing the state-of-the-art in STT-MRAM testing. Finally, we detail the thesis organization.

1

1.1. VLSI TEST PHILOSOPHY

This section introduces the VLSI test philosophy as well as some basic concepts and terminologies in this field. It first identifies the position and role of VLSI tests within the broad scope of electronic testing. Thereafter, a classification of VLSI tests is discussed. Finally, the concepts of test escape and yield loss in production tests are elaborated.

1.1.1. POSITION AND ROLE OF VLSI TESTS

With the successful advancement in *very large scale integration* (VLSI) technology for nearly half a century, semiconductor chips have become indispensable components in any modern electronic system. For example, smartphones are probably the most commonly known and used electronic system in our daily lives nowadays. Typically, a smartphones contains a large number of semiconductor chips, of which the system-on-chip (SoC) is undoubtedly the most important one. A SoC is a monolithic VLSI circuit including a variety of modules; an example is the Kirin 990 5G processor which integrates a central processing unit (CPU), neural processing unit (NPU), graphics processing unit (GPU), 5G modem, on-chip memories etc., which together are built with 10.3 billion transistors in a single chip of 113.31 mm^2 using TSMC's 7 nm process[1]. It is obvious that fabricating such a sophisticated VLSI chip is a complicated and time-consuming process which is prone to manufacturing defects. Therefore, to guarantee the quality and reliability of semiconductor chips, it is crucial to rigorously test them in different manners at different phases of lifetime.

Typically, the lifetime of a VLSI chip can be divided into three phases involving three key parties, as illustrated in Figure 1.1. The first phase is the gestation period, where the

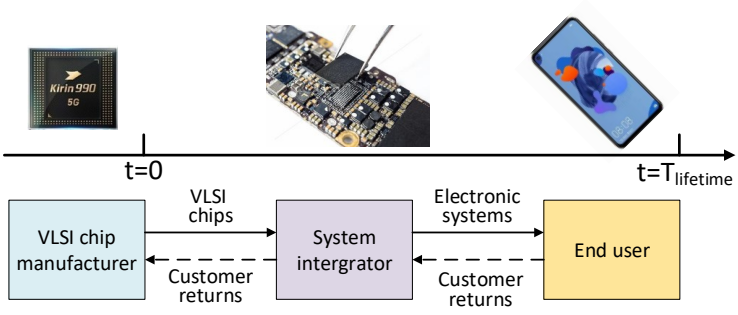


Figure 1.1: Three key phases and involved parties in the lifetime of a VLSI chip.

involved party is the VLSI chip manufacturer which defines the specifications of semiconductor chip products and subsequently designs and mass produce them. Note that the design company of a semiconductor chip may also be different from the one which eventually manufactures it. In the semiconductor industry, this is also a typical business model where a fabless company designs a product and get it fabricated in a foundry company which is dedicated to manufacture instead of designs. In the second phase, the fabricated chips are delivered to the system integrator which mounts them into electronic systems such as smartphones, laptops, and servers; these electronic systems are intended to be sold to the electronic market. The third phase mainly involves the end

user which buys these electronic devices and use them to accomplish a specific task. From a semiconductor chip's perspective, its life starts (typically referred to as t=0) when being shipped out from the manufacturer. Obviously, most of its lifetime is with the end user where it performs its designed functions in a system in the field of operation until wear-out.

In the above-mentioned three phases, VLSI chips are subjected to different tests. The VLSI chip manufacturer needs to conduct various manufacturing tests to weed out defective parts and guarantee that the outgoing parts to customers perform good functions as designed at t=0. Typically, the quality of VLSI chips is evaluated using a metric called *defective part per million* (DPPM). For instance, ten-DPPM means statistically ten parts out of one million parts shipped to the system integrator are defective. The test efforts that the manufacturer would make vary significantly depending on the chip quality requirements that are demanded by the system integrator. The chip quality requirements are in turn determined by the specific application that the system integrator expects its electronic system products to be used for. For example, a VLSI chip product targeting healthcare or aerospace applications requires much higher quality and therefore more stringent tests than that for kids' toys or consumer electronics.

In phase II, the second party, system integrator, may perform some basic tests (known as *incoming inspection*) with much less efforts and time than the previous manufacturing tests on certain number of selected samples of purchased VLSI chips, depending on the chip quality and system requirement. The purpose of incoming inspection is to avoid assembling defective chips into systems. But this practice is gradually disappearing, as companies nowadays expect the received chips to be high-quality and are often pressured by the time to market. Nevertheless, the system integrator focuses on a different type of test called *system test*. In other words, once a system composed of a large number of VLSI chips and other electronic components such as resistors, capacitors, batteries, and screens is manufactured, it also necessitates extensive tests before delivering to a customer. During this stage of testing, VLSI chips which are identified to be defective or cause system failures will be sent back to the manufacturer in the form of customer returns. The manufacturer is then expected to investigate these returned chips for failure analysis and diagnosis, which will be useful for improvements in either the test program or manufacturing process.

In phase III, the end user as the third party is not expected to conduct any testing work on the received product other than setting it up for regular usage. Similar to the customer returns from the system integrator to the chip manufacturer, the end user sends back defective products to the system integrator for reparation or replacement. However, as modern electronic systems are becoming increasingly complex and CMOS technology has entered into sub-10 ns era raising more reliability concerns, *on-line test* have become an important field for testing especially for some mission-critical industrial sectors such as satellites, automotive, and medical electronics [2]. On-line test is the test procedures running without the engagement of the end user in the field of operation, to monitor the hardware status so as to detect defective parts and enhance reliability or robustness. It can take place either concurrently during the normal operation mode or periodically during the idle mode.

Despite the fact that defective chips can be detected in all three phases of their life-

time, the first phase and the chip manufacturer should be primarily relied on to ensure the chip quality. This is because of the exponential increase in the cost of detecting a defective chip after being integrated into increasingly more complicated systems. A widely accepted rule of thumb in test economics in the electronics industry is the *rule of ten* [3]. It suggests that if a defective chip is not caught by chip-level testing, then finding it at printed circuit board level costs ten times as much as at the chip level. This cost factor continues to apply when the defective chip is incorporated into higher-level systems. Apart from the economic reason, selling defective chips to customers and receiving them back also have a negative impact on the manufacturer's reputation. In the worst case, a system failure due to a defective chip in the field may lead to a catastrophic accident or even the loss of human lives in some mission-critical applications such as automotive and healthcare.

All of the above aspects emphasize the importance of VLSI tests in phase I before t=0, which are mainly performed by the chip manufacturer. Since this test stage plays the most critical role in determining the chip quality, it incurs the biggest investment in testing, thus having the highest possibility of payback on research. Due to this reason, this thesis will be focused on this domain.

1.1.2. CLASSIFICATION OF VLSI TESTS

If a VLSI chip product is designed, fabricated, and tested, and it fails the test, then there must be a cause for the failure [3]. The cause can be the following: 1) the test is wrong, 2) the manufacturing process is faulty, 3) the design is incorrect, and 4) the specifications have a problem. Anything can go wrong. The responsibility of VLSI tests is to detect whether there is something wrong. If all chips fail, probably the first cause applies, i.e., the test is wrong. If the test is good and only a very small fraction of fabricated chips are tested negative, then we suspect 2), 3) and 4) might be the potential cause. To determine which type of cause leading to a chip failure, typically a variety of tests will be performed over the entire course of developing a VLSI chip product. Next, the classification of VLSI tests will be discussed.

VLSI tests taken by the chip manufacturer in phase I can be classified into three types as follows, depending on the test objectives and the development stage of a VLSI chip product [3].

1) Characterization: also known as design debug or verification test. This test form is performed on a new design before being sent to mass production. The first objective of characterization test is to verify that the design is correct and meets all specifications. Functional tests along with comprehensive AC and DC parametric measurements are run at this stage, to determine the limits of chip operation conditions such as supply voltage, temperature, and speed. Typically, these conditions are swept in given ranges and functional tests are performed repetitively for each combination of the above parameters. The measured results are plotted as a Shmoo plot where both the pass (P) and fail (F) regions are marked [4]. Other objectives of characterization tests include measuring chip characteristics for setting final specifications and determining a final production test program.

2) Production: every fabricated chip has to go through production tests. The objective of production tests is to enforce the quality requirements by determining whether

the chip under test meets all specifications. Production tests are go/no-go decision making processes which are less comprehensive than the previous characterization tests. The tests at this stage may not cover all possible functions, but they must guarantee a high coverage of modeled faults such that defective chips can be weeded out with a high confidence. As every chip must be tested, production test time for each chip is typically very short and the cost needs to be minimized as much as possible but without sacrificing the effectiveness of test.

3) Burn-in: passing production tests means that the passed chips meet design specifications at t=0, but it does not guarantee that they perform their functions as long as expected when getting to actual usage. Burn-in tests ensure the reliability of those chips which have passed production tests by testing either continuously or periodically over a long period of time at elevated voltage and/or temperature to force weak chips to fail at an accelerated speed [3]. Two types of failures can be isolated by burn-in tests: infant mortality and freak failures. The former are typically caused by weak defects or process variations; they can be screened out by short-term burn-in (10–20 hours) in a normal or slightly accelerated conditions. The latter occur to those chips which are as reliable as designed, thus requiring long burn-in time (100–1000 hours) in accelerated conditions. Compared to production tests, burn-in tests are much more expensive and time-consuming. Therefore, in practice, a manufacturer must take economics into account and make a trade-off between test overheads and chip reliability depending on the target applications.

1.1.3. TEST ESCAPES AND YIELD LOSS

As introduced previously, a production test is a short and go/no-go decision making process for every single fabricated chip which is intended to going to costumers. Figure 1.2a depicts the production test process where all fabricated chips need to go through the test program and end up in four sets of test results as follow.

① **Pass, OK**. Refer to chips which have passed the test and are real defect-free.

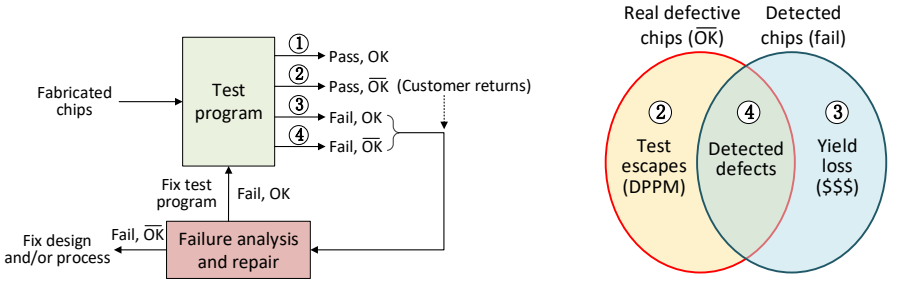
② **Pass, $\overline{\text{OK}}$**. Refer to chips which have passed the test but are defective actually.

③ **Fail, OK**. Refer to chips which have failed the test but are real defect-free.

④ **Fail, $\overline{\text{OK}}$**. Refer to chips which have failed the test and are defective actually.

Ideally, we would like to have all defect-free chips pass the test and all defective chips fail the test. In other words, only set ① and ④ are desired assuming that the test program is perfect. This maximize the interest of the manufacturer, as all chips being sold to customers would be as good as designed (i.e., 0 DPPM). However, this is almost impossible to achieve in practice, at least not achieved based on today's test technology at the point of writing this dissertation. A practical test program usually gives us a certain number of chips fallen into set ② and ③, unfortunately. Both of these two sets of chips cost money to a manufacturer.

Set ② contains defective chips that escape the test and therefore will be delivered to customers, along with other real defect-free chips in set ①. Test escapes can be caused by incomplete coverage of fault models due to high complexity or high cost. *Fault models*



(a) Production test process and four sets of test results.

(b) Venn diagram of set 2,3, and 4.

Figure 1.2: Test escapes and yield loss in production tests.

typically are the target of production tests; they are the high-level abstraction of physical defects. Another cause could be that, in practice, not all physical defects are taken into account and are well modeled and represented by existing fault models. Some of these defective chips will be mounted onto PCBs and electronic systems, and subsequently leaked to the market. As a result, they may incur user complaints and even lead to accidents or loss of human lives in the worst case. Some will be sent back the their manufacturer in the form of customer returns if identified by higher-level tests such as system and on-line tests. costumer returns have a significant influence on the business-to-business relationship and may even damage the established reputation of the chip manufacturer.

Set ③ contains good chips which however fail the test. This can be caused by excessively stringent tests; an example is I_{DDQ} test [3] which may over kills some good chips by mistakenly identifying the increased leakage current due to process variations as defects. This set of chips directly lead to yield loss, thus increasing the cost of manufacturing a chip on average. In addition, rejecting good chips also indicates that the test itself needs to be improved so that this set would be minimized as much as possible in the future production.

Set ④ contains defective chips which are captured by the test. These chips do not meet design specifications and should go to the failure analysis and repair department along with the chips in set ③ and customer returns (belonging to set ②). As illustrated in Figure 1.2a, investigating and understanding the failure mechanisms of chips in Set ④ are very important for yield learning process; the results can be used to fix design and/or manufacturing process.

Figure 1.2b shows a Venn diagram describing the relationship of set ②, ③, and ④. Set ② and ③ are mainly caused by the incompetence of the test program. Thus, investigating the failure analysis of chips in these two sets are beneficial for an enhancement in the test program. From an economic point of view, the two circles in Figure 1.2b need to be as closely overlapped as possible for the purpose of reducing test escapes and yield loss. Since these two benefits literally mean higher quality and less cost in manufacturing for a VLSI chip product, identifying the real defective chips, i.e., making these two circles overlap, is the invariable goal of R&D investment in VLSI tests.

1.2. EMERGING NON-VOLATILE MEMORY TECHNOLOGIES

Memory is an indispensable component in any computer system, and it is also one of the biggest sectors in the semiconductor industry. As existing memories such as SRAM, DRAM, and flash gradually approach their down-scaling limits, they become increasingly power hungry, and less reliable while the fabrication is more expensive due to the increased manufacturing complexity. As alternative solutions, several promising non-volatile memory (NVM) technologies have emerged and attracted extensive R&D attention for various levels in the memory hierarchy. This section starts with a brief overview of today's memory hierarchy. Then, a classification of mainstream existing and emerging memories is provided. Finally, a comparison of these memories is presented.

1.2.1. PRESENT MEMORY HIERARCHY

It is well recognized that the classical Von Neumann architecture comprises separate central processing unit (CPU) and memory unit. This means that data and instructions have to be frequently moved between these two units. Ideally, one would desire a system with its memory as fast as CPU, which maximizes the system performance. However, the reality is that the CPU speed is much higher than that of any type of existing memory nowadays. Starting from 1980 when their speeds are approximately the same, both CPU and memory have substantially evolved over the past four decades. The performance of CPU has improved tremendously by first boosting clock rate of single-core processor and subsequently incorporating multiple cores starting from around 2005. For instance, the Intel Core i7-960 processor contains 4 cores, each of which runs at 3.2 GHz (i.e., 0.3 ns per clock cycle). In contrast, the performance of main memory has not improved significantly in the past few decades, despite the fact that the density has increased considerably and the price per bit has become more and more affordable. Typically, the access latency of dynamic random access memory (DRAM) is 50-100 ns [5], which is more than three orders of magnitude slower than the speed of a high-end multi-core processor these days. This is well known as the "memory wall" [6], making memory the bottleneck of system performance. Compared to DRAM, static random access memory (SRAM) is much faster, up to ~1 ns. However, the downsides of SRAM are its high cost per bit and large area of memory cell. Other memory technologies such as flash and magnetic disk are cheap and large in volume, but they are orders of magnitude slower than DRAM. Unfortunately, there is no such an ideal memory technology which is fast, cheap, and large in volume, combining the benefits of all aforementioned memories.

To build such a desirable memory system, an economical solution is a memory hierarchy, which takes advantage of locality and cost-performance trade-offs of memory technologies. The principle of locality means that programs tend to reuse data and instructions they have used recently [6]. Figure 1.3 shows a multi-layer memory hierarchy, including the typical access speed of each level and the position of each memory technology. Traditionally, the memory hierarchy consists of three major layers: cache, main memory, and mass storage, implemented by SRAM, DRAM, and hard drive disk, respectively. Cache is fast, small, and expensive; thus, it is located the closest to the CPU. Main memory provides medium performance and cost, thus following the cache as the next memory layer in the hierarchy. With the lowest speed and largest volume, mass storage layer is the farthest layer to the CPU. In most cases, the data contained in a farther layer

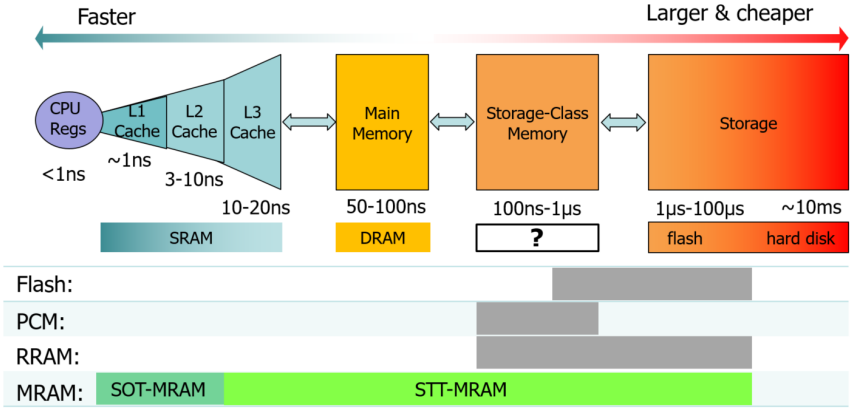


Figure 1.3: Present memory hierarchy in computer systems.

is a superset of data in the previous layer closer to the CPU. The goal of the memory hierarchy is to provide a memory system with cost per bit almost as low as the cheapest layer and speed almost as fast as the fastest layer.

With the CPU-memory performance gap becoming wider and the emergence of new memory technologies, the memory hierarchy has also been evolving over time. First, the cache layer has been split into several sub-layers to meet the ever-increasing memory access demand from the CPU. Figure 1.3 shows a three-level cache structure with the fastest L1 cache closest to the CPU and slower but larger L2 and L3 at lower levels. Second, flash memories are ubiquitous these days, serving as complementary storage media to the traditional hard disks. Thanks to their fast speed, non-volatility, and continuous bit cost reduction, they are widely used in solid-state drives (SSDs), smart phones, tablets, laptops, databases etc. Third, the gap in performance and price between the storage layer (including both flash and hard disk) and the main memory layer (DRAM) is still much wider than that between main memory and last-level cache. This has motivated the idea of adding a new memory layer which is commonly referred to as storage-class memory (SCM) [7] to fill in this gap in recent years. Flash memory, as a successful pioneer non-volatile memory technology, has the potential to adapt to the SCM layer. However, the main obstacles include its limited endurance and access speed in comparison to DRAM, making flash memory alone unable to serve as SCM. To address this issue, research attempts have been focused on hybridizing flash memory with other high-performance memory types such as DRAM [8].

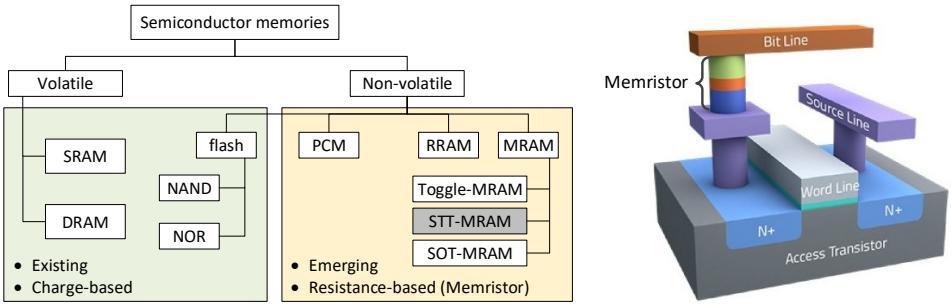
Candidates for SCM also include emerging NVM technologies such as phase-change memory (PCM), resistive random access memory (RRAM), and magnetic random access memory (MRAM). These memory technologies offers storage-class retention, relatively higher endurance than flash memory, and attractive read/write performance as high as DRAM or even SRAM but with considerably less static power consumption [9]. Due to these advantageous features, they can not only adapt to the SCM layer, but also may even revolutionize the entire memory hierarchy once they are mass produced and their cost per bit drops. Figure 1.3 shows the potential application position of each of these NVM technologies in the memory hierarchy. Limited by endurance (primary) and speed

(secondary), PCM and RRAM are predicted to be suitable for the SCM layer and below [10]. In contrast, MRAM provides an excellent tailorability by making trade-offs between retention, endurance, and speed with different programming technologies. Therefore, many believe that MRAM including its sub-classes with different flavors can be a true universal memory technology in the future [11]. Next, we will introduce all these memory technologies in more detail.

1.2.2. TYPES OF SEMICONDUCTOR MEMORIES

In general, semiconductor memories can be classified into two categories: volatile and non-volatile [10], as shown in Figure 1.4a. Volatile memories require continuous power supply to retain the stored data while non-volatile memories can retain the stored data even if the power is switched off. The mainstream volatile memories are SRAM and DRAM, which are ubiquitous in today's computer systems. Non-volatile memories include magnetic disk and flash memory conventionally. Note that magnetic disk is not considered as a type of semiconductor memory, as it is not based on transistors and involves slow mechanical movements in read and write operations. Thus, magnetic disk, through still the dominant storage medium at the moment, is not covered in our discussions in this thesis. Flash memory has two types: NAND and NOR. NAND flash features increasingly higher density and lower cost per bit, thanks multi-level cell and 3D stack technologies [8]. It is suitable for high-end storage applications in replacement of magnetic disk. In contrast, NOR flash memory is more expensive and faster in random access but lower in programming and erasing operations. These features make NOR flash more suitable for storage applications requiring fast read and occasional write (e.g., storing program code in mobile devices).

In addition, the majority of emerging memory technologies are non-volatile. For example, PCM, RRAM, and MRAM have attached large amounts of R&D attention over the past decades and have been prototyped and even commercialized in a small scale by worldwide semiconductor companies such as Intel, Samsung, Globalfoundries, and Everspin in recent years [12–15]. There are also several emerging memory technologies at early R&D stages, including ferroelectric random access memory (FeRAM), Carbon-based memory, Mott memory etc. [10]. These memories will not be discussed in this thesis.



(a) Classification of semiconductor memories.

(b) Basic 1T-1R memory cell sturcture.

Figure 1.4: Types of semiconductor memory technologies.

1

Based on the physical form in which the information is stored, the aforementioned semiconductor memories can also be categorized into charge-based and resistance-based memories. The former category include three existing mainstream semiconductor memories on the market: SRAM, DRAM, and flash memories, which utilize the quantity of electric charge to encode logic state '0' and '1'. The latter category comprises emerging memories: PCM, RRAM, and MRAM, which store data in the form of resistance; based on the write mechanism, MRAM can be further divided into first-generation Toggle-MRAM, second-generation STT-MRAMs, and SOT-MRAM as a representative of third-generation MRAM technologies. As the data-storing devices of these three types of NVM all encode logic states by exploiting the large resistance contrast in distinct physical states (e.g., amorphous and crystalline phases), they are all referred to as Memristors sometimes. These three types of Memristor are all compatible with the conventional CMOS process and are typically integrated between two adjacent metal lines in the back-end-of-line (BEOL) process. Figure 1.4b shows a schematic of the most commonly used 1T-1R memory cell structure. It consists of a transistor (access selector) at the bottom fabricated in the front-end-of-line (FEOL) process and a Memristor device (data-storing element) inserted in the subsequent BEOL process. Note that there exist several other selector candidates such as two-terminal diode or non-linear device in [10, 16], despite transistor is still the most popular one. Next, the working principle of each type of memristor will be elaborated.

PHASE-CHANGE MEMORY (PCM)

Phase-change memory (PCM) stores data by exploiting the large resistance contrast between poly-crystalline and amorphous phases in phase-change materials such as chalcogenide [17, 18]. Figure 1.5 illustrates the basic structure of mushroom-shaped PCM device and the transformation principle between the two phases. To reset the PCM device into the amorphous phase, a positive pulse with large amplitude (V_{set}) and short width (t_{set}~50 ns) is applied across the device. As a result, a current flows through the heater (resistor) which contacts the above phase-change layer in the device, generating a large amount of Joule heating. The generated Joule heating raises the temperature above the melting point, thus transforming the phase-change material in the mushroom-cap area into the amorphous phase, corresponding to the high resistance state (HRS). To set the PCM device into the crystalline phase, a medium pulse (V_{rst}) with the same polarity as

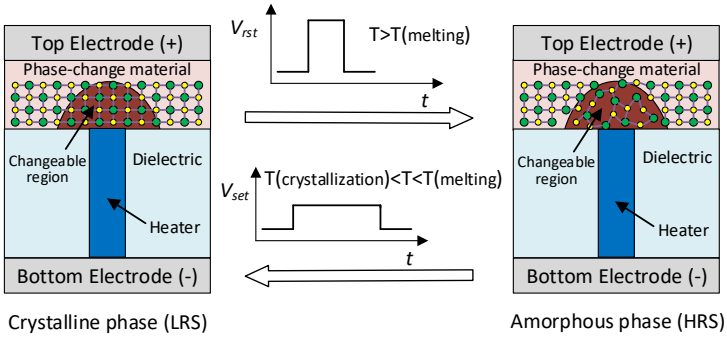


Figure 1.5: Basic mushroom-shaped structure of PCM device and its working principle.

the previous set pulse is applied for 100 ns-10 μs (t_{rst}), which anneals the changeable region at a temperature between the crystallization point and the melting point. The crystalline phase corresponds to the low resistance state (LRS), whose resistance can be three or four orders of magnitude lower than that of HRS. The large contrast in resistance is used to distinguish the two phases by applying a small bias to the device without disturbing its state.

PCM is among the most promising NVM technologies and has undergone significant academic and industrial research since the late 1960s. This has resulted numerous demonstration chips including a 1Gb chip by Micron in 2010 [19], a 8Gb chip by Samsung in 2012[12] and even a commercial product: 3D-Xpoint by Intel[20] in 2016. In the 2000s, PCM was considered to serve as a universal memory replacing both DRAM and NAND flash, as it exhibited high speed and scalability competitive to DRAM while being nonvolatile and owning higher endurance than NAND flash. However, this initial goal was not achieved due to the continuous improvement of DRAM and NAND flash as well as the limitations of PCM itself. Further innovations on PCM are needed to reduce power consumption, minimize resistance drift, improve endurance, and increase density [18]. Later on until now, the community has converged on the use of PCM as a SCM candidate complementing the traditional memory hierarchy shown in Figure 1.3.

RESISTIVE RANDOM ACCESS MEMORY(RRAM)

Resistive Random Access Memory (RRAM) stores data by exploiting the large resistance contrast between the complete *conductive filament* (CF) phase and the incomplete CF phase in the metal-oxide materials such as HfO_x [21]. Figure 1.6 illustrates the basic structure of metal-oxide RRAM device and its working principle as a non-volatile memory. A RRAM device fundamentally consists of two electrodes sandwiching a metal-oxide layer. As pure metal oxides are intrinsically dielectric, a fresh RRAM device (left one in the figure) exhibits a extremely high resistance. To make the RRAM device ready for transitions between the aforementioned binary states, a key manufacturing step known as the forming process is required. It refers to the process of creating a conductive filament, akin to a tunnel for electrons to freely move through, in the metal oxide by applying a high voltage ($V_{forming}$) across the raw RRAM device. Under the applied high electric field (>10 MV/cm), oxygen atoms are knocked out of the lattice and drift in the form of negative ions towards the top electrode (anode). This process leaves oxygen vacancies behind in the oxide layer. When enough oxygen vacancies are localized and form a conductive

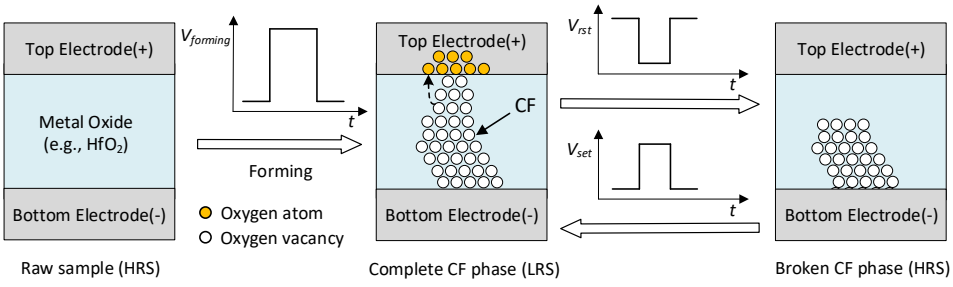


Figure 1.6: Basic structure of metal-oxide RRAM device and its working principle.

filament connecting the top and bottom electrodes, the device exhibits high conductance (i.e., LRS) as illustrated with the device schematic in the middle of the figure.

After the forming process, the RRAM device can be transformed from LRS to HRS (reset process) and from HRS to LRS (set process) by applying appropriate pulses. To reset the device, a negative pulse (V_{rst}) has to be applied. Under the induced electric field, oxygen ions migrate back to the oxide layer and recombine with a portion of oxygen vacancies near the top electrode. As the device contains an incomplete CF in its oxide layer, it transforms to the HRS (the right device schematic in Figure 1.6). Note that the resistance with an incomplete CF is much smaller than that of raw samples before the forming process, since the incomplete CF acts as a virtual bottom electrode. To set the device, a positive pulse (V_{set}) is applied to regenerate the complete CF. Due to the existence of the incomplete CF during normal switching cycles, both V_{rst} and V_{set} are smaller than V_{forming}. To read the resistive state of the device, a small bias is applied, similar to the read operation for the PCM device previously. The read window (i.e., HRS/LRS ratio) for RRAM devices is very wide, typically in the range of $10^1 - 10^4$ [21].

With the intensive R&D investment in the past decades, several RRAM test chips have been prototyped as both embedded and standalone memories [22]. Embedded RRAMs are used as IPs integrated into SoCs to replace existing e-flash memories. For example, in 2013, Panosonic announced the world's first mass-production of MCU with embedded RRAMs [23], which outperformed Flash-based MCU by five times faster and 50% less power consumption. In addition, RRAMs have demonstrated their capability to fit into SCM layer between DRAM and NAND flash as standalone memories. In 2014, Micron and Sony unveiled a 16Gb RRAM macro [24] in a 27 nm technology node with 200MB/s write speed and 1GB/s read speed. In the same year, SanDisk also demonstrated a 32Gb cross-point RRAM chip in a 24 nm process [25]. During the persistent memory summit in 2019, Sony disclosed that it aimed to commercialize 128Gb RRAM chips targeting high-end SSDs in 2020 (similar market positioning to Intel's Optane memory products).

MAGNETIC RANDOM ACCESS MEMORY (MRAM)

Magnetic Random Access Memory (MRAM) stores data by exploiting the significant resistance contrast between two different magnetic configurations in magnetic tunnel junctions (MTJs), which are the data-storing elements in MRAMs. Figure 1.7 shows the basic MTJ structure and its working principle. Fundamentally, an MTJ is composed of two ferromagnetic layers sandwiching an ultra-thin (~1 nm) dielectric layer. These three layers are named as free layer, tunnel barrier, and pinned layer respectively as illustrated in the figure. The magnetization in the free layer can be switched by [26]: 1) a perpendicular magnetic field, 2) a perpendicular electric current flowing through it under the effect of spin-transfer torque (STT), and 3): a horizontal electric current flowing through the top electrode in contact with the free layer under the effect of spin-orbit torque (SOT). These three switching methods lead to the three generations of MRAM technologies as shown in Figure 1.4a. In contrast, the magnetization in the pinned layer is strongly pinned to a certain direction. When the magnetizations in the two ferromagnetic layers are parallel, the MTJ exhibits LRS. When anti-parallel, the MTJ is in HRS. To switch between the two magnetic states, a pulse has to be applied across the MTJ; the pulse polarity determines the switching direction, as shown in the figure. To read the resistive state, a small bias is applied. Unlike the large read window in PCM and RRAM devices, MTJs have a

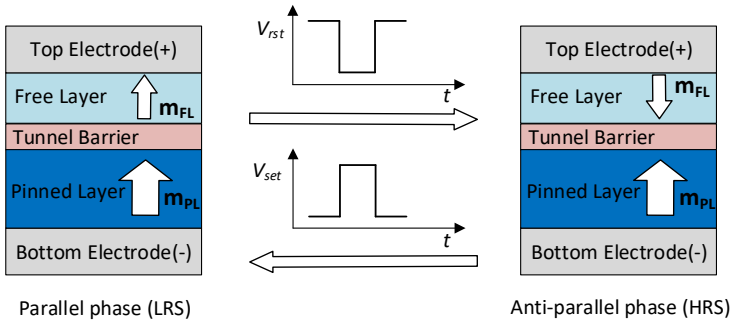


Figure 1.7: Basic structure of MRAM device (MTJ) and its working principle.

much smaller read window, typically HRS/LRS=2-3, limited by today's MRAM technologies. More details about MTJ technologies and write/read circuits will be explained later in Chapter 3.

Over the past two decades, MRAM R&D has experienced an exciting progress, which attracts a large amount of investment from most of the global semiconductor companies. Key players across the globe include Everspin, Avalanche, Honeywell, Intel, Samsung, Globalfoundry, SK hynix, and TSMC, TDK Headway, Toshiba, IBM, IMEC, SPINTEC. For example, Everspin technologies has commercialized both standalone Toggle-MRAM products since 2006 [27] and STT-MRAM products since 2015 [28] on a small scale. Avalanche offers both stand-alone and embedded STT-MRAM products [29]. SK hynix demonstrated a 4Gb STT-MRAM prototype targeting the replacement of DRAM and flash memories in 2016. Intel presented its embedded STT-MRAM solution in 2019 [30] and CMOS compatible process integration of SOT-MRAM in 2020 [31]. Samsung [14], Globalfoundries [4], and TSMC [32] demonstrated embedded STT-MRAM macros up to 1Gb in 2019. It is clear that Toggle-MRAM and STT-MRAM technologies are ready for mass production and deployment in the industry, while SOT-MRAM technology still requires further efforts in improving process and device/circuit co-optimization. According to a report from Coughlin Associates after the 2018 MRAM Developer Day, it was projected that the market for MRAM solutions will experience a fast growth from $36 million in 2017 to about $3.3 billion in 2028, and the annual shipped capacity will rise to 84PB by 2028 [33].

1.2.3. COMPARISON OF SEMICONDUCTOR MEMORIES

Table 1.1 compares the performance of semiconductor memories in Figure 1.4a using various metrics. SRAM is the fastest existing memory with high endurance and low dynamic power. Thus, it is mainly used as caches which are the closest to CPU in the memory hierarchy. However, the drawbacks of SRAM include: 1) volatility, 2) large cell size (6 transistors in a cell), 2) high static power due to leakage current, and 3) high cost per bit. DRAM features high speed, high endurance, and low cost per bit, but it is limited by its volatility and high static power induced by constantly refreshing DRAM cells to retain data. NOR flash is byte addressable and providers much faster read than NAND flash which is only accessed in page or block. However, NOR flash has lower density, high cost

Table 1.1: Performance comparison between different types of semiconductor memories.

Metric \ Memory	SRAM	DRAM	NOR FLASH	NAND FLASH	PCM	RRAM	Toggle MRAM	STT MRAM	SOT MRAM
Non-volatile	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Cell size (F^2)	120	6-10	10	5	6-12	6-10	20-30	6-10	10
Read time	~2ns	~30ns	5us	50us	20-50ns	10-50ns	~35ns	~5ns	~1ns
Write/ Erase time	~2ns	~30ns	500us/ 900ms	200us/ 2ms	100ns-10us	10ns-100us	~35ns	5-100ns	~1ns
Endurance	10^{16}	10^{16}	10^5	10^5	10^6-10^9	10^5-10^9	10^{16}	10^9-10^{12}	$>10^{12}$
Dynamic power	L	M	H	H	H	M	H	M	L
Static power	H	H	L	L	L	L	L	L	L

per bit, and lower write and erase speed, compared to NAND flash. Therefore, NOR flash is more suitable for code storage and execution while NAND flash is the best medium for bulk data storage such as solid-state drives (SSDs).

The performance of PCM and RRAM lies between DRAM and NAND flash; thus, these two types of NVMs are believed to be promising candidates to complement the existing memory hierarchy as storage-class memory. The density (cell size) of PCM and RRAM is comparable to DRAM but still much smaller than NAND flash at present. To further improve the density of these two memories, techniques such as multi-level cell and 3D stacking have to explored. The write speed of PCM varies from 100 ns to 10 μs, whereas RRAM has even larger range of write speed (10 ns-100 μs) according to the collected data of test chips [10]. The endurance of PCM and RRAM is typically limited below 10^9, as the write operations for these two memories involve physical changes (atom/ion movements) in the memory devices. This has made them incompetent to serve as main memory alone despite the fact that their access speed can be as fast as DRAM. Moreover, the dynamic power of PCM and RRAM in write operations are comparably high; thus, innovations in reducing switching power are still required.

The performance of MRAM family spans a wide spectrum in the memory hierarchy, from the fastest L1 cache down to SCM. The first-generation MRAM: Toggle-MRAM has slightly bigger cell size than later STT-MRAM and SOT-MRAM, as it requires to add current-carrying wires to generate magnetic field in the proximity of MTJ device to switch its state. The read/write latency of Toggle MRAM is ~35 ns and its endurance is almost unlimited. The downsides of Toggle-MRAM include it poor scalability towards advanced technology nodes and high dynamic power. In contrast, the second-generation MRAM technology: STT-MRAM surpasses Toggle-MRAM in most performance metrics except for endurance. The write speed of STT-MRAM is typically 5-100 ns, which means STT-MRAM can be as fast as last-level cache. A key challenge facing STT-MRAM is the high switching current, which not only incurs high dynamic power but also limits endurance. To obtain sub-ns write speed and low write power, novel write mechanisms are being explored, typically referred to as third-generation MRAM technologies. A representative third-generation MRAM is SOT-MRAM, which is still in intensive R&D. SOT-MRAM manufacturing and COMS-compatible integration solutions were demonstrated separately by first IMEC in 2019 [34] and Intel in 2020 [31].

Comparing the performance of different semiconductor memories in Table 1.1, STT-

MRAM stands out due to its advantageous features: non-volatility, high density, fast access speed, high endurance, and low standby power. The tunability of write speed, endurance and data retention makes STT-MRAM customizable for a variety of applications such as SSD buffer, last-level cache, Internet-of-Things (IoTs), and automotive. According to a report from Coughlin Associates after the 2018 MRAM Developer Day, it was projected that the market for STT-MRAM solutions will experience a fast growth from $36 million in 2017 to about $3.3 billion in 2028, and the annual shipped capacity will rise to 84PB by 2028 [33]. As STT-MRAM technology and manufacturing process get mature over time, an industrial ecosystem from equipment providers to chip manufacturers, to system integrators, is gradually getting shaped. This would undoubtedly drive the mass production and adoption of STT-MRAMs in the industry in the upcoming years. Under such a circumstance, the research carried out in this thesis aims to develop high-quality STT-MRAM test solutions with minimum test cost and time.

1.3. STATE OF THE ART IN MEMORY TESTING

Compared to logic testing, memory testing is more difficult as memory contains millions or even billions of states. Exhaustive testing of all possible states in a memory chip is an impossible task. Therefore, dedicated test techniques have to be developed and applied to guarantee the quality of memory chips. This section first reviews the milestone advancements in traditional memory testing over the past decades. Thereafter, the prior art in STT-MRAM testing is discussed.

1.3.1. TRADITIONAL MEMORY TESTING

Memory testing has gone through a long evolution process. The early memory tests (before 1980) can be classified as ad-hoc tests due to the absence of formal fault models and proofs [35]. They have a low defect coverage and a very long test time, typically in the order $\mathcal{O}(n^2)$, where n is the number of bits in a memory chip. Examples of ad-hoc tests are Zero-One test, GALPAT test, and Walking 1/0 test [35, 36].

To reduce the test time and cost per memory chip with the exponential increase in memory capacity, the focus of test development shifted to investigating the possible faults which can appear in the memory. For this reason, many functional fault models had been introduced during the early 1980's. The advantage of these models is that the fault coverage of a certain test can be provable while the test time is usually in the order $\mathcal{O}(n)$; i.e., linear with the size of the memory. Some important fault models introduced in that time were stuck-at faults and address-decoder faults [36]. These are abstract fault models not based on any actual memory design nor real defects.

In the late 1990s, experimental results based on DPPM screening of a large number of tests applied to a large number of memory chips indicated that many detected faults cannot be explained with the well-known fault models [37, 38], which suggested the existence of additional faults. This stimulated the introduction of new fault models (both static and dynamic) based on linear resistor defect injection and SPICE simulation [39, 40]: read destructive faults, write disturb faults, transition coupling faults, read destructive coupling faults, etc.

The current cell-aware test (CAT) approach [41, 42] is quite similar to the conven-

tional linear-resistor-based test approach in essence. As CMOS technology scales down to more advanced nodes, a growing number of defects occur within library cells. It was reported that less than 50% of these cell-internal defects were detected by the conventional tests targeting defects at the library cell ports [41]. This has motivated the development of CAT which explicitly targets cell-internal defects in the past decade to further reduce test escapes rate. CAT has demonstrated its value in the semiconductor industry for recent technology nodes: 45 nm, 32 nm, and even 14 nm based on FinFET technology [42]. As the technology down-scaling continues, 5 nm FinFET process technology has entered into volume production by TSMC in the first half of 2020 [43]. However, it is widely recognized that defects and variability in device characteristics during the fabrication process, and their impact on the overall quality and reliability of the system represent major challenges, especially when considering high-quality levels, e.g., in the range of *defective parts per billion* (DPPB) [44]. As CAT only targets resistive defects (e.g., opens and bridges) at the terminals and interconnects of devices (e.g., transistors), its effectiveness in detecting device-internal defects remains a question to date.

Moreover, with the advent of emerging devices such as PCM, RRAM, and STT-MRAM devices introduced previously, new materials, fabrication steps, and failure mechanisms are involved. It is shown in [45] that the fault mode of chips is dominated by transient, intermittent, and weak faults rather than hard and permanent faults in the nano-era. This shift in failure mechanisms may impact the way fault modeling and test development have to be done in the future.

1.3.2. STT-MRAM TESTING

STT-MRAM mass production is around the corner as major foundries worldwide invest heavily on is commercialization. Like all semiconductor products, STT-MRAM chips need to undergo intensive electrical tests to weed out the defective parts and guarantee outgoing product quality and reliability to customers. The STT-MRAM manufacturing process involves not only conventional CMOS process but also MTJ fabrication and integration [46]. The latter is more vulnerable to defects as it requires deposition, etch, and integration of magnetic materials with new tools [47]. A blind application of conventional tests for existing memories such as SRAMs to STT-MRAMs may lead to test escapes and yield loss. Hence, it is crucial to research on developing effective and cost-efficient test solutions for STT-MRAMs.

Testing MRAM started with toggle MRAM and thermally-assisted switching MRAM [48]. They are typically considered as the first generation of MRAM technology prior to STT-MRAM, as the switching method is based on external magnetic field. The first paper on testing MRAM was published in 2004 by Su *et al.* [49]. In this work, the authors injected resistive shorts and opens into a SPICE model of MRAM cell and found that most of the injected defects can be covered by the stuck-at fault model; they also identified two additional faults models: multi-victim and kink faults. In 2006, the same research group presented write disturbance fault model for MRAM, due to excessive magnetic field during write operations; the results were validated by designing and fabricating an MRAM chip on a 0.18 μm CMOS process [50]. In 2012, Azevedo *et al.* [51, 52] analyzed the impact of resistive bridges and opens on the read and write operations for thermally-Assisted switching MRAM; the simulation results revealed three fault models: stuck-at fault, tran-

sition fault, and state coupling fault. These early works focus on field-driven MRAM technologies; i.e., the magnetic states of MRAM cells are switched by applying external magnetic fields generated by current-carrying wires. Therefore, the above-mentioned fault models may not be applicable to STT-MRAM which is driven (written) by current instead of magnetic field.

Testing STT-MRAM is still in its infant stage with limited publications. In 2015, Chintaluri *et al.* [53, 54] studied the faulty behaviors of STT-MRAM induced by resistive opens and shorts as well as extreme process variations. Based on circuit simulations, they derived six fault models: stuck-at fault, transition fault, incorrect read fault, read disturb fault, retention fault, and coupling fault. In 2016, the same research group presented a memory built-in-self-test (BIST) to detect these faults; furthermore, this MBIST design was also claimed to have the capability of characterizing retention time of STT-MRAM cells at affordable test time [55]. In 2018, Nair *et al.* [56] performed layout-aware defect injection and fault analyses, whereby they observed dynamic incorrect read faults; a test algorithm was also proposed to detect all observed faults in the same paper. More recently, Radhakrishnan *et al.* [57] developed and implemented a Design-for-Testability (DfT) scheme for STT-MRAM parametric testing and process optimization. The CMOS-based DfT circuit replicates the electrical characteristics of MTJ devices. They also extended this DfT design to monitor electrical parameter deviations of MTJ device due to aging defects formation over time [58].

Scanning the prior works on testing STT-MRAM or MRAM reveals four major limitations. First, *linear resistors* are used to model all STT-MRAM manufacturing defects, including those in MTJ devices which are the data-storing elements in STT-MRAMs. However, linear resistors (with only electrical properties) *cannot* reflect the changes of defects on the MTJ's magnetic properties which are as important as electrical ones. Second, there is a lack of characterization data of defective STT-MRAM cells; this is needed to understand the mechanisms, causes, locations, and impact of STT-MRAM defects. Third, existing fault modeling approaches are unsystematic, and the fault model terminology is ambiguous. For instance, Chintaluri *et al.* [54] refer to a failed transition write fault as *transition fault* (TF), while Vatajelu *et al.* [59] use the term *slow write fault* (SWF) to describe the same faulty behavior. In addition, the term *read distrub fault* (RDF) is used to describe different faulty behaviors with different failure mechanisms in [54] and [60]. Finally, the proposed test solutions in the prior art have never been implemented in real-world STT-MRAM prototype chips; therefore, their effectiveness in detecting STT-MRAM-specific defects has not been justified with silicon data yet.

1.4. RESEARCH TOPICS

Conventionally, the structural approach used to develop test solutions for memories and logic circuits mainly consists of three steps: 1) defect modeling, 2) fault modeling, and 3) test development. Given the limitations of STT-MRAM testing in prior art discussed previously, innovations in all these three steps are required to develop high-quality test solutions for STT-MRAM production. Next, we elaborate research issues in each step.

1.4.1. DEFECT MODELING

Defect modeling is the first critical step in the test development process. Having an accurate defect model that is able to mimic the way a physical defect manifests itself at the electrical level is the best way to close the gap between the reality and the abstraction (fault models). To this end, the research carried out in this thesis focuses on the following three topics.

1) Complete defect space: The first research topic is to survey all possible types of physical defect that may take place during the STT-MRAM manufacturing process, especially those in MTJ devices which are the data-storing elements in STT-MRAMs. Furthermore, understanding the forming mechanisms, occurrence rates, electrical consequences of these defects are also of great importance. Preferably, silicon data on real fabricated STT-MRAM devices or prototype chips has to be collected for better understanding of the characteristics and effects of defects.

2) Accurate defect modeling approach: The traditional test approach assumes that all manufacturing defects can be modeled as linear resistors irrespective of their physical natures, as can be found in the prior art reviewed in the previous section. Although this assumption is convincing for modeling defects in interconnects, it has never been corroborated for defects in devices such as MTJs. It is well known that MTJ is a non-linear bipolar device of which its magnetic attributes (e.g., hysteresis loop) are as critical as its electrical ones. As a consequence, having linear resistors represent the MTJ-related defects may not necessarily appropriate in reflecting the defect-induced changes in the MTJ's magnetic attributes, switching mechanism, and tunneling magneto-resistance. Thus, whether or not the aforementioned assumption is applicable to MTJ-internal defects is a fundamental question which needs to be answered. If linear resistors are unqualified in modeling one or more defects in MTJ devices, an alternative device-aware defect modeling approach has to be developed to accurately present a physical defect at electrical level.

3) Accurate and realistic defect models: Once the defect modeling approach is obtained, it has to be applied to each defect in STT-MRAMs. This allows us to derive an accurate defect model which represent how the physical defect behaves at the electrical level. If possible, the obtained defect model has to be calibrated with silicon data of the defect being modeled. By repeating the same modeling process for all physical defects found in STT-MRAMs, a complete set of defect models can be obtained for subsequent fault modeling and test development. Note that inaccurate defect modeling may lead to inaccurate fault models. This in turn results in low-quality tests and/or DfT solutions, which cannot guarantee a low test escape rate, even with a claim of high fault coverage.

1.4.2. FAULT MODELING

Fault modeling is the second step in the test development process. This step is extreme important since the derived results, fault models, are typically the targets of test. Therefore, developing accurate and realistic fault models which capture the faulty behavior of a memory cell in the presence of a defect is the key to high-quality tests. To this end, the following four topics are explored in this thesis.

1) Circuit simulation platform: Typically, fault modeling is performed by SPICE-based circuit simulations. Therefore, a practical circuit simulation platform has to be

built. It has to include a complete STT-MRAM design with memory arrays and all necessary peripherals. Write and read functions in normal mode should be verified in the defect-free case. In addition, defect injection method (i.e., inserting defect models into STT-MRAM circuits) has to be developed to study and model the resultant faulty behavior of memory cell in the presence of a defect.

2) Complete fault space: Unlike SRAMs and DRAMs which store data in the form of electric charge, STT-MRAMs store data in the form of tunneling magneto-resistance. Apart from the data-storing principle, write/read methods and defect mechanisms in STT-MRAMs are also fundamentally different from those in conventional memories. Therefore, it is imperative to investigate whether existing fault models developed from conventional memories are applicable to STT-MRAMs, whether there are unique faults in STT-MRAMs. In other words, the complete fault space dedicated to STT-MRAMs has to be defined; in contains all possible faults that may occur in STT-MRAMs.

3) Fault analysis procedure: Based on the STT-MRAM circuit simulation platform, a sound fault analysis procedure has to be developed and automated to validate the defined fault space.

4) Accurate and realistic faults: Finally, the above fault analysis procedure has to be applied to all defect models in STT-MRAMs to obtain accurate and realistic faults. This step aims to create a clear mathematical graph between faults and defects in STT-MRAMs.

1.4.3. TEST DEVELOPMENT

Test generation is the last step in the test development process. In this step, the following three topics are explored to generate optimal test solutions for STT-MRAMs.

1) March algorithms and conditions: March tests are the most commonly used test solutions for memory testing. In this phase, we aim to develop March algorithms that cover all observed fault models in STT-MRAMs. Optimization of the tests for efficient test time and effective application from chip-external test equipment will also be explored. Furthermore, it is also important to take into account magnetic requirements during testing and how they can be best realized in a test environment, and to look into the need for stress/burn-in tests and conditions (voltage, temperature, duration etc.).

2) DfT and BIST/R solutions: In this phase, we will address the issue of how to provide test access to embedded STT-MRAMs (e.g., embedded in a MCU or SoC chip). We will devise test algorithms with regular structure to minimize the hardware cost of built-in-self-test (BIST). We will look into STT-MRAM repair options and built-in-self-repair (BISR). Special DfT to increase the fault coverage and/or reduce test time will be explored as well.

3) Validation of test solutions: This phase aims at validating and measuring the proposed solutions. If possible, the proposed test solutions will be implemented and evaluated with STT-MRAM demonstration chips. Experiments will be conducted to collect silicon data so as to optimize the test solutions if necessary.

1

1.5. CONTRIBUTIONS OF THE THESIS

Over the entire course of this PhD project, we are devoted to addressing the research issues at all the three phases of test development, as presented in the previous section. The main contributions of this thesis can be summarized into five items as follows.

1) Survey on STT-MRAM failure mechanisms, fault models, and tests. We first surveyed STT-MRAM failure mechanisms in the literature, and classified them into five categories: i) manufacturing defects, ii) extreme process variations, iii) magnetic coupling, iv) STT-switching stochasticity, and v) thermal fluctuation. We also investigated and classified fault models induced by these failure mechanisms. Finally, the state-of-the-art test solutions were examined and discussed. The limitations of the state of the art include: i) physical defects are all modeled as linear resistors (i.e., opens, shorts, and bridges) regardless of their physical natures; ii) existing fault modeling approaches are unsystematic, and the fault model terminologies are ambiguous and inconsistent in the literature; 3) the proposed faults and tests have never been corroborated with silicon data. This work is published in [61] and is excluded in this thesis as they belong to the prior art.

2) Define the complete STT-MRAM defect space. Defects are closely related to manufacturing imperfections. In this thesis,we introduce the manufacturing process of STT-MRAM and potential defects that may take place in each step [46]. As the fabrication of MTJ devices (data-storing elements) entails extra and unique steps inserted into the conventional CMOS process, special attentions are given to these new manufacturing steps and associated defects in MTJs (see Table 3.3). To the best of our knowledge, we are the first to open the "black box" of MTJ and look into manufacturing defects inside the device in the test community.

3) Characterize and model STT-MRAM defects based on silicon measurements. We divide STT-MRAM defects into two categories: interconnect defects and MTJ-internal defects. For interconnect defects, we model them as linear resistors using the conventional approach. For MTJ-internal defects such as pinhole defects, we demonstrated with silicon measurements and circuit simulations that modeling them as linear resistors is inaccurate [62]. A linear resistor is not qualified to mimic the way an MTJ-internal defect behaves at the functional level, leading to non-existent fault models. This in turn results in poor-quality test solutions and a waste of test time and resources. To address this issue, we propose a device-aware defect modeling approach, which specifically targets MTJ-internal defects [47]. This approach has been applied to pinhole defects [62], synthetic anti-ferromagnet flip (SAFF) defects [63], intermediate (IM) state defects [64]. All of these defects are integrated into a parameterized defective MTJ model, which is calibrated by the measured silicon data at imec.

4) Characterize, model, and evaluate the magnetic coupling effect based on silicon measurements [65]. The magnetic coupling effect is a unique mechanism in STT-MRAM; it poses a critical constraint when designing MTJ devices and STT-MRAM arrays. In this thesis, we present magnetic characterization results of MTJ devices with various sizes ranging from 35 nm to 175 nm. We propose an analytical intra-cell magnetic coupling model, which is calibrated and validated by the measured silicon data. Thereafter, we extrapolate this model to study inter-cell magnetic coupling on a memory array with varying pitches. We also introduce the inter-cell magnetic coupling factor Ψ to quantify

the coupling strength. The impact of magnetic coupling on the MTJ's write characteristics and retention time is also evaluated.

5) Develop a field-aware compact model of pMTJ in Verilog-A for electrical/magnetic co-simulation of STT-MRAM: With the fast development of spintronics and STT-MRAM commercialization, there is an urgent need for magnetic/electrical co-simulations of hybrid MTJ/CMOS circuits. This is because spintronic circuits such as STT-MRAMs exploit both charge and spin properties of electron, thus they are very sensitive to magnetic fields. We propose a field-aware compact model of pMTJ, implemented in Verilog-A. With this compact MTJ model, electrical/magnetic co-simulation of STT-MRAM circuits can be performed under PVT variations and various magnetic configurations, for fast and robust device/circuit co-design of STT-MRAM using existing commercial CAD tools.

6) Develop accurate and realistic fault models for STT-MRAM using SPICE-based circuit simulations: We propose a device-aware fault modeling framework, which consists of complete fault space definition and fault analysis [66]. We extend the conventional fault primitive notation to describe all memory faults in emerging non-volatile memories. To our knowledge, we are the first to propose undefined ('U'), extreme low ('L'), and extremely high ('H') resistive states, and corroborate their existence with silicon data of defective MTJ devices [67]. To obtain realistic faults in the presence of a defect, we propose a systematic fault analysis procedure. We are also the first to observe intermittent faults in STT-MRAM; they are intermittent passive neighborhood pattern sensitive fault ($PNPSF1_i$) caused by SAFF defects and intermittent write transition faults ($W1TFU_i$ and $W0TFU_i$) caused by IM state defects.

7) Propose optimal test solutions for STT-MRAM: With the clear mapping relations between physical defects and fault models in the previous phases, optimal test solutions can be derived depending on the target applications with different IC quality requirements.

- Resistive defects in interconnects result in a set of easy-to-detect faults, which can be detected by conventional March tests such as March C- [46].
- The detection of pinhole defects depends on the defect size. Large pinhole defects can simply be detected by March tests, while small pinhole defects require stress tests with hammering write '1' operation sequence at elevated voltage or prolonged pulse [62].
- SAFF defects cause an intermittent fault $PNPSF1_i$; conventional March tests cannot guarantee the detection of such a fault. To detect it, we propose a hybrid March test incorporating a magnetic write operation ($W0_H$ or $W0_H$). To our knowledge, we are the first to introduce magnetic write operations in STT-MRAM testing [63].
- IM state defects lead to intermittent faults $W1TFU_i$ and $W0TFU_i$. To detect them, we propose and implement a test with weak write operations ($\widehat{w}0$ and $\widehat{w}1$) [68].

1

1.6. THESIS ORGANIZATION

The aforementioned contributions advancing the state of the art in STT-MRAM testing will be elaborated in detail in the remainder of this thesis, which is organized as follows.

Chapter 2 introduces STT-MRAM behavior and architecture. First, a hierarchical modeling approach is described; it covers different abstraction levels from chip external behavior down to the physical buildup of internal components. Thereafter, the behavioral STT-MRAM model is discussed; it describes the external behavior of STT-MRAM using Everspin's latest 1Gb discrete STT-MRAM product as an example. This chapter also presents the STT-MRAM functional model, covering all functional blocks inside the chip, memory organization, and internal behavior.

Chapter 3 covers STT-MRAM fundamentals and implementation. It starts with introducing MTJ device technologies. Then, electrical and layout models for STT-MRAM are discussed. Thereafter, STT-MRAM manufacturing process and potential defects are detailed. Finally, this chapter ends with a brief review of milestone breakthroughs in the development of STT-MRAMs, potential applications, and remaining challenges along the road towards mass production and deployment in the semiconductor industry.

Chapter 4 presents STT-MRAM testing using the conventional approach based on linear resistors (i.e., opens, shorts, and bridges). First, a Verilog-A compact model for defect-free MTJs is developed can calibrated with silicon data. Thereafter, STT-MRAM manufacturing defects are modeled as linear resistors at all possible locations in a single memory cell. This is followed by a comprehensive fault modeling process based on circuit simulations, which derive appropriate fault models. The derived fault models are then used to develop a March algorithm, which covers all considered resistive defects.

Chapter 5 is concerned with a magnetic-field-aware compact model of pMTJ. First, the motivation and prior work is examined, followed by a detailed explanation of three sources of magnetic field disturbance. Thereafter, magnetic fields are physically modeled and calibrated with silicon data. Then, the implementation of the magnetic-field-aware compact MTJ model is detailed. Finally, SPICE-based circuit simulations are performed to demonstrate electrical/magnetic co-simulation for robust STT-MRAM designs.

Chapter 6 elaborates the device-aware test approach. First, the motivation behind DAT is presented. Thereafter, the three key steps of DAT: device-aware defect modeling, fault modeling, and test generation are detailed respectively.

Chapter 7 applies DAT to pinhole defects in MTJ devices. First, the pinhole defect mechanism is discussed. Second, comprehensive characterization of pinhole defects at both t=0 and t>0 is presented. This is followed by device-aware defect modeling to obtain a pinhole-parameterized MTJ compact model, which is calibrated by the measured data. Subsequently, device-aware fault modeling and test generation are performed to develop appropriate test solutions for pinhole defects.

Chapter 8 and Chapter 9 are another two case-studies, which apply DAT to synthetic anti-ferromagnet flip and intermediate state defects respectively, in the same way as the above pinhole defects.

Chapter 10 concludes this thesis and provides an outlook to future research directions.

2

STT-MRAM Behavior and Architecture

The first stand-alone STT-MRAM product was commercialized by Everspin in 2015. Since then, numerous start-ups and foundries worldwide had joined the race in improving and commercializing this new memory technology in both embedded and stand-alone forms with various interfaces. In 2019, Everspin announced its 1 Gb *stand-alone STT-MRAM chip with DDR4 interface using GlobalFoundries's* 28 nm *CMOS technology node, targeting at DRAM replacement for applications which require low access latency, high data persistence, and high endurance. It is expected that STT-MRAM is going to gradually penetrate into the semiconductor memroy market as it becomes mature and cheaper over time. This chapter employs a hierarchical modeling approach to describe a stand-alone STT-MRAM chip in a top-down manner. We first introduce the generic modeling hierarchy for a semiconductor chip covering five abstraction levels: behavioral, functional, logical, electrical, and layout levels. Thereafter, we describe in detail the STT-MRAM behavioral model, which is concerned with chip package, input/output pins, and operations with the associated timing diagrams. Finally, we present the STT-MRAM functional model, covering all functional blocks inside the chip, memory organization, and internal behavior. The logical model is generally not applicable for memory chips; the electrical and layout models will be presented in the next chapter for balanced contents and length in each chapter.*

2.1. STT-MRAM MODELING HIERARCHY

Today's IC chips are extremely complex systems which may contain billions of transistors, on-chip memories, and analog/mixed signal components. It is almost impossible for a single person or a small team to design and implement them. Therefore, it becomes increasingly important to have collaborations between colleagues with different expertise and between companies worldwide in the semiconductor industry. To facilitate seamingless and effective interaction between different designers, a hierarchical design and modeling methodology is crucial. It describe an IC at different abstraction levels and hide unnecessary design and implementation details.

Figure 2.1 depicts five different modeling levels for an IC chip. The lowest level, represented by the largest block, is the layout model; it is the one which is closest to the actual physical system and contains the most implementation details. As we move from the layout model (lowest level) towards the behavioral model (highest level) in the figure, the models become less representative of the physical buildup of the IC and more related to the way the IC behaves, or in other words, less physical and more abstract. In the figure, each modeling level is called a level of abstraction. It is possible to have a model that contains components from different levels of abstraction, an approach refereed to as mixed-level modeling. With mixed-level modeling, one may focus on low-level details only in the area of interest in the system, while maintaining high-level models for the rest of the system. Next, we elaborate each modeling level in more detail [69].

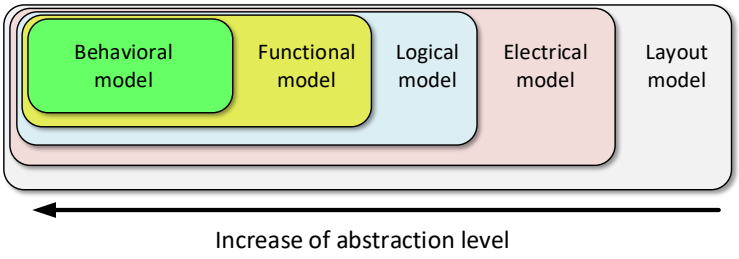


Figure 2.1: IC modeling hierarchy.

1) Behavioral model: This is the highest modeling level in the figure and it is based on the specifications of the system. At this level, there is practically no information given about the internal structure of the system or possible implementations of the performed functions. The only information given is the relation between input and output signals while treating the system as a black box. A model at this level usually makes use of timing diagrams to convey information about the system's behavior. It describes how the model interacts with the external world, such as memory read and write operations. In this chapter, the behavioral STT-MRAM model is presented in Section 2.2.

2) Functional model: This model describe the functions of the system that it needs to fulfill in order to operate properly. At this abstraction level, the system is divided into several interacting subsystems, each of which has a specific function. Each subsystem is basically a black box called a *functional block* with its own behavioral model. The collective operation of the functional blocks result in the proper operation of the system as a whole. In this chapter, the functional STT-MRAM model is presented in Section 2.3.

3) Logical model: This model is based on the logic gate representation of the system. At this level, simple Boolean relations are used to establish the desired system functionality. It is a common model to describe digital circuits. However, for memory circuits, it is not common to model and present them exclusively using logic gates. This is because logic gates constitute peripheral circuits and only account for a small part in a memory chip. The majority area is occupied by memory arrays where information is stored. Therefore, no exclusive STT-MRAM logical model is given in this chapter.

4) Electrical model: This model is based on the basic electrical components that make up the system. In semiconductor memories such as SRAM and DRAM, the basic electrical components are transistors, resistors, and capacitors. For STT-MRAM, MTJ device is the most important component as it serves as the data-storing element. At this level, we are not only concerned with the logical interpretation of an electrical signal but also the actual electrical value of it (e.g., voltage, current, and resistance). Since this thesis is primarily concerned with electrical-level simulations (SPICE-based) of STT-MRAM circuits, this memory model is presented in depth in Section 3.2.

5) Layout model: This is the lowest modeling level shown in the figure and the one with the most implementation details about the system. It is directly related to the actual physical structure with information about location and dimension etc. Section 3.3 briefly discusses the layout model of STT-MRAM.

2.2. BEHAVIORAL STT-MRAM MODEL

The behavioral model of STT-MRAM describes how the STT-MRAM interacts with the external world. Therefore, the adopted interface standard, input and output signals, and the associated timing relations have to be provided. Typically, this information can be found in the datasheet of an STT-MRAM product, provided by its vendor. Most of the time, a Verilog HDL model will also be provided to facilitate the design of a more complicated system integrating STT-MRAM components. Since STT-MRAM is suitable for a large variety of applications in both stand-alone and embedded forms, different STT-MRAM products will have different behavioral models (i.e., different interface, different write/read timings etc.). For example, Everspin offers 1Gb stand-alone STT-MRAM chip with DDR4 interface and it has been commercialized [70]. Avalanche lists five STT-MRAM product families up to 32 Mbit in both stand-alone and embedded forms on its website [29]. The supported interfaces include serial peripheral interface (SPI) and parallel interface (x8/x16). But these STT-MRAM products are not commercially available at the moment of writing this thesis, although the device features, datasheets, Verilog models are provided on Avalanche's website. Next, we will elaborate the behavioral model of Everspin's 1Gb stand-alone STT-MRAM product (X8) as an example.

2.2.1. STT-MRAM PACKAGE AND BLOCK DIAGRAM

As STT-MRAM offers competitive access speed and high density, it is considered as a promising candidate to replace DRAM for some applications. In 2020, Everspin released its newest and highest capacity STT-MRAM product: EMD4E001G [70]. This STT-MRAM product targets enterprise and computing applications which need high capacity, low latency, data persistence, and high endurance. Its interface complies with ST-DDR4,

a modified DDR4 version by Everspin. ST-DDR4 is physically compatible with DDR4, meaning that the AC/DC characteristics and ball/signal assignments are identical to that in DDR4 specification: JEDEC's JESD79-4A [71]. This allows the STT-MRAM product to be used as a DRAM replacement. However, the underlying technology is absolutely different from DRAM. For example, DRAM is volatile and requires periodical refresh operations, while STT-MRAM is non-volatile and does not require refresh operations. Therefore, the the ST-DDR4 specification has some deviations or enhancements when compared to the standard DDR4 specification defined for DRAM. More details about ST-DDR4 can be found in [70], which is openly accessible.

Figure 2.2a shows the package of the Everspin's 1Gb STT-MRAM chip and Figure 2.2b shows its block diagram. To command the chip to perform certain operations (e.g., read and write) or to change settings, a set of control signals are deployed, as illustrated in Figure 2.2b. For example, CK_t and CK_c are differential clock inputs; this chip runs at a clock rate of 667 MHz (i.e., clock period t_{CK}=1.5 ns). All addresses and control input signals are sampled on the rising edge of CK_t. CKE is the clock enable signal, which is active HIGH (logic '1'). RESET_n provides the chip with active LOW (logic '0') asynchronous reset command. CS_n represents the chip select signal; when it is asserted LOW, this chip is selected and commands are acceptable. ACT_n means activation command input. RAS_n is the row address strobe signal. A detailed description of all control signals can be found in the datasheet of the STT-MRAM chip [70].

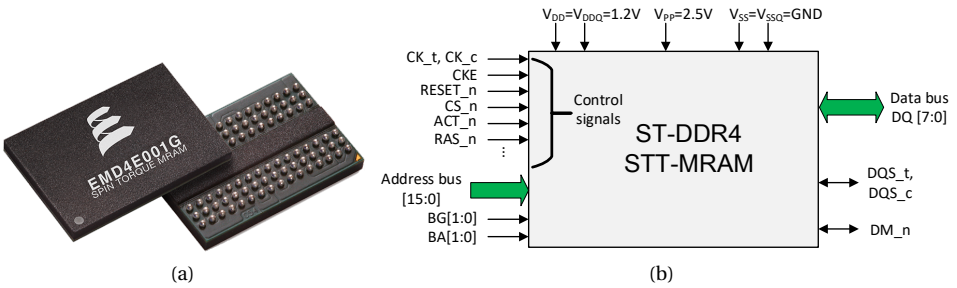


Figure 2.2: Everspin's 1Gb ST-DDR4 STT-MRAM [70]: (a) package and (b) block diagram.

To access a specific word (eight bits) in an STT-MRAM array, a row address and column address have to be provided via the input address bus which is 16-bit wide in a time division multiplexing manner, along with a 2-bit bank group address BG and 2-bit bank address BA within the addressed bank group. Note that the minimum amount of addressable data for this chip is a byte consisting of eight bits. A detailed explanation of memory organization and addressing scheme will be covered later in Section 2.3. Furthermore, the chip has a bi-directional data bus to exchange data with external devices. The data bus is 8-bit wide (named as DQ[7:0]); it is synchronized to a differential data strobe signal pairs: DQS_t and DQS_c, which run at the same frequency (667 MHz) as the input clock. In each cycle of DQS, two words are transferred on the data bus, one at the rising edge and the other on the falling edge. This is known as double data rate (DDR) mode, an effective way to double data transfer bandwidth without increasing clock frequency. The DM_n signal means input data mask; when it is sampled LOW on the rising

or falling edge of DQS, the input data is masked on the date bus during a write operation.

The power supply pins are listed at the top of the block diagram. V_{DD} is the power supply voltage: 1.2 V. V_{DDQ} is the DQ power supply with the same input voltage as V_{DD}. V_{PP} is the STT-MRAM activating power supply, which is 2.5 V. Note that an activating operation means an internal read operation to move a row/page of data from the addressed STT-MRAM array to the corresponding sense amplifier (i.e., open a page). V_{SS} and V_{SSQ} should be both grounded. Figure 2.3 shows the ball assignments of a fine ball grid array (FBGA) package, corresponding to the STT-MRAM chip shown in Figure 2.2a. The pins are located at the bottom of the chip; each signal in Figure 2.2b and its location can be found in the figure.

Row	1	2	3	4	5	6	7	8	9	Row
A	VDD	VSSQ	TDQS_c				DM_n / TDQS_t	VSSQ	VSS	A
B	VPP	VDDQ	DQS_c				DQ1	VDDQ	ZQ	B
C	VDDQ	DQ0	DQS_t				VDD	VSS	VDDQ	C
D	VSSQ	DQ4	DQ2				DQ3	DQ5	VSSQ	D
E	VSS	VDDQ	DQ6				DQ7	VDDQ	VSS	E
F	VDD	NC	ODT				CK_t	CK_c	VDD	F
G	VSS	NC	CKE				CS_n	NC	TEN	G
H	VDD	WE_n / A14	ACT_n				CAS_n / A15	RAS_n	VSS	H
J	VREFCA	BG0	A10 / AP				A12 / BC_n	BG1	VDD	J
K	VSS	BA0	A4				A3	BA1	VSS	K
L	RESET_n	A6	A0				A1	A5	ALERT_n	L
M	VDD	A8	A2				A9	A7	VPP	M
N	VSS	A11	PAR				NC	A13	VDD	N
	1	2	3	4	5	6	7	8	9	

Figure 2.3: Pin assignment of 78-ball x8 FBGA package (reprinted from [70]).

Table 2.1 summarizes the key features of the Everspin's 1Gb STT-MRAM x8 chip. Apart from the features covered previously, the following ones are worth mentioning. The chip consists of 16 banks, each of which is an STT-MRAM array of 64 Mb. The page size is 1024 bits. The bit error rate (BER) is expected to be 1×10^{-11}, considering soft errors. Cyclic redundancy check (CRC) is not supported for this product. The data retention time is three months at 70 °C. The endurance of write cycles reaches 1×10^{10}. This chip operates at a temperature range of 0 °C to 85 °C. This limits the use of this STT-MRAM product at low-temperature situations.

Table 2.1: Key features of the Everspin's 1Gb STT-MRAM x8 chip.

Data bus	8 bits	CRC	Not supported
Capacity	1 Gb (128 Mb×8)	Burst length	8
#Bank	16	Clock frequency	667 MHz
Page size	1024 bits	Interface standard	Everspin ST-DDR4
Bit error rate	1×10^{-11}	Power supply	V_{DD}=V_{DDQ}=1.2 V, V_{PP}=2.5 V
Data retention	3 months @ 70 °C	Package	FBGA (78 balls)
Endurance	1×10^{10}	Operating Temp.	[0, 85]°C

2.2.2. ST-DDR4 OPERATIONS AND TIMING DIAGRAMS

As mentioned previously, ST-DDR4 STT-MRAM has the same physical interface as DDR4 DRAM (e.g., same pin assignments). This allows STT-MRAM to gradually enter into the memory market where data persistence and low power are required, in replacement of DRAM with minimum changes in hardware. However, STT-MRAM is a absolutely different technology from DRAM, meaning that some operations and their timing constraints have to be redefined or modified. This implies that DRAM controllers are not compatible with STT-MRAM. In order to deploy STT-MRAM chips in a system, the memory controller has to be re-designed.

This section introduces some key operations and the associated timing diagrams, as defined in the ST-DDR4 specification [70]. Figure 2.4 shows a state machine, describing all possible states of the memory chip and commands that invoke certain operations to transition between these states. Each command represents a unique combination of digital values on the control signals such as CS_n, ACT_n, and RAS_n. The detailed de-

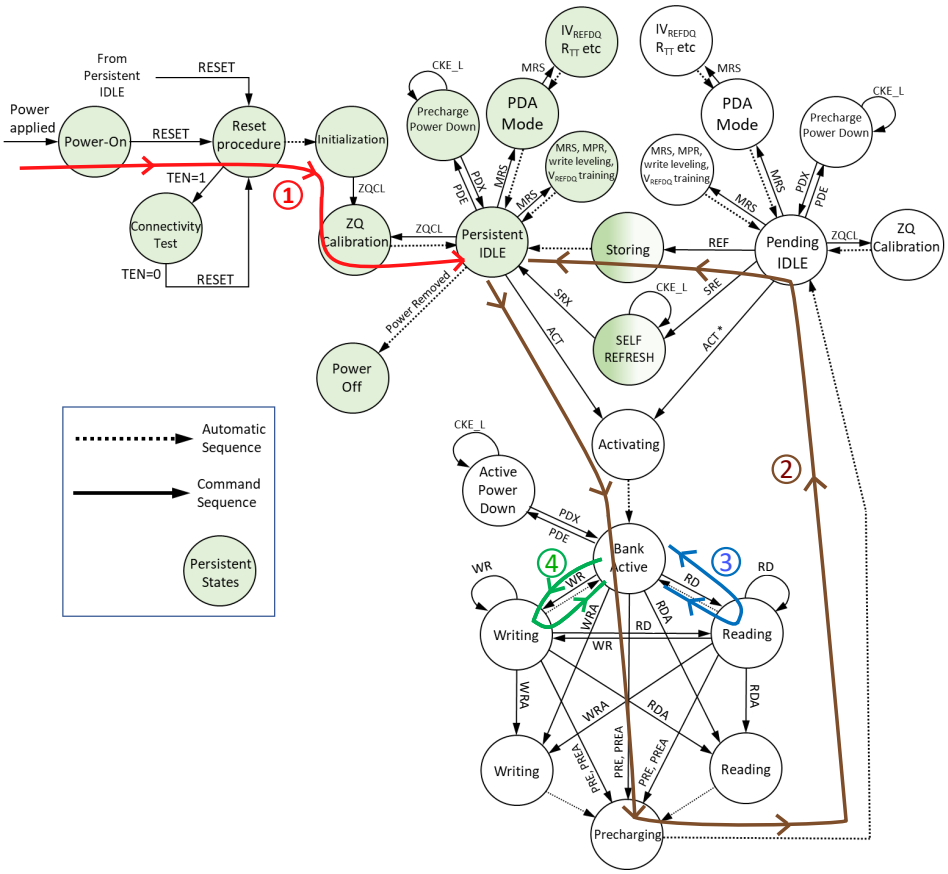


Figure 2.4: Simplified state machine for operations in the Everspin's ST-DDR4 STT-MRAM chip (reprinted from [70] with some added marks).

scription of these signals and the command truth table can be found in [70, 71]. Note that as long as a legal command is issued and the corresponding signal timings comply with specifications, the chip will transition from one state to another as marked with the arrows in the figure. Next, we will select four operation sequences to elaborate some key operations, as marked with colored paths and circled numbers in the figure. We assume that readers have a background on DRAM fundamentals. For those who find this section difficult to understand, you are directed to some DRAM books or technical materials such as [69, 71–73].

Operation sequence ①: reset and initialize

After power-up, the ST-DDR4 STT-MRAM chip needs a reset operation to initialize it to a known default state (i.e., "Persistent IDLE" as shown in Figure 2.4), prior to normal operations. Figure 2.5 illustrates the timing diagram of all relevant signals for this operation sequence.

The first command that the chip should receive is "RESET" after the power supply is applied. This is done by pulling down the RESET_n signal for at least 200 μs (i.e., $t_{PW\_RESET\_L}$>200 μs) with stable power. Then, CKE has to be pulled LOW before RESET_n being de-asserted (T(MIN)=10 ns) to disable the input clock signals. After RESET_n is de-asserted, wait for 500 μs until CKE becomes active. During this time period, the chip will start internal initialization, which is a process independent on the clock. After the asynchronous reset operation, clock signals CK_t and CK_c need to be started and stabilized for at least t_{CKSRX}=10 ns before CKE goes HIGH. Since CKE is a synchronous signal, the corresponding setup time (t_{IS}=115 ps) to the clock edges must be met. In addition, a "DES" (chip deselect) command must be registered at the first enabled clock edge, as shown in the figure. Once the CKE is asserted after reset, it needs to continuously stay HIGH until the chip enters into "Persistent IDLE" state. The on-die termination (ODT)

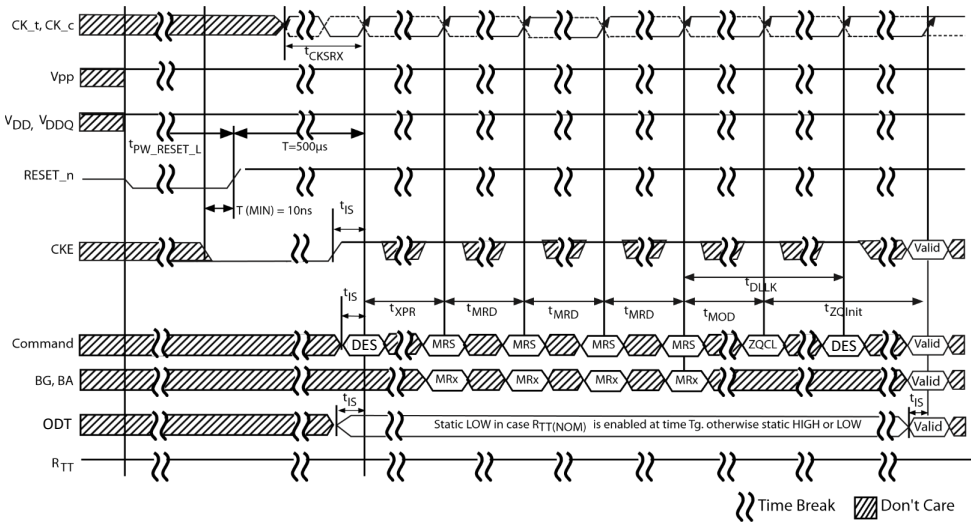


Figure 2.5: Timing diagram of reset and initialization after power-up [70].

input signal may be in undefined state until t_{IS} before CKE is registered HIGH. ODT is a feature that enables the chip to change termination resistance (R_{TT}) for each DQ, DQS, and DM_n signal by setting a value in the mode register MR5 [70]. When CKE is registered HIGH, the ODT input signal must be statically held LOW if R_{TT} is to be enabled. Otherwise, ODT may be statically held at either LOW or HIGH. In both cases, ODT has to remain static until reaching the "Persistent IDLE" state.

After CKE is registered HIGH, seven "MRS" (mode register set) commands have to be registered consecutively to set the mode registers: MR0-MR6. Before the arrival of the first "MRS" command, a minimum time t_{XPR}=5·t_{CK} is required after CKE is registered HIGH. For each two consecutive "MRS" commands, MRS command cycle time t_{MRD}>8·t_{CK} is required.

Finally, a "ZQCL" (ZQ calibration long) command is issued to start ZQ calibration, which is an automatic process to tune the pull-up and pull-down resistors to exact 240 Ω for the DQ driver [74]. After delay-locked loop locking time t_{DLLK}=597·t_{CK} and ZQ calibration time t_{ZQInit}=1024·t_{CK}, the chip will be in the "Persistent IDLE" state, ready for normal operations.

Operation sequence ②: activate, precharge, and refresh

The second operation sequence consists of an activating operation, a precharging operation, and a refreshing operation. The timing diagram is shown in Figure 2.6. When the chip is in the "Persistent IDLE" state, the bank activate ("ACT") command is used to open a row (also referred to as a page) in a particular bank for subsequent accesses. In other words, a row of data is moved from an STT-MRAM array to its sense amplifier. The BG inputs (2 bits) select a bank group, the BA inputs (2 bits) select a bank with the selected bank group, and the row address provided on the address bus (16 bits) selects a row within the selected bank. Note that each bank in the chip is independent and different banks may stay in different states in Figure 2.4; this is a key feature of DRAM to increase data throughput via bank interleaving [73]. With the activating operation, the addressed bank transitions to the "Bank active" state, as shown in Figure 2.4. The "Bank active" state is non-persistent (volatile), since the sense amplifier is built with transistors. In contrast, the "Persistent IDLE" state is persistent (non-volatile), since STT-MRAMs are intrinsically non-volatile. This is a key difference from DRAM. After a particular STT-

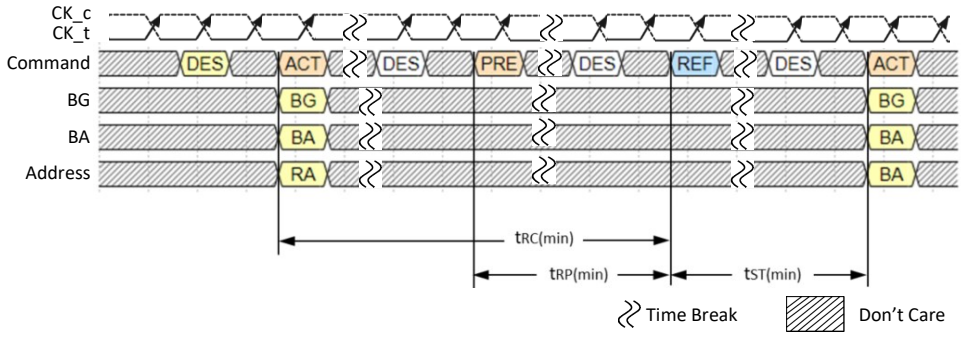


Figure 2.6: Timing diagram of activating, precharging, and refreshing operations (reprinted from [70]).

MRAM page is open, it is ready for read or write accesses from the outside. This will be explained in the third and fourth operation sequences, respectively.

The single bank precharge ("PRE") command is used to deactivate an open row. Note that a "PRE" command has to be issued if a different row in the same bank needs to be opened. With the precharging operation, the chip transitions from "Bank active" to "Precharging", and automatically goes to the "Pending IDLE" state after the precharing process completes (see Figure 2.4). Before the next command can be issued, a minimum time period t_{RP}(min)=143 ns for the precharging operation is required. In addition, another timing constraint t_{RC}(min)=143 ns (RAS_n cycle time) has to be met before a follow-up command in the "Pending IDLE" state. When the chip is in the "Pending IDLE" state, an "ACT" command is required to activate the row again prior to any read or write operations.

The refreshing operation for the ST-DDR4 STT-MRAM chip is a big difference from that for DDR4 DRAM chips. It is well known that DRAM requires periodical refreshing operations (~64 ms) to retain the stored data in capacitors. In contrast, STT-MRAM is intrinsically non-volatile, thus it does not need any refreshing operations. To maximize the compatibility with the DDR4 specification, the ST-DDR4 specification keeps the refreshing operation. But what it does internally in STT-MRAM chips is absolutely different from the refreshing operation in DRAM chips. If a "REF" command is issued and MR3[8]=1 when the selected bank is in the "Pending IDLE" state, an internal store operation will be executed. Depending on the signal on A10, the store operation takes different actions. If A10 is sampled LOW, the store operation moves the open page of the addressed bank (determined by BG and BA) from the sense amplifier to the STT-MRAM array. If A10 is sampled HIGH, the store operation moves all open page of all banks to the corresponding STT-MRAM arrays. If MR3[8]=0, the "REF" command is ignored. By adding an internal store operation, the refreshing operation resembles a write-back operation in the cache coherence protocol. Here, the store operation writes back the cached data in the sense amplifier (page buffer) to the STT-MRAM array. After the refreshing operation, the selected bank or the entire STT-MRAM chip transitions from "Pending IDEL" to "Persistent IDLE", and thus ensures data persistence. Note that an internal store operation takes minimum execution time t_{ST}(min)=380 ns.

Operation sequence ③: burst read

The read operation defined in the ST-DDR4 specification is burst-oriented. A burst read operation reads data at multiple addresses of a particular open page of a selected bank in one shot. The burst length is 64 bits (8 consecutive bytes) for Everspin's 1Gb ST-DDR4 STT-MRAM X8 chips. Figure 2.7 shows the timing diagram of a burst read operation. It starts with issuing a read ("RD") command. Meanwhile, the address on the bank group BG and bank BA signals are sampled to determine the target bank which has an open page ready for read. The column address CA[6:0] on the address bus [6:0] determines the eight consecutive bytes to be read in the addressed open page. If Address[10]=1, an auto precharging operation is added following the burst read operation. If Address[10]=0, no precharging operation is added and the bank returns to the "Bank active" state after the bust read operation completes (see path ③ in Figure 2.4). After the "RD" command is registered, a read latency t_{RL}=10·t_{CK} is expected before the first byte of data appears on

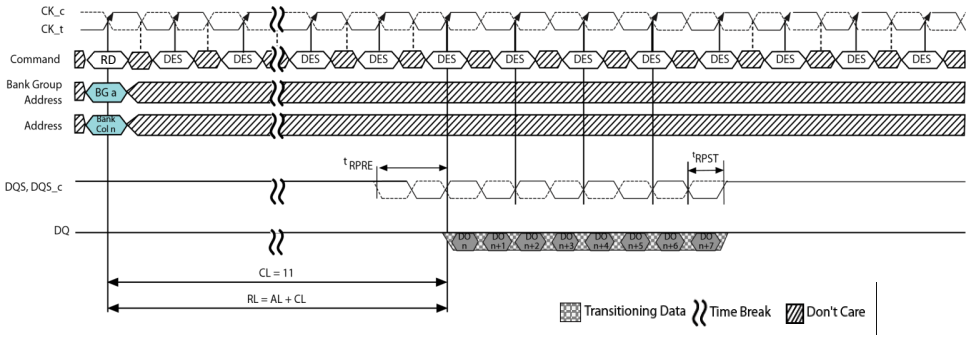


Figure 2.7: Timing diagram of a burst read operation (reprinted from [70]).

the DQ data bus. In read mode, the DQ strobe signals DQS_t and DQS_c are outputs from the STT-MRAM chip. They run at the same frequency and phase as the chip clock signals CK_c and CK_t. DQS_t and DQS_c precede the DQ signals with a clock cycle (i.e., "RD" preamble t_{RPRE}=1·t_{CK}) and they are edge-aligned, as shown in the figure. For this chip, the addressed open page (1kb) is divided into 16 chucks, each of which is 64 bits (i.e., 8 consecutive bytes). Depending on the value of CA[6:3], one of the 16 chucks is selected and read out sequentially from the first byte. Note that the lowest three bits of the row address CA[2:0] defines the data output order with a burst, as defined in the ST-DDR4 standard. But for this chip, this option is not supported, and the output order is fixed (i.e., CA[2:0] is ignored). Within a burst read of 8 bytes, the first byte outputs first and the last byte outputs last. The addressing scheme will be further examined in the next section. The minimum pulse width of "RD" postamble t_{RPST}(MIN)=0.33·t_{CK}.

Operation sequence ④: burst write

The write operation defined in the ST-DDR4 specification is also burst-oriented. A burst write operation program data to multiple addresses of a particular open page of a selected bank in one shot. The burst length is also 64 bits (8 consecutive bytes) for Everspin's 1Gb ST-DDR4 STT-MRAM X8 chips. Figure 2.8 shows the timing diagram of a burst write operation. It starts with issuing a write "WR" command. Meanwhile, the address on the bank group BG and bank BA signals are sampled to determine the target bank

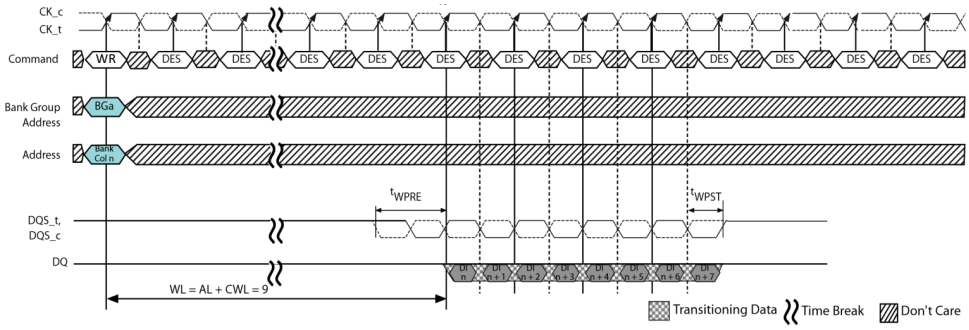


Figure 2.8: Timing diagram of a burst write operation (reprinted from [70]).

which has an open page. The column address CA[6:0] on the address bus [6:0] determines the eight consecutive addresses to be written to in a selected open page. If Address[10]=1, an auto precharging operation is added following the burst write operation. If Address[10]=0, no precharging operation is added and the bank returns to the "Bank active" state after the bust write operation completes (see path ④ in Figure 2.4). After the "WR" command is registered, a write latency t_{WL}=9·t_{CK} is expected before sending the first byte of data on the DQ data bus. In write mode, the DQ strobe signals DQS_t and DQS_c are inputs to the STT-MRAM chip. They should be set at the same frequency and phase as the chip clock signals CK_c and CK_t. DQS_t and DQS_c should precede the DQ signals with a clock cycle (i.e., "WR" preamble t_{WPRE}=1·t_{CK}) and they are center-aligned, which is different from the alignment in a read operation. For this chip, the addressed open page (1kb) is divided into 16 chucks, each of which is 64 bits (i.e., 8 consecutive bytes). Depending on the value of CA[6:3], one of the 16 chucks is selected and overwritten sequentially by the 8 bytes of data on DQ. Again, the lowest three bits of the row address CA[2:0] is ignored. The minimum pulse width of "WR" postamble t_{WPST}(min)=0.33·t_{CK}.

Comparison between JEDEC DDR4 and ST-DDR4 specifications

From the above brief introduction into some key operations and timing diagrams of ST-DD4 STT-MRAM, we can see that the Everspin ST-DDR4 specification is actually a variant of the JEDEC DDR4 specification. All commands in the ST-DDR4 specification are borrowed from the JEDEC DDR4 specification for DRAM. However, there are some key differences that worth paying attention to. First, some features in the JEDEC DDR4 specification are not supported in the ST-DDR4 specification. For example, the JEDEC DDR4 specification offers four speed bin options while the ST-DDR4 specification only supports one, as listed in the first row of Table 2.2. The speed bin indicates the data transfer speed, that is, the number of transfers per second per pin (in units of MT/s). "-1333" means 1333M transfers per second per pin; due to the DDR data transfer mode (i.e., two transfers per clock cycle), this number also implies the clock frequency is 667 MHz and clock period is 1.5 ns. Similar interpretations can be derived for other speed bins in the table. Second, the JEDEC DDR4 specification only defines a single idle state, while the ST-DDR4 defines a "Pending IDLE" state for page buffer which is volatile and a "Permanent IDLE" state for STT-MRAM cells which are non-volatile (see Figure 2.4). Third,

Table 2.2: Comparison of timing parameters defined in Everspin's ST-DDR4 and JEDEC's DDR4 specifications.

Parameter	Description	JEDEC DDR4 (DRAM)	ST-DDR4 (STT-MRAM)
Speed bin	data transfer speed	-1600, -1866, -2133, -2400	-1333 only
t_{CK} (ns)	clock period	1.25, 1.071, 0.938, 0.833	1.5
t_{RCD} (ns)	ACT to internal read or write delay time	min=12.5	min=135
t_{RC} (ns)	ACT to ACT or REF command period	min=44.5	min=190
t_{RAS} (ns)	ACT to PRE command period	min=32	min=143
t_{RP} (ns)	PRE command period	min=12.5	min=7.5
t_{ST} (ns)	Internal store operation period	Not applicable	min=380

the internal actions for executing the "REF" command are different. For DRAM, this command refreshes the charge stored in cell capacitors to retain data. In contrast, this command evokes an internal store operation which moves a row (page) of data from the page buffer to the persistent STT-MRAM array. Fourth, the values of most timing parameters are diffident. Table 2.2 compares the required minimum values of some key timing parameters. It can be seen that t_{RCD}, t_{RC}, and t_{RAS} are all defined much larger in the STT-DDR4 specification than the JEDEC DDR specification, but t_{RP} is smaller in the STT-DDR4 specification. This suggests that the Everspin' ST-DDR4 STT-MRAM chip is still much slower than DDR4 DRAM products in terms of write/read latency.

2.3. FUNCTIONAL STT-MRAM MODEL

A memory chip can be internally divided into a number of function blocks, each of which fulfills a specific function. These functional blocks are interconnected and compactly distributed in the memory chip. They interact with each other and together achieve the chip-level behavior interacting with the external world as discussed in the previous section. This section presents the functional model of the Everspin's 1Gb ST-DDR4 STT-MRAM chip (Figure 2.2a). At this abstraction level, we open the black box in Figure 2.2b to examine all internal functional blocks of the chip and how they behave individually and interact with each other.

2.3.1. FUNCTIONAL BLOCK DIAGRAM

Figure 2.9 illustrates a simplified functional block diagram for STT-MRAM. The functional block diagram distinguishes several functional blocks needed for the STT-MRAM to operate properly, including memory arrays, control logic, address decoders, data buffers, IO circuits. Next, we explain these functional blocks in detail.

1) Memory arrays: This functional block is responsible for storing data and occupies the majority area of an STT-MRAM chip. Therefore, it is considered as the most important part in an STT-MRAM chip. A memory array is organized in the form of a matrix with n rows and m columns of STT-MRAM cells. Since each STT-MRAM cell typically stores 1 bit of data, an $n \times m$ memory array contains $n \times m$ bits of data. Due to the relatively large size of this functional block and its importance as data-storing medium, it become the main research focus for fault analysis and test development.

2) Control logic: This functional block accepts the external control signals (e.g., RESET_n, CKE, CS_n, ACT_n) and translates them into internal commands (e.g., "ACT", "WRITE", and "RD"). Depending on the generated internal command, the control logic activates certain functional blocks and generates internal control signals meeting all timing constraints to execute the command. The control logic also contains mode registers and address register. The mode registers accept settings of the chip via the control signals as well as the address bus during the initialization process, as shown in Figure 2.5. The address register is used to hold the row or column address, provided on the address bus when "ACT", "WR", or "RD" command is received.

3) Address decoders: This functional block decodes the provided address and selects a particular word of data for subsequent write or read operations. The address decoders can be generally divided into two classes: row address decoder and column address de-

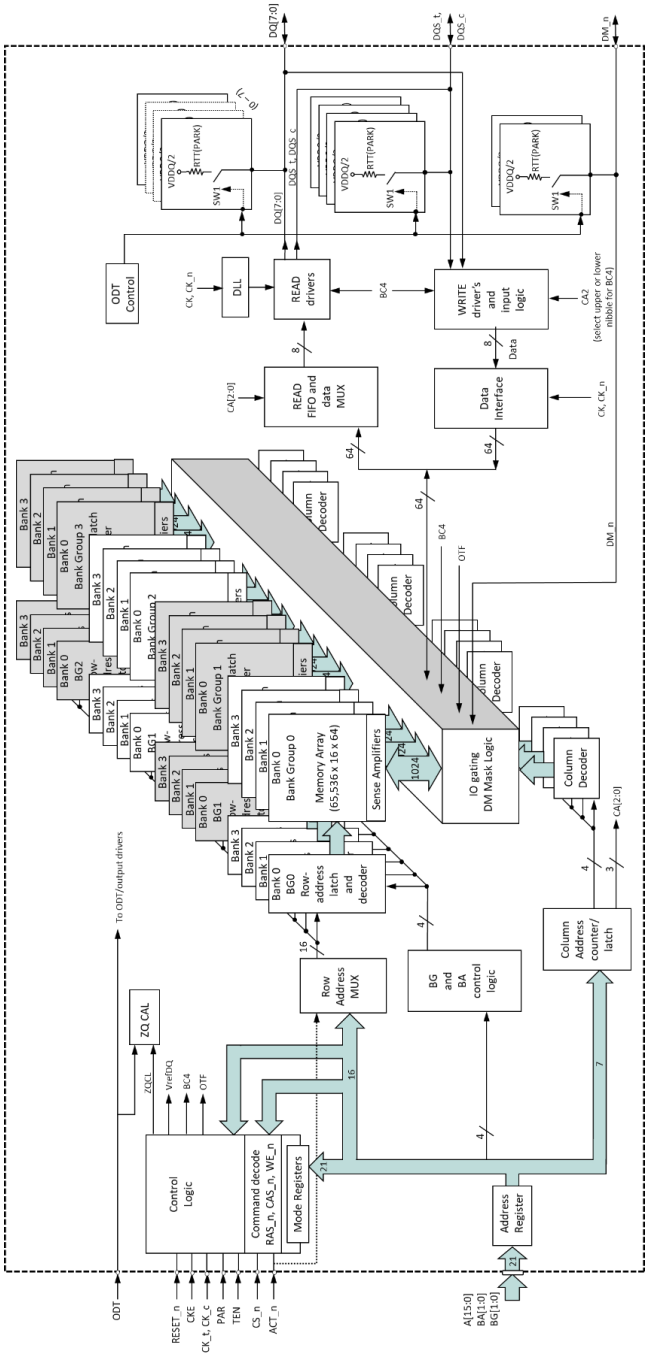


Figure 2.9: Simplified STT-MRAM functional block diagram (reprinted from [70]).

coder. A row address decoder decodes the row address and activates a particular word line (WL) for the "ACT" command. Each WL is connected to a row of STT-MRAM cells. A column address decoder decodes the column address and selects a specific word out of the active page in the page buffer selected by the row address. Accordingly, the widths of row address and column address are correlated with the organization of memory array. Assuming an memory array of $n \times m$ bits and the word size is s bits, the width of row address is $W_{\mathrm{RA}} = \log_2(n)$ and the width of column address is $W_{\mathrm{CA}} = \log_2(\frac{m}{s})$.

4) Data buffers: This functional block buffers the input data to the memory and output data from the memory. Typically, the data buffers are implemented as first-in-first-out (FIFO) buffers. Sometimes, the data buffers also plays a critical role in moving data between two different clock domains and involve the conversion of data bus width. For example, the input of a FIFO buffer accepts data chucks with width of 64. The FIFO buffer converts the 64-bit data chunks into 8-bit data chunks and output them at a faster clock rate. The output order of the 8 bytes corresponding to a 64-bit input data chunk can be configured. More commonly, the output order is sequential and the first byte is the lowest 8 bits of the 64-bit data chunk.

5) Date IO circuits: This functional block controls the data input and output interface, which reliably transmits data to and receives data from the external devices via the pins DQ[7:0], DQS_t, DQS_c, and DM_n. The data IO circuits includes: a delay-locked loop (DLL), write&read drivers, and ZQ calibration circuits. The DLL is responsible for adjusting the timing relations between: 1) CK_t and CK_c, 2) DQS_t, DQ_c, and 3) DQ[7:0] under the influence of process, voltage, and temperature variations. The DLL ensures that the output data strobe signals DQS_t and DQ_c are always edge-aligned with the DQ data bus when transmitting data out of the chip, as illustrated in Figure 2.7. The write and read drivers drive data transmission between the internal data buffers and the chip pins. The ZQ calibration circuit performs ZQ calibration used for the the output read drivers and the on-die termination (ODT) [74].

2.3.2. ORGANIZATION OF MEMORY ARRAYS

In this part, we delve more into the organization of memory arrays in the Everspin's 1Gb STT-MRAM chip. We first discuss how the 1Gb data is divided and stored in different STT-MRAM arrays, and how each array is internally structured. Thereafter, we elaborate how STT-MRAM cells are interconnected in an array and how they are connected to the peripheral circuits.

Memory organization

As shown in Figure 2.9, the STT-MRAM chip comprises four bank groups, each of which includes four banks. Each bank is a rectangular STT-MRAM array. The four bank groups are addressed by a 2-bit bank group address BG[1:0], and the four banks in each bank group are addressed by a 2-bit bank address BA[1:0]. It can be seen in the figure that each bank is an independent memory unit which has its own peripherals such as row address decoder, column address decoder, and sense amplifier. This allows for bank interleaving to hide the latency of accessing each bank and thus significantly boost data throughput in the data bus.

Each bank stores 64 Mb of data; it is structured as a matrix of 64k rows (i.e., n=64k)

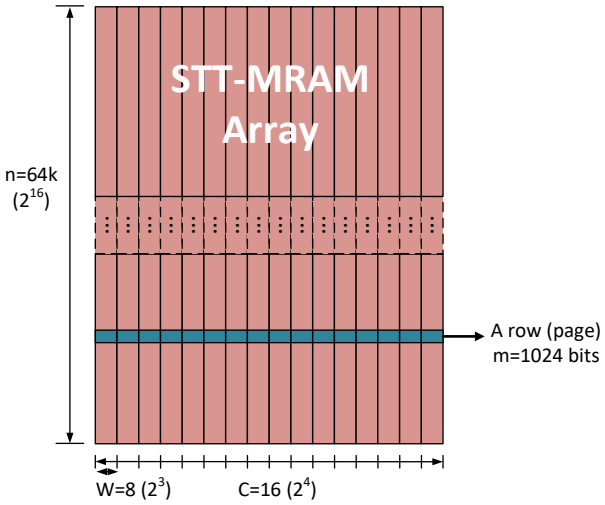


Figure 2.10: A bank of STT-MRAM cells (64 Mb).

and 1k columns (i.e., m=1k) of STT-MRAM cells, as shown in Figure 2.10. Therefore, the width of row address W_{RA} is 16 and the width of column address W_{CA} is 7. The reason why n is much larger than m is that each column requires a sense amplifier and a write driver while the peripheral circuit on the row side is much simpler. The rectangular array shape maximizes the area of STT-MRAMs while minimizing the area of peripheral circuits. A row of STT-MRAM cells is also referred to as a memory page, which contains 1024 bits of data. Each page is divided into 16 data chunks (C=16) with equal size (W=64 bits=8 bytes). The 16 data chucks are addressed by the 4 most significant bits of the column address CA[6:3] and the 8 bytes within each data chunk are addressed by the 3 least significant bits CA[2:0], as illustrated in Figure 2.9.

Memory connection

Each STT-MRAM cell in an array is connected to the peripheral circuits via three different types of connecting line: word line (WL), bit line (BL), and source line (SL). Figure 2.11a shows the way a single STT-MRAM cell is connected to a WL, a BL, and a SL. The WL controls the access to the cell. When the WL is asserted to select the cell, the voltages on the BL and SL determine which operation is performed on the cell. There are three basic operations: write '1', write '0', and read for accessing an STT-MRAM cell; this will be detailed in the next chapter. Figure 2.11b shows how an $n \times m$ STT-MRAM array is connected to the peripheral circuits via WLs, BLs, and SLs. It can be seen that each row of STT-MRAM cells are connected to an individual WL. For a given row address RA, a specific WL is activated by the row address decoder. When a WL is activated, all STT-MRAM cells connected to that WL are selected simultaneously and will be accessed together. For the "ACT" command, the entire row of STT-MRAMs (1 kb) are read out to the page buffer in the addressed bank. This is achieved by the sense amplifiers which generate appropriate voltages on the BLs and SLs. This leads to a small sensing current flowing through each cell. By comparing the cell current to a fixed reference current going through a reference cell, the stored data (logic 0 or 1) in each cell can be sensed. For the store operation

evoked by the "REF" command, the 1 kb data in the page buffer of the addressed bank is moved back to the addressed row in the STT-MRAM array. This is achieved by the write drivers shown in Figure 2.11b. They generate appropriate voltages on the BLs and SLs, leading to a relatively large spin-polarized current flowing through each memory cell. The write current switches the state of the memory cell under the spin-transfer torque effect. More details about the write '1', write '0', and read operations on the STT-MRAM cell will be explained in more details in the next chapter.

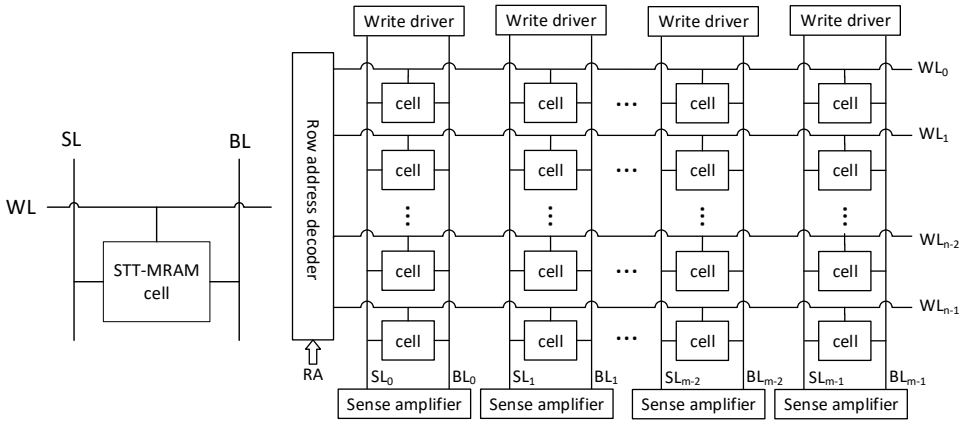


Figure 2.11: Memory connection: (a) a single STT-MRAM cell and (b) an $n \times m$ array of STT-MRAM cells.

2.3.3. INTERNAL BEHAVIOR

As discussed in Section 2.2, Today's high-speed DRAM chips and the Everspin's STT-MRAM chips support a variety of mode settings and operations. From a memory chip point of view, it communicates with the external world via control signals, address bus, and data bus. Internally, the memory chip samples the logic values appeared on these signals, all synchronized to the input clock. The sampled logic values on the control signals are translated into internal commands by the control logic block. Depending on the translated command, the control logic block selectively activates relevant function blocks to execute the command. The execution of internal commands induces data flows between the persistent STT-MRAM arrays, page buffers, and data bus. All internal activities can be described by the state machine shown in Figure 2.9.

Internal commands

Next, we summarize all internal commands that are mentioned in this chapter; the corresponding mapping relationships with the control signals and addresses are listed in Table 2.3.

1) Chip deselect (DES): This command is used to deselect an STT-MRAM chip; it can also used to mask other control signals when the chip is in processing of an existing command and has not finished yet. When the chip select signal CS_n is pulled HIGH and the clock enable signal CKE is also HIGH, this command will be registered by the control logic at the rising edge of the clock. Other control signals and address bus do not

Table 2.3: Command truth table.

Command	Abbr.	CKE	CS_n	ACT_n	RAS_n	CAS_n/ A15	WE_n/ A14	BG[1:0]	BA[1:0]	BC_n/ A12	A13, A11	AP/ A10	A[9:7]	A[6:0]
Chip deselect	DES	H	H	X	X	X	X	X	X	X	X	X	X	X
Mode register set	MRS	H	L	H	L	L	L	BG	BA	Opcode				
ZQ calibration long	ZQCL	H	L	H	H	H	L	V	V	V	V	H	V	V
Bank activate	ACT	H	L	L	V	RA[15:14]		BG	BA	RA [13:0]				
Single bank precharge	PRE	H	L	H	L	H	L	BG	BA	V	V	L	V	V
Precharge all banks	PREA	H	L	H	L	H	L	V	V	V	V	H	V	V
Refresh	REF	H	L	H	L	L	H	BG	BA	V	V	V	V	V
Read (fixed BL8 or BC4)	RD	H	L	H	H	L	H	BG	BA	V	V	L	V	CA[6:0]
Write (fixed BL8 or BC4)	WR	H	L	H	H	L	L	BG	BA	V	V	L	V	CA[6:0]

Notes:
1. "H" stands for logic "1" and "L" stands for logic "0".
2. "V" means a defined logic state, either "H" or "L".
3. "X" means "don't care".

play a role in determining the "DES" command. Therefore, they are all marked as "X", meaning "don't care", as shown in the first row of the table.

2) Mode register set (MRS): This command is used to configure the mode registers MR0-MR6. The corresponding values of control signals and address bus are listed in the second row of the table. CKE and CS_n are both asserted. The row activate signal ACT_n must be de-asserted at HIGH. The row address strobe signal RAS_n, column strobe signal CAS_n, and write enable signal WE_n are all asserted at LOW. Note that CAS_n and WE_n are both multi-functional. They also represent two bits of the address bus A15 and A14 for the command "ACT". The values of bank group address BG[1:0] and bank address BA[1:0] designate which mode register to write into. The values on A[12:0] are configuration opcode. The detailed introduction of these mode registers and the associated configurations can be found in [70].

3) ZQ calibration long (ZQCL): This command performs ZQ calibration during chip reset and initialization as mentioned in Section 2.2.2.

4) Bank activate (ACT): This command activates a specific row in a target bank by asserting the WL of that row. It also moves 1kb data stored in the selected row to the sense amplifier (page buffer). To generate an "ACT" command, CKE, CS_n, and ACT_n should all be asserted. RAS_n can be in 'H' or 'L' (denoted as 'V' in the table). BG and BA select a bank group and a bank within the selected group. The row address in the selected bank is provided on A[15:0].

5) Single bank precharge (PRE): This command precharges a single bank to close its open page. To generate such a command, ACT_n and CAS_n are de-asserted. CKE, CS_n, RAS_n, and WE_n are asserted. The auto-precharge signal AP, shared with A10, has to be at 'L'. BG and BA together select a target bank. The rest of address bits are at 'V', that is, either 'H' or 'L' is OK. Note that if the AP signal is at 'H', the "PRE" command transforms to a precharge all banks (PREA) command (i.e., close all open pages). Therefore, BG and BA are ignored; they can be at 'V'.

7) Refresh (REF): This command for STT-MRAM is a key difference from the refresh command for DRAM. Unlike the periodical refresh operation to retain charges in DRAM cell capacitors, the "REF" command here evokes different actions depending on the setting of MR3[8] and the value on AP/A10. If MR3[8]='H' and A10='L', a store operation is performed, which moves the open page of the addressed bank (determined by BG and

BA) from the page buffer to the persistent STT-MRAM array. If MR3[8]='H' and A10='H', the store operation moves all open pages of all banks to STT-MRAM array. If MR3[8]='L', the "REF" command is ignored. Note that executing the store operation takes at least 380 ns, which is slowest operation among all.

8) Read (RD): This command reads a burst of four bytes (BC4 mode) or eight bytes (BL8 mode) from the 1kb page buffer of a selected bank and transmits the readout data out of the chip via the DQ data bus. The read mode can be configured by setting MR0[1:0]. If MR0[1:0]=00, it means the burst read is fixed at 8 bytes (i.e., fixed BL8 mode). In other words, each "RD" command returns a fixed amount of data, which is 8 bytes. If MR0[1:0] = 10, it means the burst read is fixed at 4 bytes (i.e., fixed BC4 mode). If MR0[1:0] = 01, it indicates that the burst length is not fixed and can be selected on the fly (i.e., BL8/BC4 OTF mode). In this case, the burst chop signal BC_n, shared with A12, determines the burst length. If BC_n='L', it means BC4; 'H' means BL8. In addition to the setting of read burst length, an auto-precharge command can be merged to the "RD" command. This is done by setting AP/A10='H'. In this case, the 'RD' command becomes 'RDA', meaning that a precharge operation is automatically performed once the read operation completes, as illustrated in Figure 2.4. For the ease of illustration, we limit our discussion to the simplest read mode: fixed BL8, no auto-precharge. Therefore, MR0[1:0] should be set to 00, BC_n at 'V', and AP at 'L'. BG and BA designate the target bank which is supposed to be at the "Bank active" state. The column address CA[6:0] determines the address of the first byte of data within the selected open page. The values of the other signals are shown in Table 2.3.

9) Write (WR): This command write a burst of four bytes (BC4 mode) or eight bytes (BL8 mode) into 4 or 8 sequential addresses in the 1kb page buffer of a selected bank. The burst write modes include fixed BL8, fixed BC4, and BL8/BC4 OTF; the configuration method is the same as the burst read operation. The values of all relevant control signals and addresses are shown in the table.

Apart from the above 8 commands, there are many other commands defined in the ST-DDR4 and the DDR4 specifications. The complete list of internal commands and command truth table can be found in [70, 71]. It is important to note that these internal commands are not independent from each other. In other words, they cannot be issued any time or any order as desired. They need to be issued in an order described by the state machine shown in Figure 2.9. In addition, timing contrarians should be strictly met as discussed in Section 2.2.2. For example, a complete write operation which programs data into a target STT-MRAM array consists of a series of internal commands shown in Figure 2.12. It starts with issuing an "ACT" command for a bank at the "Persistent IDLE"

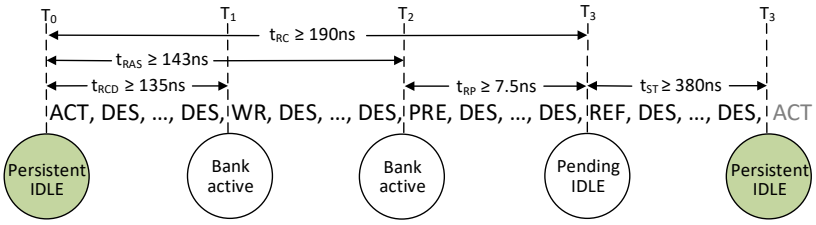


Figure 2.12: A complete write operation programing data into an STT-MRAM array.

state at time T_0. This command opens a particular row of the selected bank. Next, a "WR" command is issued at T_1 to write a burst of data (8 bytes or 4 bytes) into the addressed locations in the open page. After completing the "WR" command, the bank returns to the "Bank active" state. Next, a "PRE" command is issued at T_2 to make the active bank transition to "Pending IDLE". Finally, a "REF" is issued at T_3; it evokes a store operation which moves the dirty open page back to the STT-MRAM array. After the store operation completes, the bank transitions from "Pending IDLE" to "Persistent IDLE". In this process, the following timing constraints need to be met: 1) $T_1 - T_0 = t_{RCD} \leq 135\,ns$; 2) $T_2 - T_0 = t_{RAS} \leq 143\,ns$; 3) $T_3 - T_2 = t_{RP} \leq 7.5\,ns$; 4) $T_3 - T_0 = t_{RC} \leq 190\,ns$; 5) $T_3 - T_0 = t_{RC} \leq 380\,ns$. The detailed description of these timing parameters can be found in Table 2.2.

Internal data flow

Internally in the Everspin's 1Gb STT-MRAM chip, there are three places where data can reside. They are FIFO data buffers, page buffers (sense amplifiers), and STT-MRAM arrays, as shown in Figure 2.13a. Note that FIFO data buffers and page buffers are volatile while STT-MRAM arrays are non-volatile; the memory size relation is: STT-MRAM array > page buffers > FIFO data buffers. In response to different internal commands, data flows between these three memory components. Let us discuss the data flow involved in the command sequence in Figure 2.12 as an example. The first command "ACT" moves a row of data (1 kb) from the addressed bank (i.e., STT-MRAM array) to its page buffer. Subsequently, the "WR" command accepts data byte-wise via the DQ data bus and buffers

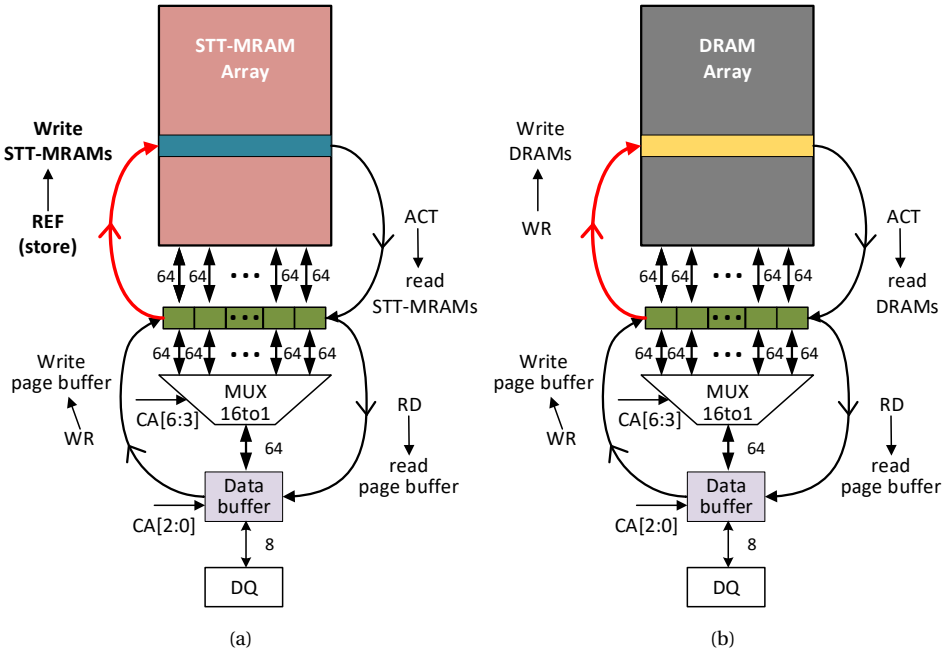


Figure 2.13: Comparison of internal data flows with different chip commands between (a) STT-MRAM and (b) DRAM.

the received data in the FIFO data buffer. The FIFO data buffer converts the received 8 bytes of data into a 64-bit data chuck and then send it to the right location in the addressed page buffer. After precharging the active bank, the "REF" command writes the entire open page back to its corresponding row in the STT-MRAM array. If a "RD" command is issued when the addressed bank is in "Bank active" state, it reads out data from the right location in the page buffer and sends the data to the FIFO data buffer, as depicted in Figure 2.13a.

Similarly, Figure 2.13b shows the internal data flow in a typical DDR4 DRAM chip. It can be seen that the "ACT" and "RD" commands on the right side take the same actions as those in the STT-MRAM chip in Figure 2.13a. The key differences between the data flows in STT-MRAM and DRAM lie in the "WR" and "REF" commands. For DRAM, the "WR" command writes the received data to both the open page buffer and the DRAM array. The "REF" command does something else; it periodically refreshes the charges stored in DRAM cell capacitors to retain data. In contrast, the "WR" command for STT-MRAM only writes the received data to the open page buffer. The data movements from the open page buffer to the STT-MRAM array is actually done by the internal store operation which is evoked by the "REF" command.

3

STT-MRAM TECHNOLOGY AND IMPLEMENTATION

STT-MRAM technology has gone through a long evolution process since the discovery of TMR effect by Julliere in 1975. Since then, significant breakthroughs have been made at various aspects: device engineering, switching mechanisms, manufacturing process, and peripheral circuit designs. Thanks to technology advancements in these fields, STT-MRAM has entered into a mature phase and has been commercialized by several semiconductor companies. This chapter focuses on STT-MRAM technology and implementation. First, we introduce the basic organization of MTJ, the data storage element in STT-MRAMs, and its working principles. Second, we elaborate the electrical STT-MRAM model that is adopted in SPICE-based circuit simulations in this dissertation. Third, we briefly discuss the STT-MRAM layout model. Fourth, we examine the STT-MRAM manufacturing process and the associated defects in each step. Manufacturing defects related to the fabrication and integration of MTJs are emphasized since they are unique to STT-MRAMs and typically do not occur in the conventional memories, thus requiring special attention from a test perspective. Finally, we concisely review the development history of MTJ and the commercialization attempts of four generations of MRAM. The potential applications of STT-MRAM and its remaining challenges are also discussed.

Parts of this chapter have been published in TETC'19 [46].

3.1. MTJ TECHNOLOGIES

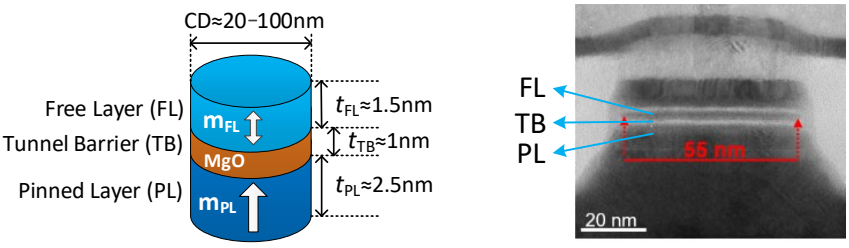
This section introduces the basic structure of STT-MRAM device and its working principles to store, write, and read data.

3.1.1. MTJ ORGANIZATION

Magnetic Tunnel Junction (MTJ) is the most important component in STT-MRAMs, as it is the data-storing element which contains one-bit of data in the form of binary magnetic configurations. The MTJ device is in the shape of cylinder, as shown with the schematic in Figure 3.1a. The diameter of MTJ is commonly referred to as *Critical Diameter* (CD) in the MRAM community; CD is typically in the range of 20-100 nm. Figure 3.1b shows a cross-sectional transmission electron microscopy (TEM) image of a MTJ device (CD=55 nm) fabricated at IMEC. It can be seen that the sidewall of fabricated MTJ device has an angle smaller than 90 degree. This is caused by the high-energy ion beam etch and redeposition of peeled materials along the sidewall. This also leads to the inertness of electrical and magnetic properties at the edge of MTJ, compared to the inner part of the device. As a result, electrical CD (eCD) is also commonly used to represent the effective CD with uniform properties across the device. The *cross-sectional area* $A_0 = \frac{1}{4}\pi \cdot \text{eCD}^2$ is a key technology parameter of the device [46]. The MTJ structure is fundamentally composed of three layers as follows[75].

1) Free Layer (FL). The top layer is called free layer, which is typically made of CoFeB material (t_{FL}=~1.5 nm [76]). The magnetization ($\boldsymbol{m}_{\text{FL}}$) in the FL is engineered towards the easy axis (an energetically favorable direction), and it can be switched to the opposite direction by applying a spin-polarized current flowing through the device. The easy axis lies in the thin film (i.e., horizontal direction) if the FL has in-plane magnetic anisotropy (IMA)[75]. In contrast, the easy axis points perpendicular to the FL for MTJs with perpendicular magnetic anisotropy (PMA). Compared to IMA-MTJ devices, PMA-MTJ devices have many advantages such as: 1) better scalability to smaller sizes, 2) better manufacturability due to the symmetric shape, and 3) smaller switching current [26]. Therefore, PMA-MTJ devices are more favorable and adopted in all STT-MRAM prototypes demonstrated in recent years. Due to this reason, we limit our focus to PMA-MTJ devices in this thesis.

2) Tunnel Barrier (TB). The MgO dielectric layer in the middle is called tunnel barrier. As the TB layer is ultra-thin, typically t_{TB}=~1 nm [76], electrons have chance to



(a) (b)

Figure 3.1: (a) Simplified MTJ device organization, (b) cross-section TEM image of a device with CD=55 nm.

tunnel through it overcoming its *potential barrier height* $\bar{\varphi}$ [47]. This makes the device behave as a tunneling-like resistor. To compare the sheet resistivity of different MTJ designs, the *resistance-area* (*RA*) product (in units of $\Omega{\cdot}\mu m^2$) [75] is used. This is a figure of merit which is commonly used in MRAM community, and it is independent on device size. *RA* can be measured by specific characterization techniques such as current-in-plane tunneling (CIPT) and conducting atomic force microscopy (CAFM) at various processing stages [77], typically in the range of 5-15 $\Omega{\cdot}\mu m^2$.

3) Pinned Layer (PL). The bottom ferromagnetic layer is referred to as pinned layer; typically its thickness is t_{PL}=2.5 nm [76]. The magnetization ($\boldsymbol{m}_{PL}$) of the PL is strongly pinned to a certain direction by an inner synthetic anti-ferromagnet (iSAF) [76]. With the fixed magnetization in PL as a reference, the magnetization in FL is either *parallel* (P state) or *anti-parallel* (AP state) to that of PL, as illustrated with the left and right device schematics, respectively, in Figure 3.2.

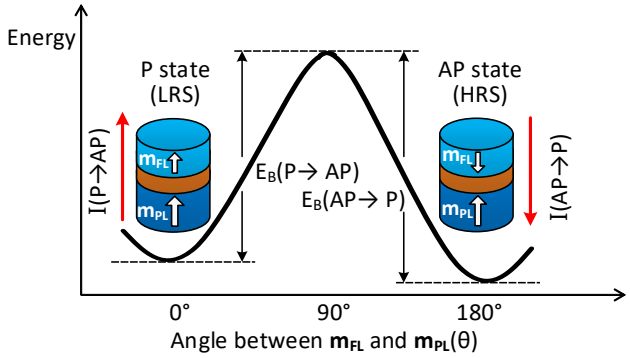


Figure 3.2: Energy barrier E_B between the binary MTJ states: P and AP.

Note that the real-world MTJ stack designs are much more complicated; the MTJ stack may consist of more than 10 thin films to enhance the device performance, as can be found in [76, 78–80]. For example, it is found that adding an MgO capping layer, the white line above the FL in Figure 3.1b is very effective in enhancing the thermal stability [78]. Furthermore, the PL consists of a *reference layer* (RL), composed of Co/Ru/CoFeB, and a *hard layer* (HL), composed of $[Co/Pt]_x$; The RL and HL are anti-ferromagnetically coupled [80].

3.1.2. WORKING PRINCIPLES

To work properly as memory elements, MTJ devices need to provide three basic functionalities: 1) data storing, 2) data retrieving, and 3) data recording. Next, we will elaborate them in detail and the associated physical mechanisms.

1) Data storing. This functionality means that data has to be retained in a memory cell in the standby mode for a certain period of time, which is known as retention time. The retention time of MTJ is mainly determined by the thermal stability factor Δ. It is defined as the energy barrier E_B that the magnetization in the FL has to overcome to switch to the opposite direction divided by the thermal activation energy at the operating

temperature T [26].

$$\Delta = \frac{E_B}{k_B T} = \frac{\mu_0 M_s V H_k}{2 k_B T}, \tag{3.1}$$

where μ_0 is the vacuum permeability, k_B the Boltzmann constant, M_s the *saturation magnetization*, $V = A \cdot t_{FL}$ the FL volume, and H_k the *magnetic anisotropy field*. It is worth noting that the retention time in P state Δ_P can be different from Δ_{AP}, since E_B(P→AP) is not the same as E_B(AP→P) (see Figure 3.2) if there exists stray fields from the PL at the FL [65].

Conventionally, the following static model is used to roughly estimate the retention time (RT) of AP or P state for a given Δ [81]:

$$RT = \tau_0 \exp(\Delta), \tag{3.2}$$

where τ_0 is the inverse of the attempt frequency (~1 ns). However, the retention time for STT-MRAMs has intrinsic stochasticity, as the magnetization flip induced by thermal fluctuation is unpredictable. This static model fails to capture the stochastic property. Actually, the calculated retention time using Equation (3.2) corresponds to the time after which the MTJ state flips at a probability of 63%, as pointed out in [82]. As an alternative, a statistic model derived from the switching model in thermal-activation regime is widely used, as can found in [75, 82, 83]:

$$RT = \tau_0 \exp(\Delta) \cdot \left(\frac{1}{1 - P_{RT}}\right), \tag{3.3}$$

where P_{RT} is the switching probability of a certain MTJ state due to thermal fluctuation after time RT (i.e., the confidence in the estimation of RT). Based on Equations (3.2–3.3), it is clear that the higher thermal stability factor, the more robust against thermal perturbation, and thus the longer retention time. For an MTJ with Δ=40, the retention time is approximately 7.4 years [82]. The requirement of retention time (or Δ) depends on the target application. Typically, for data storage applications, $\Delta > 80$ is needed for a 1Gb STT-MRAM array to meet the industrial requirement, i.e., a retention time larger than 10 years [26]. In contrast, cache applications only necessitate ms-scale retention time, corresponding to Δ =~30 [84, 85].

Due to the stochastic switching property, charactering RT (or Δ) of STT-MRAM devices or arrays is tedious and takes much more efforts, compared to other NVMs such flash, PCM, and RRAM. Tillie *et al.* [85] introduced and compared four different retention extraction methods; all of them are statistical methods over a large number of transition cycles and/or memory cells. The most direct and simplest method is called switching time probability (STP) at elevated temperature. It collects the statistics of switched cells among a large array at different elevated temperatures and thereafter extrapolates RT (or Δ) to a target temperature; examples are [15, 86, 87]. Readers who are interested in details about the retention characterization methods and applications are directed to these papers.

2) Data retrieving. This functionality means that the stored data in a memory cell can be read out and decoded into logic values in a certain form. For STT-MRAMs, this is realized by the *Tunneling Magneto-Resistance* (TMR) effect [75, 88]. The TMR effect means that the resistance of MTJ depends on not only the TB thickness t_{TB}, but also

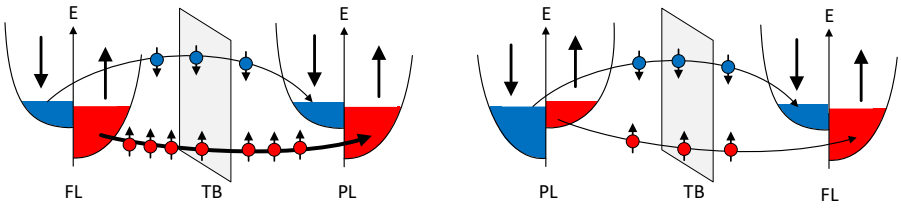
the relative direction of magnetization in the FL and PL, i.e., P or AP state. As shown in Figure 3.2, when the MTJ device is in P state, its resistance is relatively low. By contrast, the device's resistance is high in AP state. To quantitatively evaluate the TMR effect, the TMR ratio is used; it is defined as:

$$TMR = \frac{R_{\mathrm{AP}} - R_{\mathrm{P}}}{R_{\mathrm{P}}} \times 100\%, \tag{3.4}$$

where R_{AP} and R_{P} are the resistances in AP and P states, respectively. They are key electrical parameters of MTJ. Obviously, the higher the TMR ratio, the easier to distinguish between P and AP states during read operations. In the past decades, the TMR ratio has increased significantly thanks to device stack innovations and process improvement. For example, the TMR ratio of AlO_x-based MTJs was 70% in 2004 [89]. With the introduction of crystalline MgO-based tunnel barrier, the TMR ratio exceeded 100% and reached 200% at labs [90]. Most STT-MRAM test chips demonstrated between 2010-2020 show TMR ratio 150%-200%. In late 2019, Samsung demonstrated MTJ devices with *TMR*=220% [14]. For commercially-feasible STT-MRAM products, *TMR*>200% is desired for reliable reading of a Meg-bit array [26].

From a physics perspective, the origin of TMR effect can be explain by the band structure model shown in Figure 3.3. It is well established that an electron has two attributes: charge and spin. An electron carries a negative elementary charge, and moving electrons are the source of electric current which contributes to the conductance in NMOS. The spin attribute lays the foundation of Spintronics. Naturally, an electron can be either spin-up or spin-down, and the number of spin-ups and spin-donws are even in non-magnetic materials such as copper. However, in a magnetized ferromagnetic material, the number of spin-ups and spin-downs are totally different, leading to different contributions to electrical conductance. In other words, the ferromagnetic material polarizes the incoming electrons making the number of spin-ups and spin-downs uneven. For example, when electrons arrive at the FL of an MTJ in P state (see the schematic in Figure 3.2), the FL polarizes the incoming electrons to align with the magnetization direction. This results in the number of spin-ups much larger than spin-downs. In this case, spin-up is referred to as majority spin, while spin-down is minority spin. The spin polarization P is defined as [91]:

$$P = \frac{n\uparrow - n\downarrow}{n\uparrow + n\downarrow} \tag{3.5}$$



(a) Good band match in P state (LRS). (b) Poor band match in AP state (HRS).

Figure 3.3: Band structure model which explains the physical origin of the TMR effect.

where, $n\uparrow$ and $n\downarrow$ are the numbers of spin-ups and spin-downs, respectively. The electrons of a certain spin direction can tunnel through the MgO barrier to the PL only if there are empty states to accept the electrons in the PL. Thus, assuming that spin is conserved during tunneling, the majority spins in the FL will fill majority states in the PL, while minority spins will fill the minority states, as shown in Figure 3.3a. This results in large conductance and small resistance R_P in P state. Compared to the good band match in P state, AP state has poor band match (see Figure 3.3b) since the majority spin becomes spin-down, which has the same direction as the magnetization in the FL. Therefore, the majority spins in the FL have to fill the minority states in the PL, while the minority spins fill the majority states. This leads to small conductance and high resistance R_{AP} in AP state. Based on this physical model, it can be deduced that the TMR ratio is determined by the spin polarization of the two ferromagnetic layers [91]:

$$TMR = \frac{2P_{FL}P_{PL}}{1 - P_{FL}P_{PL}}, \tag{3.6}$$

where P_{FL} and P_{PL} are the spin polarization of the FL and PL, respectively.

3) Data Recording. This functionality means that data can be written into a memory cell in a certain. It is realized by the *spin-transfer torque* (STT) effect [26, 75]. To understand the fundamental physics behind the STT-effect, we can use the free-electron model [26] to explain the switching mechanism. Figure 3.4 depicts the STT-induced switching mechanism from AP state to P state for an IMA-MTJ as an example. When a voltage is applied across the junction, electrons flow through it from the RL to the FL. The RL polarizes the incoming electrons to align the magnetization of this layer, making spin-up the majority spin and spin-down the minority spin. As electrons with spin-up tunnel through the TB, the spin of transmitted electrons precesses incoherently around the local exchange field which is along the magnetization (pointing down) of the FL. As a result, the electrons quickly become repolarized along the magnetization of the FL. Due to momentum conservation, the difference between the momentum of the incoming and outgoing electrons yields a torque acting on the FL magnetization. This torque is known as spin-transfer torque. If the spin-transfer torque is high enough, it flips the magnetization in the FL, making its direction consistent with that of the magnetization in the RL (i.e., P state). The switching process for the P→AP direction and PMA-MTJ devices are similar to the above explanation, thus omitted here in this thesis.

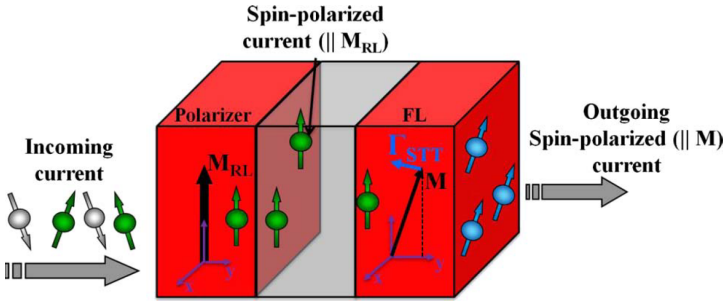


Figure 3.4: Free-electron model for STT-induced switching from AP state to P state (reprinted from [26]).

To induce a complete transition between the P and AP states, a high write current is required to provide an energy larger than the energy barrier E_B (see Figure 3.2). Obviously, the smaller the E_B, the easier to switch, but the shorter the retention time. If the current is larger than the *critical switching current* (I_c), the magnetization in the FL may switch, depending on the pulse width, to the other direction. By definition, I_c is the current to switch the device's state within infinitely long time and at zero temperature [75]. It is a key electrical parameter to characterize the switching capability by current. Due to the bias dependence of STT efficiency and stray fields H_{stray} [75], I_c(P→AP) can be significantly different from I_c(AP→P) in practice. In addition, the *switching time* (t_w) [47] is another critical parameter, which is inversely correlated with the actual write current. In other words, the higher the write current over I_c, the less time required for the magnetization in FL to flip. Note that t_w(P→AP) can also differ from t_w(AP→P) depending on the write current magnitude and duration. To achieve higher t_w (faster write speed), one has to boost the write current. However, a higher write current also results in a reduction in the endurance, defined as the write cycles that a memory cell can bear before it wears out. This is due to the large electric field across the ultra-thin TB layer and Joule heating, which together accelerate the breakdown of the TB. Therefore, it is well known that retention, write speed, and endurance pose a dilemma for designing STT-MRAMs.

In summary, the data-storing function (retention) of MTJ is determined by the thermal stability factor Δ. The data retrieving (read) and recording (write) functions are realized by the TMR and STT effects, respectively. It is important to note that these two effects are closely correlated, as illustrated in Figure 3.5. In the TMR effect, the magnetic state P or AP determines the MTJ resistance state LRS or HRS. In the STT effect, it is the electric current flowing through the MTJ that changes its magnetic state. Therefore, both the magnetic and electrical properties are crucial for MTJs as memory elements with write and read functions. Table 3.1 lists the MTJ's key technology and electrical parameters. The MTJ modeling process will be presented in the next Chapter; it abstracts the MTJ device from the physical level to electrical level. The essence of this modeling process is to map the device's technology parameters to electrical ones.

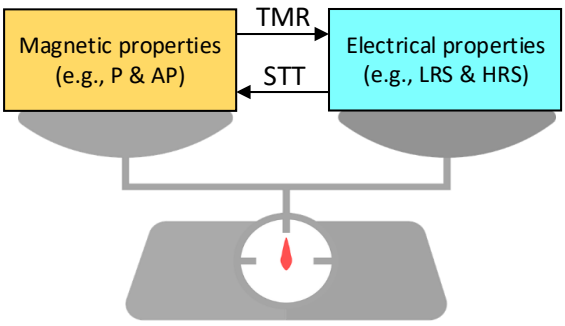


Figure 3.5: Equal importance of MTJ's magnetic and electrical properties.

Table 3.1: Key technology and electrical parameters of MTJ.

Technology Parameters		Electrical Parameters	
A_0	Cross-sectional area of MTJ	R_P	Resistance in P state
M_s	Saturation magnetization of the FL	R_{AP}	Resistance in AP state
H_k	Magnetic anisotropy field of the FL	I_c(P→AP)	P→AP critical switching current
$\bar{\varphi}$	Potential barrier height of the TB	I_c(AP→P)	AP→P critical switching current
RA	Resistance-area product	t_w(P→AP)	P→AP switching time
TMR	Tunneling magneto-resistance ratio	t_w(AP→P)	AP→P switching time
H_{stray}	Stray field at the FL		

3.2. ELECTRICAL STT-MRAM MODEL

This section presents STT-MRAM test circuits used in the research work in this thesis. The circuits integrate MTJ devices into CMOS-based circuits and implement memory functions such as bit-cell selection, write/read operation on the selected cell. Figure 3.6 shows the STT-MRAM circuit organization which consists of an STT-MRAM array and peripheral circuits such as address decoders, write drivers, and sense amplifiers. Next, we discuss each of them in detail.

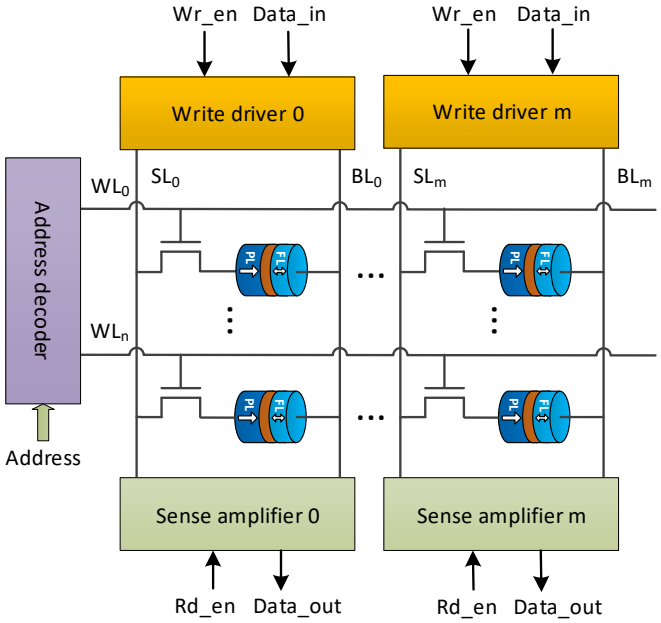


Figure 3.6: STT-MRAM circuit organization.

3.2.1. STT-MRAM BIT CELL

Memory arrays are the entities where data is stored; they occupy the majority area in a memory chip [14]. A memory array is a matrix of memory cells, each of which stores an atomic amount of data. Although multi-level cell (MLC) designs also exist for STT-

MRAM [92], single-level cell (SLC) is predominant in STT-MRAM designs. An SLC STT-MRAM cell stores one bit of data; the bottom-pinned 1T-1MTJ bit cell design is the most widely-adopted SLC design, comprising an MTJ device connected serially with a selector device [93, 94], as shown in Figure 3.7a. The MTJ in this structure serves as a storage element, while the selector is responsible for selective access to this cell. Since the MTJ has to be switched by spin-polarized electrons as introduced previously, the selector is typically implemented using NMOS where the majority carriers for current transport are electrons. The NMOS gate is connected to a *word line* (WL), which determines whether a row is accessed or not. The other two terminals are connected to a *bit line* (BL) and a *source line* (SL), respectively. They control write and read operations on the internal MTJ device depending on the magnitude and polarity of voltage applied across them. It is worth noting that the 1T-1MTJ cell size is mainly limited by the size of NMOS, which has to drive a large switching current for the MTJ device. To further boost the density of STT-MRAM array, 3D stacking techniques and other types of selector are under research; e.g., a two-terminal bi-direction threshing selector in [16].

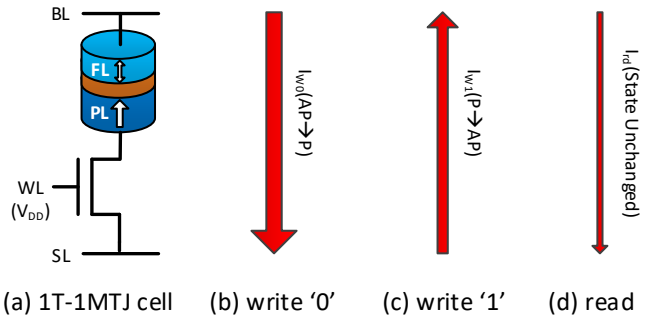


Figure 3.7: Bottom-pinned 1T-1MTJ bit cell and its write and read operations.

Figure 3.7b and 3.7c show write '0' and write '1' operations for the 1T-1MTJ bit cell design, respectively. During a write '0' operation, WL is pulled up to V_{DD} to enable access to this cell. In addition, a positive pulse is applied across the BL and SL, thus leading to a current I_{w0} flowing from BL to SL (electrons flow in the opposite direction from the SL to BL). When electrons arrive at the PL, they get spin-polarized. The spin-polarized electrons then tunnel through the MgO barrier layer and exert a torque on the magnetization in the FL. If the MTJ is in AP (1) state and I_{w0} is larger the critical switching current I_c, the MTJ state switches to P (0) state by means of the STT effect after a certain period of time t_{w0}, which has to be shorter than the pulse width t_p. The higher the I_{w0} over I_c, the shorter the t_w. For a MTJ which is in P (0) state, the state remains under the write '0' current I_{w0}. In contrast, a write '1' operation requires an opposite current I_{w1} going through the MTJ device. Note that the write '0' and '1' operations are typically asymmetric in STT-MRAMs, meaning that I_{w0} (t_{w0}) differs from I_{w1} (t_{w1}) in practice. This is caused by the difference in I_c and resistance between P and AP states, as well as the source degeneration issue in the access NMOS [95, 96].

Figure 3.7d shows the read operation, where a small read pulse V_{read} is applied. It leads to a read current I_{rd} with the same direction as I_{w0} to sense the resistive state (R_{AP}

or R_P) of MTJ. To avoid an inadvertent state change during read operations, known as *read destructive fault* [60], I_{rd} should be as small as possible; typically $I_{rd} < 0.5I_c$ for MTJs with a thermal stability $\Delta = 65$ [97]. However, a too small I_{rd} may lead to *incorrect read faults* [54]. A read operation requires a sense amplifier to determine the resistive state. The sense amplifier may be implemented using a current sensing scheme, where the read-out value is determined by comparing the current of the accessed cell ($I_{cell} = I_{rd}$) with the current of a reference cell I_{ref}. The sensing result is logic '0' if $I_{cell} < I_{ref}$; otherwise, it outputs logic '1'.

3

3.2.2. STT-MRAM PERIPHERAL CIRCUITS

As introduced previously, a 1T-1MTJ bit cell is selected in a large STT-MRAM array by asserting its corresponding WL. This is realized by an address decoder. The selected cell is written or read by putting appropriate voltages on the BL and SL. This is realized by a write driver and a sense amplifier, respectively. Next, we will elaborate these three key peripheral circuits.

ADDRESS DECODER

Address decoders are used to select particular memory cells in a memory array. Typically, a row address decoder decodes the row address and activate a specific WL while a column address decoder decodes the column address and activate a specific pair of BL and SL in a large STT-MRAM array. This allows subsequent read/write operations to target the selected cell in the memory array. To reduce simulation overhead, we implemented a small STT-MRAM array (e.g., 3×3) with necessary peripheral circuits. Figure 3.8 illustrates how the row address decoder works as an example. Figure 3.8a shows a straightforward implementation of the row decoder using NOT and AND gates; it has two address inputs and four WLs as outputs. Figure 3.8b and Figure 3.8c show the CMOS implementations of these two gates, respectively.

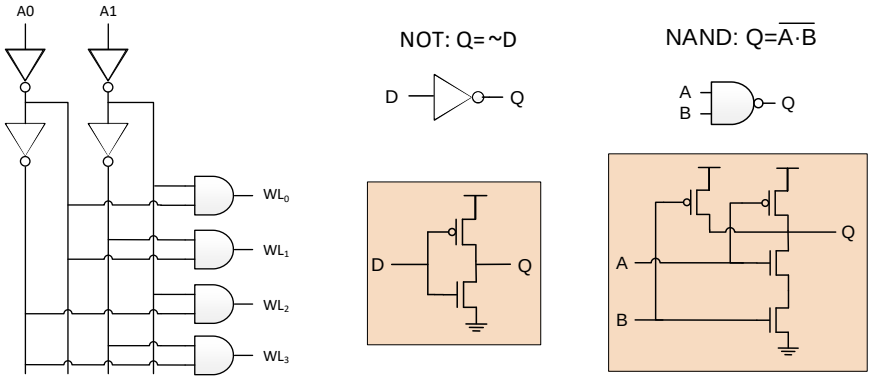


Figure 3.8: 2-input static address decoder: (a) gate-level schematic, (b) NOT gate based on a CMOS inverter, and (c) NAND gate and its CMOS implementation.

WRITE DRIVER

Write drivers are the circuits used to program data into selected memory cells. Figure 3.9 shows the write driver design used in this thesis. It consists of two parts: control logic and driving circuit. The control logic comprises three NOT gates and two NAND gates; it translates two inputs: Wr_en and Data_in into internal control signals: P1, N1, P2, and N2 , which are connected to the driving circuit in the second part. A write operation is enabled by pulling up the voltage on Wr_en to V_{DD}; the data to be written is put on the other input port: Data_in. The truth table is shown in Table 3.2. For example, when Wr_en=1 and Data_in=1, it launches a write '1' operation under the synchronized clock. These two signals are translated to P1=0, N1=0, P2=1, and N2=1, which turn on MP1 and MN2 while turn off MN1 and MP2 in the driving circuit. As a result, a write '1' current I_{w1} flows through the MTJ device, as illustrated in the figure. Similarly, a write '0' operation can be performed by setting Wr_en=1 and Data_in=0.

Figure 3.10 shows the simulation waveforms of some key signals for the write sequence: 0w1w0. The Cadence Spectre simulator is used in our circuit simulations. V_{DD} is set at 1.6 V; WL and Wr_en are set to 1.8 V to boost the switching current. The MTJ model used in the simulations is a Verilog-A MTJ compact model with CD=60 nm. More details about our MTJ model will be presented in Section 4.1. All transistors in the netlist are built with the 90 nm predictive technology model (PTM) [98]. It can seen that the MTJ is initialized to state '0'. A write '1' operation with a pulse t_p=18 ns is applied. After $t_w \approx 9.6$ ns, the MTJ switches from '0' to '1'. Subsequently, a write '0' operation is performance, which switches the MTJ state back to '0'. The switching process takes approximately 12.7 ns.

There also exist more sophisticated write driver designs for STT-MRAMs in the literature for the purpose of enhancing write robustness or saving write energy. Examples are write-verify-write scheme and self-write-termination scheme. Readers who are interested in this topic are directed to works in [30, 99–102].

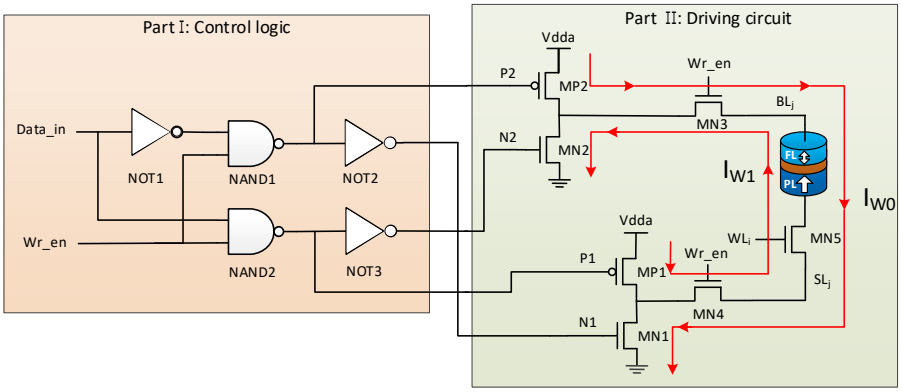


Figure 3.9: Write driver for STT-MRAMs.

Table 3.2: Truth table of translating input signals into internal control signals.

Wr_en	Data_in	P1	N1	P2	N2	write operation
0	0	1	0	1	0	—
0	1	1	0	1	0	—
1	0	1	1	0	0	w0(AP→P)
1	1	0	0	1	1	w1(P→AP)

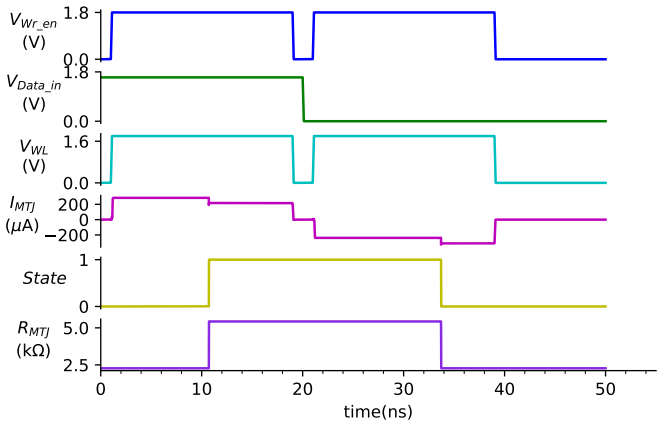


Figure 3.10: SPICE-based circuit simulation waveforms of write sequence 0w1w0.

SENSE AMPLIFIER

Sense amplifiers refer to the circuits that read the stored information in the form of logic levels (i.e., '0' and '1') from selected memory cells. For STT-MRAM chips, a sense amplifier senses the magnetic states of MTJ (i.e., P and AP), transform them into voltages, and then amplify the voltages to full voltage low and high as logic '0' and '1', respectively.

In this thesis, we use a pre-charged sense amplifier (see Figure 3.11), similar to designs in [101, 103]. In essence, the sense amplifier compares the currents going through the memory cell under sensing (I_{cell}) and a reference cell (I_{ref}). The resistance of the reference cell is typically set in the middle of R_{P} and R_{AP}; i.e., $R_{\text{ref}} = (R_{\text{P}} + R_{\text{AP}})/2$. This can be implemented by configuring four MTJs and connecting them in a certain order [101] or by fabricating a thin-film resistor directly [30]. If I_{cell}<I_{ref}, the sense amplifier outputs logic '1' on the Q node; otherwise, the output is logic '0'.

Figure 3.12 shows the simulation waveforms of some key signals in a complete sensing operation on an STT-MRAM cell with an MTJ in P state. The sensing operation takes 4 ns, during which the Rd_en signal is asserted. It consists of three phases as follows.

I. **Pre-charge:** In this phase, Q and $\bar{Q}$ are pre-charged to V_{dd}. The pre-charge circuit includes three PMOS transistors as shown in the figure. MP1 and MP2 connect Q and $\bar{Q}$ to V_{dd}, respectively. MP3 equalizes the potential of Q and $\bar{Q}$ when the PC signal is pulled down to GND. This ensures that Q and $\bar{Q}$ are always pre-charged to the same potential in the presence of process variations. The WL is

deactivated so that no static current flows through the memory cell and reference cell in the pre-charge phase.

II. **Voltage development:** This phase starts by pulling up the voltage on PC and WL. As Q and $\bar{Q}$ are disconnected to the power supply and all NMOS transistors in the sense amplifier are turned on, they start to discharge simultaneously along the two paths marked in Figure 3.11. Since the memory cell is in the P(0) state, which exhibits lower resistance than that of the reference cell, i.e., $R_{\text{cell}} = R_{\text{P}} < R_{\text{ref}}$. As a result, the potential on the Q node drops faster than the $\bar{Q}$ node, as shown with the inset in Figure 3.12. Thus, a voltage difference (ΔV) develops between the two nodes and keeps increasing over the discharging process.

III. **Voltage amplification:** MN1, MP4, MN2, and MP5 form two cross-coupled inverters. They start amplifying the small voltage difference ΔV, once it reaches a certain threshold. The potential of Q node continues to drop down to GND, while the $\bar{Q}$ node is pulled back to V_{dd}. Similarly, if the STT-MRAM cell under sensing stores a '1' (high resistance state), Q node will be at V_{dd} and $\bar{Q}$ node will

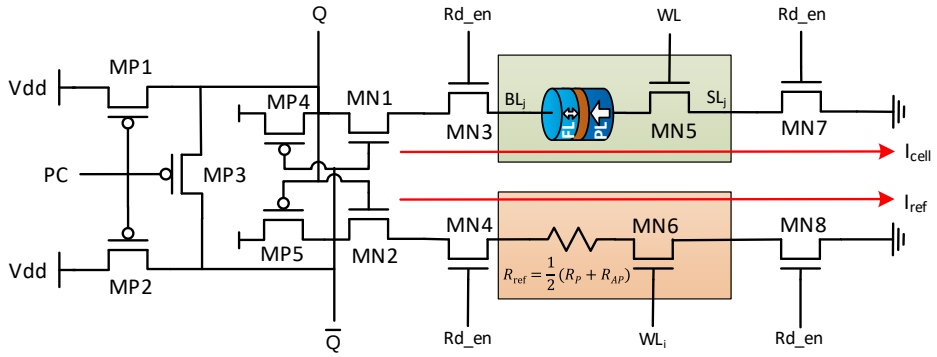


Figure 3.11: Pre-charge based sense amplifier for STT-MRAMs.

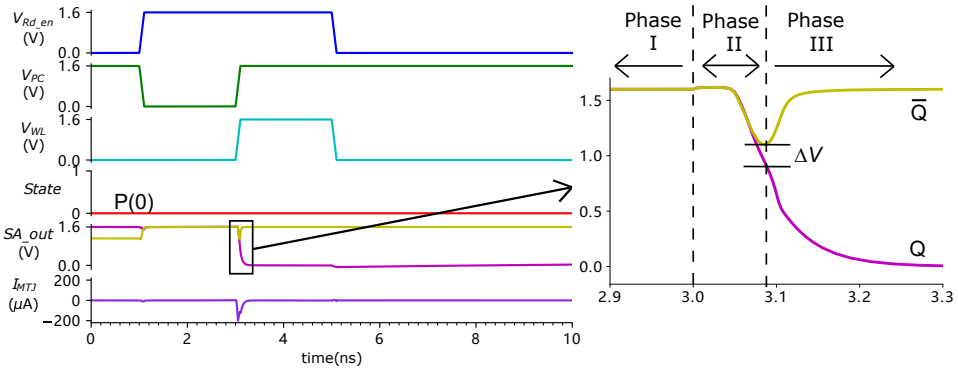


Figure 3.12: SPICE-based circuit simulation waveforms of reading an STT-MRAM cell in state '0'.

be at GND at the end of this phase. Similar to SRAM cell designs, the two cross-coupled inverters here ensures that the voltages on the Q and $\bar{Q}$ nodes are always at complimentary levels after the completion of a sensing operation.

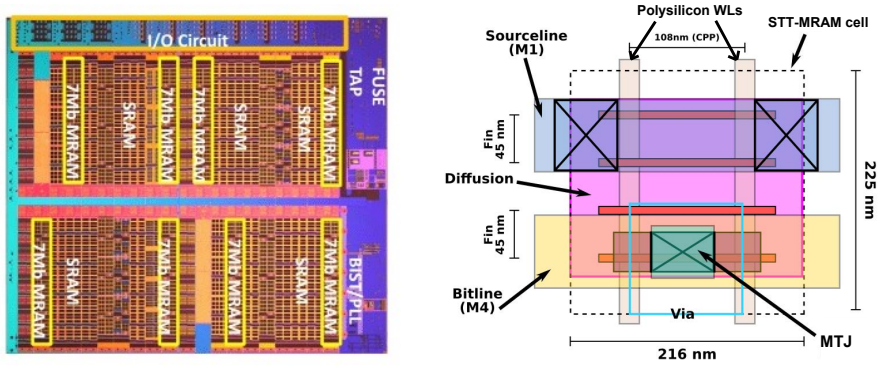
The above pre-charge sense amplifier design features high speed and low power consumption. It can be seen in Figure 3.12 that the pre-charge phase can finish within 200 ps and the subsequent two phases (the zoomed-in part) also take around 200 ps before stabilization. Therefore, the read latency with this sense amplifier can reach sub-nm level by tightening the margins of different phases. In addition, the working principle of this sense amplifier design is based on charging and discharging operations. Thus, there is no static current going through any memory cells during a sensing operation, thereby contributing to a low-power read operation.

Based on this sense amplifier, different variants have been proposed in the literature to incorporate various functionalities such as reliability enhancement and offset cancellation. More works about sense amplifier designs for STT-MRAMs can be found in [99, 101, 102, 104–107].

3.3. STT-MRAM LAYOUT MODEL

The layout model is the representation of an integrated circuit in terms of planer geometric shapes which correspond to the patterns of metal, oxide, or semiconductor layers that make up the components of an integrated circuit. The layout model describes the physical structure, location, and dimensions of all components, as they are manufactured on silicon. Due to the proprietary nature of this information and it high complexity at this modeling level, semiconductor manufacturers rarely disclose the layout models of their chips. Obviously, this is also the case for STT-MRAM products. Although the behavioral and functional models of the Everspin's 1Gb STT-MRAM chip introduced in the previous chapter are exposed to the public in the form of chip datasheet, the electrical and layout models are not publicly available. Therefore, in this section, we will illustrate the STT-MRAM layout model using Intel's STT-MRAM design as an example, as some of its implementation details were revealed at IEDM'18 [79] and ISSCC'19 [30].

Figure 3.13a shows the die photo of Intel's STT-MRAM test chip, which contains eight STT-MRAM arrays, each of which is 7 Mb. The chip also implements SRAM arrays, a PLL, and eFuse, a BIST and DDR IOs. The chip was fabricated on Intel's 22 nm FinFET low-power (22FFL in short) platform. When presenting this test chip at ISSCC in early 2019, Intel claimed that its embedded STT-MRAM technology was production ready for eFlash and even eDRAM replacement [30]. Figure 3.13b shows the layout of STT-MRAM cell, which consists of an PMA-MTJ device and a FinFET selector [79]. The cell area is $0.0486\,\mu m^2$ (216 nm×225 nm); the size of MTJ is between 60-80 nm in diameter. It can be seen that the FinFET size is much larger than the MTJ device. This is because today's MTJ technology requires a large switching current, which is driven by the underlying transistor. Therefore, it is well recognized that the transistor size becomes the bottomneck of boosting STT-MRAM density. Furthermore, the cell is connected to a Metal-4 BL, a Metal-1 SL, and a WL which is split into two polysilicon lines. The WLs are finally connected to the Metal-5 layer. More details about the Intel's eSTT-MRAM design and implementation can be found in [30, 79].



(a) (b)

Figure 3.13: STT-MRAM layout model: (a) die photo of Intel's STT-MRAM test chip, and (b) its 1T-1MTJ cell layout (source: Intel [30, 79, 108]).

3.4. STT-MRAM MANUFACTURING DEFECTS AND CLASSIFICATION

A defect is a physical imperfection in manufactured chips (i.e., an unintended difference from the intended design) [3]. To guarantee a high-quality test solution and improve the manufacturing process itself so as to improve yield, understanding all potential defects is of great importance. The STT-MRAM manufacturing process mainly consists of the standard CMOS fabrication steps and the integration of MTJ devices into metal layers (e.g., between M4 and M5 layers [85, 110]). Figure 3.14 shows the bottom-up manufacturing flow and the vertical structure of STT-MRAM cells [109]. Based on the manufacturing phase, STT-MRAM defects can be classified into front-end-of-line (FEOL) and

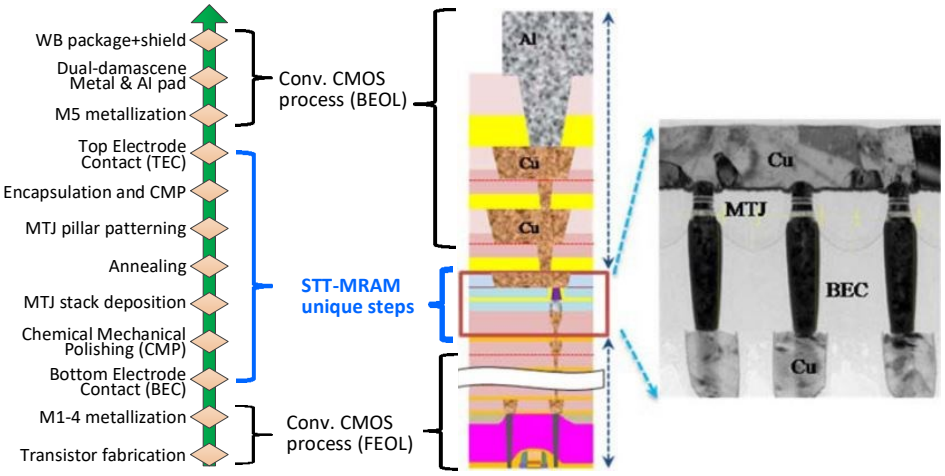


Figure 3.14: General manufacturing process of STT-MRAM (right part reprinted from [109]).

back-end-of-line (BEOL) defects. As MTJs are integrated into metal layers during BEOL processing, BEOL defects can be further categorized into interconnect defects and MTJ-related defects. All potential defects are listed in Table 3.3. Next, we will examine them in detail along with their corresponding processing steps, with a particular emphasis on those introduced during MTJ fabrication.

Table 3.3: STT-MRAM defects and classification.

FEOL	BEOL	
Transistor	Interconnect	MTJ Device
Material impurity Crystal imperfection Pinholes in gate oxides Shifting of dopants Patterning proximity etc.	Open vias/contacts Irregular shapes Big bubbles Small particles etc.	Pinholes in TB Redepositions on MTJ sidewalls Synthetic anti-ferromagnet flip Intermediate states Back-hopping Extreme thickness variation of TB MgO/CoFeB interface roughness Atom inter-diffusion Magnetic layer corrosion etc.

3.4.1. CONVENTIONAL DEFECTS IN FEOL

The first step of the STT-MRAM manufacturing process is the FEOL process where transistors are fabricated on the wafer. In this phase, typical defects may occur such as semiconductor impurities, crystal imperfections, pinholes in gate oxides, and shifting of dopants [111]. These are the conventional defects which have been sufficiently studied and are generally modeled by resistive opens, shorts and bridges [112–114].

3.4.2. CONVENTIONAL DEFECTS IN BEOL

After FEOL, M1-M4 metal layers are stacked on top of the transistors followed by a bottom electrode contact (BEC), as illustrated in the zoomed-in part of Figure 3.14. M1-M4 metalization does not differ from traditional CMOS BEOL steps. The BEC step is used to connect bottom Cu lines with MTJ stacks [76, 109]. During this phase, typical interconnect defects may take place, such as open vias/contacts, irregular shapes, big bubbles, etc. [112]. For instance, Figure 3.15 shows a TEM image of an open contact defect between the BEC and the underlying Cu line due to polymer leftovers [109].

To obtain a super-smooth interface between the BEC and the MTJ stack, a chemical mechanical polishing (CMP) step is required. The smoothness of the interface between layers is key to obtaining a good *TMR* value. CMP processing minimizes the surface roughness with a root-mean-square average of 2Å[85]. At this stage, both under-polishing and over-polishing of the surface can introduce defects. Specifically, under-polishing causes issues such as orange peel coupling or offset fields which affect the hysteresis curve, while over-polishing may result in dishing or residual slurry particles that are left behind [59].

After the CMP step, MTJ devices are fabricated, which will be detailed in the following

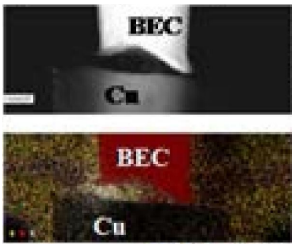


Figure 3.15: An open contact defect between the BEC and the underlying Cu layer (reprinted from [109]).

subsection. Next, MTJ pillars are connected to the top electrode contact (TEC), followed by M5 metallization. The rest of manufacturing process is the same as the BEOL steps of CMOS technology. Typical defects such open contact/vias, small particles etc. can occur in this phase as well. It is worth-noting that a package-level magnetic shield can be added to enhance the stand-by magnetic immunity of STT-MRAMs, as proposed in [86]. The magnetic shield was reported to be effective in protecting STT-MRAMs against external magnetic fields.

3.4.3. MTJ-RELATED DEFECTS IN BEOL

As mentioned previously, the next critical step following the CMP step is the fabrication of the MTJ stack. The latest published MTJ design includes more than 15 layers in pursuit of better performance [115]. However, the increasingly sophisticated design of the MTJ also makes it more vulnerable to manufacturing defects. Figure 3.16 shows the measured TMR vs. R_P of 450 MTJ devices with CD=60 nm. Each point in the figure represents a MTJ device with its TMR on the y-axis and R_P on the x-axis. Clearly, there is a large device-to-device variation in these two parameters due to process variations, manufacturing defects, and measurement errors. We classified the measured devices into three bins based on their TMR values; TMR>130% is marked as good TMR (green circle), TMR < 20% is marked as poor TMR (red diamond), and TMR in between is marked as inter-

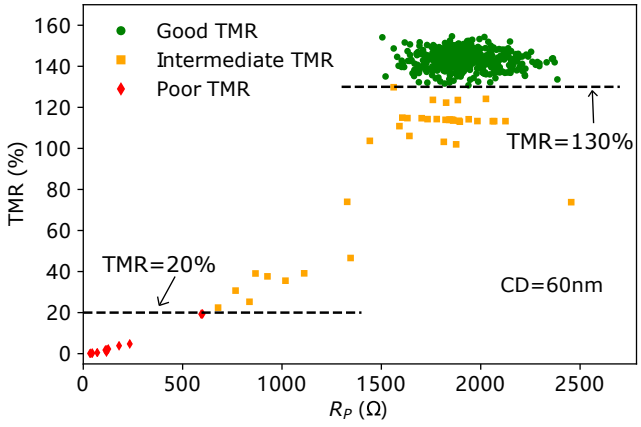
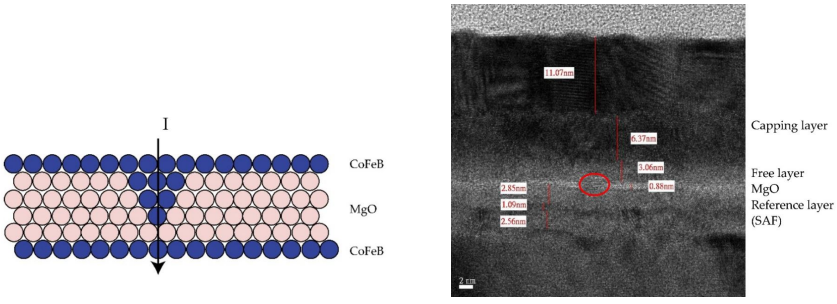


Figure 3.16: Measured TMR vs. R_P of 450 MTJ devices with CD=60 nm.

mediate TMR (yellow square). It can be seen that the majority of MTJ devices have good TMR values above 130%. In this cluster of devices, the device-to-device variation in TMR and R_P is mainly caused by process variations. Today's STT-MRAM foundries can fabricate MTJ devices with TMR around 200% and process variation in R_P with $\sigma(R_P)/\mu(R_P)$ below 10% [4]. But it is also worth noting that some devices in this cluster may also have some weak defects; although these weak defects do not cause a severe degradation in device parameters such as TMR and R_P (i.e., masked by process variations) at the point of this measurement, they may deteriorate very fast when entering into the field. For MTJs with TMR < 130%, it is expected that they have manufacturing defects. For those device with poor TMR due to certain defects, it is almost impossible to distinguish the P and AP states, leading to a stuck-at-0 fault (SA0). For those devices with intermediate TMR, they may function as well as intended, or exhibit some faulty behaviors in some manner, or deteriorate very fast over time. Therefore, all possible manufacturing defects related to MTJ devices need to be fully studied to understand their faulty behaviors.

The processing step following the CMP is MTJ stack deposition, as illustrated in Figure 3.14. In this step, several manufacturing defects may arise. For example, pinholes in the tunneling barrier (e.g., MgO) could be introduced in this phase. Figure 3.17a shows a schematic of a pinhole defect, and Figure 3.17b shows a vertical cross-section TEM image of a deposited MTJ stack with a pinhole in its 0.88 nm tunnel barrier [116]. In this defective MTJ device, a pinhole forms in the tunnel barrier due to the rough deposition of MgO. As the CoFeB free layer is deposited on top of the tunnel barrier, the pinhole is filled with CoFeB material, as indicated by the red cycle in Figure 3.17b. Therefore, the pinhole filled with CoFeB material forms a defective high-conductance path across the two ferromagnetic layers. It severely degrades the resistance and *TMR* values, and may even lead to breakdown due to the ohmic heating when an electric current passes through the barrier [117, 118]. Furthermore, the MgO barrier thickness variation and interface roughness result in degradation of resistance and TMR values as well. TEM images in [116] show that the MgO barrier thickness varies from 0.86 nm to 1.07 nm, leading to a huge difference in resistance. Figure 3.18 shows with images of atomic force microscopy (top two) and high-resolution transmission electron microscopy (bottom two) that a complicated iSAF pinned layer design elevates interface roughness from 0.5 Å to



(a) (b)

Figure 3.17: Pinhole defect in the MgO tunnel barrier of MTJ: (a) Schematic and (b) Cross-sectional TEM (both graphs reprinted from [116]).

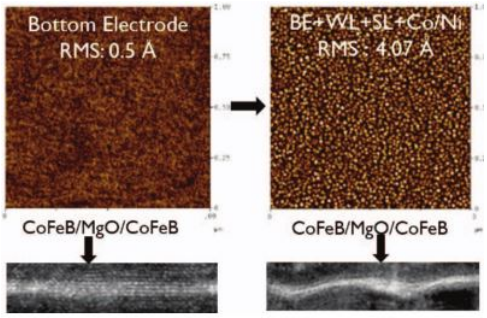


Figure 3.18: The MgO/CoFeB interface is rougher for an advanced MTJ stack design with an iSAF pinned layer on the right side than a simple MTJ stack design on the left side (reprinted from [76]).

4.07Å. The increased interface roughness leads to significant TMR degradation [76].

After the MTJ stack deposition, annealing is applied to obtain crystallization in MgO barrier as well as in CoFeB PL and FL layers [119, 120]. At this stage, the PMA originating from the MgO/CoFeB interface and TMR value are strongly determined by the annealing conditions such as temperature, magnetic field and annealing time. With appropriate annealing conditions, the PMA can be considerably enhanced, leading to higher thermal stability. Under-annealing can lead to lattice mismatch between the body-centered bubic (bcc) CoFeB lattice and the fcc MgO lattice, whereas over-annealing introduces atom inter-diffusion between layers. As illustrated in Figure 3.19, oxygen atoms can diffuse out of MgO, leaving behind oxygen vacancies, thus severely degrading TMR value [78]. Worse still, diffusion of Ta from the seed layer to MgO layer has been reported in several papers [121, 122], which scavenges O from MgO.

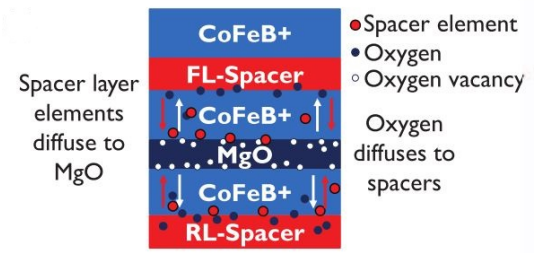
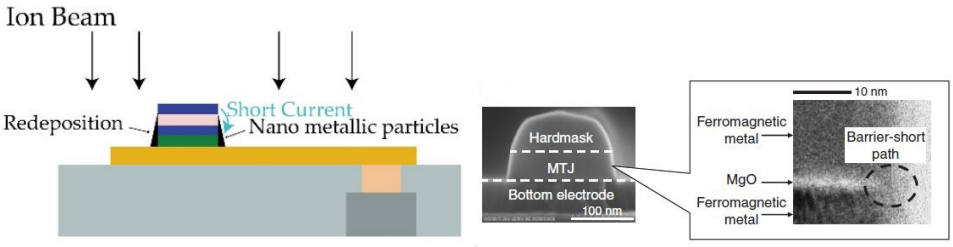


Figure 3.19: Schematic of atom inter-diffusion mechanism showing that oxygens diffuse out of the MgO barrier into neighboring layers while spacer layer materials diffuse into the MgO layer (reprinted from [78]).

After MTJ multi-layer deposition, annealing and optical lithography processing, the next crucial step is to pattern individual MTJ nanopillars [123]. Typically, Ion beam etching (IBE) is widely used to pattern MTJ nanopillars [124, 125]. Figure 3.20a illustrates the etching process, where Ar ion beams are ionized and accelerated in a chamber and subsequently irradiate the wafer underneath, leading to selective etching of the area where a hard mask does not cover. During the MTJ etching process, it is extremely difficult to obtain desired MTJ nanopillars with steep sidewall edges, while avoiding sidewall rede-



(a) Schematic of ion beam etching with redeposition on the MTJ sidewall (reprinted from [116]).

(b) Vertical cross-section TEM images of an MTJ device with sidewall redeposition (reprinted from [124]).

Figure 3.20: Magnetic material redeposition defect on the sidewall of MTJ devices.

position and magnetic layer corrosion. The redeposition phenomenon on MTJ sidewall may significantly deteriorate the electrical property of the MTJ device, and even cause a barrier-short defect shown in Figure 3.20b. In order to mitigate the redeposition effect, a side-etching step combined with the Halogen-based reactive ion etching (RIE) and inductively-coupled plasma (ICP) techniques [126–128] is needed by rotating and tilting the wafer. Nevertheless, other concerns arise. For instance, the shadowing effect (limited etching coverage at the lower corner of the MTJ profile due to insufficient spacing between MTJs) [116, 124] limits a high-density array patterning, and magnetic layer corrosion degrades the reliability of MTJ devices due to the non-volatile chemicals attached to the CoFeB layers. Another critical issue is magnetic coupling effect [65] between different ferromagnetic layers after the MTJ nanopillars are patterned. Many prior works [65, 79, 129, 130] show that stray fields at the FL from underlying ferromagnets have a significant impact on the switching characteristics and retention time of MTJ devices. As the magnetic couple effect applies globally to all cells on a wafer with the same design, it is generally not considered as a spot defect.

After the MTJ etching process, encapsulation and CMP are required to separate individual MTJ pillars. In this step, an oxygen showering post-treatment (OSP) can be applied to recover patterning damage so as to improve the electrical and magnetic properties of MTJ devices [131]. The oxygen showering process selectively oxidizes the perimeter (damaged by previous ion beam etching) of the MTJ pillar with non-reactive oxygen ions. However, over-oxidization into the MTJ device also causes degradation in key device parameters such as *TMR*. Thus, the OSP condition needs to be carefully tuned to maximize the damage suppression while protecting the inner undamaged parts.

Other manufacturing defects that may arise during the fabrication of MTJ devices include synthetic ferromagnetic flip (SAFF), intermediate (IM) state, and back-hopping. These defects are not strongly linked to a specific manufacturing step introduced previously. Instead, they are more related to thin film materials and MTJ stack designs. For example, we experimentally observed the SAFF defect in some MTJ devices; it means that the magnetization in both the hard layer and reference layer undergoes an unintended flip to the opposite direction. Due to such a defect, the polarity of stray field at the free layer reverses, leading to intermittent passive neighborhood pattern sensitive fault within a STT-MRAM array. A possible cause is an initial flip of hard layer with reduced

coercivity due to inhomogeneities arising during the device fabrication. More details about SAFF defect will be presented in Chapter 8. The IM state defect manifests itself as an abnormal resistive state between R_P and R_{AP}. It results in intermittent transition faults due to the undefined resistive state. The root causes of IM state can be attributed to some physical imperfections such as unreversed magnetic bubbles [132], inhomogeneous distribution of stray field [79] or even skyrmion generation [133]. More details about IM state defect will be presented in Chapter 9. The back-hopping phenomenon is another failure mechanism that afflicts STT-MRAM. It means that an MTJ hops back to its initial state after a transition write operation with a relatively high write current, leading to write failure [134]. A key characteristic of back-hopping phenomenon is that the switching probability typically increases with the write voltage, but it abnormally decreases after the write voltage reaches certain threshold, as shown in Figure 3.21. The occurrence of back-hopping during a write operation originates from the flip of reference layer under spin-transfer torque and consequent alternating reversal of both reference layer and free layer [134, 135]. A solution to address this issue is to improve the reference properties so as to increase its switching voltage under spin-transfer torque [135].

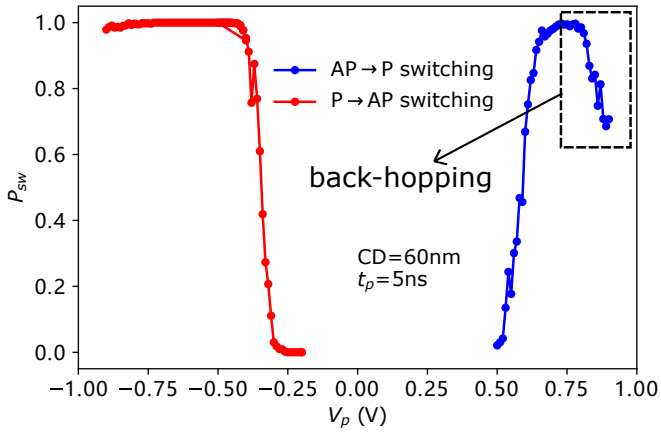


Figure 3.21: AP→P switching probability decreases at high write voltage due to back-hopping phenomenon.

3.5. STT-MRAM PAST, PRESENT, AND FUTURE

In physics, charge and spin are two intrinsic properties of electron. The exploitation of electron charge has driven extraordinary prosperity of the semiconductor industry in the past decades since the invention of transistor in 1947. As the transistor size shrinks over time, the number of transistors in a given IC area doubles about every two years, which is well known as Moore's law. Benefiting from the Moore's law, we have been enjoying cheaper and faster IC chips. However, the increased manufacturing complexity and leakage power in the sub-45 nm era has significantly slowed down the continuation of Moore's law.

The development of spintronics which utilizes both the spin and charge properties

of electron brings us an exciting alternative solution. Spintronics has the potential to deliver high performance and low power simultaneously, meeting the stringent requirements in emerging applications such as edge AI, IoTs, and automotive. Undoubtedly, STT-MRAM is a representative technology in spintronics. With extensive R&D activities in the past decades, STT-MRAM has reached at a production-ready stage.

In this section, we first retrospect key breakthroughs over the development course of MTJ device, the data-storing element in any type of MRAM. Thereafter, we review the MRAM test chips and commercial products that were demonstrated in the past. As a relatively mature and superior MRAM technology, STT-MRAM will be highlighted, and its potential applications as well as remaining challenges will be discussed separately.

3.5.1. MTJ EVOLUTION COURSE

As introduced previously, MTJ is the core component (data storing element) in MRAMs. Thus, the performance improvement in MTJ is the key driving force to MRAM development. Figure 3.22 shows the key milestones in the MTJ evolution course from three dimensions: MTJ boost, Write technique, and MTJ stack innovation.

A significant part of MTJ innovations have been dedicated to boosting the TMR ratio, since it directly determines the data retrieving (reading) capability in MRAM. The TMR effect was first discovered by Julliere in a Fe-Ge-Co junction in 1975 [136]. It was until in 1995 that the experimental demonstration of the TMR effect in a $CoFe/Al_2O_3/Co$ junction at room temperature (300 K or 25 °C) was reported [137]; the TMR ratio was 11.8%. This major breakthrough was followed by intensive work among researchers around the world on increasing TMR ratio. Although the TMR ratio was boosted gradually over time and up to 70% at room temperature was presented in 2004 [89], this value of TMR in MTJs based on AlO_x barrier was too small for realizing working memories. The obstacle of low TMR in MTJ designs was overcome by replacing AlO_x with MgO as materials of the tunnel barrier, which was theoretically predicted in 2001 [138]. Subsequently, Yuasa *et al.* [139] in 2004 experimentally demonstrated Fe/MgO/Fe MTJs with TMR ratio reach-

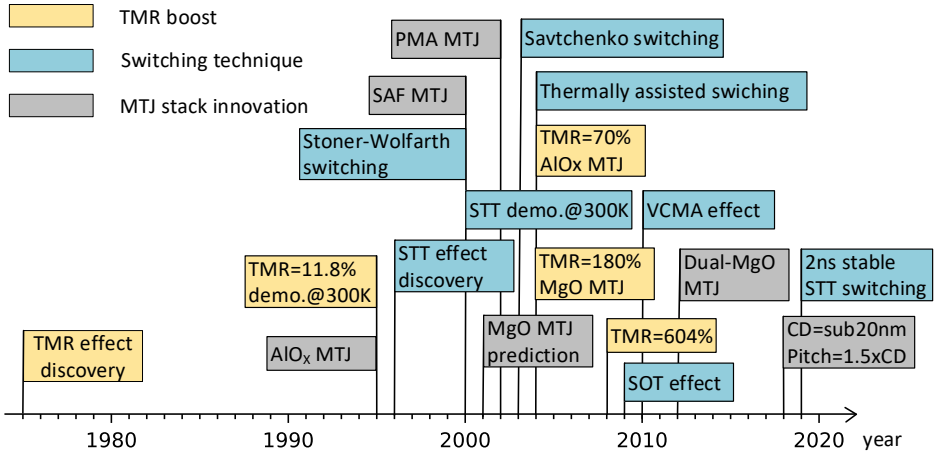


Figure 3.22: Key milestones in the MTJ evolutionary process.

ing 180% at room temperature. Four years later, Ikeda *et al.* [90] unprecedentedly observed a TMR ratio as high as 604% at room temperature by suppression of Ta diffusion in CoFeB/MgO/CoFeB junctions. Despite the TMR ratio in MgO-based MTJs can be over 1000% theoretically at device level, all MRAM test chips demonstrated in recent years have a TMR below 250% due to the high thermal budget (400K) at the CMOS BEOL process. Compared to the large read window (several orders of magnitude difference in the resistance between the LRS and HRS states) in PCM or RRAM, the low TMR value results in a narrow read window in MRAM mega-bit arrays, posing a big challenge for circuit and system designs.

The second dimension of MTJ evolution lies at the improvement of switching technique between P and AP states in terms of robustness, energy efficiency, speed, and manufacturability. Initially, the MTJ state in early MRAM designs was typically switched by applying external magnetic fields generated by two current-carrying metal lines above and below a MTJ device [26]. In 2000, Scheuerlein *et al.* demonstrated a MRAM circuit with a write functionality implemented using the Stoner-Wolfarth switching technique [140]. However, this type of field switching method is inherently flawed with the "half-select" issue. It means that memory cells in the proximity of the addressed cell are also inevitably exposed to the writing fields, thus leading to a non-negligible disturb probability. This problem was then solved by the Savtchenko switching method by Leonid Savtchenko at Motorola in 2003 [141]. This field switching method is also known as toggle switching, which employed a SAF free layer and a sequence of pulses that creates a rotating field to switch the MTJ state. This new type of field switching technique directly led to the first generation of MRAM commercial product, which began production in 2006 by Freescale Semiconductor, which eventually spun off their MRAM sector as Everspin Technologies in 2008. Despite the fact that toggle MRAM is a mature technology and commercial products are available on the market, toggle MRAM is afflicted with high power consumption (~10 mA write current to generate switching fields) and poor scalability to advanced technology nodes. To reduce the power consumption, thermally assisted switching was proposed by Prejbeanu *et al.* in 2004 [142]. In this method, a current first flows through the addressed MTJ to heat it up. As the temperature goes up, the thermal stability is reduced considerably. Therefore, a much smaller switching field is required compared to that at normal temperature. After a write operation, the MTJ cools down for information storage. This method is effective in reducing write current (~1 mA), but the scalability issue was still unsolved and the write speed is slow due to the added heating and cooling steps.

STT switching method emerged as an alternative solution to address the power consumption and scalability issues in the aforementioned field switching methods. The discovery of STT effect dates back to 1996, in which Berger [143] and Slonczewski [144] independently predicted that passing an electric current through an MTJ results in a flip of the magnetization in the FL. Experimental observations of the STT effect were achieved in 2000 by Katine [145] and Albert [146], independently. Due to the use of all-metallic MTJ structure in these two works, the switching current is very high (~5 mA). In 2005, Diao *et al.* from Grandis Inc. presented STT switching on MTJs with AlO_x and MgO barriers; the switching current was reduced to 750 μA and 220 μA, respectively due to the improved spin polarization [147]. Later MTJ stack optimizations such as replacing IMA-

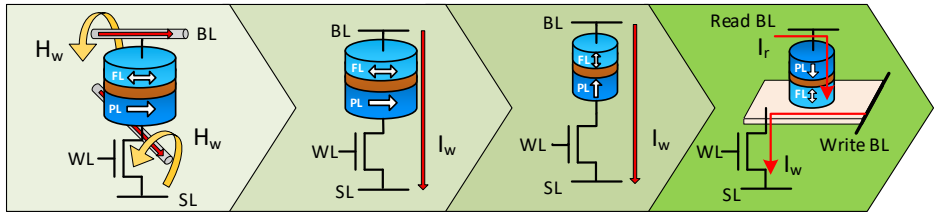
MTJ with PMA-MTJ has cut the switching current down below 100 μA. Since there is no magnetic fields involved in the write operation, STT-MRAM offers much better performance, energy efficiency, and scalibility, compared to the previous toggle MRAM. Due to its intrinsic limitations, the switching speed of STT effect cannot reach sub-ns level, which eliminates the possibility of replacing SRAMs at all cache levels. The fastest stable STT switching that has been demonstrated is 2 ns at a switching current of around 110 μA in 2019 [148]. To further reduce the write power consumption and reach sub-nm speed, novel switching techniques beyond the STT effect such as spin-orbit torque (SOT) effect [34, 149, 150] and voltage-controlled magnetic anisotropy (VCMA) effect [151, 152] are under intensive R&D, which is out of the scope of this thesis.

The third dimension of MTJ evolution is the advancement in the physical MTJ stack towards higher performance, higher reliability, and better CMOS compatibility. The MTJ stack has evolved from the simply three-layer sandwich structure at the early stage to a pillar with more than 15 layers [115]. First, the material for the tunnel barrier changes from metal to AlO_x, and to MgO eventually. In the early MTJ stack designs, metallic materials were used for the tunnel barrier. For example, when the STT effect was first experimentally observed in 2000, the MTJ was a Co/Cu/Co sandwich, which has very low spin-transfer efficiency and thus require a very high switching current [145]. Later on, it was found that the AlO_x-based barrier showed better TMR and switching performance. After around 2004, MTJ designs based on MgO barrier became prevalent. As a MgO capping layer on top of the FL showed better thermal stability without lowering the switching current [153], dual-MgO MTJ designs were widely adopted after 2012. Second, the transition from in-plane magnetic anisotropy (IMA) to perpendicular magnetic anisotropy (PMA) for the ferromagnetic layers is also a critical milestone. The cross-section of IMA-MTJs is an ellipse with an aspect ratio of 2 or more, since the IMA originates from the MTJ shape. In 2002, Nishimura *et al.* first presented a PMA-MTJ device with a circular cross-section [154]. The circular shape significantly reduces the manufacturing complexity, thus enabling higher density of MRAM array. In addition, the PMA originates the CoFeB/MgO interface, and allows a much smaller switching current to flip the FL magnetization. Third, the introduction of synthetic anti-ferromagnetic (SAF) structure in MTJ designs is another key breakthrough [155]. It enables appropriate cancellation of stray fields from the RL at the storage FL. Fourth, the MTJ size and array pitch keeps scaling down in order to reduce the switching current and increase density competing with DRAM. With advanced sputtering and etching techniques, the MTJ's critical diameter (CD) can reach below 20 nm at a pitch of 1.5×CD [115, 156].

3.5.2. MRAM COMMERCIALIZATION

Stimulated by the promising features of MRAM technology, a number of research institutes and semiconductor companies across the world have contributed to its development and commercialization in the past decades. To date, four generations of MRAM technology have been classified and demonstrated, as shown in Figure 3.23. For each MRAM generation, various test chips and/or commercial products have been reported, as illustrated in Figure 3.24. Next, we will elaborate them in detail.

The first generation of MRAM was developed using the magnetic field switching method on IMA-MTJs. In 2000, IBM reported a 1 kb MRAM array based on AlO_x MTJs with



Gen. 1: IMA + Field:
- Not scalable below 130nm
- High write power
- High write disturb
- Complex cell geometry

Gen. 2: IMA + STT:
- Not scalable below 65nm
- Elliptical MTJ cross-section
- Low STT switching efficiency
- Low retention for small MTJs

Gen. 3: PMA + STT:
- Fully scalable
- Circular MTJ cross-section
- Trilemma between speed, endurance, and retention
- High STT switching efficiency

Gen. 4: PMA + SOT/VCMA:
- Fully scalable
- Sub-ns speed
- Infinite endurance
- Low write power

Figure 3.23: MRAM technology commercialization roadmap.

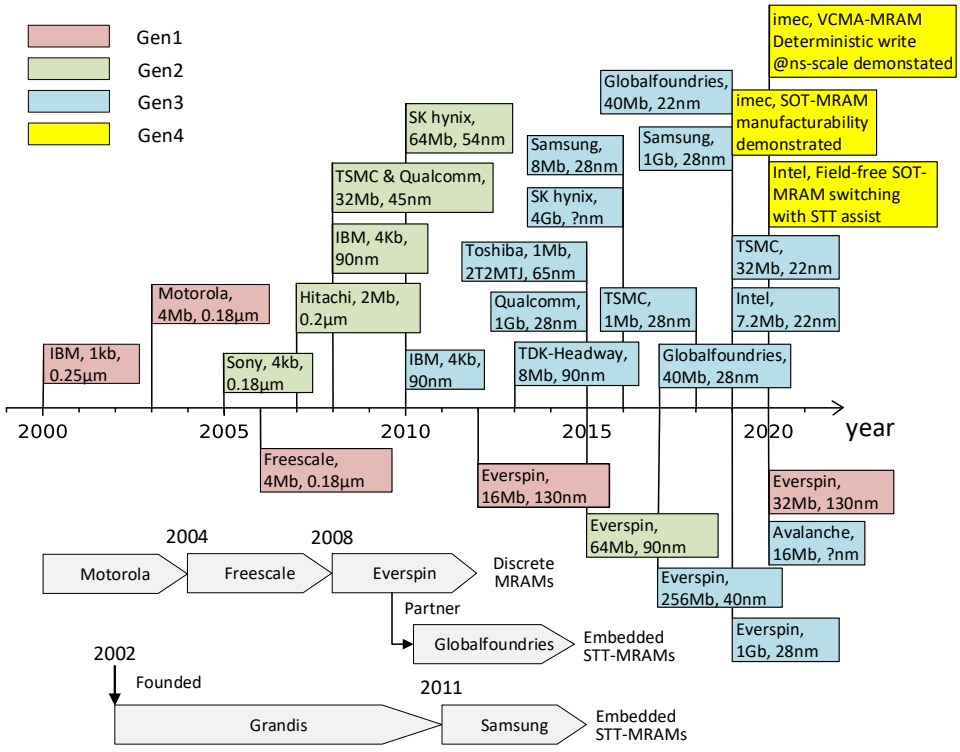


Figure 3.24: MRAM test chips or commercial products over time.

basic read and write circuits [140]. In 2003, Motorola demonstrated a 4 Mb MRAM circuit with 1T-1MTJ cell design and the Savtchenko switching technique which cleverly addressed the “half-select” issue in the previous Stoner-Wolfarth switching method [157]. This chip was later commercialized in 2006 by Freescale [158], a spin-off from Motorola

in 2004. In 2008, Freescale span off its MRAM business as Everspin Technologies, which upgraded its standalone toggle MRAM product to 16 Mb in 2012 and 32 Mb in 2020 [27, 28]. Everspin's toggle MRAM products provide parallel and series peripheral interfaces, which are competitive memory for applications that must store and retrieve data and programs quickly using a minimum number of pins. The first generation of MRAM (i.e., Toggle MRAM) features high speed (~35 ns symmetric read and write), unlimited endurance as the field switching method does not cause any damage to MTJ devices over time. In addition, it is radiation immune, suitable for aerospace applications. However, the biggest limitation facing the first-generation MRAM is its poor down-scaling ability. As MTJ shrinks and density increases, higher write power and error rate are incurred. The cell geometry with current-carrying wires above and below MTJs to generate switching fields also adds to manufacturing and scaling complexity.

STT-MRAM emerged as a better solution for scalability. Initial STT-MRAM development focused on IMA-MTJs, which have an elliptical cross-sectional shape. This type of STT-MRAM is typically referred to as the second generation of MRAM. In 2005, Sony first demonstrated a 4 kb STT-MRAM chip on a 0.18 μm process [159]. Two years later, Hitachi presented an STT-MRAM test chip with 2 Mb capacity and 0.2 μm process [160]. Later, IBM [161], TSMC & Qualcomm [93], and SK hynix [162] also separately disclosed their STT-MRAM solutions, as shown in Figure 3.24. The first STT-MRAM product arrived in 2015, released by Everspin; this 64 Mb chip was based on IMA-MTJ technology and was built on 90 nm CMOS technology [28, 163]. With the advent of PMA-MTJ technology, commercialization attempts shifted to perpendicular STT-MRAM, also known as the third generation of MRAM. The first step of commercializing perpendicular STT-MRAM was taken by IBM, which demonstrated a 4 kb test chip in 2010 [164]. Later on, most of semiconductor companies across the globe recruited their own STT-MRAM teams focusing on STT-MRAM R&D. The momentum in pursuit of STT-MRAM technology as a future memory is greater than ever before. With a number of test chips demonstrated in the past decade [92, 109, 110, 165–168] (see Figure 3.24), STT-MRAM R&D reached a peak in 2019. In this year, Everspin launched a game-changing 1 Gb standalone STT-MRAM product with DDR4 interface, targeting the replacement of DRAM in some applications such as enterprise SSDs [28]. Samsung [14], TSMC [32], Globalfoundries [4], and Intel [30] demonstrated embedded STT-MRAM macros up to 1Gb; all of these foundries claimed that their perpendicular STT-MRAM technology is production ready. In addition, Avalanche, a start-up founded in 2006, also offers STT-MRAM chips up to 64Mb in both stand-alone and embedded forms [29]. Other start-ups around the globe include Spin Memory, Numem, Hikstor Tech., HFC Semiconductor etc. With an ecosystem around STT-MRAM gradually getting shaped, it is believed that STT-MRAM mass production and deployment in industry is around the corner.

To further cut down write power and provide sub-ns speed, several novel switching mechanisms beyond the STT effect have been intensively studied targeting ultra-low power and low-level cache applications. MRAM technologies exploiting these new switching mechanisms are the fourth generation of MRAM. Two most promising representatives are spin-orbit torque (SOT) [26, 28, 169] and voltage-controlled magnetic anisotropy (VCMA) [170, 171]. SOT devices have three terminals with an SOT track below a top-pinned MTJ, as illustrated in Figure 3.23. The SOT switching is archived by

applying an in-plane current through the SOT track, generating an SOT at the interface between the free layer and the SOT track. The SOT then switches the magnetization of the FL as fast as <1 ns. Since the write current does not pass through the MTJ device, the reliability and endurance of SOT device are highly improved in comparison to the STT-switching method. Despite its promising prospect, SOT-MRAM is still facing many challenges. One of the biggest challenges is that an in-plane magnetic field is required for deterministic switching. In 2019, imec proposed a field-free switching SOT-MRAM concept, which was demonstrated on a 300 mm using COMS-compatible processes [34]. In 2020, Intel also demonstrated a CMOS-compatible process of field-free SOT device; the deterministic switching speed is 10 ns with the assistance of STT effect [31]. VCMA switching is another promising fourth-generation switching mechanism. It manipulates the magnetization in the FL with voltage (i.e., electric field) instead of electric current. Since no charge flow is required, VCMA-MRAM is in principle more energy efficient than STT-MRAM. However, fundamental innovations (e.g., deterministic switching) are still required before it becomes a practical working memory.

3.5.3. STT-MRAM POTENTIAL APPLICATIONS

With the advent of first-version STT-MRAM products, finding the right position in a market place where existing memories are still dominant in the foreseeable future is extreme important. Depending on its physical form, STT-MRAM can be classified into two categories: discrete and embedded. Discrete STT-MRAMs are packaged into standalone chips, which are intended to be mounted into printed circuit board. Embedded STT-MRAMs are integrated into logics, forming SoCs along with other components such as digital, analog, mixed-signal circuits. Thanks to its flexible tunability in speed/power, endurance, and retention, STT-MRAM can be tailored with three flavors: SRAM-like, DRAM-like, and flash-like, as depicted in Table 3.4. As these three flavors span a large range in the memory hierarchy, STT-MRAM is dubbed as the future "universal memory".

Obviously, different applications have different requirements on memory and/or storage. For cache applications, performance and endurance are the two most critical metrics. The demonstrated STT switching with peripherals typically cannot reach below

Table 3.4: STT-MRAM potential applications and the associated requirements.

Flavor:	SRAM-like	DRAM-like	Flash-like
Requirements:	• R/W: 5-15 ns • endurance>10^{13} • Retention: D/M	• R/W: 25-50 ns • endurance>10^{10} • Retention: >10Y	• R: 25ns, W: 50-500ns • Endurance: 10^{6}-10^{8} • Retention: >10Y
Specific targets:	• Battery-backed cache • Last-level cache	• Power-fail protected write buffer • Persistent memory	• Data/code storage in MCUs • eFlash replacement
Applications:	Enterprise SSDs, MCUs, AIoT, aerospace, neuromorphic computing, automotive		

5 ns, making today's STT-MRAM technology infeasible for L1/L2 caches. Nevertheless, STT-MRAM is qualified to serve as last level caches, by trading retention (relaxing to days or months) for high access speed at ~10 ns. Compared to SRAM, STT-MRAM has much smaller cell size and lower leakage power. Given these features, SRAM-like STT-MRAM has the potential to replace SRAM in IoT, mobile, and wearable devices where power is backed by battery and speed is not the biggest consideration.

By relaxing write speed to 25-50 ns, STT-MRAM becomes a good candidate for DRAM replacement in some specific applications. For example, Everspin is currently shipping 1Gb STT-MRAM standalone parts with DDR3 interface, with the aim of replacing DRAM in enterprise SSDs [172]. Unlike consumer-class SSDs, enterprise SSDs put a high priority on protecting in-flight data in the event of power loss or interruption. Conventionally, volatile DRAM is used for data buffers, and super capacitors are deployed for energy storage to flush all in-flight data to non-volatile flash array when power fails. By employing STT-MRAM, the power-fail protection is considerably simplified. Furthermore, STT-MRAM can replace DRAM completely for storing flash translation layer (FTL) tables and alleviating write amplification. IMB and Buffalo both disclosed that they employ Everspin's STT-MRAM chips as data caches in their high-end SSDs [173]. With a high performance comparable to DRAM and data retention higher than ten years, STT-MRAM is considered as the most promising persistent memory. However, cost per bit and density are the two biggest weaknesses for STT-MRAM when competing with DRAM.

By further relaxing write performance and boost data retention for a wide range of operating temperature (-40-125 °C), STT-MRAM enters into the storage domain (i.e., flash-like flavor). Despite the fact that STT-MRAM is no rival to NAND flash for standalone bulk storage in terms of density and cost per bit, embedded STT-MRAM is a perfect candidate to replace on-chip eFlash memories. It is well recognized that eFlash is getting prohibitively complex and expensive to scale below the 28 nm node [32]. As an alternative, STT-MRAM offers much better performance in speed, scalability, and endurance. In addition, STT-MRAM is friendly to CMOS processes, as MTJ devices can be plugged into BEOL metal layers with minimal process changes. Due to the promising prospect of embedded STT-MRAM, most foundries including TSMC, Globalfoundries, Intel, Samsung have announced their embedded STT-MRAM solutions up to 1 Gb capacity [14].

Moving eyes from the memory hierarchy and considering from a perspective of high-level applications, STT-MRAM is also very suitable for AIoT, aerospace, neuromorphic computing, and automotive applications. When artificial intelligence meets internet of things (i.e., AIoT), a trade-off between compute, power, and cost has to be made. Conventionally, on-chip SRAM is used for accelerating convolutional neural network (CNN) in AI processors. However, SRAM suffers from its large cell area and leakage power at advanced technology nodes, which pose a bottleneck of performance for edge AI chips. In 2019, Sun *et al.* presented a CNN accelerators fabricated using 22 nm CMOS technology for mobile and IoT applications [174]. This chip employs embedded STT-MRAM instead of SRAM to achieve better power efficiency (9.9 TOPS/W). In addition, aerospace applications require high reliability and radiation immunity. MRAM technologies are intrinsically resistant to radiation. For example, AAC Microtec used Everspin's MRAM chips replacing both flash and battery-backed SRAM in an earth observation satellite [175].

3.5.4. STT-MRAM REMAINING CHALLENGES

Although STT-MRAM has attracted extensive R&D attentions from both academia and industry, there exists many challenges that need to be addressed before it becomes a ubiquitous memory technology like SRAM and DRAM in the semiconductor industry.

1) High-quality and cost-efficient test solutions for mass production. Currently, one of biggest barriers for STT-MRAM to penetrate into memory markets is its high cost per bit, compared to incumbent memories. To reduce cost, the most effective way obviously is mass production. To this end, high-quality test solutions are paramount in a bid to weed out defective chips and ensure quality chips being shipped to end customers in a cost-efficient manner. The work carried out in this PhD dissertation mainly focus on this topic, including: (a) understanding STT-MRAM-specific defects and propose accurate defect models; (b) developing accurate and realistic fault models; and (c) developing high-quality and cost-efficient STT-MRAM tests.

2) Advanced process technology to achieve high yield. As introduced previously, the STT-MRAM manufacturing process require not only the conventional CMOS processes, but also STT-MRAM unique processes to fabricate and integrate MTJ devices into two adjacent metal layers. The latter necessitate special sputtering and etching tools etc., which are currently under development by equipment companies such as Applied Materials as well as start-ups such as Hprobe. The quality of these manufacturing tools has a direct impact on the process yield. Challenges on this topics include: (a) minimizing the MTJ edge damage induced by high-energy ion beam etch; (b) sputtering ultra-smooth surface for thin films in the MTJ stack; (c) suppressing atom inter-diffusion between different layers in MTJ devices; and (d) controlling process variations especially in R_P and R_{AP}, and (e) ensuring magnetic immunity to external fields to certain degree in accordance with an industrial standard to be established.

3) Enlarge read window for Gigbit arrays. Typically, the measured TMR in demonstrated chips in recent years is 150%-200%. This results in a very small read window when compared to other NVM technologies such as PCM and RRAM where the difference between LRS and HRS can be several orders of magnitude. Worse still, TMR reduces as temperature arises [176]. To enlarge the read window for Gigbit STT-MRAM arrays, the TMR ratio at device level needs to be further increased as much as possible. At circuit level, PVT variations and parasitic resistance and capacitance also need to be taken into account. A solution to tolerate these variation sources is optimizing the resistance of reference cell. For example, Yun *et al.* [177] proposed an STT-MRAM BIST which is capable of automatically trimming the reference resistance, so as to set it in the middle of the R_P and R_{AP} distributions for an STT-MRAM array.

4) Robust memory array and peripheral designs. STT-MRAM has several unique mechanisms such as stochastic switching, magnetic coupling, and thermal fluctuation. These mechanisms pose a large challenge for robust STT-MRAM designs. For example, to mitigate write failure caused by the STT-switching stochasticity, write-verify-write scheme [30] and self-write-termination scheme [102] have been proposed and adopted in the industry. In addition, magnetic coupling effects from both internal and external sources [65, 86] should be taken into account when designing STT-MRAM arrays. Robust sense amplifier designs which are capable to tolerate PVT variations also play a critical role in enhancing robustness.

5) Break endurance/speed/retention trilemma. Technically speaking, the biggest barrier that impedes STT-MRAM from revolutionizing the conventional memory hierarchy and becoming a true universal memory is the trilemma between endurance, speed, and retention [26]. It means that the improvement in one metric leads to a degradation on the other one or two metrics. For example, an increase in retention typically results in higher write power and slower write speed. This is the main reason why there exists three flavors of STT-MRAMs at the moment, as shown in Table 3.4. To unleash full potential of STT-MRAM, fundamental innovations are still required to break the trilemma.

4

TESTING STT-MRAM WITH CONVENTIONAL APPROACH

As STT-MRAM technology becomes more and more mature, a high-quality test solution is a key enabler for its mass production. The STT-MRAM manufacturing process involves not only the conventional CMOS process but also MTJ fabrication and integration. The latter is more vulnerable to defects as it requires deposition, etch, and integration of magnetic materials with new tools. A blind application of conventional tests for existing memories such as SRAMs and DRAMs to STT-MRAMs may lead to test escapes and yield loss. Thus, special attention needs to be paid to testing STT-MRAM defects especially those in MTJ devices which are the data-storing elements. This chapter explores STT-MRAM testing using the conventional memory test approach where a physical defect is always modeled as a linear resistor (i.e., open, short, and bridge) irrespective of its physical nature. We develop a Verilog-A compact model for defect-free MTJ devices, and calibrate it with measurement data of MTJ devices fabricated at imec. Thereafter, STT-MRAM defects are modeled with linear resistors at all possible locations in a single STT-MRAM cell. After injection of each resistor, a systematic fault modeling process is conducted based on SPICE circuit simulations of the STT-MRAM netlist. The simulation results suggest that resistive defects only lead to two types of fault: transition faults and incorrect read faults. To detect these faults, March tests such March C- can be used.

The contents of this chapter have been published in ETS'19 [62] and TETC'19 [46].

4.1. VERILOG-A COMPACT MODEL FOR DEFECT-FREE MTJS

To facilitate STT-MRAM designs empowered by computer-aided design tools, an accurate and computation-efficient MTJ model is a must. In the literature, there are many MTJ models that have been proposed by different researchers. Typically, MTJ models can be classified into three categories: micro-magnetic models, macro models, and behavioral models [178]. Micromagnetic models feature high simulation accuracy, as they model the movement of individual magnetic moment using micromagnetic simulation tools such as OOMMF [179]. This type of model is useful for understanding the physical switching process of a single MTJ device, but inappropriate for simulating a large STT-MRAM array due to the complexity. Macro models (e.g., the model in [180]) are composed of basic circuit elements such as resistors, capacitors, and voltage sources. They are beneficial for their compatibility with circuit simulators. But the number of circuit elements increases with the complexity of the MTJ's dynamic characteristics. Behavioral models (e.g., the model in [181]) are often written in hardware description languages such as Verilog-A. They provide friendly compatibility with circuit simulators and have a good balance between simulation accuracy and speed.

Given the different features of the above three types of MTJ model, we use a behavior MTJ model for the work in this thesis, as our fault modeling framework has to efficiently run on circuit simulators. Our PMA-MTJ compact model is built and improved based on the work in [181]; it is also calibrated with measurement data of fabricated MTJ devices at imec. In Chapter 3, we introduced the key technology and electrical parameters of MTJ. The essence of the MTJ compact model is to map device technology parameters to electrical ones (i.e., R_P, R_{AP}, I_c(P→AP), I_c(AP→P), t_w(P→AP), and t_w(AP→P)). Next, first derive and calibrate the modeling results of MTJ resistance at various bias voltages with measured R-V hysteresis loops. Thereafter, we repeat the same thing for I_c(AP→P), t_w(P→AP), and t_w(AP→P) by modeling and measuring the switching current for various pulse widths.

4.1.1. BIAS DEPENDENCE OF MTJ RESISTANCE

We consider CoFeB/MgO/CoFeB MTJ devices [78] for our work. Despite this choice, our approach is generic and can be applied to any type of MTJ device. The device tunneling conductance is bias-voltage dependent, as shown in Figure 4.1 by the measured R-V hysteresis loop for a sample device with CD=60 nm. The physical model in [182] shows that the resistance is mainly determined by the MgO barrier thickness and the interfacial effects between the barrier and neighboring CoFeB layers. We use two simplified Equations (4.1) and (4.2) from [91] to model R_P at varying bias voltage.

$$R_P(V) = \frac{R_0}{1 + s \cdot |V|} \tag{4.1}$$

$$R_0 = \frac{t_{ox}}{F \cdot \sqrt{\bar{\varphi}} \cdot A} \exp(\text{coef} \cdot t_{ox} \cdot \sqrt{\bar{\varphi}}) \tag{4.2}$$

where t_{ox} is the MgO barrier thickness, $\bar{\varphi}$ the potential barrier height of MgO, A the horizontal cross-section of the MTJ device. F, coef, and s are fitting coefficients depending on the RA product as well as the material composition of the MTJ layers. TMR decreases

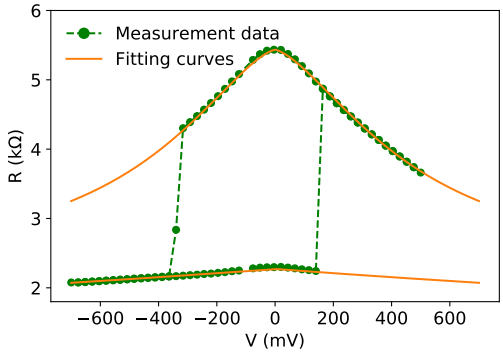


Figure 4.1: Curve fitting of R_P and R_{AP}.

with bias voltage; the relation is modeled with Equation (4.3) [91]:

$$TMR(V) = \frac{TMR(0)}{1 + \frac{V^2}{V_h^2} + b \cdot V^{\frac{4}{3}}} \tag{4.3}$$

It is worth noting that we added a correction term (i.e., $b \cdot V^{\frac{4}{3}}$) in the denominator to get a better fitting result in comparison to the original equation in [91]. $TMR(0)$ is the TMR ratio at 0 V, and V_h is the bias voltage when $TMR(V_h) = 0.5\,TMR(0)$. Based on Equations (4.2–4.3), R_{AP} at certain bias voltage can be derived with Equation (4.4).

$$R_{AP}(V) = R_0 \cdot (1 + TMR(V)) \tag{4.4}$$

The solid curves in Figure 4.1 show our fitting results of R_P and R_{AP}, which match the measurement data.

4.1.2. SWITCHING CURRENT AT VARIOUS PULSE WIDTHS

Since the switching behavior of the MTJ state is intrinsically stochastic, we measured the switching voltage V_c in steps of 10 mV from 0% to 100% switching probability P_{sw} for a given pulse width. For example, we observed that V_c spans from −0.7 V at P_{sw}=0% to −0.9 V at P_{sw}=100% for the AP→P transition, at a pulse width of 12 ns. Based on the measured V_c at various switching probabilities, we extracted V_c at P_{sw}=50% as the average switching voltage. Thereafter, we derived the switching current I_w based on the above-mentioned R-V fitting curves. Figure 4.2 shows the derived I_w data for both P→AP and AP→P transitions at various pulse widths from 4 ns to 100 ns.

The Landau-Lifshitz-Gilbert equation under the macrospin assumption is commonly used to model the switching dynamic of the magnetization in the FL [75]. Depending on the mechanism which dominates the switching event, the entire switching spectrum can be divided into two regimes: (1) precessional, (2) thermal activation regimes.

In the *precessional* regime, the STT effect is the main driving force flipping the magnetization in the FL with a pulse width less than ~40 ns. To switch the state, I_w has to be

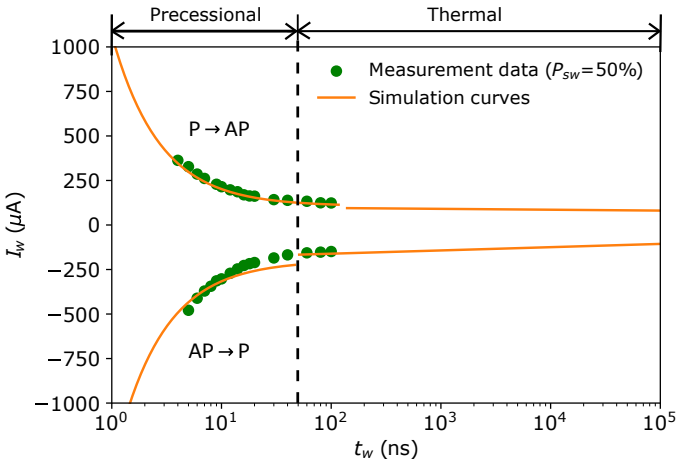


Figure 4.2: Measured vs. simulated results of I_{w} at varying pulse width.

larger than the critical switching current I_{c} defined as [91]:

$$I_{\mathrm{c}} = 2\alpha \frac{\gamma e}{\mu_{\mathrm{B}} \cdot g} E_{\mathrm{B}} \tag{4.5}$$

$$E_{\mathrm{B}} = \frac{\mu_0 \cdot t_{\mathrm{FL}} \cdot M_{\mathrm{s}} \cdot A \cdot H_{\mathrm{k}}}{2} \tag{4.6}$$

$$g = \frac{\sqrt{TMR \cdot (TMR+2)}}{2(TMR+1)} \tag{4.7}$$

where α is the magnetic damping constant, γ the gyromagnetic ratio, e the elementary charge, μ_{B} the Bohr magneton, μ_0 the vacuum permeability, t_{FL} the thickness of the FL, M_{s} the saturation magnetization, H_{k} the magnetic anisotropy field, and g the spin polarization efficiency factor which can be estimated by *TMR*. The switching time $t_{\mathrm{w}}^{\mathrm{pr}}$ in this regime can be estimated using Sun's model [183] as follows.

$$\frac{1}{t_{\mathrm{w}}^{\mathrm{pr}}} = \frac{2}{C + \ln(\frac{\pi^2 \Delta}{4})} \cdot \frac{\mu_{\mathrm{B}} P}{e \cdot m(1+P^2)} \cdot (I_{\mathrm{w}} - I_{\mathrm{c}}) \tag{4.8}$$

where $C \approx 0.577$ is Euler's constant, $\Delta = \frac{E_{\mathrm{B}}}{k_{\mathrm{B}} T}$ the thermal stability, P the spin polarization of the FL and the PL, and m the FL magnetization.

In the *thermal activation* regime where the pulse width increases above 40 ns, observed in our devices, a small current less than I_{c} is able to flip the magnetization due to the increased thermal fluctuation. The thermal fluctuation plays a main role in determining the switching behavior. In this regime, the Neel-Brown model can be used to describe the switching time $t_{\mathrm{w}}^{\mathrm{T}}$ [184]:

$$t_{\mathrm{w}}^{\mathrm{T}} = \tau_0 \exp(\Delta(1 - \frac{I_{\mathrm{w}}}{I_{\mathrm{c}}})) \tag{4.9}$$

Our model is based on combining the model of the precessional regime and the thermal activation regimes. Figure 4.2 shows clearly that by appropriately combining these

regimes, we obtain simulation results which are in line with data measured on actual MTJ wafers. Note that the boundary between the two switching regimes is not strictly demarcated. It is significantly impacted by Joule heating. Given that R_{AP} is more than twice as large as R_P, the heat generated during an AP→P transition is much higher than the opposite direction. Therefore, the thermal activation regime of an AP→P transition shifts towards the left compared to a P→AP transition.

4.2. DEFECT MODELING WITH LINEAR RESISTORS

Defect modeling is the first critical step in the test development process. It abstracts physical defects and presents them at electrical level so as to be processed by circuit simulators such as SPICE. Therefore, having an accurate defect model that is able to mimic the way the physical defect manifests itself at the electrical level is the best way to close the gap between the reality and the abstraction (fault models). Next, we will discuss the defect models for interconnects/contacts.

Traditionally, a spot defect in an electronic circuit is modeled as a linear resistor, and the defect strength is represented by its resistance value [36, 55, 56]. For instance, missing material is modeled as a disconnection, while extra material is modeled as an undesired connection. These undesired connections and disconnections can be typically classified into three groups as follows. [36, 185].

- Open: An undesired extra resistor (R_{op}) within a connection; $0\,\Omega < R_{op} \leq \infty\Omega$.
- Short: An undesired resistive path (R_{sh}) between a node and power supply (either V_{DD} or GND); $0\,\Omega \leq R_{sh} < \infty\Omega$.
- Bridge: A parallel resistor (R_{br}) between two connections; $0\,\Omega \leq R_{br} < \infty\Omega$.

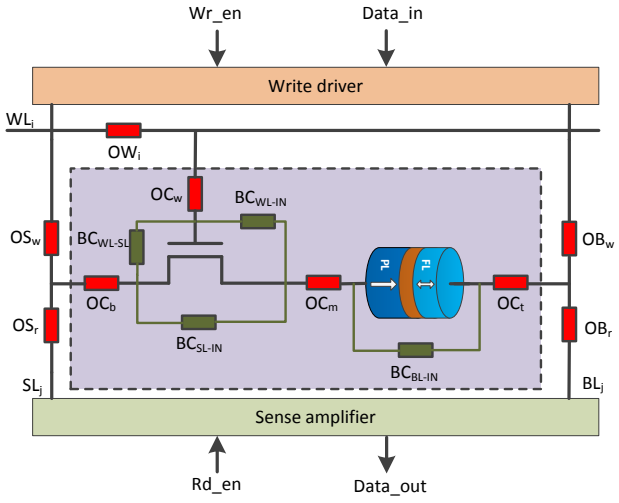


Figure 4.3: Resistive defects in a single 1T-1MTJ memory cell.

Figure. 4.3 illustrates how the above models are used to model some defects in interconnects and contacts of a single-cell STT-MRAM. For instance, OC_m denotes an open between the NMOS selector and the MTJ device; it can be used to model the missing material defect on the contact shown in Figure. 3.15. BC_{BL-IN} denotes a bridge bypassing the MTJ device; it can be used to model the extra material redeposited on the MTJ sidewalls. Theoretically, there are four opens, six bridges, and eight shorts within a single STT-MRAM cell. Outside the memory cells, resistive defects can also occur in/between the WL, BL, and SL. For instance, OB_w denotes an open in the bit line disconnecting the memory cell with the write driver, while OB_r denotes an open in the bit line disconnecting the memory cell with the sense amplifier. It is worth noting that some resistive defects are not realistic when considering the physical layout of the design, as also emphasized in [56]. For example, shorts connecting the inner node (between the MTJ and NMOS) to V_{DD} or GND and bridges between the BL and WL are not possible, since they reside in different metal layers which are far away from each other [56].

4

4.3. FAULT MODELING

The resistive defect models in Figure 4.3 are used to develop appropriate fault models, which are the targets of a test. A fault model describes the faulty behavior of a specific memory cell in the presence of a given defect. Typically, the fault modeling process consists of two steps: 1) fault space definition and 2) fault analysis/validation. The former defines all possible faults theoretically. The latter validates realistic faults in the presence of the defect under investigation in the pre-defined fault space using SPICE-based circuit simulations. Next, we will elaborate these two steps and present the fault modeling results corresponding to all the restive opens and bridges in Figure 4.3.

FAULT SPACE DEFINITION

Depending on the number of cells involved, memory faults can be classified into three classes as shown in Figure 4.4 [186]: single-cell faults, two-cell faults (i.e., coupling faults), and multi-cell faults (i.e., neighborhood pattern sensitive faults). These faults can be systematically described by *fault primitive* (FP) notation [39]. An FP describes the deviation

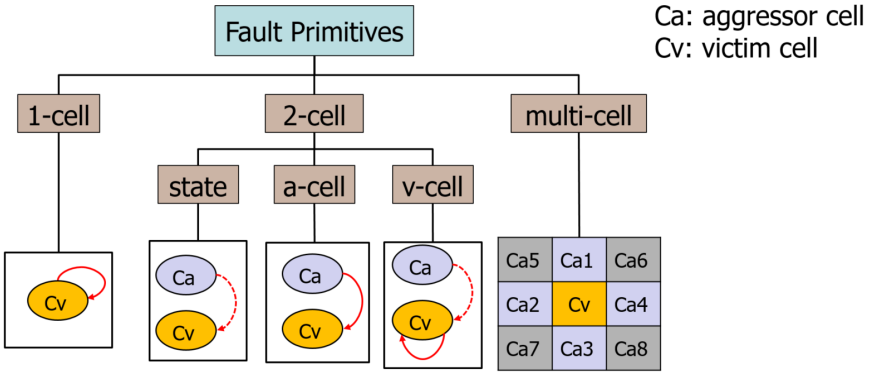


Figure 4.4: Memory fault space and classification.

of the observed memory behavior from the expected one. For a *single-cell fault*, an operation on the addressed cell which is considered as the victim sensitizes this fault irrespective of neighboring cells. A single-cell fault can be denoted as a three-tuple $\langle S/F/R\rangle$, where

- S (sensitizing sequence) denotes an operation sequence that sensitizes a fault. It takes the form of $S{=}x_0O_1x_1\ldots O_mx_m$, where $x_i{\in}\{0,1\}$ ($i{\in}\{0,1,...,m\}$) and $O{\in}\{\mathrm{r},\mathrm{w}\}$. Here, '0' and '1' denote the logic values of memory cells, while 'r' and 'w' denote a reading and a writing operation, respectively. m is the number of operations involved in the sensitizing sequence. If $m \leq 1$, the fault is static, otherwise it is dynamic.
- F (faulty effect) describes the value of the faulty cell after S is performed; $F{\in}\{0,1\}$.
- R (readout value) describes the output of a read operation in case the last operation in S is a read. $R \in \{0,1,-\}$ where '$-$' denotes that R is not applicable.

Table 4.1 lists all single-cell static FPs along with their names and corresponding fault models. Note that a fault model is a non-empty set of fault primitives with similar or complementary properties. For example, TF0: $\langle 0\mathrm{w}1/0/-\rangle$ denotes that an up-transition operation to a cell containing '0' (S=0w1) fails, the cell remains in its initial value '0' (F=0), and the read output is not applicable (R=$-$). TF1: $\langle 1\mathrm{w}0/1/-\rangle$ shares a similar fault behavior but with a down-transition operation (S=1w0). Therefore, both TF1 and TF0 belong to the fault model: *transition fault*. An example of read-related FPs is IRF0: $\langle 0\mathrm{r}0/0/1\rangle$, which denotes a r0 operation on a cell that holds '0' (S=0r0), where the cell remains in its correct state '0' (F=0) while the read output is '1' (R=1) instead of the expected '0'. It belongs to the fault model: *incorrect read fault*. In summary, there are 12 single-cell static FPs, which can be divided into 6 fault models. They are *static fault, write*

Table 4.1: Single-cell static faults.

#	S	F	R	⟨S/F/R⟩	FP Name	Fault Model
1	0	1	–	⟨0/1/–⟩	SF1	State Fault
2	1	0	–	⟨1/0/–⟩	SF0	
3	0w0	1	–	⟨0w0/1/–⟩	WDF1	Write Disturb Fault
4	1w1	0	–	⟨1w1/0/–⟩	WDF0	
5	0w1	0	–	⟨0w1/0/–⟩	TF0	Transition Fault
6	1w0	1	–	⟨1w0/1/–⟩	TF1	
7	0r0	0	1	⟨0r0/0/1⟩	IRF0	Incorrect Read Fault
8	1r1	1	0	⟨1r1/1/0⟩	IRF1	
9	0r0	1	0	⟨0r0/1/0⟩	DRDF1	Deceptive Read Destructive Fault
10	1r1	0	1	⟨1r1/0/1⟩	DRDF0	
11	0r0	1	1	⟨0r0/1/1⟩	RDF1	Read Destructive Fault
12	1r1	0	0	⟨1r1/0/0⟩	RDF0	

disturb fault, transition fault, incorrect read fault, read destructive fault, and *deceptive read destructive fault.*

For dynamic faults which are sensitized by more than one operation (i.e., m>1), their names get the prefix md− where m denotes the number of operations in S. Note that the naming scheme follows the same rules of static FPs using the last operation and its preceding state in S, For example, ⟨1r1w0/1/–⟩ and ⟨1w1w0/1/–⟩ are both named as 2d-TF1. Note that the total number of dynamic faults depends on the number of operations in S (i.e., m value).

Coupling Faults (CFs) can be denoted as $\langle S_a; S_v/F/R\rangle$, where S_a denotes the sensitizing sequence or the state (i.e., 0 or 1) of the aggressor cell (Ca) while S_v denotes the sensitizing sequence or the state of the victim cell (Cv). CFs can be further divided into three groups as shown in Figure 4.4: 1) state CF, 2) a-cell accessed CF, and 3) v-cell accessed CF.

A state CF has the property that the state of Ca (rather than an operation applied to Ca) pins Cv at a faulty state. As shown in Table 4.2, there are in total four state CFs. For example, CFst1: ⟨0;0/1/–⟩ means that a state '0' in Ca forces Cv which is initialized to state '0' to state '1'. These four FPs are compiled to a single fault model, which is named as *state coupling fault.*

Table 4.2: State coupling faults.

#	Sa	Sv	F	R	⟨Sa;Sv/F/R⟩	FP Name	Fault Model
1	x	0	1	–	⟨x;0/1/–⟩	CFst1	State Coupling Fault
2	x	1	0	–	⟨x;1/0/–⟩	CFst0	

Note: x can be 0 or 1.

An a-cell accessed CF indicates that an operation to Ca causes a fault in Cv. Table 4.3 lists all FPs in this group. For example, CFdst1: ⟨1w0;0/1/–⟩ means that a down-transition on Ca disturbs Cv at initial state '0' and forces it to flip to state '1'. Similar interpretations can be derived for non-transition write operations (#3 and #4) and read operations (#5 and #6). In total, there are 12 FPs in this group and they can be together compiled to a single fault model: *disturb coupling fault.*

Table 4.3: Aggressor cell accessed coupling faults.

#	Sa	Sv	F	R	⟨Sa;Sv/F/R⟩	FP Name	Fault Model
1	xw$\bar{x}$	0	1	–	⟨xw$\bar{x}$; 0/1/–⟩	CFdst1	Disturb Coupling Fault
2	xw$\bar{x}$	1	0	–	⟨xw$\bar{x}$;1/0/–⟩	CFdst0	
3	xwx	0	1	–	⟨xwx; 0/1/–⟩	CFdsn1	
4	xwx	1	0	–	⟨xwx;1/0/–⟩	CFdsn0	
5	xrx	0	1	–	⟨xrx; 0/1/–⟩	CFdsr1	
6	xrx	1	0	–	⟨xrx;1/0/–⟩	CFdsr0	

Note: x can be 0 or 1, and $\bar{x}$=~x

A v-cell accessed CF means that an operation applied to Cv while Ca is in a certain state induces a fault in Cv itself. Table 4.4 lists all FPs in this group. For example, CFwd1: ⟨0;0w0/1/–⟩ means that when Ca and Cv are both in state '0', a w0 operation on Cv flip its state from '0' to '1'. This FP belong to the fault model: *write disturb coupling fault*. In total, there are 20 FPs in this group; they are divided into five fault models: *write disturb CF*, *transition CF*, *incorrect read CF*, *read destructive CF*, and *deceptive read CF*.

Table 4.4: Victim cell accessed coupling faults.

#	Sa	Sv	F	R	⟨Sa;Sv/F/R⟩	FP Name	Fault Model
1	x	0w0	1	–	⟨x;0w0/1/–⟩	CFwd1	Write Disturb Coupling Fault
2	x	1w1	0	–	⟨x;1w1/0/–⟩	CFwd0	
3	x	0w1	0	–	⟨x;0w1/0/–⟩	CFtr0	Transition Coupling Fault
4	x	1w0	1	–	⟨x;1w0/1/–⟩	CFtr1	
5	x	0r0	0	1	⟨x; 0r0/0/1⟩	CFir0	Incorrect Read Coupling Fault
6	x	1r1	1	0	⟨x;0r0/1/0⟩	CFir1	
7	x	0r0	1	1	⟨x; 0r0/1/1⟩	CFrd1	Read Destructive Coupling Fault
8	x	1r1	0	0	⟨x;1r1/0/0⟩	CFrd0	
9	x	0r0	1	0	⟨x;0r0/1/0⟩	CFdr1	Deceptive Read Coupling Fault
10	x	1r1	0	1	⟨x;1r1/0/1⟩	CFdr0	

Note: x can be 0 or 1.

For *Neighborhood Pattern Sensitive Faults* (NPSFs) which involve m cells (m>2), the above FP can be extended to $\langle S_{a_0}; ...; S_{a_{m-2}}; S_v/F/R\rangle$, where S_{a_i} ($i \in [0, m-2]$) indicates the sensitizing sequence or state of the aggressor cell a_i and S_v describes the sensitizing sequence or state of the Cv. NPSF is also known as m-coupling faults, meaning that the victim cell is coupled the $m-1$ aggressor cells. Most commonly, these aggressor cells are physically adjacent to the victim cell. For example, in a 3×3 memory array, the central cell Cv is coupled to its direct neighbors Ca1, Ca2, Ca3, and Ca4, as shown in Figure 4.4. Cv may be also coupled to all nine neighboring cells including the four diagonal ones Ca5, Ca6, Ca7, and Ca8. Similar to the above two-cell CFs, NPSFs can also be divided into three categories: state NPSF, a-cell accessed NPSF, and v-cell accessed NPSF. In case of 8 aggressor cells, state NPSF means that certain neighborhood pattern(s) in Ca 1-8 lead to a faulty state in the central Cv. In this category, there are in total 12288 FPs. For a-cell accessed NPSFs, there are in total 12288 FPs. For v-cell accessed NPSFs, there are in total 2560 FPs.

With the above FP theory, the *entire* fault space can be obtained. It can be easily derived that the total number of possible *static* faults consist of 12 single-cell faults, 36 CFs and 15360 NPSFs. With this pre-defined fault space, we can then validate the realistic FPs that may take place in the presence of a specific defect by means of circuit simulations.

FAULT ANALYSIS METHODOLOGY

After the complete fault space is defined, the STT-MRAM netlist or layout with an injected defect model is simulated in a SPICE-based circuit simulator to validate corre-

sponding memory faults. Our fault analysis consists of the following seven steps: 1) circuit generation, 2) defect injection, 3) stimuli generation, 4) circuit simulation, 5) fault analysis, 6) fault primitive identification, and 7) defect strength sweep and repetition of step 2 to 6 until all defects are covered.

To this end, we built up a 3×3 STT-MRAM array along with all necessary peripheral circuits, as introduced in Section 3.2. In our simulations, we used our Verilog-A MTJ compact model with CD=60 nm presented previously for MTJ devices; this model has been calibrated with silicon data. The predictive technology model (PTM) [98] on 45 nm node was adopted to build peripheral circuits along with the NMOS selectors in STT-MRAM cells. In terms of defect injection, we considered resistive opens, resistive bridges, as shown in Figure 4.3. Each time, a specific defect (e.g., $BC_{BL\text{-}IN}$) was injected into the simulation circuit and the faulty behavior of the memory cell was analyzed. The resistance value of each resistor was swept from 1 Ω to 100 MΩ to represent the defect strength in our simulations.

We first simulated the obtained netlist in Cadence's circuit simulator Spectre to verify the design as a defect-free case. Thereafter, we performed static fault analysis and validation of the static fault space defined previously (i.e., single-cell faults, CFs and NPSFs). We assume that the defective cell is located in the center of a 3×3 memory array; the other eight surrounding cells are defect-free. The data pattern in these eight cells were swept from 0 to 255 (in decimal form) to investigate NPSFs. Each time, we injected a resistor as a defect model into our netlist, as shown in Figure 4.3. The resistance was swept from 10^0 to 10^9 Ω using 45 steps which are equally distributed on a logarithmic scale. The same simulation was repeated for all sensitizing sequences before moving to the next resistive model. This above simulation procedure was fully automated by a fault modeling controller written in Python3.

FAULT MODELING RESULTS

Table 4.5 lists the fault modeling results of all resistive opens (see Figure 4.3) in a single 1T-1MTJ cell. For each defect in the table, the sensitized FPs depend on the defect strength (i.e., resistance value in this case). For a given resistance range, a single FP or multiple FPs can be sensitized. Detecting any one among them can guarantee the detection of the corresponding defect range. For example, the fault analysis results of OC_t (representing an open defect between the BL and the MTJ device) results in four different fault groups which depend on the defect resistance. (1) If the resistance of OC_t is below 466 Ω, no FPs are sensitized; thus, it results in a weak fault. In this case, March tests cannot not guarantee the detection and extra efforts are needed to detect it. (2) If the resistance is between 466 Ω to 870 Ω, a single FP IRF0: ⟨0r0/0/1⟩ is sensitized; its detection condition is simply a read operation on the cell which is in logic '0', irrespective of the addressing direction. We denote the detection condition as ⇳(...0, r0, ...). (3) If the resistance is between 870 Ω and 1.6 kΩ, two FPs are sensitized including TF1: ⟨1w0/1/−⟩ and the previous IRF0. In this case, any March test which is able to detect one of the sensitized FPs can guarantee the detection of the open defect OC_t in such resistance range. Considering the detection condition for TF1 is ⇳(...1, w0, r0, ...) which requires more operations than the detection condition ⇳(...0, r0, ...) for IRF0 (marked with bold font), we select ⇳ (...0, r0, ...) as the detection condition for OC_t in this specific resistance range. (4) If the resistance is above 1.6 kΩ, three FPs are sensitized as shown in the table. Again, the

occurrence of IRF0 makes ⇳(...0,r0,...) the simplest detection condition for this defect range.

Similarly, Table 4.6 presents the fault modeling results for all resistive bridges in a single 1T-1MTJ cell. For instance, the resistive bridge $BC_{SL\text{-}IN}$ (which connects the SL to the internal cell node, as shown in Figure 4.3) results in IRF1=⟨1r1/1/0⟩ when the resistance is below 13 kΩ; The detection condition of IRF1 is ⇳ (...1,r1,...). If the resistance is larger than 13 kΩ, it leads to a weak fault. The detection condition for each FP is shown in the last column, and the FP which is easiest to be detected corresponding to each resistance range is marked in bold.

Table 4.5: Single-cell static fault modeling results of resistive opens.

Defect	Resistance (Ω)	Sensitized FP	Fault Model &FP Name	Detection Condition
OC_t & OC_m & OC_b	(466, 870]	⟨0r0/0/1⟩	**Incorrect Read Fault: IRF0**	⇳ (...0,r0,...)
	(870, 1.6k]	⟨0r0/0/1⟩	**Incorrect Read Fault: IRF0**	⇳ (...0,r0,...)
		⟨1w0/1/–⟩	Transition Fault: TF1	⇳ (...1,w0,r0,...)
	(1.6k, +∞]	⟨0r0/0/1⟩	**Incorrect Read Fault: IRF0**	⇳ (...0,r0,...)
		⟨1w0/1/–⟩	Transition Fault: TF1	⇳ (...1,w0,r0,...)
		⟨0w1/0/–⟩	Transition Fault: TF0	⇳ (...0,w1,r1,...)
OS_w	(870, 2k]	⟨1w0/1/–⟩	**Transition Fault: TF1**	⇳ (...1,w0,r0,...)
	(2k, +∞]	⟨1w0/1/–⟩	**Transition Fault: TF1**	⇳ (...1,w0,r0,...)
		⟨0w1/0/–⟩	Transition Fault: TF0	⇳ (...0,w1,r1,...)
OS_r	(180, +∞]	⟨0r0/0/1⟩	**Incorrect Read Fault: IRF0**	⇳ (...0,r0,...)
OB_w	(870, 1.6k]	⟨1w0/1/–⟩	**Transition Fault: TF1**	⇳ (...1,w0,r0,...)
	(1.6k, +∞]	⟨1w0/1/–⟩	**Transition Fault: TF1**	⇳ (...1,w0,r0,...)
		⟨0w1/0/–⟩	Transition Fault: TF0	⇳ (...0,w1,r1,...)
OB_r	(570, +∞]	⟨0r0/0/1⟩	**Incorrect Read Fault: IRF0**	⇳ (...0,r0,...)
OC_w & OW_i	(870, 14M]	⟨0r0/0/1⟩	**Incorrect Read Fault: IRF0**	⇳ (...0,r0,...)
	(14M, +∞]	⟨0r0/0/1⟩	**Incorrect Read Fault: IRF0**	⇳ (...0,r0,...)
		⟨1w0/1/–⟩	Transition Fault: TF1	⇳ (...1,w0,r0,...)
		⟨0w1/0/–⟩	Transition Fault: TF0	⇳ (...0,w1,r1,...)

Table 4.6: Single-cell static fault modeling results of resistive bridges.

Defect	Resistance (Ω)	Sensitized FP	Fault Model &FP Name	Detection Condition
$BC_{SL\text{-}IN}$	[0, 13k)	⟨1r1/1/0⟩	**Incorrect Read Fault: IRF1**	⇕(...1,r1,...)
$BC_{BL\text{-}IN}$	[0, 1.1k)	⟨1r1/1/0⟩	**Incorrect Read Fault: IRF1**	⇕(...1,r1,...)
		⟨0w1/0/–⟩	Transition Fault: TF0	⇕(...0,w1,r1,...)
		⟨1w0/1/–⟩	Transition Fault: TF1	⇕(...1,w0,r0,...)
	[1.1k, 3.1k)	⟨1r1/1/0⟩	**Incorrect Read Fault: IRF1**	⇕(...1,r1,...)
		⟨1w0/1/–⟩	Transition Fault: TF1	⇕(...1,w0,r0,...)
$BC_{WL\text{-}SL}$	[0, 5.6k)	⟨0r0/0/1⟩	**Incorrect Read Fault: IRF0**	⇕(...0,r0,...)
		⟨1w0/1/–⟩	Transition Fault: TF1	⇕(...1,w0,r0,...)
	[5.6k, 56.1k)	⟨0r0/0/1⟩	**Incorrect Read Fault: IRF0**	⇕(...0,r0,...)
$BC_{WL\text{-}IN}$	[0, 7.7k)	⟨0r0/0/1⟩	**Incorrect Read Fault: IRF0**	⇕(...0,r0,...)
		⟨1w0/1/–⟩	Transition Fault: TF1	⇕(...1,w0,r0,...)
	[7.7k, 13.1k)	⟨0r0/0/1⟩	**Incorrect Read Fault: IRF0**	⇕(...0,r0,...)

4.4. TEST DEVELOPMENT

Based on the previous fault analysis results, appropriate test solutions can be developed. All easy-to-detect faults can be detected by March tests. To minimize the test cost, the minimal detection condition for each resistance (defect strength) range is first identified. Thereafter, all the detection conditions for all resistance ranges are merged to obtain an optimal test algorithm. For example, Table 4.5 and 4.6 list all sensitized fault primitives and detection conditions for considered resistive defects in STT-MRAMs. By combining all the detection conditions in the two tables, March algorithms can be derived. For instance, the March element ⇕(w1,r1,w0,r0) or March C- [187, 188] can be used to detect all these easy-to-detect faults.

5

MAGNETIC-FIELD-AWARE COMPACT MODEL OF PMTJ

The performance of STT-MRAM is very sensitive to magnetic fields including both internal stray fields and external disturbance fields. This chapter presents a magnetic-field-aware compact model of pMTJ, which is the data-storing element for STT-MRAM, for magnetic/electrical co-simulation of MTJ/CMOS circuits. We propose a magnetic coupling model for internal stray fields existing in STT-MRAM arrays. Magnetic measurement data of MTJ devices with diameters ranging from 35 nm to 175 nm is collected and used to calibrate the model. We also propose the inter-cell magnetic coupling factor Ψ to indicate coupling strength. This magnetic coupling model is subsequently integrated into our SPICE-compatible compact MTJ model, implemented in Verilog-A. We demonstrate the power of the proposed compact MTJ model for device/circuit co-design of STT-MRAM, by simulating a single MTJ as well as STT-MRAM full circuits. The design space is explored under PVT variations and various configurations of magnetic fields, for the purpose of robustness enhancement of STT-MRAM designs.

Parts of this chapter have been published in DATE'20 with the best paper award [65].

5.1. MOTIVATION AND PRIOR WORK

STT-MRAM is considered as the next-generation non-volatile memory technology for a variety of applications such as enterprise SSD, industrial-grade MCU, automotive, and AIoT [4]. In recent years, its commercialization progress towards both discrete and embedded memories has accelerated with heavy investments from major semiconductor companies worldwide. For example, Everspin first commercialized discrete STT-MRAM (64 Mb) chips in 2015 and started shipping 1 Gb parts in 2019 [28]. SK hynix [167], Samsung [14], Globalfoundries [4], and TSMC [32] all revealed their STT-MRAM solutions in recent years and claimed they are production ready. Similar to the development process of all semiconductor products, STT-MRAM circuit design, design automation and verification using commercial computer-aided design tools play a critical role. Since MTJ devices are the data-storing elements in STT-MRAMs, an accurate and computation-friendly compact MTJ model is required for SPICE-based circuit simulations.

Unlike conventional semiconductor devices such as MOSFETs where only the *charge property* of electrons is exploited, MTJ devices leverage *both the charge and spin properties* of electrons. Therefore, the magnetic properties of MTJ are as important as electrical ones. The performance of MTJ is known to be very sensitive to magnetic fields including both internal stray fields and external disturbance fields. First, an MTJ device contains multiple ferromagnetic layers; each of them generates a stray field, which has a significant impact on the device's performance. It has been shown that the stray field increases as the MTJ dimension shrinks [189], which makes *intra-cell magnetic coupling* a critical constraint for STT-MRAM designs at advanced technology nodes. Second, to compete with DRAM and flash memories, high-density STT-MRAM arrays are required. It was reported that the STT-MRAM array pitch can be made as small as 1.5× the MTJ diameter at sub-20 nm nodes, using advanced nano-patterning techniques [156]. As the pitch decreases, MTJ devices are pushed closer to each other. This makes *inter-cell magnetic coupling* between neighboring cells become increasingly evident, which may lead to write errors [190]. Third, external fields originating from the operating environment may also disturb STT-MRAM operations [4, 191]. Hence, it is crucial to develop a SPICE-compatible compact MTJ model which accurately captures both the magnetic and electrical characteristics of MTJ and is aware of magnetic fields.

Several MTJ models with different features and implementation methods have been introduced in the literature [62, 178, 180, 181, 192–196]. Generally, they can be classified into four categories: 1) micro-magnetic models, 2) TCAD models, 3) macro models, 4) behavioral models [178]. Micro-magnetic MTJ models are implemented using micro-magnetic simulation tools such as OOMMF, which offers high simulation accuracy and is suitable for studies of the switching dynamics of a single MTJ [192]. TCAD MTJ models are implemented using commercial TCAD simulators such as Sentaurus Device, which provides decent simulation accuracy of a single MTJ as well as small MTJ/CMOS circuits [193]. Macro MTJ models are composed of SPICE inbuilt circuit elements such as resistors, capacitors, and voltage-/current-dependent voltage/current sources [180]; this type of MTJ model owns good compatibility with circuit simulators, but the number of circuit elements dramatically increases with the complexity of MTJ's dynamic characteristics. Behavioral MTJ models describe the analog behaviors of MTJ using a hardware description language such as Verilog-A; they gain popularity for circuit-level simulations

due to several advantages including: 1) good compatibility with circuit simulators, 2) fast simulation, 3) flexible configuration with input parameters, and 4) easiness of designing, sharing, and upgrading. In view of this, many Verilog-A MTJ models have been presented and improved over the past decade [62, 181, 194–196]. Nevertheless, these MTJ models were not capable of simulating magnetic coupling effects and external field disturbance on MTJ's performance, which poses a critical constraint for STT-MRAM designs as reported with silicon characterization data in [4, 65, 191].

5.2. THREE SOURCES OF MAGNETIC FIELD DISTURBANCE

Magnetic tunnel junction (MTJ) devices are the data-storing elements in STT-MRAMs. Each MTJ device stores one-bit data in the form of binary magnetic configurations. Figure 5.1a shows the MTJ stack which essentially consists of four layers: FL/TB/RL/HL. The *hard layer* (HL) is composed of $[Co/Pt]_x$, which is used to pin the magnetization in the upper *reference layer* (RL). The RL is generally built up with a Co/spacer/CoFeB multilayer, which is anti-ferromagnetically coupled to the HL. These two layers form a *synthetic anti-ferromagnetic* (SAF) structure, providing a strong fixed reference magnetization in the RL. The *tunnel barrier* (TB) layer is made of dielectric MgO, typically ~1 nm. The *resistance-area* (RA) product is commonly used to evaluate the TB resistivity, as it depends on the TB thickness but not the device size. The CoFeB-based *free layer* (FL) is the data-storing layer where the magnetization can be switched by a spin-polarized current. Note that the magnetization is perpendicular to the FL of MTJ (i.e., pMTJ); pMTJ offers better scalability towards smaller sizes and less write power, as opposed to the counterpart with in-plane magnetization [46]. Therefore, we limit our discussions to pMTJ which dominates today's STT-MRAM designs in industry.

To work properly as memory devices, MTJs need to provide read and write mechanisms, which are realized by the *tunneling magneto-resistance* (TMR) effect and the

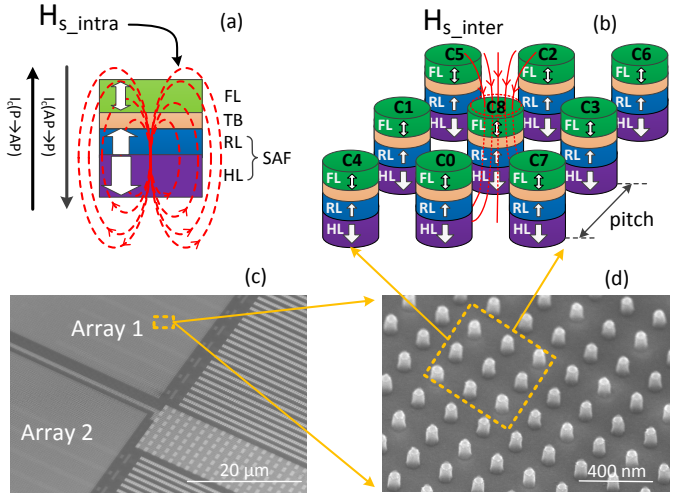


Figure 5.1: (a) MTJ stack and the intra-cell stray fields from the RL and HL, (b) 3×3 MTJ array and the inter-cell stray fields from neighboring cells, (c) SEM image of the 0T0R wafer floor plan, and (d) SEM image of MTJ array.

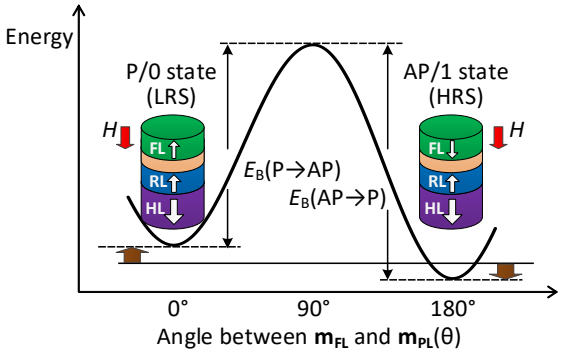


Figure 5.2: Energy barrier E_B between AP and P states is bifurcated into E_B(P→AP) and E_B(AP→P) due to magnetic field H at the FL.

spin-transfer-torque (STT) effect [75]. Due to the TMR effect, the MTJ's resistance is low (R_P) when the magnetization in the FL is parallel to that in the RL, while the resistance is high (R_{AP}) when in anti-parallel state (see Figure 5.2). For STT-MRAM, the low resistance state (LRS) represents logic '0', while the high resistance state (HRS) represents logic '1'. If the write current magnitude (with sufficiently long pulse width) is larger than the *critical switching current* I_c, the magnetization in the FL can switch to the opposite direction. It is a fundamental parameter to characterize the switching capability by current. The STT-induced switching behavior also depends on the current direction, as shown in Figure 5.1a. I_c(AP→P) can be significantly different from I_c(P→AP) due to the bias dependence of STT efficiency and external field disturbance [75]. In addition, the *average switching time* t_w [47] is another critical parameter, which is inversely correlated with the write current. In other words, the higher the write current over I_c, the less the time required for the magnetization in FL to flip. In practice, t_w(AP→P) can also differ from t_w(P→AP) depending on the write current magnitude and duration.

In addition, enough retention time is required for STT-MRAMs depending on the target application. Storage applications require >10 years typically, while cache applications only necessitate ms-scale retention time [84]. an STT-MRAM retention fault occurs when the magnetization in the FL of the MTJ flips spontaneously due to thermal fluctuation. Thus, the STT-MRAM retention time is generally characterized by the *thermal stability factor* (Δ) [75]. The higher the Δ, the longer the retention time.

STT-MRAM performance is vulnerable to magnetic fields, which may arise from the following three sources.

1) Intra-cell stray field $\boldsymbol{H}_{s\_intra}$: To obtain high TMR and strong interfacial perpendicular magnetic anisotropy (iPMA), our MTJ devices were annealed at 375 °C for 30 min in a vacuum chamber under the perpendicular (out-of-plane) magnetic field of 20 kOe. Once the ferromagnetic layers (i.e., FL, RL, and HL) in the MTJ stack are magnetized, each of them inevitably generates a stray field in the space. Figure 5.1a illustrates the intra-cell stray field $\boldsymbol{H}_{s\_intra}$ perceived at the FL, generated by the RL and HL together; its in-plane component $H^{x-y}_{s\_intra}$ is marginal [129], while its out-of-plane component $H^{z}_{s\_intra}$ at the FL has a significant influence on the energy barrier E_B between the P and AP states

[81]. For example, if $H^z_{s\_intra}$ has the same direction as the magnetization in the FL in AP state, it leads to an increase in E_b(AP→P) and a decrease in E_b(P→AP), as illustrated in Figure 5.2. The bifurcation of E_B along the two switching directions has a significant impact on the retention and the STT-switching characteristics of MTJ devices, as reported in [79, 129, 130]. In the extreme case where $H^z_{s\_intra}$ exceeds the FL *coercivity* H_c, defined as the reverse field needed to drive the magnetization of a ferromagnet to zero, the bistable states will disappear [189].

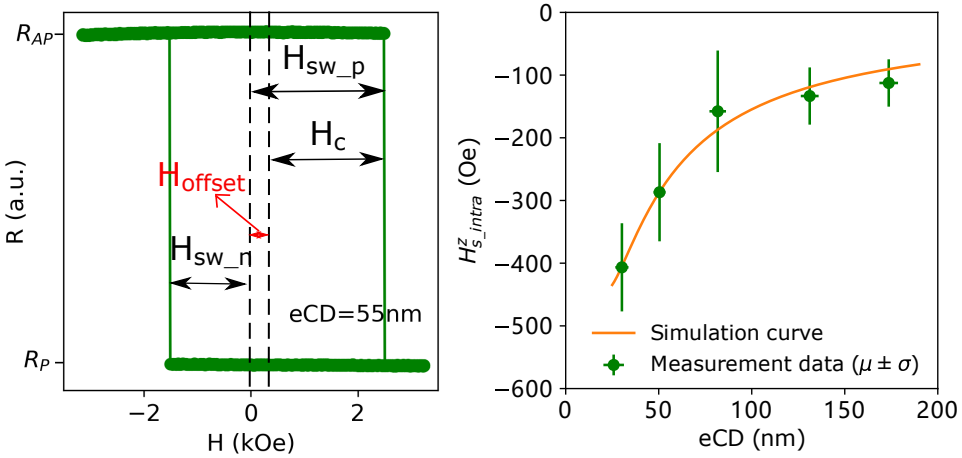
2) Inter-cell stray field H_{s_inter}: As the density of STT-MRAMs increases, the spacing between neighboring MTJ devices becomes narrower (i.e., smaller pitch). This makes stray fields from neighboring cells not negligible any more [81, 197]. Figure 5.1b shows a 3×3 MTJ array, where the eight cells C0-C7 (aggressors) surrounding cell C8 (victim) in the center inevitably generate an inter-cell stray field $\boldsymbol{H}_{s\_inter}$ acting on the victim cell. Figure 5.1c and Figure 5.1d show the scanning electron microscope (SEM) images of our 0T1R wafer floorplan and MTJ array, respectively.

3) External disturbance field H_{ext}: When being deployed in the field, STT-MRAM products may be subject to external magnetic fields unintentionally or maliciously in the operating environment. These unexpected disturbance fields further bifurcate E_B shown in Figure 5.2, thus causing data retention and write errors when $\boldsymbol{H}_{ext}$ reaches a certain extent [191]. Lee *et al.* [86] observed with silicon measurements that the sensitivity of switching voltage V_c to $\boldsymbol{H}_{ext}$ was ~8%/500Oe; with a 300 μm-thick shield at package level, the V_c sensitivity was reduced to ~3%/500Oe. Naik *et al.* [4] demonstrated STT-MRAM with 500 Oe magnetic immunity by boosting write voltage and adding 2-bit ECC.

5.3. CHARACTERIZATION OF INTRA-CELL STRAY FIELDS

In this section, we detail how we measure the out-of-plane component $H^z_{s\_intra}$ of the intra-cell stray field at the FL of isolated MTJ devices (i.e., without any neighboring cells) with various sizes.

$H^z_{s\_intra}$ can be extracted from R-H hysteresis loops. Figure 5.3a shows a measured R-H hysteresis loop for a representative MTJ device with the HL/RL configuration shown in Figure 5.3a. During the measurement, an external field was applied perpendicularly to the device under test. It was ramped up from 0 Oe to 3 kOe, then it went backwards to −3 kOe and finished at 0 Oe. In total, we measured 1000 field points, each of which was followed by a read operation to read out the device resistance with a voltage of 20 mV. It can be seen that the MTJ device switches from AP state (high resistance) to P state (low resistance) when the field reaches at H_{sw_p}, and it switches back to AP state at a negative field H_{sw_n}. The device coercivity can be obtained by $H_c = (H_{sw\_p} - H_{sw\_n})/2$. Due to the existence of stray fields at the FL, the loop is always offset to the positive side for the device configuration in Figure 5.3a. The offset field H_{offset} is equal to $(H_{sw\_p} + H_{sw\_n})/2$, as shown in the figure. Since H_{offset} is essentially equivalent to the extra external field applied to cancel out H_{s_intra}, the relation of these two parameters is $H_{s\_intra} = -H_{offset}$. Given the fact that the *resistance-area product* (*RA*) does not change with the device size, the *electrical Critical Diameter* (eCD) of each device can be derived by [80]: $eCD = \sqrt{\frac{4}{\pi} \cdot \frac{RA}{R_P}}$, where RA=4.5 Ω·μm^2 (measured at blanket stage) for this wafer, and R_P can be extracted from the R-H loop (i.e., the lower horizontal line in Figure 5.3a). The calculated



(a) (b)

Figure 5.3: (a) Measured R-H hysteresis loop, (b) device size dependence of $H^z_{\text{s\_intra}}$: measured vs. simulated.

eCD=55 nm for the device shown in Figure 5.3a.

In this way, we can obtain $H^z_{\text{s\_intra}}$ and eCD for MTJ devices with different sizes on the same wafer. The measurement results are shown in Figure 5.3b. The error bars indicate the device-to-device variation in the the measured values due to process variations and the intrinsic switching stochasticity. It can be seen that the smaller the device size (i.e., smaller eCD), the higher $H^z_{\text{s\_intra}}$; the trend even tends to grow exponentially for eCD<100 nm. The solid curve in the figure represents simulation results which will be explained in the next section.

5.4. MODELING OF INTERNAL STRAY FIELDS

To analyze and quantify the effects of magnetic fields on the MTJ's performance, we need to first develop an accurate model to cover all the three sources of magnetic field disturbance as mentioned in Section 5.2. $\boldsymbol{H}_{\text{ext}}$ originates from the external surroundings thus is independent on any STT-MRAM design; it can be directly fed into a Verilog-A MTJ model as an input parameter. In contrast, $\boldsymbol{H}_{\text{s\_intra}}$ and $\boldsymbol{H}_{\text{s\_inter}}$ both depend on STT-MRAM designs. Therefore, this section focuses on analytical modeling of these two internal stray fields using Python3. To this end, we first model and calibrate $\boldsymbol{H}_{\text{s\_intra}}$ for isolated MTJ devices, based on the measurement data presented in the previous section. Thereafter, we extrapolate this model to derive $\boldsymbol{H}_{\text{s\_inter}}$ for an memory array with various pitches.

5.4.1. INTRA-CELL STRAY FIELD

Under the assumption that each ferromagnetic layer (i.e., FL, RL, and HL) in MTJ devices is uniformly magnetized, the produced field is identical to the field that would be pro-

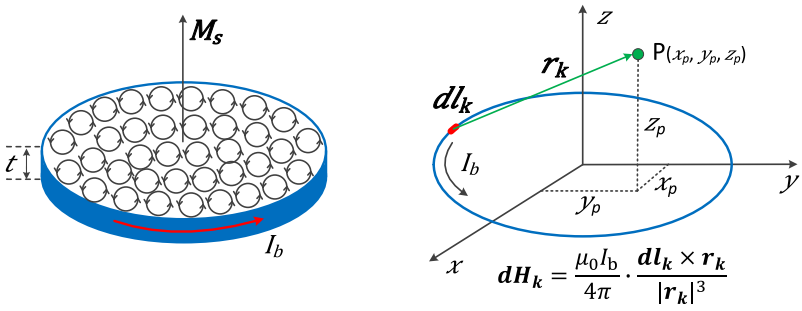


Figure 5.4: (a) Bound current, (b) Biot-Savart law.

duced by the bound current [198]. Figure 5.4a depicts a thin ferromagnet with tiny current loops representing dipoles. All internal currents cancel each other while there is no adjacent loop at the edge to do the canceling. As a result, the net effect is a macroscopic current I_b (referred to as bound current) flowing around the boundary. The magnetic moment of this ferromagnet can be expressed as $\boldsymbol{m} = \boldsymbol{M_s} \cdot A \cdot t$ [198], where $\boldsymbol{M_s}$ is the saturation magnetization, A is the cross-sectional area, and t is the thickness of this ferromagnet. Considering the bound current I_b, $\boldsymbol{m}$ can also be written as $I_\mathrm{b} \cdot A \cdot \hat{\boldsymbol{n}}$ where $\hat{\boldsymbol{n}}$ is the unit vector along the direction of $\boldsymbol{M_s}$ [198]. Therefore, one can easily derive $I_\mathrm{b} = M_s t$. For each ferromagnet in the MTJ stack, the $M_s t$ product is measured at blanket film level by vibrating sample magnetometry (VSM) measurements.

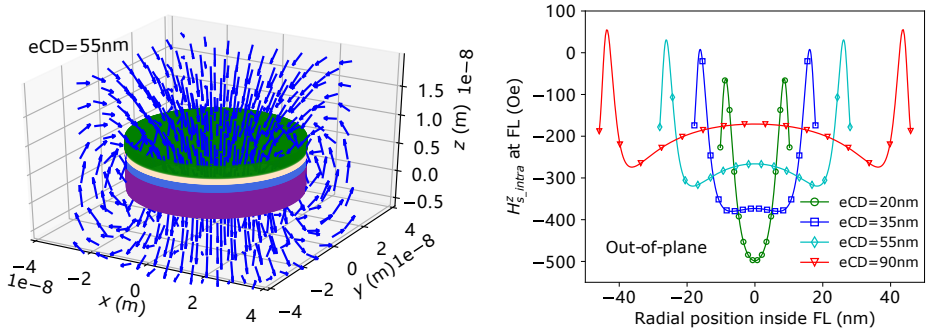
With the derived bound current I_b for each ferromagnet in the MTJ stack, the generated stray field in the space can be modeled as the field of a current loop with current I_b, as shown in Figure 5.4b. In this way, the stray field at any point P(x_p,y_p,z_p) in the space can be calculated by the Biot-Savart law [198]:

$$\boldsymbol{H}(\boldsymbol{r}) = \frac{\mu_0}{4\pi} \oint \frac{I_\mathrm{b} \boldsymbol{dl} \times \boldsymbol{r}}{|\boldsymbol{r}|^3}, \tag{5.1}$$

where $\boldsymbol{dl}$ is an infinitesimal length of the current loop, $\boldsymbol{r}$ the vector distance from $\boldsymbol{dl}$ to the point P, and μ_0 the vacuum permeability. To calculate the above integral in a discrete form, we can divide the current loop into a large number of small segments, thereafter sum up the fields of all segments at point P as an approximation of $\boldsymbol{H}(\boldsymbol{r})$.

Assume the current loop is cut into N segments. For the k^{th} segment $\boldsymbol{dl_k}$ ($k \in [0, N-1]$), we derive:

$$\begin{aligned} \boldsymbol{dl_k} &= (x_{k+1} - x_k, y_{k+1} - y_k, z_{k+1} - z_k), \\ \boldsymbol{r_k} &= (x_p - x_k, y_p - y_k, z_p - z_k). \end{aligned}$$



(a) (b)

Figure 5.5: (a) intra-cell stray field $\boldsymbol{H}_{\text{s\_intra}}$ from the HL and RL for an MTJ with eCD=55 nm, and (b) the out-of-plane component $H^z_{\text{s\_intra}}$ distribution over the cross-section of the FL, with respect to various eCDs.

5

Therefore, $\boldsymbol{dl_k} \times \boldsymbol{r_k} = (S_k^x, S_k^y, S_k^z)$, where

$$S_k^x = (y_{k+1} - y_k) \cdot (z_p - z_k) - (z_{k+1} - z_k) \cdot (y_p - y_k),$$
$$S_k^y = (z_{k+1} - z_k) \cdot (x_p - x_k) + (x_{k+1} - x_k) \cdot (z_p - z_k),$$
$$S_k^z = (x_{k+1} - x_k) \cdot (y_p - y_k) - (y_{k+1} - y_k) \cdot (x_p - x_k).$$

The field generated by the tiny segment $\boldsymbol{dl_k}$ is

$$\boldsymbol{dH_k} = (dH_k^x, dH_k^y, dH_k^z) = \frac{\mu_0}{4\pi} \cdot \frac{I_b}{|\boldsymbol{r_k}|^3} \cdot (S_k^x, S_k^y, S_k^z).$$

By summing up the fields of all N segments, we derive the overall field at the spot P generated by the entire current loop:

$$\boldsymbol{H} = \sum_{k=0}^{N-1} \boldsymbol{dH_k} = (\sum_{k=0}^{N-1} dH_k^x, \sum_{k=0}^{N-1} dH_k^y, \sum_{k=0}^{N-1} dH_k^z)$$

In this way, we can calculate the intra-cell stray field from the HL ($\boldsymbol{H}_{\text{s\_HL}}$) and intra-cell stray field from the RL ($\boldsymbol{H}_{\text{s\_RL}}$), respectively. The overall intra-cell stray field is the vector sum of these two fields (i.e., $\boldsymbol{H}_{\text{s\_intra}} = \boldsymbol{H}_{\text{s\_HL}} + \boldsymbol{H}_{\text{s\_RL}}$), which is visualized in Figure 5.5a for an MTJ device with eCD=55 nm. Figure 5.5b shows the distribution of the z-component $H^z_{\text{s\_intra}}$ (i.e., the out-of-plane component) over the horizontal cross-section of the FL. It can be seen that $H^z_{\text{s\_intra}}$ is not uniformly distributed at the FL; its magnitude is smaller at the edge than at the center. We took the values at the center (i.e., at radial position=0 nm) and calibrated them with the measured data. Figure 5.3b presents the simulation results of $H^z_{\text{s\_intra}}$ vs. eCD, which match the silicon data.

5.4.2. INTER-CELL STRAY FIELD

To model the inter-cell stray field, we extrapolate the intra-cell stray field model from a single MTJ device to a 3×3 MTJ array in Cartesian Coordinates. The nine devices are

named C0 to C8, as illustrated in Figure 5.1b. Cell C8 in the center is considered as the victim whereas the four direct neighbors (C0-C3) and four diagonal neighbors (C4-C7) are aggressor cells. In this way, the inter-cell magnetic coupling effect is translated to the impact of net stray field from the eight neighboring cells (denoted as $\boldsymbol{H}_{\text{s\_inter}}$) on the FL of the victim C8. $\boldsymbol{H}_{\text{s\_inter}}$ can be calculated by:

$$\boldsymbol{H}_{\text{s\_inter}} = \sum_{i=0}^{7} (\boldsymbol{H}_{\text{s\_HL}}(\text{C}i) + \boldsymbol{H}_{\text{s\_RL}}(\text{C}i) + \boldsymbol{H}_{\text{s\_FL}}(\text{C}i)).$$

Since the HL and RL are both fixed layers after the fabrication of MTJ devices, $\boldsymbol{H}_{\text{s\_HL}}$ and $\boldsymbol{H}_{\text{s\_RL}}$ are fixed, given an eCD and a pitch node. However, the direction of $\boldsymbol{H}_{\text{s\_FL}}$ changes dynamically depending on the data stored in the MTJ device though its magnitude remains the same. As a result, $\boldsymbol{H}_{\text{s\_inter}}$ depends on the *neighborhood pattern* in the eight neighboring cells (i.e., C0-C7), which we denote as NP_8. In the binary form, NP_8 can be expressed as: $[d_0, d_1, d_2, d_3, d_4, d_5, d_6, d_7]_2$, where $d_i \in \{0, 1\}$ represents the data stored in Ci. In addition, NP_8 can also be transformed to the decimal form: $[n]_{10}$, where $n \in [0, 255]$.

Figure 5.6 shows the resultant $H^z_{\text{s\_inter}}$ values at the FL of victim C8 as a function of the number of 1s in direct neighbors C0-C3 (marked in yellow) and the number of 1s in diagonal neighbors C4-C7 (marked in skyblue). Since C0-C3 are in symmetric positions and C4-C7 are also in symmetric positions, there are 25 distinct combinations as shown in the figure. For this example, we set eCD=55 nm and pitch=90 nm (design spec. from the SK hynix high-density STT-MRAM design in [167]). It can be seen that $H^z_{\text{s\_inter}}$ reaches its lowest point (−16 Oe) when C0-C7 are all in 0 (P) state (i.e., NP_8=0). In this case, the magnetization in the FL of every aggressor cell is in parallel with that of the RL; together, they generate a stray field which is stronger enough to compensate the stray field from the HL. As the bit number of 1s increases, $H^z_{\text{s\_inter}}$ increases; it increases in a step of 15 Oe

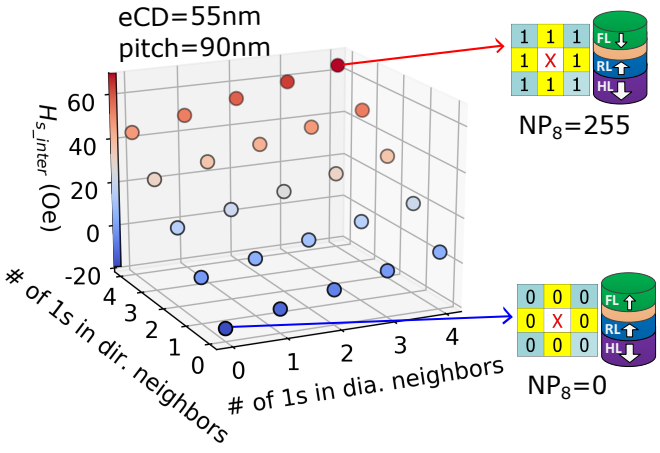


Figure 5.6: $H^z_{\text{s\_inter}}$ at the FL of victim C8 under various combinations of the number of 1s in direct neighbors and diagonal neighbors.

with the number of 1s in direct neighbors and in a step of 5 Oe with the number of 1s in the diagonal neighbors. When C0-C7 are all in 1 (AP) state (i.e., NP_8=255), $H^z_{s\_inter}$ reaches the peak (64 Oe). Therefore, the maximum variation in $H^z_{s\_inter}$ among the 256 neighborhood patterns is 80 Oe in this case. If the value is too large compared to the device coercivity (H_c=2.2 kOe for the measured devices in this paper), it may result in a significant variation in the device performance. To quantitatively evaluate the inter-cell magnetic coupling strength, we defined *inter-cell magnetic coupling factor* Ψ as the ratio of the maximum variation in $H^z_{s\_inter}$ due to different NPs to H_c. Ψ will be used as an indicator of inter-cell magnetic coupling strength in the remaining part of this paper.

The Ψ value varies with device size and array pitch, as shown in Figure 5.7. In our simulations, we set the minimum pitch to 1.5×eCD according to [156] for high-density STT-MRAMs and the maximum pitch to 200 nm, which is adopted by both Samsung and Intel [30, 109]. It can be seen that $\Psi \approx 0\%$ at pitch=200 nm for all three device sizes, indicating the inter-cell magnetic coupling is negligible due to the far distance between devices. As the pitch decreases, Ψ increases gradually until reaching a threshold point after which it goes up exponentially. For our devices, Ψ=2% (marked with the dashed line) can be considered as the threshold point, where the array density is maximized with negligible inter-cell magnetic coupling. For a device with eCD=35 nm, this corresponds to pitch=80 nm, approximately.

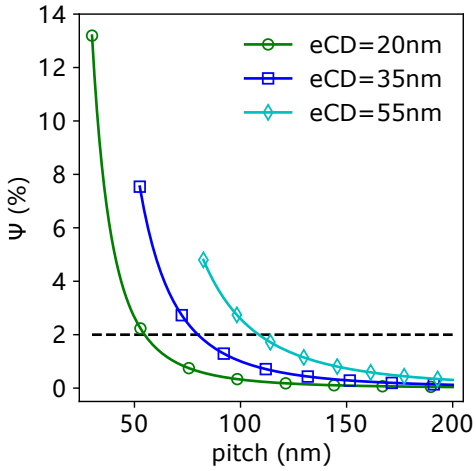


Figure 5.7: Ψ vs. pitch with respect to three MTJ sizes.

5.5. IMPACT OF INTERNAL STRAY FIELDS ON MTJ PERFORMANCE

In this section, we evaluate the impact of internal stray fields on the critical switching current I_c and the average switching time t_w, using the proposed model. Thereafter, we investigate the impact on the thermal stability factor Δ in a similar way. Simulation results for MTJ devices with eCD=35 nm are presented as an example.

5.5.1. IMPACT ON THE CRITICAL SWITCHING CURRENT

Under the influence of stray field, I_c can be expressed as follows [75]:

$$I_c(H^z_{\text{stray}}) = \frac{1}{\eta}\frac{2\alpha e}{\hbar} M_s \cdot V \cdot H_k \cdot (1 \pm \frac{H^z_{\text{stray}}}{H_k}), \tag{5.2}$$

where η is the STT efficiency, α the magnetic damping constant, e the elementary charge, $\hbar$ the reduced Planck constant, M_s the saturation magnetization, V the volume of the FL, H_k the magnetic anisotropy field. The sign in the parentheses is '+' for I_c(P→AP) and '−' for I_c(AP→P), given the definition of coordinates in this paper. In Equation (5.2), $H^z_{\text{stray}} = H^z_{\text{s\_intra}} + H^z_{\text{s\_inter}}$ can be calculated with our proposed magnetic coupling model taking into account both intra-cell and inter-cell stray fields, while H_k needs to be extracted from measurement data. the other parameters in the equation are measured at blanket stage before etch. Since the switching points (i.e., $H_{\text{sw\_p}}$ and $H_{\text{sw\_n}}$ in Figure 5.3a) are intrinsically stochastic, we measured the R-H loop of the same device for 1000 cycles to obtain a statistical result of the switching probability at varying fields. With the technique proposed in [199], we are able to extract H_k and Δ_0 by performing curve fitting. Δ_0 is the intrinsic thermal stability factor without any stray field at the FL; it will be used in the next subsection. By doing this for a large number of devices, we obtained $\Delta_0 = 45.5$ and $H_k = 4646.8$ Oe (both in median) for devices with eCD=35 nm.

Figure 5.8 shows the critical switching current I_c for C8 (for both P→AP switching and AP→P switching) at different pitches with respect to various stray fields. For isolated devices without any stray field (i.e., ideal case, $H^z_{\text{stray}} = 0$), the intrinsic I_c for the two switching directions is supposed to show no difference; $I_c = 57.2\,\mu$A. When taking into account the intra-cell stray field (i.e., $H^z_{\text{stray}} = H^z_{\text{s\_intra}}$), a static shift in I_c is introduced, making I_c(AP→P)=61.7 μA (i.e., 7% above the intrinsic I_c) and I_c(P→AP)=52.8 μA (i.e., 7% below). When considering both intra-cell and inter-cell stray fields (i.e., $H^z_{\text{stray}} = H^z_{\text{s\_intra}} + H^z_{\text{s\_inter}}$) for different neighborhood patterns NP_8, the impact on I_c shows a clear dependence on

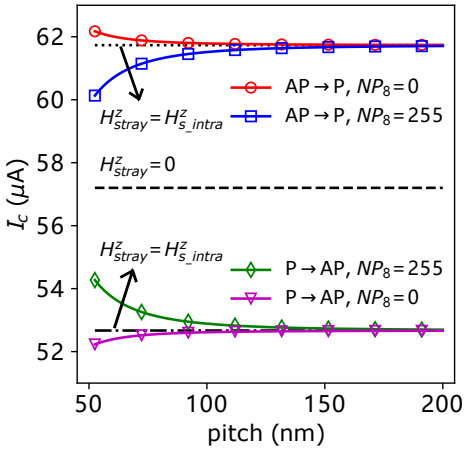


Figure 5.8: I_c vs. pitch under the circumstance of different stray fields.

the array pitch. I_c(AP→P) becomes larger at smaller pitches when NP_8=0, while it shows an opposite trend when NP_8=255. This indicates that the variation in I_c(AP→P) between different neighborhood patterns increases as the pitch goes down. It can be seen that at pitch≈80 nm (corresponding to $\Psi = 2\%$), the variation is marginal. Similar observations can be seen on the P→AP switching direction.

5.5.2. IMPACT ON THE AVERAGE SWITCHING TIME

The average switching time t_w in the presence of H^z_{stray} in the precessional regime (namely, switched by the STT-effect) can be estimated using Sun's model as follows [62]:

$$t_w(H^z_{stray}) = \left(\frac{2}{C+\ln(\frac{\pi^2\Delta}{4})}\cdot\frac{\mu_B P}{em(1+P^2)}\cdot I_m\right)^{-1}, \tag{5.3}$$

$$I_m = \frac{V_p}{R(V_p)} - I_c(H^z_{stray}). \tag{5.4}$$

Here, C≈0.577 is Euler's constant, μ_B the Bohr magneton, P the spin polarization, e the elementary charge, and m the FL magnetic moment. V_p is the voltage applied on the MTJ device to switch its state. $R(V_p)$ is the resistance of the MTJ device as a function of the applied voltage V_p; it shows a non-linear dependence on V_p [62].

Figure 5.9a–5.9c show the voltage dependence of the average switching time from AP state to P state (t_w(AP→P)) for MTJs with eCD=35 nm at pitch=3×eCD, 2×eCD, and

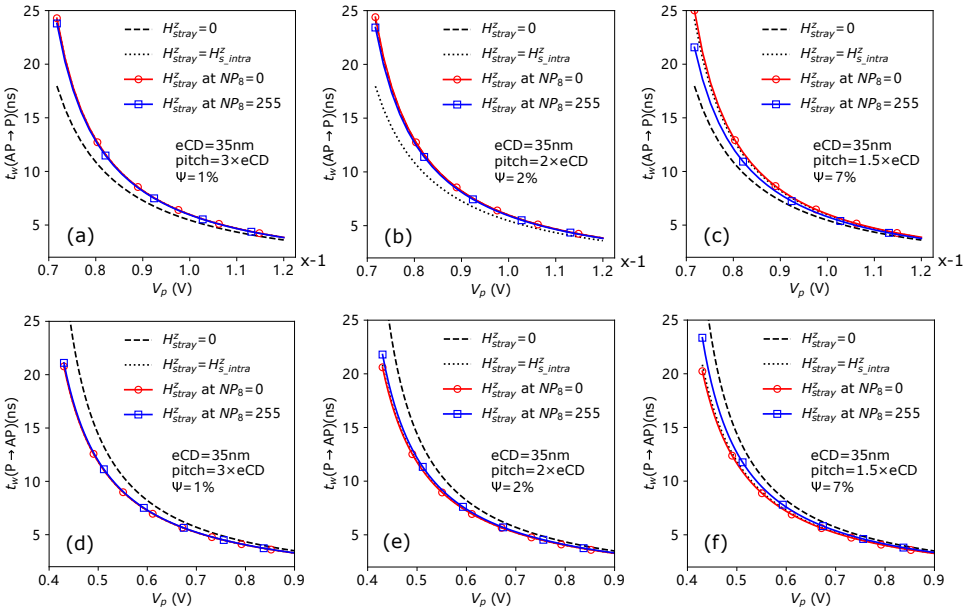


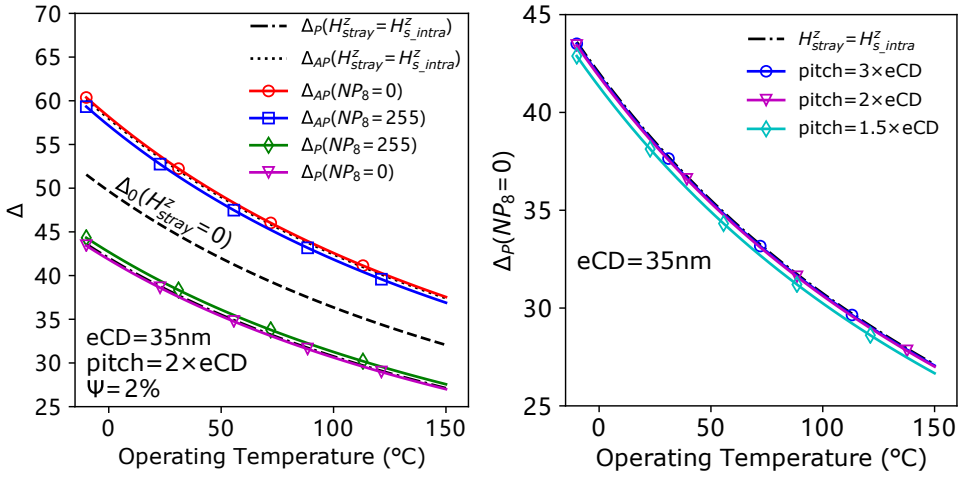
Figure 5.9: Impact of internal stray fields on the voltage dependence of t_w with eCD=35 nm at various pitches: (a) 3×eCD, AP→P switching, (b) 2×eCD, AP→P switching, (c) 1.5×eCD, AP→P switching, (d) 3×eCD, P→AP switching, (e) 2×eCD, P→AP switching, and (f) 1.5×eCD, P→AP switching.

1.5× eCD. Due to the space limitation, the simulation results of t_w(P→AP) are excluded. It can be seen that t_w(AP→P) becomes larger for MTJ devices in the presence of H^z_{stray} (solid lines), comparing to devices without any stray field (dashed lines). It is worth noting that the larger the voltage, the smaller the impact of the stray field on t_w(AP→P). However, an increase in the switching voltage V_p also results in more power consumption and a higher vulnerability to breakdown. In addition, when the pitch goes from 3×eCD (Figure 5.9a) to 2×eCD (Figure 5.9b), the inter-cell magnetic coupling factor Ψ increases from 1% to 2% and the change in t_w(AP→P) is negligible. However, when the pitch goes down to 1.5×eCD (Figure 5.9c), Ψ increases to 7% and the variation in t_w(AP→P) between different NPs (i.e., $H^z_{s\_inter}$) becomes very visible. For example, at a voltage of 0.72 V, t_w(AP→P) under NP_8=0 is ~4 ns slower than NP_8=255, as shown in Figure 5.9c. This indicates that a larger write margin (e.g., a longer pulse) is required to avoid write failure in the worst-case (i.e., NP_8=0). Similarly, Figure 5.9d–5.9f show the simulation results of the other switching direction: P→AP, under the same Python simulation setup. It is clear that H^z_{stray} exerts an inverse influence on t_w(P→AP), in comparison to t_w(AP→P). When pitch=1.5×eCD (see Figure 5.9f), NP_8=0 facilitates P→AP switching to the highest extent, whereas the same data pattern impedes AP→P switching the most (see Figure 5.9c).

5

5.5.3. IMPACT ON THE THERMAL STABILITY FACTOR

The intrinsic thermal stability factor Δ_0 (without any stray field at the FL) of the MTJ device is given by [75]: $\Delta_0 = \frac{H_k M_s V}{2 k_B T}$, where k_B is the Boltzmann constant and T is the absolute temperature. However, in the presence of stray fields, the thermal stability factor in AP state deviates from that in P state, i.e., $\Delta_{AP} \neq \Delta_P$. The Δ value in the presence of



(a) (b)

Figure 5.10: Impact of magnetic coupling on Δ with eCD=35 nm at: (a) pitch= 2×eCD and (b) worst-case Δ for pitch=3×eCD, 2×eCD, and 1.5×eCD.

H^z_{stray} is given by [75]:

$$\Delta(H^z_{stray}) = \Delta_0(1 \pm \frac{H^z_{stray}}{H_k})^2, \tag{5.5}$$

where the sign in the parentheses is '+' for Δ_P and '−' for Δ_{AP} for the devices considered in this paper. H^z_{stray} can be calculated with our proposed magnetic coupling model, while H_k and Δ_0 are extracted from measurement data.

Figure 5.10a shows the thermal stability factor Δ at varying temperature for eCD=35 nm and pitch=2×eCD, corresponding to $\Psi = 2\%$. It can be seen that the intra-cell stray field $H^z_{s\_intra}$ introduces a static shift in Δ_{AP} and Δ_P; Δ_{AP} is ~30% smaller than Δ_P comparing the dash-dotted line to the dotted one. The solid lines represent the thermal stability factors considering both intra-cell and inter-cell magnetic coupling. It can be seen that the MTJ device has the smallest Δ (highest vulnerability to a retention fault) when the victim cell is in P state and all neighboring cells are also in P state (i.e., NP_8=0). Figure 5.10b compares the worst-case Δ, i.e., $\Delta_P(NP_8=0)$, at pitch=3×eCD, 2×eCD, and 1.5×eCD. One can observe that $\Delta_P(NP_8=0)$ shows a marginal degradation when the array pitch goes down to 1.5×eCD, in comparison to pitch=2×eCD.

5

5.6. IMPLEMENTATION OF MTJ MODEL IN VERILOG-A

Robust and fast STT-MRAM designs require an accurate MTJ model for efficient circuit simulations. After verifying the proposed physics-based model of internal stray fields and its impact on MTJ performance in Python, we then integrated this model into our Verilog-A compact MTJ model. In this section, we first overview the block diagram of the MTJ compact model. Thereafter, we delve into each internal functional module and elaborate its functions and modeling principles.

5.6.1. OVERVIEW OF THE COMPACT MTJ MODEL

Figure 5.11 illustrates the block diagram of our compact MTJ model. The model has two terminals and meets Ohm's law: i.e., $V(T1,T2) = I_{MTJ} \cdot R_{MTJ}$. The MTJ resistance R_{MTJ} depends on the magnetic state AP or P, the bias voltage $V(T1,T2)$, and the ambient temperature T; R_{MTJ} can also be switched between R_P and R_{AP}, depending on the current I_{MTJ} and its duration. In essence, the compact MTJ model describes the complex relationships between these three electrical variables. It abstracts MTJ devices from physical level to electrical level via compact behavioral modeling, described in an analog circuit description language: Verilog-A. In other words, the inputs of the MTJ model are physical and technology parameters (e.g., eCD and RA) and the outputs are MTJ's electrical parameters (e.g., R_P and I_c); the mapping relationships from the inputs to the outputs are analytically described by physical equations such as Equation (5.2).

The internal implementation of the MTJ compact model consists of different functional modules, as shown in Figure 5.11. We divide them into three groups. First, the R_P, TMR, and R_{AP} modules are all concerned with the modeling of MTJ resistance. Second, the Δ, I_c, stochastic switching, and state machine modules are related to the modeling of MTJ switching behavior. Third, MTJ devices are never fabricated perfectly in practice. The MTJ resistance and switching behavior are significantly influenced by several factors such as magnetic fields, process variations, and manufacturing defects. These

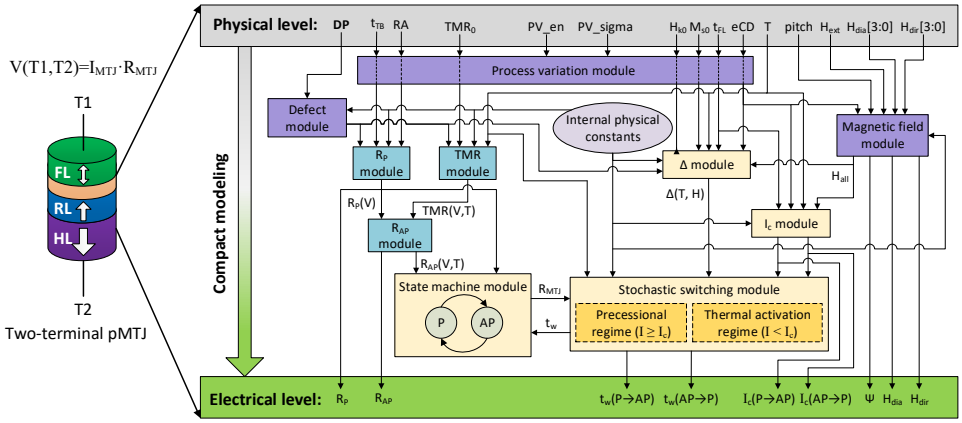


Figure 5.11: Block diagram of the proposed magnetic-field-aware compact MTJ model for simulations of hybrid MTJ/CMOS circuits.

factors have a large impact on MTJ performance, thus requiring special attention. Next, we elaborate these three groups of functional modules in detail.

5.6.2. MODELING OF MTJ RESISTANCE

R_P MODULE

The physical model of MTJ's tunneling magneto-resistance originates from [182], where it indicates that the resistance is mainly determined by the TB thickness and the interfacial effects between TB and adjacent CoFeB layers. The resistance in P state R_P decreases slightly with bias voltage V and it can be approximately considered independent on temperature [91]. We adopted the following two equations to model R_P at varying bias voltage and fitted the modeling results to our measurement data in [62]:

$$R_P(V) = \frac{R_0}{1 + s \cdot |V|}, \tag{5.6}$$

$$R_0 = \frac{t_{ox}}{F \cdot \sqrt{\bar{\varphi}} \cdot A} \exp(\text{coef} \cdot t_{TB} \cdot \sqrt{\bar{\varphi}}). \tag{5.7}$$

t_{TB} is the TB thickness, $\bar{\varphi}$ the potential barrier height of MgO, $A = \frac{1}{4} \cdot \text{eCD}^2$ the horizontal cross-section of the MTJ device. F, coef and s are fitting coefficients depending on the RA product as well as the material composition of the MTJ layers.

TMR MODULE

TMR ratio plays a critical role in determining the difficulty of distinguishing R_P and R_{AP} in read operations. Thus, a high TMR ratio, preferably above 180%, is expected in practice for commercially feasible STT-MRAM products. Experimental results have showed that TMR ratio decreases with both temperature T and bias voltage V [200]. We model

the dependence of V and T on TMR ratio as follows [62, 91].

$$TMR(T) = \frac{TMR_0 + 1}{1 + 2Q \cdot \beta_{AP} \cdot \ln(\frac{k_B T}{E_c})} - 1, \tag{5.8}$$

$$TMR(T,V) = TMR(T) \cdot (1 + \frac{V^2}{V_h^2} + b \cdot V^{\frac{4}{3}})^{-1}. \tag{5.9}$$

In the above equations, TMR_0 is the TMR ratio at T=0 K and V=0 V. Q describes the probability of a magnon involved in the tunneling process. β_{AP}=$Sk_B T/E_m$, where S is the spin parameter, k_B is the Boltzmann constant, and E_m is related to the Curie temperature T_c of the ferromagnetic materials: $E_m = 3k_B T_c/S + 1$. E_c is the magnon cutoff energy. V_h and b are both fitting parameters.

R_{AP} MODULE

Based on Equations (5.6–5.9), R_{AP} at certain T and V can be derived accordingly:

$$R_{AP}(T,V) = R_P(V) \cdot (1 + TMR(T,V)). \tag{5.10}$$

5.6.3. MODELING OF MTJ SWITCHING BEHAVIOR

Δ MODULE

The thermal stability factor Δ is a figure of merit for MTJs. Δ directly determines the retention time of data stored in an MTJ and it also has an impact on the switching behavior between AP and P states. Under the macrospin assumption (i.e., the magnetization in the FL switches uniformly as a whole), Δ can be expressed as [75, 91]:

$$\Delta = \frac{E_B}{k_B T} = \frac{\mu_0 \cdot t_{FL} \cdot M_s(T) \cdot A \cdot H_k(T)}{2k_B T}, \tag{5.11}$$

$$M_s(T) = M_{s0} \cdot (1 - \frac{T}{T^*})^{\frac{3}{2}}, \tag{5.12}$$

$$H_k(T) = f_1 \cdot T + f_2. \tag{5.13}$$

In Equation (5.11), μ_0 is the vacuum permeability and the other physical parameters have been introduced previously. Note that M_s and H_k are both dependent on T, as suggested by the experimental and modeling results in [201]. From the same paper, we extracted Equations (5.12–5.13) for modeling the temperature dependence of M_s and H_k in our compact MTJ model. M_{s0} is the saturation magnetization of the FL at 0 K; T^*, f_1, and f_2 are all fitting parameters.

I_c MODULE

The magnetization dynamics in the STT switching process is typically described by the Landau-Lifshitz-Gilbert (LLG) equation with the addition of STT-related terms, under the assumption of macrospin approximation [62, 75]. Solving the LLG equation results in Equation (5.2) for the critical switching current I_c.

STOCHASTIC SWITCHING MODULE

The switching behavior between AP and P states is a complex process, which is intrinsic stochastic and dependent on the applied pulse width t_p. Depending on the mechanism which dominates the switching behavior, the entire switching spectrum can be divided into two regimes: 1) *precessional regime* and 2) *thermal activation regime*. In the precessional regime where t_p <~40 ns, the STT effect is the main driving force which flips the magnetization of FL. In this regime, the average switching time t_w can be estimated using Sun's model, namely Equations (5.3–5.4). The actual switching time varies from pulse to pulse (i.e., switching stochasticity). The root cause can be attributed to the variation of incubation time after the pulse onset, due to thermal fluctuation [202]. We model the switching stochasticity by assigning a normal distribution to t_w, which has a fair agreement with measurement data [63, 91]. In the thermal activation regime where the pulse width increases above 40 ns, observed in our devices, a small current less than I_c is able to flip the magnetization due to the increased thermal fluctuation. The thermal fluctuation plays a main role in determining the switching behavior. In this regime, the Neel-Brown model can be used to describe the average switching time t_w [62]:

$$t_w = \tau_0 \exp(\Delta(1 - \frac{I_{MTJ}}{I_c})), \tag{5.14}$$

where τ_0 is the attempt period (~1 ns). The actual switching time in this regime is modeled as an exponential distribution with its mean value at the calculated t_w in Equation (5.14) [196, 203]. As a result, the switching probability $Pr(t_p)$ under a long pulse t_p with a small current I_{MTJ} is [61]:

$$Pr(t_p) = 1 - \exp(-\frac{t_p}{t_w}). \tag{5.15}$$

Equations (5.14–5.15) are commonly used to estimate the read disturb rate, as the read current shares the same path and direction with the write current in w0 operations [47].

STATE MACHINE MODULE

The state machine controls the transition between P and AP states at run time. It outputs the MTJ resistance $R_{MTJ} \in \{R_P, R_{AP}\}$ to the stochastic switching module for calculating I_{MTJ} under the voltage bias V(T1,T2) applied across the MTJ device. Meanwhile, the stochastic switching module sends t_w to the state machine to activate a transition between P and AP states when meeting all switching conditions.

5.6.4. MODELING OF OTHER KEY CHARACTERISTICS

MAGNETIC FIELD MODULE

Analog circuit simulators such as Cadence Spectre and HSPICE are intended for simulations of electrical circuits. With the emergence and fast development of spintronics, there is a need of simulating hybrid MTJ/CMOS circuits such as STT-MRAM, magnetic flip-flop, and magnetic full adder. Unlike MOSFETs where only electrical properties matter, MTJ devices own both electrical and magnetic properties. These two types of property are typically interacted exploiting the spin and charge properties of electron, and

they are very sensitive to magnetic fields, as mentioned in the previous sections. Therefore, it is paramount to consider and evaluate the effects of magnetic fields when simulating and designing MTJ-based circuits. As a solution, we implemented our magnetic field module presented in the previous sections using Verilog-A and then integrated it into our compact MTJ model. The magnetic field module takes into three sources of magnetic fields:

$$H_{\text{all}} = H^z_{\text{s\_intra}} + H^z_{\text{s\_inter}} + H^z_{\text{ext}}. \tag{5.16}$$

In the above equation, $H^z_{\text{s\_intra}}$, $H^z_{\text{s\_inter}}$, and H^z_{ext} are the out-of-plane components of intra-cell stray field, inter-cell stray field, and external stray field, respectively. $H^z_{\text{s\_intra}}$ is calculated internally in the compact MTJ model, depending on eCD and pitch. $H^z_{\text{s\_inter}}$ consists of $H_{\text{dia}}[3{:}0]$ and $H_{\text{dia}}[3{:}0]$, standing for the inter-cell stray fields from four direct and four diagonal neighbors. Note that $H_{\text{dia}}[3{:}0]$, $H_{\text{dia}}[3{:}0]$, and H^z_{ext} are all defined as electrical input ports, which connect to other MTJ devices or circuit elements. Together, these three magnetic fields result in a net overall field H_{all} acting on a specific MTJ device in an STT-MRAM array. H_{all} is then fed into the Δ, I_c, and stochastic switching modules, as described by Equations (5.2–5.5). The magnetic field module also outputs Ψ, H_{dir}, and H_{dia} at run time (depending eCD, pitch, and MTJ state) via three electrical ports of the compact MTJ model in the form of voltage.

DEFECT MODULE

MTJ devices are typically fabricated and integrated between two adjacent metal layers (e.g., M4 and M5) in the BEOL of CMOS process; this process is unique to STT-MRAM and is susceptible to manufacturing defects [47]. We have successfully designed and integrated models for MTJ-internal defects such as pinhole [46], synthetic anti-ferromagnetic layer flip [63], and intermediate state [64] into our compact MTJ model. The effects of these defects are first incorporated into the physical parameters of MTJ and thereafter intothe electrical parameters; these defect models were also corroborated and calibrated by silicon data of defective MTJ devices fabricated at imec.

PROCESS VARIATION MODULE

Process variation (PV) is inevitable when fabricating integrated circuits. The impact of PV on the performance and reliability of integrated circuits becomes increasingly pronounced as the CMOS technology node scales down. To design robust STT-MRAM circuits, PV related to MTJ devices should also be taken into account. The PV module is implemented by assigning normal distributions to key MTJ dimension parameters such as eCD, t_{FL}, and t_{TB}, as well as key physical parameters such as RA and TMR_0. "PV_en" and "PV_sigma" are two input parameters of the compact MTJ model, controlling the internal PV module.

5.7. MTJ ELECTRICAL CHARACTERISTICS UNDER VARIOUS MAGNETIC CONFIGURATIONS

After obtaining the magnetic-field-aware compact MTJ model, we verified it with Cadence Spectre, a commercial analog circuit simulator. In this section, we first present DC simulation results of the model. Thereafter, we present transient simulation results in the form of write error rate (WER) statistics.

5.7.1. DC SIMULATIONS: R-V LOOPS

Measuring R-V loops is a common practice to characterize the voltage dependence of MTJ resistance at P and AP states. We have calibrated the DC simulation results of R-V loops of our compact model with silicon data, as can be found in [62]. Figure 5.12a shows the DC simulation results of R-V loops for MTJs with eCD=35 nm and 55 nm. For each size, we simulated three configurations of HL by modifying its saturation magnetization M_{HL}; the change of M_{HL} resulted in different stray fields at the FL. $M_{\mathrm{HL}}^{\mathrm{bl}}$ means the baseline M_{HL} from experimental results. It can be seen in Figure 5.12a that both R_{P} and R_{AP} increase significantly as eCD decreases. A change in M_{HL} has no impact on MTJ resistance, but it affects the switching voltage V_{c}. Reducing M_{HL} by 40% of $M_{\mathrm{HL}}^{\mathrm{bl}}$ leads to an increase in V_{c} for P→AP switching and a decrease in V_{c} during AP→P switching. In contrast, increasing M_{HL} by 40% of $M_{\mathrm{HL}}^{\mathrm{bl}}$ results in an opposite effect on V_{c}, as shown in the figure. Similarly, we also simulated three values of saturation magnetization of RL (M_{RL}) and three values of the external magnetic field at FL (H_{ext}). The simulation results are shown in Figure 5.12b and Figure 5.12c, respectively.

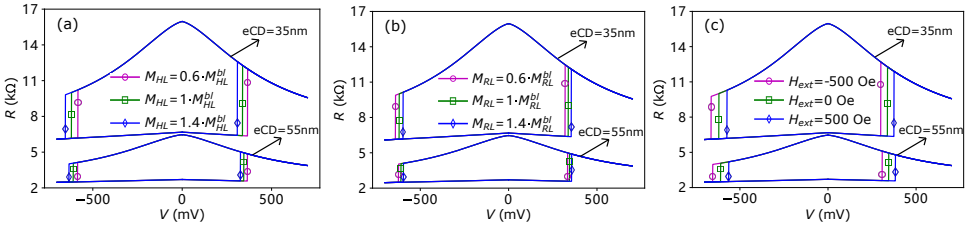


Figure 5.12: DC simulation results of R-V loops for MTJs with eCD=35 nm and 55 nm, with respect to different configurations of magnetic fields at the FL.

5.7.2. TRANSIENT SIMULATIONS: WER STATISTICS

The MTJ switching behavior is intrinsically stochastic and is significantly dependent on the applied pulse width t_{p} and amplitude V_{p}. This characteristic directly affects STT-MRAM circuit designs such as cell selector and write driver. Therefore, it is very important to experimentally characterize WER vs. V_{p} at varying t_{p}, meanwhile providing a capability of simulating this characteristic to facilitate and verify circuit designs. Figure 5.13a–5.13c present the simulation results of WER vs. V_{p} at t_{p}=10 ns with respect to different magnetic configurations, using our compact MTJ model with eCD=35 nm. It is clear that increasing the V_{p} magnitude is very effective in reducing WER for both switching directions; note that here a negative V_{p} results in AP→P switching whereas a positive V_{p} results in P→AP switching. For the original MTJ design where M_{HL}=$M_{\mathrm{HL}}^{\mathrm{bl}}$, M_{RL}=$M_{\mathrm{RL}}^{\mathrm{bl}}$, and H_{ext}=0 Oe, the WER curve is asymmetric; $|V_{\mathrm{p}}(\mathrm{AP}\rightarrow\mathrm{P})|$ is much larger than $|V_{\mathrm{p}}(\mathrm{P}\rightarrow\mathrm{AP})|$ for a given WER value. By reducing $M_{\mathrm{HL}}^{\mathrm{bl}}$ by 40%, the WER curve in Figure 5.13a shifts to the right side, indicating an approximately one order of magnitude decrease in WER at a fixed V_{p} for AP→P switching and an opposite effect on P→AP switching. Figure 5.13b–5.13c depict how the WER curve is affected when modifying M_{RL} and H_{ext}, respectively. Figure 5.13d–5.13f present similar simulation results, but at t_{p}=40 ns. It is worth noting that the slope of WER curve when t_{p}=40 ns is much larger than

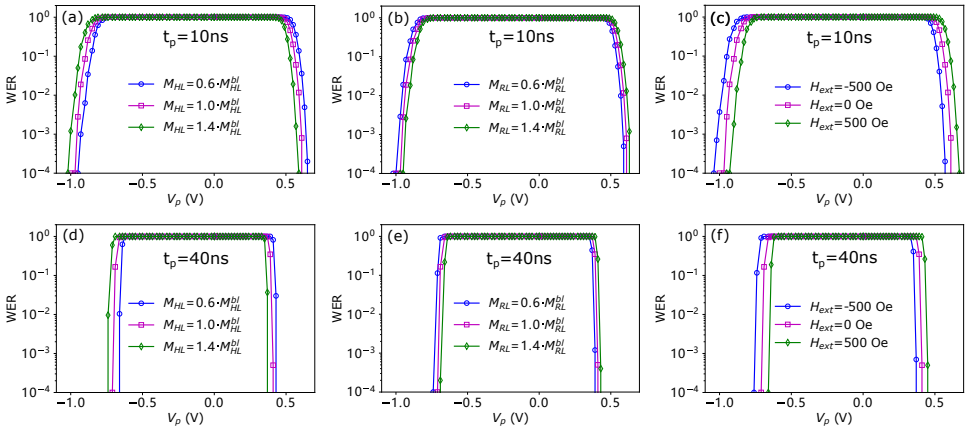


Figure 5.13: Transient simulation results of write error rate (WER) vs. bias voltage t_p for MTJs with eCD=35 nm at pulse width t_p=10 ns and 40 ns.

that when t_p=10 ns. This is because the switching variation is smaller at longer pulses, which is consistent with the measurement data of our devices [63] and others' devices [148, 204].

In summary, the above DC and transient simulation results suggest that our magnetic-field-aware compact MTJ model is qualified for emulating MTJ devices for SPICE-based circuit simulation. By manipulating stray fields at the FL, which is achieved by adjusting the SAF design of the MTJ device, we can adjust the WER curve to the position that we desire when designing STT-MRAM cell and peripheral circuits. Hence, our compact MTJ model enables device/circuit co-design for STT-MRAM.

5.8. ROBUSTNESS ANALYSIS OF STT-MRAM DESIGNS

Apart from a single MTJ device, we also simulated a 3×3 STT-MRAM array with peripheral circuits such as write driver and sense amplifier, as shown in Figure 5.14a. Each STT-MRAM cell consists of an MTJ device and an NMOS as a selector; Figure 5.14b shows the memory cell and three basic operations [47]. The details of simulation circuits can be found in [46]; all transistors in the netlist were built with the 45 nm predictive technology model (PTM) [98]. In this section, we first present transient simulation results of the STT-MRAM full circuit under different eCDs and pitches. Thereafter, we explore the design space under PVT variations and different magnetic configurations.

5.8.1. TRANSIENT SIMULATIONS UNDER DIFFERENT ECDS AND PITCHES

To demonstrate the capability of our compact MTJ model for electrial/magnetic co-simulation under SPICE-based circuit simulation environment, we simulated the STT-MRAM full circuit in Figure 5.14 under two eCDs (35 ns and 55 ns) and two pitches (3×eCD and 1.5×eCD). During the simulations, we set the data background in C0-C7 at 255 (i.e., NP_8=255) and applied the operation sequence: 0w1r1w0r0 to the central cell C8 as a

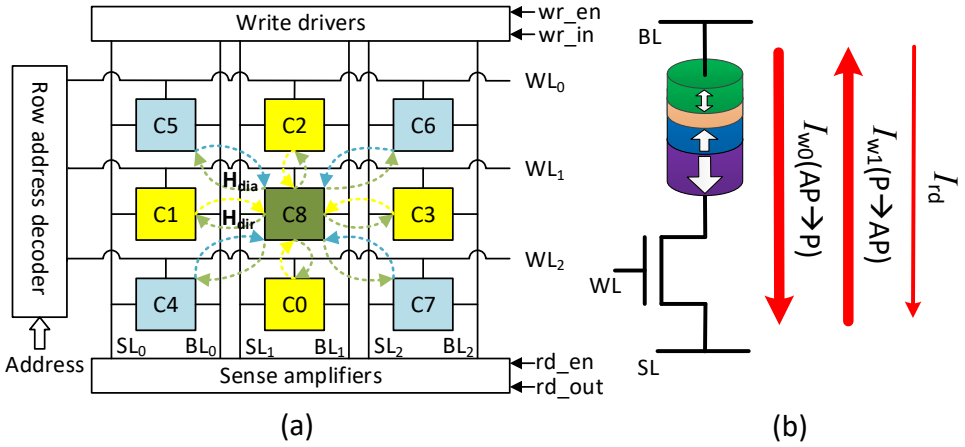


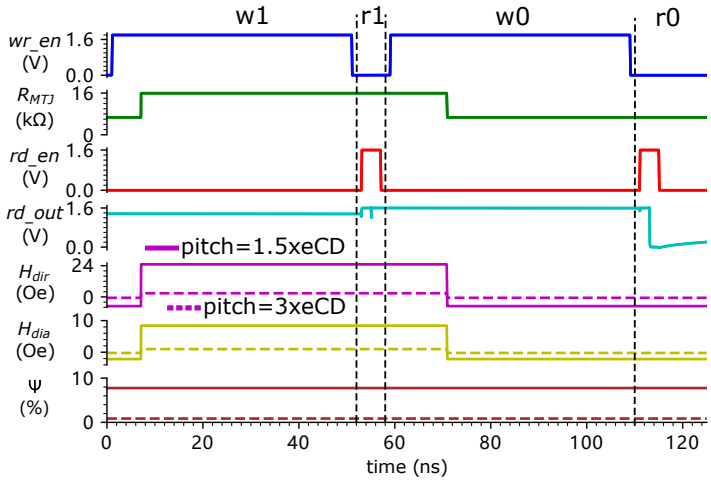
Figure 5.14: (a) 3×3 STT-MRAM array with peripheral circuits, and (b) 1T-1MTJ memory cell and the associated cell operations.

5

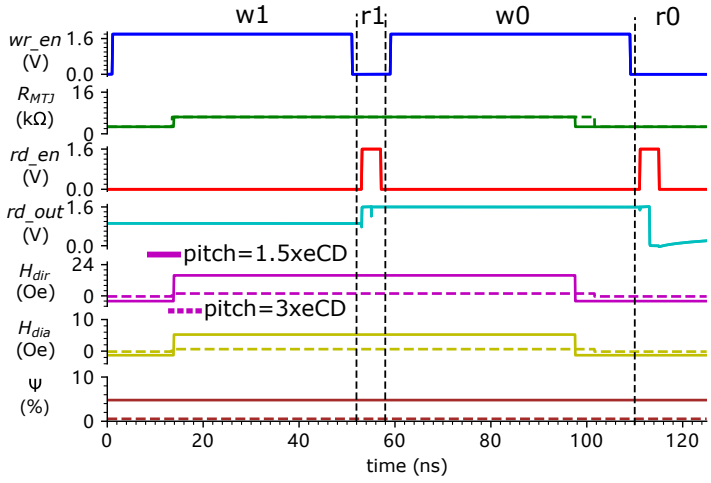
case-study. Figure 5.15a shows the simulation waveforms of seven key signals related to C8 when eCD=35 nm. It can be seen that C8 is initialized to 0; it outputs H_{dir}=7.06 Oe to its direct neighbors C0-C3 and H_{dia}=−2.17 Oe to its diagonal neighbors C4-C7 at pitch= 1.5×eCD. Ψ is 7.8% at this pitch value. In contrast, H_{dir}, H_{dia}, and Ψ are all close to 0 at pitch=3×eCD. During the w1 operation, the state of C8 transitions to 1 (see R_{MTJ}); H_{dir} and H_{dia} are changed to 24.27 Oe and 8.16 Oe, respectively, when at pitch=1.5×eCD. Following the w1 operation, a r1 is applied, which outputs 1 on the signal rd_out. Similar observations can be seen for the following w0 and r0 operations. Figure 5.15b shows the simulation waveforms when eCD=55 nm. The following three differences from Figure5.15a are worth noting: 1) the switching time in both w1 and w0 operations become longer, as larger MTJ devices require larger switching current; 2) H_{dir}, H_{dia}, and Ψ are different due to the change of eCD; 3) when the pitch changes from 3×eCD to 1.5×eCD, the switching time during the w1 operation larger while it becomes smaller in the w0 operation, due to the inter-cell magnetic coupling effect.

5.8.2. DESIGN SPACE WITH VARIOUS VARIATION SOURCES

It is well known that STT-MRAM designs are significantly influenced by the following sources of variations: 1) process variation (device-to-device variation), 2) supply voltage variation, 3) operating temperature variation, 4) MTJ switching stochasticity (cycle-to-cycle variation), and 5) magnetic field variation. We explored the design space considering the aforementioned five variation sources in our circuit simulations. The process variation was modeled by assigning normal distributions to key parameters of both transistors and MTJs. For transistors, it was lumped into the variation in the threshold voltage V_{th} with 10% away from its nominal value at 3σ corners. For MTJs, we assigned the same normal distribution to key input parameters shown in Figure 5.11. In terms of supply voltage V_{DD} variation, we assigned a uniform distribution to V_{DD} with its minimum at 1.5 V and maximum at 1.7 V. The typical industrial standard of operating temperature



(a) eCD=35 nm.



(b) eCD=55 nm.

Figure 5.15: Waveforms of key signals during the transient simulation of operation sequence: 0w1r1w0r0, under four different combinations of eCD and pitch.

$T \in$ [-40, 125]°C [4]. The MTJ switching stochasticity was implemented in our compact MTJ model and it can be enabled or disabled as required. The magnetic field variation includes $\boldsymbol{H}_{\text{s\_intra}}$, $\boldsymbol{H}_{\text{s\_inter}}$, and $\boldsymbol{H}_{\text{ext}}$ as mentioned in the previous sections.

We performed 10k-cycle Monte Carlo simulations of 0w1 and 1w0 operations while sweeping two variables: pulse width t_{p} and voltage on the WL V_{WL}. This is based on the fact that boosting V_{WL} is required to deliver sufficient switching current going through MTJ devices due to the source degeneration issue [47]; this has been a common practice in industry. Figure 5.16a shows a contour plot of WER of 1w0 operation with respect to t_{p}

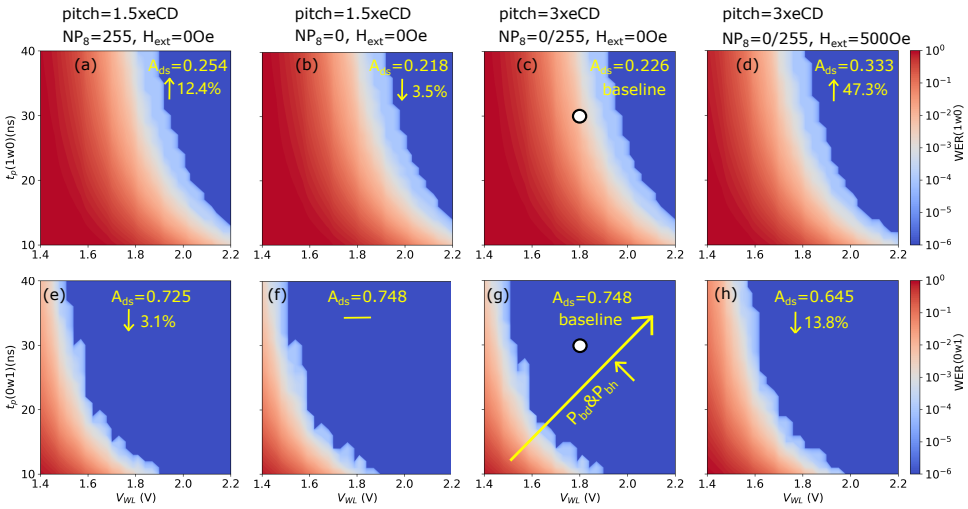


Figure 5.16: WER at different combinations of t_p and V_{WL} for 1w0 and 0w1 operations under different simulation set-ups about pitch, NP_8, and H_{ext}.

and V_{WL}, when eCD=35 nm, pitch=1.5×eCD, NP_8=255, and H_{ext}=0 Oe at room temperature T=27 °C. It can be seen that WER(1w0) gradually decreases from the lower-left corner to the upper-right corner. When the 1w0 operations were all successful among the 10k Monte Carlo simulations, we marked the WER value at 10^{-6} (i.e., the deep blue area). We define the *area of design space* A_{ds} as the normalized area where WER=10^{-6} with respect to the entire area of the contour plot. In Figure 5.16a, A_{ds}=0.254. This is 12.4% larger than the baseline A_{ds} value in Figure 5.16c where pitch=3×eCD and NP_8 has no influence. Figure 5.16b shows the simulation results when pitch=1.5×eCD, NP_8=0, and H_{ext}=0 Oe; A_{ds} decreases by 3.5% in comparison to the baseline setup, due to the inter-cell magnetic coupling effect. In addition, we also studied the impact of H_{ext} on A_{ds}; the result is shown in Figure 5.16d. When the STT-MRAM design is subject to an external magnetic field of 500 Oe, A_{ds}(1w0) increases by 47.3%.

Similarly, the simulation results for 0w1 operations are shown in Figure 5.16e–5.16h. It is clear that A_{ds}(0w1) is much larger than A_{ds}(1w0) under the same simulation conditions, which suggests a critical design challenge facing STT-MRAM: write asymmetry. For example, when fixing t_p=30 ns and V_{WL}=1.8 V, the resultant WER(0w1) has already reached the center of the deep blue area in Figure 5.16g (see the white circle). In contrast, WER(0w1) has not entered into the deep blue area (see the white circle in Figure 5.16c). Worse still, the deeper the white circle enters into the deep blue area, the probability of breakdown (P_{bd}) or back-hopping (P_{bh}) become larger, as illustrated with the yellow arrow in Figure 5.16g. Moreover, the effects of NP_8 and H_{ext} are always opposite for 0w1 operations, compared to 1w0 operations. This implies that the write asymmetry can be adjusted by manipulating magnetic fields. For example, applying H_{ext}=500 Oe increases A_{ds}(1w0) by 47.3% (see Figure 5.16d), whereas it reduces A_{ds}(0w1) by 13.8% (see Figure 5.16h).

Figure 5.17a shows the dependence of A_{ds} on H_{ext}. It can observed that A_{ds}(0w1) decreases by ~12%/500 Oe while A_{ds}(1w0) increases by ~32%/500 Oe. When H_{ext}=~1 kOe, a symmetric design space for 0w1 and 1w0 operations is achieved. On one hand, this suggests that we can design the SAF layer to generate the desired stray field at the FL (same effect as H_{ext}), meeting the requirements of circuit-level designs. On the other hand, we need to pay attention to external magnetic disturbance, requiring package-level magnetic shield or other measures to enhance magnetic immunity [32].

Figure 5.17b shows the dependence of A_{ds} on the operating temperature T. It can be observed that A_{ds} for both 1w0 and 1w0 significantly increases with T. The sensitivity of A_{ds}(0w1) to T is ~13% while the number for A_{ds}(1w0) is ~5%. Although high temperature is in favor of STT-MRAM write operations, it also bring side effects: 1) retention time reduction, 2) degraded read reliability due to TMR drop, and 3) increased vulnerability to breakdown and back-hopping.

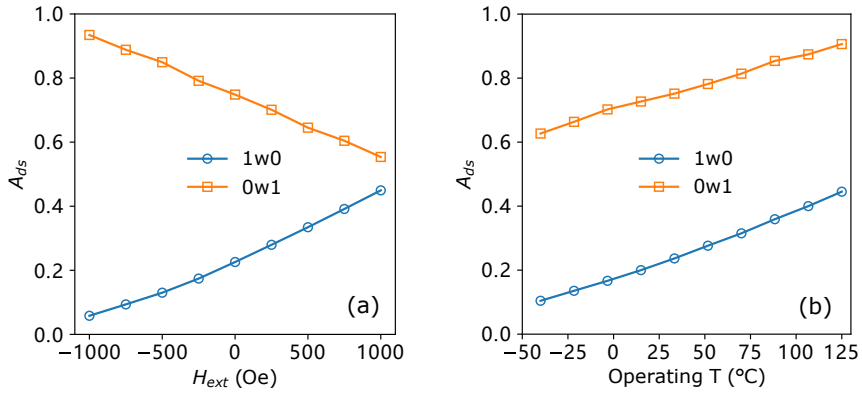


Figure 5.17: STT-MRAM write design space A_{ds} vs. (a) external magnetic field H_{ext} and (b) operating temperature T.

6

DEVICE-AWARE TEST APPROACH

The traditional memory test approach assumes that any physical defect in a semiconductor device can be modeled as a linear resistor. However, it has been shown that this approach is inaccurate at least for emerging memory technologies such as RRAM and STT-MRAM, resulting in unrealistic fault models and thus test escapes. To address this issue, we propose a new test approach: device-aware test (DAT). It goes beyond cell-aware test and sets up a step towards high-quality ICs at defective-part-per-billion level. DAT consists of three steps: DA defect modeling, DA fault modeling, and DA test development. The defect modeling does not assume that a defect in a device can be modeled electrically as a linear resistor, but it rather incorporates the impact of the physical defect on the technology parameters of the device and thereafter on its electrical parameters. Once the defective electrical model is obtained, a systematic fault analysis based on SPICE circuit simulations is performed to derive accurate fault models within a pre-defined complete fault space. Finally, the derived fault models corresponding to this specific defect are used to develop test solutions. By applying DAT to all possible defects and merging the results, we are able to optimize and customize tests with minimal cost, meeting the requirements of specific applications. In this chapter, we start with elaborating the motivation for DAT. Thereafter, we introduce DAT and its three steps in detail. Finally, we discuss DAT implications.

The contents of this chapter have been published in ITC'18 [47], ITC'19 [67], and a patent is pending [205].

6.1. MOTIVATION AND PRIOR WORK

In the conventional defect modeling approach, all defects irrespective their physical natures are modeled as linear resistors, as can be found in prior works in [49, 50, 52, 54–56]. The current cell-aware test (CAT) methodology targets cell-internal defects, which are still modeled as linear resistors (opens and shorts) at the terminals and interconnects of devices in each memory cell; the defect strength is represented by the resistance value [42, 206]. Thus, CAT makes no difference to the conventional defect modeling approach. Although this approach can be convincing for modeling opens and shorts in interconnects, it has never been validated for device-internal defects. In addition, it is well recognized that new failure mechanisms in the nano-era are causing the fault mode of chips to shift from hard and permanent faults to transient, intermittent, and weak faults; these faults may not necessarily be modeled by linear resistors [45]. Recently, it has been shown that the resistor-based defect modeling approach leads to wrong fault models for resistive random access memory (RRAM) devices [207]. Hence, it is incapable of developing high-quality test solutions for RRAMs.

To evaluate the effectiveness of modeling MTJ-internal defects as linear resistors for STT-MRAMs, we performed circuit simulations and compared the simulation results with measurement data of defective MTJs. Conventionally, each defect in an MTJ device is assumed to manifest itself as either a resistor R_{sd} in series with or a resistor R_{pd} parallel to the MTJ device, as illustrated in Figure 6.1. To investigate the effect of this conventional resistor-based defect approach on the R-V hysteresis loop, we simulated an MTJ device using the Verilog-A compact model in [62] for three cases: (1) defect-free case, (2) MTJ defect manifests itself as a series resistor R_{sd}=1 kΩ, and (3) MTJ defect manifests itself as a parallel resistor R_{pd}=10 kΩ. Figure 6.2a compares the three cases, represented by green solid curve, blue dashed curve, and red dash-dot curve respectively. The figure shows that the R-V hysteresis loop enlarges for Case (2); the switching voltage V_c increases because there is a voltage division between the series resistor and the MTJ device. For Case (3), the R-V hysteresis loop moves downwards, as the overall resistance is pulled down. In this case, the switching voltage V_c across the device does not change, as the voltage over the MTJ device is not affected by the parallel resistor.

Figure 6.2b presents the measured R-V hysteresis loops of four MTJ devices on the

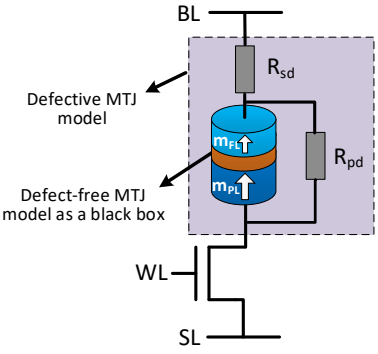


Figure 6.1: Resistive models for MTJ-internal defects.

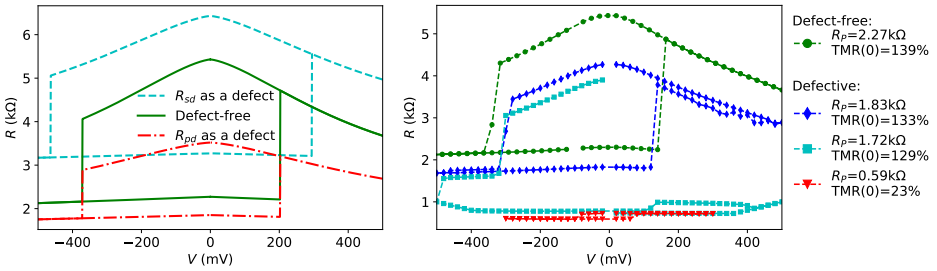


Figure 6.2: Defective MTJ: (a) simulated R-V hysteresis loops with resistive models vs. (b) measured R-V hysteresis loops of fabricated MTJs.

same wafer; the designed diameter is 60 nm, with a nominal RA=4.5 $\Omega{\cdot}\mu m^2$. The green curve (with the widest loop) represents a defect-free device, while the other three curves show defective devices with decreasing TMR and R_P. Clearly the switching voltage of defective devices decreases depending on the defect size, compared to that of a good device. This trend is not captured by the injection of resistive defects, as Figure 6.2a reveals. This is because the resistor-based model fails to accurately incorporate the relationship between the six key electrical parameters of an MTJ device: R_P, R_{AP}, I_c(P→AP), I_c(AP→P), t_w(P→AP), and t_w(AP→P), as explained in Chapter 3. Although the parallel resistor is qualified to model the decreasing trend of R_P and R_{AP}, the impact of defects on the other four parameters is not captured. In order to capture the change of magnetic properties which are critical to the switching behavior, we need a new method to accurately model MTJ-internal defects.

In conclusion, linear resistors are unable to capture defect-induced changes in magnetic properties, which are as important as electrical ones for MTJ devices.

6.2. DEVICE-AWARE TEST FLOW

To overcome the limitations in the conventional test approach, we propose a new test approach, which we name as *Device-Aware Test* (DAT). The DAT flow is shown in Figure 6.3, which fundamentally consists of three steps as follows.

- **Device-aware defect modeling**. First, a defect needs to be physically analyzed and characterized to understand its forming mechanism, location, occurrence rate, and the key technology parameters that are impacted. Thereafter, the effects of the defect are quantitatively incorporated into these technology parameters. Sec-

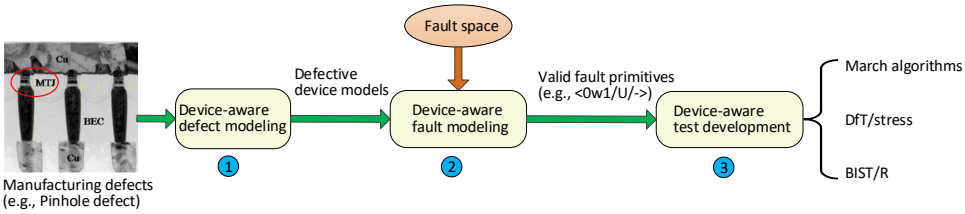


Figure 6.3: Device-aware test flow.

ond, the defect-induced changes in the technology parameters are mapped into the device's electrical parameters. This allows us to convert the defect-free device model into a parameterized defective model. Third, the obtained model can be further calibrated by fitting to silicon data if available.

- **Device-aware fault modeling**. First, a complete fault space which describes all possible faults in emerging memories is defined. This is achieved by extending the conventional *fault primitive* (FP) notation: $\langle S/F/R \rangle$ [186]. Based on the extended FP definition, all memory faults are classified into two categories: *easy-to-detect* (EtD) faults and *hard-to-detect* (HtD) faults [67]. EtD faults are those which can be detected by applying normal write and read operations, i.e., March tests, while HtD faults refer to those which cannot be guaranteed by March tests in their detection. Second, a systematic fault analysis based on circuit simulations for each targeted defect is conducted; this is to derive realistic faults that can be sensitized by such a defect within the pre-defined fault space.

- **Device-aware test development**. The accurate and realistic faults obtained from the previous step are used to develop test solutions for DPPB level. Specifically, EtD faults can simply be detected by March tests. HtD faults, however, need special Design-for-Testability (DfT) or stress tests. The clear mapping between physical defects and fault models enables us to not only reduce test escapes and time but also speed up yield learning [67].

Next, we elaborate the above-mentioned three DAT steps in more detail.

6.3. DEVICE-AWARE DEFECT MODELING

Inaccurate defect modeling may result in poor fault models, thereby limiting the effectiveness of proposed test solutions and DfT designs, not only in terms of defect coverage but also in terms of test time. For example, a test targeting a fault model that does not represent any real defect will not increase the defect coverage while still consuming test time. To accurately model physical defects, the device model should incorporate the way the defect impacts the technology parameters and thereafter the electrical parameters of the device; this is exactly what device defect modeling of DAT does. Figure 6.4 shows the flow of such modeling approach using MTJ as an example; its inputs are 1) the defect-free MTJ compact model and 2) the defect under investigation. The output is an optimized (parameterized) defective MTJ compact model. Note that the device can also be an RRAM device, a PCM device, a planar or FinFET transistor etc. The approach consists of the following three steps.

1) Physical defect analysis and modeling. Given a set of physical defects $\mathbf{D} = \{d_1, d_2, ..., d_n\}$ that may occur during MTJ fabrication, each defect d_i has to be physically analyzed and modeled. The effect of defect d_i can be reflected by modifying one or more technology parameters listed in Table 3.1; e.g., RA and TMR. This results in an *effective technology* parameter (Tp_{eff}) that can be described by the following abstract function:

$$Tp_{\text{eff}}(\boldsymbol{S_i}) = f_i(Tp_{\text{df}}, \boldsymbol{S_i}) \tag{6.1}$$

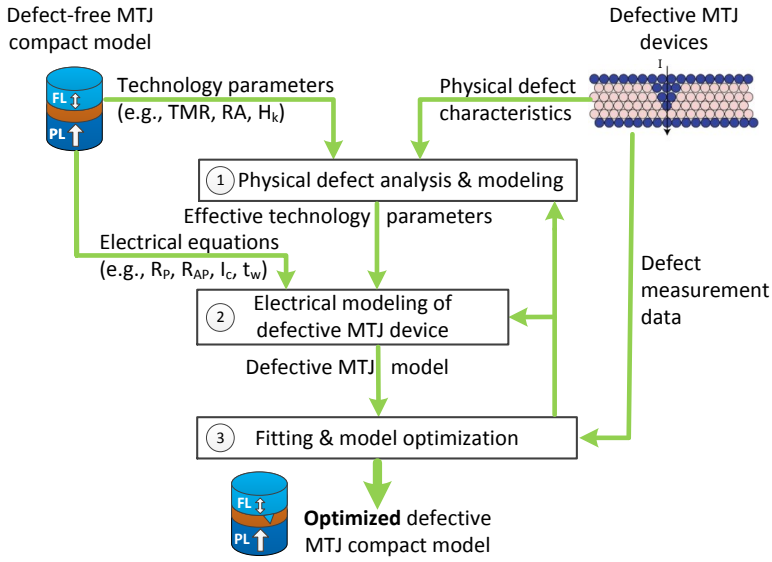


Figure 6.4: Generic device-aware defect modeling flow.

where Tp_{df} is the defect-free technology parameter, f_i is a mapping function corresponding to defect d_i ($i \in [1, \text{n}]$), $\boldsymbol{S_i} = \{x_1, x_2, \cdots, x_t\}$ is a set of parameters representing the size or strength of defect d_i. Note that each defect may impact one or more technology parameters.

2) Electrical modeling of the defective MTJ device. In this step, the impact of the updated technology parameters from Step 1 on the electrical parameters is identified; it reflects the way such defect d_i influences the electrical parameters of the MTJ device. This can be done for example by updating the electrical parameters (see Table 3.1) of the defect-free MTJ model (e.g., the Verilog-A MTJ compact model calibrated with measurement data in [62]). Note that the electrical parameters are the ones needed for accurate circuit simulation for fault modeling. This step enables us to obtain a raw defective MTJ model.

3) Fitting and model optimization. To validate the effectiveness of the defective MTJ model, it is suitable to fit the defective model to measurement data of real defective MTJ devices. If the behavior of the defective model (either its physical or electrical parameters) does not match the characterization data, the fitting parameter adjustment is necessary until an acceptable accuracy is obtained. Finally, we derive an optimized *defect-parameterized* compact model for defective MTJ devices.

6.4. DEVICE-AWARE FAULT MODELING

In order to obtain appropriate fault models, the defect models that can be generated on the approach discussed in the previous section should be used to analyze the behavior of a memory in the presence of defects. The results from this analysis are used to develop a high-quality test. Fault modeling process consists of two steps: 1) *fault space* that de-

scribes *all possible* faults and a classification of them; 2) fault analysis methodology that determines which faults from the fault space are *realistic* for the defect under consideration, i.e., which faults are sensitized in the presence of such a defect. These steps will be explained next.

6.4.1. FAULT SPACE AND CLASSIFICATION

In this work, we limit the analysis to single-cell faults [39]. If only one cell is involved, the fault is called single-cell fault. If multiple cells are involved, the fault is a multi-cell fault, which is out of the scope of this paper. Memory faults can be systematically described by *fault primitives* (FPs) [39]. An FP describes the deviation of the observed memory behavior from the expected. The FP notation is denoted as a three-tuple $\langle S/F_n/R\rangle$, which is explained as follows.

1) S (sensitizing sequence) denotes an operation sequence that sensitizes a fault. It takes the form of $S{=}x_0O_1x_1\ldots O_mx_m$, where $x_i{\in}\{0,1\}$ ($i{\in}\{0,1,...,m\}$) and $O{\in}\{\mathrm{r},\mathrm{w}\}$. Here, '0' and '1' denote the logic values of memory cells, while 'r' and 'w' denote a reading and a writing operation, respectively. m is the number of operations involved in the sensitizing sequence. For example, $S{=}0$ means the addressed cell is initialized to logic '0' state and no write/read operations are applied, while $S{=}$1w0r0 means that the addressed cell is initialized to '1' state followed by write '0' and read '0' operations.

2) F_n (faulty effect) describes the value that is stored in the cell after S is performed. For traditional charge-based memories, e.g., SRAM, there exists only two digital states, i.e., $F{\in}\{0,1\}$. However, data in STT-MRAM cells is stored in MTJ devices whose predefined resistance ranges determine the logic states '0' and '1'. Due to defects or extreme process variations, the MTJ resistance can be outside these ranges. Hence, it is necessary to define other (faulty) resistance states to cover defective MTJ devices. Figure 6.5 presents the measured resistance distribution of a large number of CD=60 nm

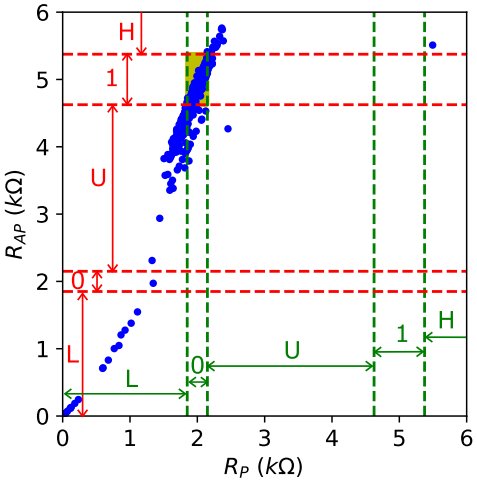


Figure 6.5: Measured resistance distribution of R_P and R_{AP} for MTJ devices with CD=60 nm, suggesting the existence of states 'L', '0', 'U', '1', and 'H'.

MTJ devices; it shows that $F \in \{0, 1, U, L, H\}$, as will be explained next. Each point in the figure represents a device whose R_P is shown on the x-axis and R_{AP} on the y-axis. From a design perspective, the nominal R_P is 2 kΩ and the nominal R_{AP} is 5 kΩ; this assures a good read reliability with $TMR = 150\%$. A 3σ variation of the nominal values is used to define the resistance ranges of the two state '0' and '1'. As shown in the figure, the points inside the shaded box represent good devices in accordance with the above design specifications. However, there are also a large number of devices outside the specification due to some defects or extreme process variations. These are: 1) extreme low resistance state 'L', (2) extreme high resistance state 'H', and (3) undefined state 'U'. The subscript 'n' specifies the nature of the faulty effect. $n \in \{p, i, t\}$, where 'p', 'i', and 't' denote permanent, intermittent, and transient faults, respectively [3]. When n=p, it is omitted as a compatibility measure to the conventional notation.

3) R (readout value) describes the output of a read operation if the last operation in S is a read operation. Here, $R \in \{0, 1, ?, -\}$. '?' denotes a random readout value in case the sensing current is very close to sense amplifier's reference current (e.g., the cell under read is in a 'U' state). '$-$' denotes that R is not applicable, i.e., when the last operation in S is not a read operation. Note that a read operation on a cell in 'L' state returns a logic '0' while the 'H' state returns a logic '1'.

Depending on the number of operations involved in the sensitizing operation S, FPs can be classified into *static* and *dynamic faults* [186]. A static fault is a fault which can be sensitized by at most one operation (i.e., $m \leq 1$), while a dynamic fault requires more than one operations (i.e., $m > 1$) to be sensitized. The FP names comply with the following format, where the fields in curly braces {} are required while the fields in square brackets [] are optional.

$$\text{FP} = \begin{cases} \text{S}\{ini\}F\{fin\}_{[n]}, & m = 0; \\ [out]\{opn\}\{opd\}\{eff\}\text{F}\{fin\}_{[n]}, & m = 1; \\ \{m\text{d}-\}[out]\{opn\}\{opd\}\{eff\}\text{F}\{fin\}_{[n]}, & m > 1. \end{cases}$$

If no read/write operation is involved in S (i.e., m=0), the FP name complies with the format: $\text{S}\{ini\}F\{fin\}_{[n]}$, where

- ini describes the initial state of the faulty cell; $ini \in \{0, 1\}$.
- fin describes the final state of the faulty cell; $fin \in \{L, 0, U, 1, H\}$.
- n describe the fault nature; $n \in \{p, i, t\}$. By default, n=i meaning a permanent fault and it is omitted.

For example, the fault primitive S1FU=$\langle 1/U/- \rangle$ means a *permanent state fault* with initialized state '1', but it ends up in state 'U' due to the existence of a defect. The *intermittent state fault*: S0FFU$_i$=$\langle 0/U_i/- \rangle$ indicates that an initialization of state '0' intermittently puts the cell at state 'U'.

If an FP involves only one sensitizing operation in S (i.e., m=1), then its name complies with the format: $[out]\{opn\}\{opd\}\{eff\}\text{F}\{fin\}_{[n]}$. Apart from the $\{fin\}$ and $[n]$ fields already introduced previously, the remaining fields are explained as follows.

Table 6.1: Complete single-cell permanent static fault primitives.

#	S	F	R	Notation	Name	#	S	F	R	Notation	Name
1	0	1	-	⟨0/1/–⟩	S0F1	27	0r0	1	0	⟨0r0/1/0⟩	dR0DF1
2	0	L	-	⟨0/L/–⟩	S0FL	28	0r0	1	?	⟨0r0/1/?⟩	rR0DF1
3	0	U	-	⟨0/U/–⟩	S0FU	29	0r0	1	1	⟨0r0/1/1⟩	iR0DF1
4	0	H	-	⟨0/H/–⟩	S0FH	30	0r0	L	0	⟨0r0/L/0⟩	dR0DFL
5	1	0	-	⟨1/0/–⟩	S1F0	31	0r0	L	?	⟨0r0/L/?⟩	rR0DFL
6	1	L	-	⟨1/L/–⟩	S1FL	32	0r0	L	1	⟨0r0/L/1⟩	iR0DFL
7	1	U	-	⟨1/U/–⟩	S1FU	33	0r0	U	0	⟨0r0/U/0⟩	dR0DFU
8	1	H	-	⟨1/H/–⟩	S1FH	34	0r0	U	?	⟨0r0/U/?⟩	rR0DFU
9	0w1	0	-	⟨0w1/0/–⟩	W1TF0	35	0r0	U	1	⟨0r0/U/1⟩	iR0DFU
10	0w1	L	-	⟨0w1/L/–⟩	W1TFL	36	0r0	H	0	⟨0r0/H/0⟩	dR0DFH
11	0w1	U	-	⟨0w1/U/–⟩	W1TFU	37	0r0	H	?	⟨0r0/H/?⟩	rR0DFH
12	0w1	H	-	⟨0w1/H/–⟩	W1TFH	38	0r0	H	1	⟨0r0/H/1⟩	iR0DFH
13	1w0	1	-	⟨1w0/1/–⟩	W0TF1	39	1r1	0	0	⟨1r1/0/0⟩	iR1DF0
14	1w0	L	-	⟨1w0/L/–⟩	W0TFL	40	1r1	0	?	⟨1r1/0/?⟩	rR1DF0
15	1w0	U	-	⟨1w0/U/–⟩	W0TFU	41	1r1	0	1	⟨1r1/0/1⟩	dR1DF0
16	1w0	H	-	⟨1w0/H/–⟩	W0TFH	42	1r1	1	0	⟨1r1/1/0⟩	iR1NF1
17	0w0	1	-	⟨0w0/1/–⟩	W0DF1	43	1r1	1	?	⟨1r1/1/?⟩	rR1NF1
18	0w0	L	-	⟨0w0/L/–⟩	W0DFL	44	1r1	L	0	⟨1r1/L/0⟩	iR1DFL
19	0w0	U	-	⟨0w0/U/–⟩	W0DFU	45	1r1	L	?	⟨1r1/L/?⟩	rR1DFL
20	0w0	H	-	⟨0w0/H/–⟩	W0DFH	46	1r1	L	1	⟨1r1/L/1⟩	dR1DFL
21	1w1	0	-	⟨1w1/0/–⟩	W1DF0	47	1r1	U	0	⟨1r1/U/0⟩	iR1DFU
22	1w1	L	-	⟨1w1/L/–⟩	W1DFL	48	1r1	U	?	⟨1r1/U/?⟩	rR1DFU
23	1w1	U	-	⟨1w1/U/–⟩	W1DFU	49	1r1	U	1	⟨1r1/U/1⟩	dR1DFU
24	1w1	H	-	⟨1w1/H/–⟩	W1DFH	50	1r1	H	0	⟨1r1/H/0⟩	iR1DFH
25	0r0	0	?	⟨0r0/0/?⟩	rR0NF0	51	1r1	H	?	⟨1r1/H/?⟩	rR1DFH
26	0r0	0	1	⟨0r0/0/1⟩	iR0NF0	52	1r1	H	1	⟨1r1/H/1⟩	dR1DFH

- out describes the readout effect of the read operation in S if applicable; $out \in \{i, r, d\}$, where 'i' means an incorrect readout, 'r' a random readout, and 'd' a deceptive readout. Note that a deceptive readout implies that the read operation returns a correct value while making the final state fin different from the one before reading. The out field is omitted when there is no read operation in S.
- opn describes the operation in S; $opn \in \{W, R\}$, where 'W' means a write operation while 'R' means a read operation.
- opd describes the operand of the operation opn; $opd \in \{0, 1\}$.
- eff describes the operational effect on the faulty cell; $eff \in \{T, D, N\}$, where 'T' means a transition operation, 'D' a destructive operation, 'N' non-destructive operation.

Table 6.1 lists all single-cell permanent static FPs with their notations and names. For intermittent and transient faults, the 'n' subscript (i.e., 'i' or 't') needs to be added to the FP notations and names. For instance, W0TFH=⟨1w0/H/–⟩ represents a *permanent*

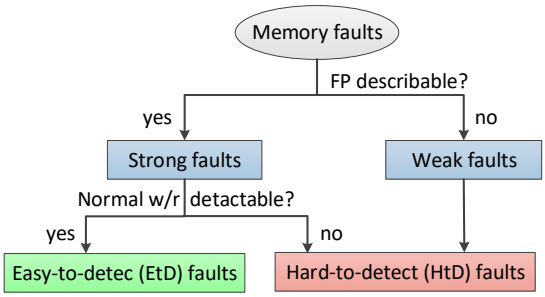


Figure 6.6: Fault classification.

Write Transition Fault where a write '0' operation forces the addressed cell with the initial state '1' to state 'H'. W0TFU$_i$= $\langle 1w0/U_i/-\rangle$ represents an *intermittent Write Transition Fault* where a down-transition write operation intermittently turns the addressed cell to state 'U' with a probability. rR1DFU=$\langle 1r1/U/?\rangle$ represents a *random Read Destructive Fault* where a read '1' operation forces the cell with initial state '1' to state 'U' and returns a random readout value. Similarly, other FPs can be interpreted according the above FP nomenclature.

It is worth noting that a *fault model* is a non-empty set of fault primitives with similar or complementary properties. For example, *State Fault* (SF) is a set of FPs from #1 to #8 in Table 6.1, whereas *Write Transition Fault* (WTF) includes FPs from #9 to #16. Similarly, one can also find the FPs belonging to *Write Destructive Fault* (WDF), *Read Non-destructive Fault* (RNF), and *Read Destructive Fault* (RDF) in the table.

For dynamic faults which are sensitized by more than one operation (i.e., m>1), their names get the prefix md− where m denotes the number of operations in S. Note that the naming scheme follows the same rules of static FPs using the last operation and its preceding state in S, e.g., $\langle 1r1w0/L/-\rangle$ is named as 2d-W0TFL.

As shown in Figure 6.6, memory faults can be classified into *strong faults* and *weak faults* depending on whether or not the fault can be described by fault primitives. Strong faults are faults that can always be sensitized by applying a sequence of operations and therefore can be described by fault primitives. Table 6.1 lists all static strong faults that may occur in a single memory cell. In contrast, weak faults *cannot* be described by fault primitives. However, they cause parametric changes in the circuits, e.g., a small reduction in the read current flowing through the cell under read. Although weak faults do not lead to any functional errors right after manufacture, they may cause severe reliability issues (e.g., shorter lifetime, higher in-field failure rate). Therefore, weak faults need to be detected as well when the target market has a strict quality requirement.

Depending on whether or not the fault is detectable by normal write or read operations, strong faults can be further divided into *easy-to-detect* (EtD) and *hard-to-detect* (HtD) faults. Although all strong faults can be sensitized by a sequence of operations S, their detection conditions may not necessarily be equal to S. EtD faults refer to those faults that can be easily detected by applying write and read operations (i.e., a March

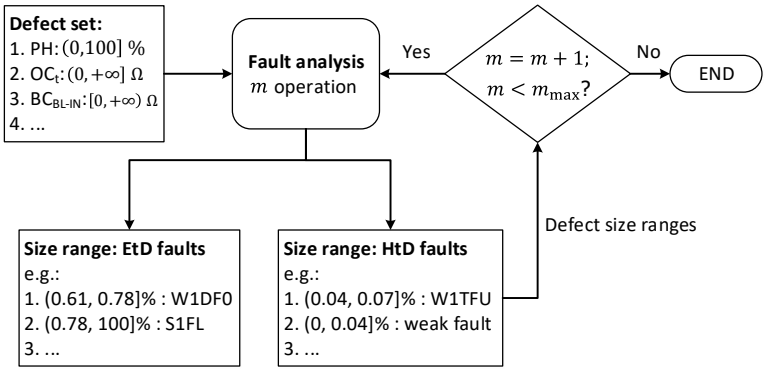


Figure 6.7: Fault analysis methodology.

test [36]). *Write Destructive Fault* W1DFL=⟨1w1/L/–⟩ and *incorrect Read Non-destructive Fault* iR1NF1=⟨1r1/1/0⟩ are two examples of EtD faults. The detection condition for the former is ⇳(...1,w1,r1,...). ⇳ denotes that the detection condition is independent on the addressing direction; (...1,w1,r1,...) denotes that the cell under test is initialized in logic '1', followed by a consecutive w1 and r1 operations, applied to each address before moving to the next address. Any March test meeting the above detection condition can guarantee the detection of the corresponding fault. In contrast, the detection of HtD faults *cannot* be guaranteed by just March tests; they require additional effort such as a special *Design-for-Testability* (DfT) circuit or a stress test in order to be detected. Note that strong faults consist of EtD and HtD faults, while weak faults are all HtD faults. Examples of strong HtD faults are *Write Transition Fault* W0TFU=⟨1w0/U/–⟩ and *random Read Non-destructive Fault* rR1NF1=⟨1r1/1/?⟩. For these two faults, March tests cannot guarantee their detections since a read operation on the faulty cell returns a random value.

6.4.2. FAULT ANALYSIS METHODOLOGY

Once STT-MRAM defects are modeled and the fault space is defined, the validation of the faults can be performed using a systematic circuit simulation approach. In this paper we restrict ourselves to single-cell fault analysis as only defects in a single 1T-1MTJ cell are considered in our simulations. Our fault analysis consists of seven steps: 1) circuit generation, 2) defect injection, 3) stimuli generation, 4) circuit simulation, 5) fault analysis, 6) fault primitives identification, and 7) defect strength sweeping and repetition of steps 2 to 6 until all defects and their sizes are covered. Note that in our simulations, defect injection means adding a specific resistor to the defect-free memory cell for interconnect defects, but it means replacement of the defect-free MTJ model with the defective MTJ model for MTJ defects (see Figure 6.4). In addition, defect size sweeping means changing resistance for the resistor model while it means changing the pinhole area A_{ph} for a pinhole defect in MTJ devices. Each time only one specific defect (e.g., an open OC_m or a pinhole PH) with certain size is analyzed in our simulations.

Figure 6.7 shows the fault analysis methodology that illustrates how we validate faults in the defined fault space due to the injection of defects. Given a set of defects and their size ranges, the seven steps of the fault analysis should be first performed for the vali-

dation of static single-cell FPs in Table 6.1 (i.e., $m \leq 1$). The simulation results are a list of {size range : EtD faults} pairs and a list of {size range : HtD faults} pairs, as shown in the figure. In case that no FP is sensitized in the presence of a defect with certain size range, the fault is considered as a weak fault belonging to HtD faults. Next, all defect size ranges resulting in HtD faults will be further analyzed using dynamic fault analysis with two sensitizing operations (i.e., m=2). In this way, some defect size ranges which lead to HtD faults from the previous static analysis may trigger EtD dynamic faults now; e.g., S=0w0 sensitizes a weak fault for a cell with a small defect, while S=0w0w0 may sensitize an EtD fault for this defective cell with the same defect size. Once two-operation single-cell dynamic fault analysis is done, we can redo similar fault analysis for m=3 for the remaining defect size ranges that result in HtD faults with two sensitizing operations. This simulation process can be iterated by extending S with one more operation each time until the pre-defined maximum number of operations ($m_{\max}$) is reached.

The aim of increasing the sensitizing operations is to reduce the defect size ranges which cause HtD faults meanwhile enlarging the ranges which lead to EtD faults. This is because EtD faults can simply be detected by March tests while HtD faults require DfT designs or stress tests to detect them. This fault analysis methodology is useful to optimize the ultimate test solution with a trade-off between the test quality and test overhead.

6.5. DEVICE-AWARE TEST DEVELOPMENT

The results of the fault analysis facilitate the development of high-quality yet efficient test solutions, as illustrated in Figure 6.8. All EtD faults can be detected by applying appropriate test algorithms. To minimize the test cost, the minimal detection conditions for each of the faults are first identified, and thereafter compiled into test algorithms. To further optimize the test time, one can also incorporate DfT; e.g., DfT that enables the test of many faults simultaneously, parallel testing, etc. [208–210].

HtD faults, however, require special attention. Special DfT schemes and tests are required. Examples are: DfT schemes that may directly measure the bit line swing [211],

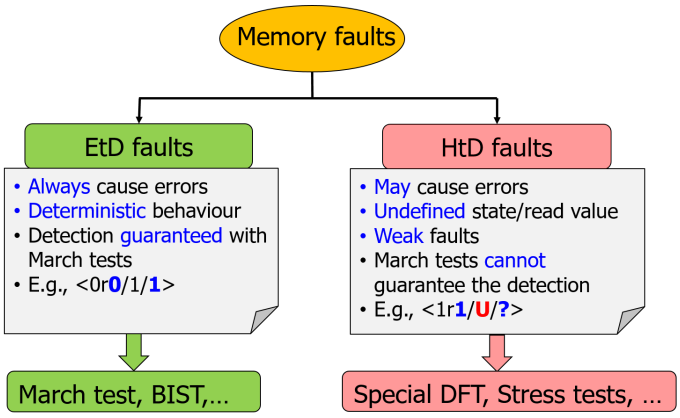


Figure 6.8: Device-aware test development.

modify the operation conditions such as weak write operations [209], stress tests [62], etc. The aim is to *maximize* the fault coverage for these faults while keeping the test cost affordable.

6.6. DAT ADVANTAGES AND CHALLENGES

In this chapter, we have presented the device-aware test approach which specifically targets device-internal defects. Compared to conventional test approaches such as cell-aware test where all defects are modeled as linear resistors (opens and shorts), DAT has the following advantages.

- **Test Escape Reduction and Quality Improvement:** As mentioned previously, linear resistors are not qualified to accurately model device-internal defects due to their non-linear and magnetic properties. Therefore, the fault models obtained using this approach cannot cover these defects at least partially. This inevitably results in poor-quality test solutions, thus leading to test escapes. To reduce test escapes, more stringent test program including I_{DDQ} and burn-in test has to be utilized. However, this comes with higher test cost and runs a risk of killing good chips (yield loss). In contrast, device-aware test is a superior solution to develop high-quality tests, thus reducing test escapes and yield loss, as illustrated in Figure 6.9. With our DAT approach, each type of manufacturing defect is characterized, analyzed, and modeled accurately. This allows us to obtain accurate fault models which appropriately represent the underlying defects at the functional level. This in turn leads to optimal test solutions, thereby reducing yield loss and test cost.

- **Efficient Yield Learning:** Modeling the defects accurately and creating a fault dictionary for them may speed up the yield learning process significantly. As each defect can be modeled separately using device-aware testing, instead of using resistive defect models for all defects, unique fault signatures can be created for each defect. This improves the yield learning curve, as the defects can be more accurately diagnosed based on their fault signatures.

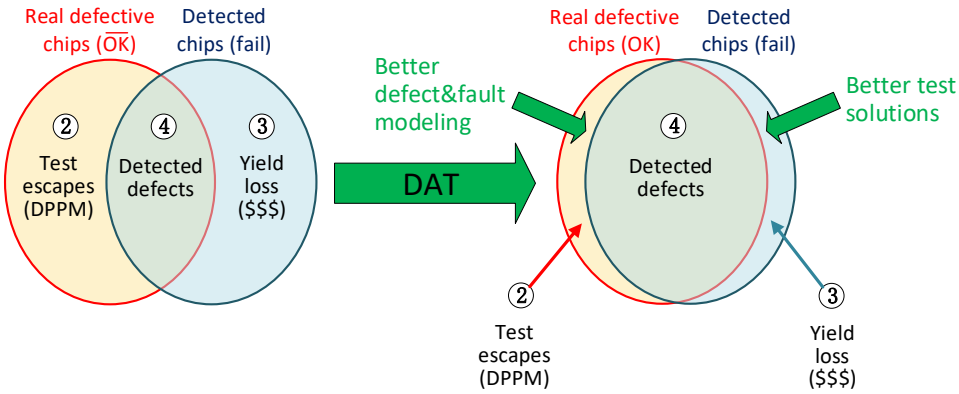


Figure 6.9: Benefits of device-aware test.

- **Test Time Optimization:** Nowadays, companies are spending a lot of time on functional test (or system test) to compensate for the fault coverage due to the limitations of traditional fault modeling and testing. The DAT approach allows for the development of appropriate and efficient structural tests, which can be applied at manufacturing stage; hence, DAT can significantly reduce the expensive test time spent on board testing.

  General Applicability: DAT has demonstrated its superiority in developing test solutions towards DPPB level for RRAM [207] and STT-MRAM [62]. It can also be applied to any kind of memories including advanced volatile technologies (e.g., SRAM, DRAM) as well as non-volatile ones (e.g., Flash, PCM). Moreover, it can be also applied to logic circuits especially for technology nodes below 10 nm, where it has been shown that many failure mechanisms cannot be modeled with linear resistors [212].

Despite the above-mentioned advantages of DAT over the conventional test approach, there exists challenges as follows.

- **Interdisciplinary collaboration:** Understanding and modeling physical defects require significantly more efforts than simply modeling them as linear resistors. But this is worth investing from an economic perspective, since it is a one-off action with long-term gain. To perform research work on DAT, it is necessary to have interdisciplinary collaboration between the device, processing technology, and test communities. Researchers at technology level are good at understanding and modeling the effects of defects on physical and technology parameters of the device and thereafter the electrical parameters, whereas test researchers are skilled with fault analysis and test development. Clearly, the fault modeling and test paradigm is changing for emerging technologies such as STT-MRAM.

- **Defect measurements data:** To obtain a good defect model, measurement data of real defective devices is crucial to calibrate the model. In addition, collecting and analyzing silicon data are also helpful to understand the defect mechanism, occurrence rate, location, etc. However, researchers in academia or in fabless companies rarely have access to silicon data, while test engineers in foundries may not necessarily have the required expertise or motivation to perform research work related to this topic. Therefore, joint research projects are required to bridge this gap.

7

DAT FOR PINHOLE DEFECTS

Understanding the manufacturing defects in magnetic tunnel junctions (MTJs), which are the data-storing elements in STT-MRAMs, and their resultant faulty behaviors are crucial for developing high-quality test solutions. Pinhole defects in the MgO tunnel barrier of MTJ are seen as a key type of STT-MRAM manufacturing defects. This chapter applies our proposed device-aware test (DAT) approach to pinhole defects. We start with introducing the pinhole defect mechanism including defect location, root causes, and potential impact. Thereafter, we identify pinhole defects in fabricated MTJ devices and characterize them both during manufacturing test ($t = 0$) and in the field ($t = 0$). The measurement data then is used to extend our defect-free MTJ compact model to a pinhole-parameterized defective MTJ model. By applying device-aware fault modeling to pinhole defects, the simulation results show that a large pinhole defect results in easy-to-detect faults (together equivalent to the traditional stuck-at-0 fault), while a small pinhole defect leads to hard-to-detect faults. The easy-to-detect faults can be detected by applying March tests. However, detecting the hard-to-detect faults require stress tests with hammering write 1 operations under elevated voltage and/or prolonged pulses.

The contents of this chapter have been published in ITC'18 [47] and ETS'19 [62].

7.1. PINHOLE DEFECT MECHANISM

The fabrication and integration process of MTJ devices is vulnerable to several defects, as introduced in Section 3.4. A pinhole defect in the tunnel barrier is seen as one of the most important manufacturing defects that may take place during the multi-layer deposition [116, 117, 213]. In [116], Zhao *et al.* present a schematic and a transmission electron microscope (TEM) image of the cross-section of an MTJ device with a small pinhole in its MgO tunnel barrier, as shown in Figure 7.1. A pinhole defect can form due

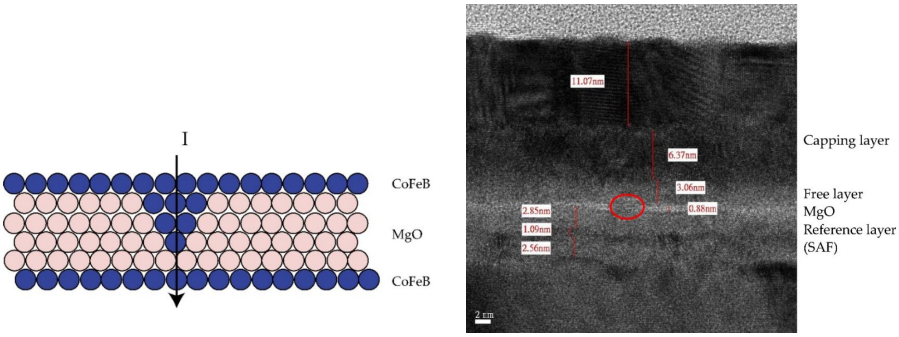


Figure 7.1: Pinhole defect mechanism: (a) schematic of an MTJ stack with a pinhole in the MgO tunnel barrier, (b) cross-section TEM of a MTJ with a pinhole defect (both graphs reprinted from [116]).

to unoptimized deposition processes [116]. This can cause the formation of metallic shorts in the MgO tunnel barrier, probably due to diffusion of Boron into the MgO barrier or other metallic impurities [213]. With a small pinhole filled with CoFeB material from the layer above, the tunneling current across the MgO barrier is shunted by a high-conductance path via the pinhole. As a result, it leads to a degradation of both *RA* and *TMR* parameters or even breakdown due to elevated Joule heating. Moreover, Oliver *et al.* [117] observed that pre-existing pinhole defects in the AlO_x-based barrier of an MTJ device grow in area over time because of Joule heating and/or an electric field across the pinhole circumference. Therefore, if even small pinhole defects are not detected during manufacturing tests (t=0), they might cause an early breakdown in the field (t>0).

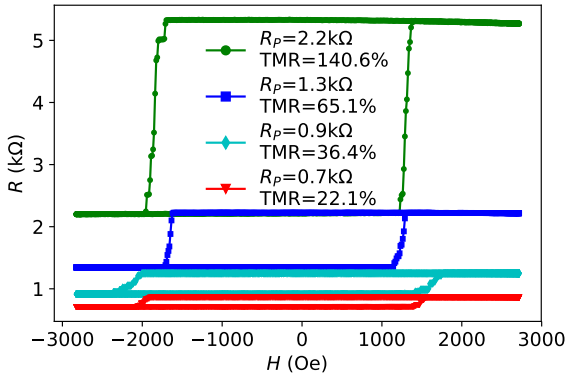
7.2. PINHOLE DEFECT CHARACTERIZATION

To develop an accurate compact model for pinhole defects, we need to characterize how they behave in our MTJ devices, both at t=0 (manufacturing stage) and t>0 (in the field). Base on the preliminary results observed in the prior work, it is clear that pinhole defect resides in the ultra-thin MgO barrier, while the FL is undamaged. In addition, small pinhole defects deteriorate over time, manifested as a decrease in both RA and TMR parameters. These unique features allow us to identify MTJ devices with pinhole defects among all fabricated devices.

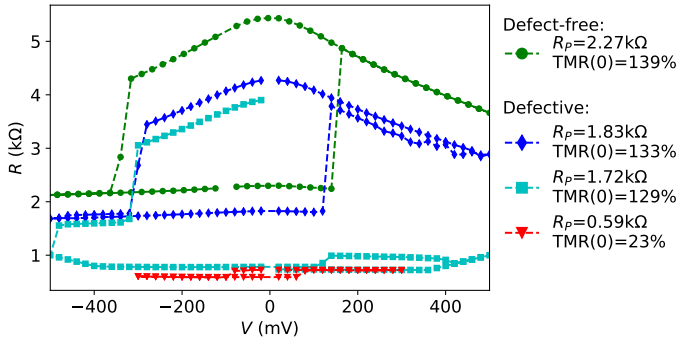
7.2.1. CHARACTERIZATION AT t=0

To characterize the MTJ devices at t=0, we measured the R-H hysteresis loop to extract R_P, R_{AP}, and switching field H_{sw} of hundreds of virgin devices with diameter 60 nm. During these measurements, ramped external fields were applied to the device under test; the magnetization in the FL flips when the external field reaches H_{sw}. After each field point, the resistive state was read out with a voltage of 20 mV. As the measured devices were not subjected to any electrical operation before, we considered the measured parameters to be representatives for the MTJ state at t=0.

Figure 7.2a shows the R-H hysteresis loops of four selected devices from the same wafer; each was measured ten times and the data was averaged to one loop. The widest green loop with R_P=2.2 kΩ and TMR=140.6% represents a good device, while the other three loops represent three defective devices. It can be seen that the resistance and TMR of the three defective devices are significantly smaller than the good one. For example, the red loop illustrates that R_P and TMR of that device decreases to 0.7 kΩ and 22.1%, respectively. However, H_{sw} does not show the same trend; the small H_{sw} variation between the four devices is caused by device process variations. This indicates that the defects reside in the MgO barrier or at the MgO/CoFeB interface, whereas the FL is *undamaged*.



(a) R-H hysteresis loops.



(b) R-V hysteresis loops.

Figure 7.2: Characterization of MTJ devices with pinhole defects at t=0.

In addition to R-H hysteresis loops, we also measured R-V hysteresis loops. Figure 7.2b shows the results of four other devices. Again, the biggest green loop indicates a good device, whereas the other three are measured from defective ones. Obviously, the loops of defective devices shrink, i.e., smaller R_P, *TMR*, and V_c. Note that the resistance of the cyan loop dives when the DC voltage reaches around −500 mV. This is because the existence of pinholes leads to an increase of current flow through them and in turn, a consequent increase in current-induced heating effects in the pinhole regions.

7.2.2. CHARACTERIZATION AT $t>0$

To study how R_P, R_{AP}, *RA*, and *TMR* parameters of defective devices change over time ($t>0$), we stressed a large number of MTJ devices (CD=60 nm) with the following two test sequences.

First, we stressed hundreds of virgin MTJ devices with 400k cycles of P→AP switching (i.e., hammering of reset operations) to track how R_{AP} changes over time. During this test, pulse amplitude V_p=−0.8 V and pulse width t_p=50 ns; note that the pulse width is more than twice the nominal value. After each pulse, we read back the MTJ resistance with a small (V_p=10 mV) but long (t_p=0.7 ms) pulse. We observed that all devices survived this stress test, except three devices broke down. Figure 7.3 shows the results of four selected devices: one defect-free device A (green wide line on the top) and three devices which broke down within the first 40 cycles (denoted as B, C, D). This suggests that probably these three devices have pinhole defects in the MgO barrier, which caused the early breakdown due to the increased Joule heating.

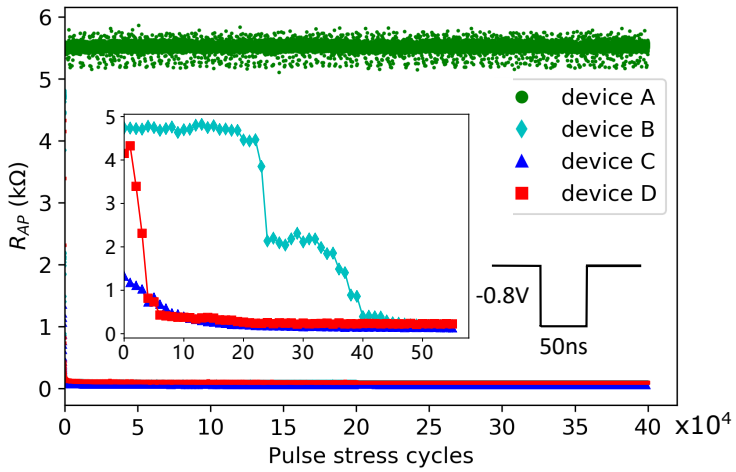


Figure 7.3: Characterization of MTJs with pinhole defects: R_{AP} degrades under pulse stress.

Second, we selected a device with a suspected large pinhole (R_P=451 Ω and *TMR*=9.1%) to investigate the impact on the effective *RA* and *TMR* over time. We increased the stress pulse width to 1 μs to speed up the degradation process, and measured R-H hysteresis loops after every 1k pulses. From the measured R-H hysteresis loop, we extracted the effective *RA* and *TMR*. Figure 7.4 shows that the effective *TMR* decreases linearly with

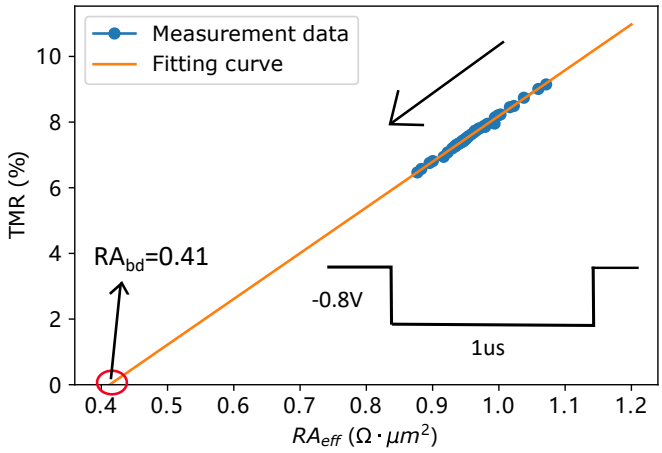


Figure 7.4: Characterization of an MTJ with a pinhole defect: TMR decreases linearly with RA under stress.

RA_{eff}. With a linear curve fitting, we obtained the breakdown resistance-area product RA_{bd}=0.41 $\Omega{\cdot}\mu\text{m}^2$ by extrapolating the curve to the crossing point at x-axis.

In conclusion, small pinhole defects grow over time into larger pinholes which cause an early/soft/extrinsic breakdown [214] at certain point. For devices with a small pinhole, the resistance and the TMR ratio drop dramatically with the applied pulses. As the pinhole grows up, their decrease rate becomes smaller.

7.3. LIMITATIONS OF THE CONVENTIONAL TEST APPROACH

As already mentioned in the previous chapter, the conventional test approach models any defect in an MTJ device as a linear resistor either in parallel to (R_{pd}) or in series with (R_{sd}) a defect-free MTJ model. The physical mechanism of defect is never taken into account and manifested as a difference in the defect model. This can be found in the prior works on STT-MRAM testing [49, 50, 52, 54–56, 61]. Applying the conventional fault modeling approach to the pinhole defect results in four FPs: iR1NF1 = $\langle 1r1/1/0\rangle$, iR0NF0 = $\langle 0r0/0/1\rangle$, W1TF0 = $\langle 0w1/0/-\rangle$, W0TF1 = $\langle 1w0/1/-\rangle$, as shown in Table 7.1. These four FPs can be used to generate test solutions such as March algorithms. First, each sensitized FP is assigned its own detection condition. For instance, iR0NF0 requires a read operation on the faulty cell at state '0' to guarantee its detection, denoted as $\Updownarrow(...0,r0,...)$, where $\Updownarrow$ means that the detection condition does not depend on the addressing direction. The detection condition for W0TF1 is $\Updownarrow(...1,w0,r0,...)$, meaning that a down-transition write followed by a read is enough to detect this fault, regardless of the addressing direction. The detection conditions of all sensitized FPs are compiled into the following optimal March test with three march elements:

$$\{\Updownarrow(w0);\Uparrow(w1,r1);\Downarrow(w0,r0)\}.$$

Note that different versions of March tests can be generated (e.g., with two march elements) as long as the test satisfies all the detection conditions.

Table 7.1: Static fault modeling results of pinhole defect using resistive models.

Defect model	Resistance (Ω)	Sensitized FP	Fault Model &FP Name	Detection Condition
Series resistor R_{sd}	(466, 870]	⟨0r0/0/1⟩	incorrect Read Non-destructive Fault: iR0NF0	⇕ (...0,r0,...)
	(870, 1.6k]	⟨0r0/0/1⟩	incorrect Read Non-destructive Fault: iR0NF0	⇕ (...0,r0,...)
		⟨1w0/1/–⟩	Write Transition Fault: W0TF1	⇕ (...1,w0,r0,...)
	(1.6k, +∞]	⟨0r0/0/1⟩	incorrect Read Non-destructive Fault: iR0NF0	⇕ (...0,r0,...)
		⟨1w0/1/–⟩	Write Transition Fault: W0TF1	⇕ (...1,w0,r0,...)
		⟨0w1/0/–⟩	Write Transition Fault: W1TF0	⇕ (...0,w1,r1,...)
Parallel resistor R_{pd}	[0, 1.1k)	⟨1r1/1/0⟩	incorrect Read Non-destructive Fault: iR1NF1	⇕ (...1,r1,...)
		⟨1w0/1/–⟩	Write Transition Fault: W0TF1	⇕ (...1,w0,r0,...)
		⟨0w1/0/–⟩	Write Transition Fault: W1TF0	⇕ (...0,w1,r1,...)
	[1.1k, 3.1k)	⟨1r1/1/0⟩	incorrect Read Non-destructive Fault: iR1NF1	⇕ (...1,r1,...)
		⟨1w0/1/–⟩	Write Transition Fault: W0TF1	⇕ (...1,w0,r0,...)

Based on our measurement results in the previous section, one can easily observe that the sensitized four FPs using the conventional fault modeling approach cannot cover the faulty behaviors of pinhole defects with different sizes. This is because a pinhole defect may turn an MTJ device into state 'U', while the MTJ device is considered as an *ideal black box* (only state '0' and '1') in the conventional fault modeling approach. In addition, linear resistors fail to capture the pinhole-induced changes on the device's magnetic properties and switching behavior (see the switching voltages in Figure 7.2b). As the four FPs are inappropriate in presenting pinhole defects, March tests that target these faults obviously cannot guarantee the detection of pinhole defects. Therefore, we need to apply DAT to pinhole defects for accurate defect and fault models, which will eventually lead to high-quality test solutions that we desire.

7.4. DEVICE-AWARE DEFECT MODELING FOR PINHOLES

Next, we apply the three steps of device-aware defect modeling (see Figure 6.4) to the pinhole defect to obtain an accurate defective MTJ model.

1) Physical defect analysis and modeling. With the comprehensive characterization of pinhole defects in the previous section, we model the impact of pinhole defects on *RA* and *TMR* as follows [47]:

$$RA_{\text{eff\_ph}}(A_{\text{ph}}) = \frac{A_0}{\frac{A_0(1-A_{\text{ph}})}{RA_{\text{df}}} + \frac{A_0 \cdot A_{\text{ph}}}{RA_{\text{bd}}}}, \tag{7.1}$$

$$TMR_{\text{eff\_ph}}(A_{\text{ph}}) = TMR_{\text{df}} \cdot \frac{RA_{\text{eff\_ph}}(A_{\text{ph}}) - RA_{\text{bd}}}{RA_{\text{df}} - RA_{\text{bd}}}. \tag{7.2}$$

In the above two equations, $A_{\text{ph}} \in [0,1]$ is the normalized pinhole area with respect to the cross-sectional area A_0 of the MTJ device. RA_{df} and TMR_{df} are *RA* and *TMR* parameters of a defect-free MTJ (i.e., when A_{ph}=0), respectively. RA_{bd} is the resultant *RA* after breakdown. These parameters depend on MTJ designs and can be extracted by measuring defect-free MTJs and defective MTJ with pinhole defects. Note that the location of the pinhole defect has negligible effects on the electron transportation in the two-terminal

MTJ device, as electrons either tunnel through the pinhole area or the undamaged parts [117, 215]. Apart from the pinhole location, its shape also plays little role as the MgO layer is ultra-thin, typically ~1 nm which is equivalent to a few atoms in thickness.

We simulated the effective technology parameters in MATLAB. We replaced the initial defect-free RA and *TMR* parameters with Equations (7.1–7.2) to observe how they change with the pinhole defect. Figures 7.5 shows the impact of pinhole defects on the *TMR* and *RA* parameters; clearly the effective *RA* (left y-axis) decreases exponentially with the pinhole area when less than ~20% of the MTJ's cross-section. This means that the tunneling magneto-resistance dominates the MTJ resistance for small pinhole defects. When A_{ph} is larger than 20%, the resistance of the MTJ behaves like a metal resistor. The *TMR* parameter (right y-axis) degrades in a similar way with the normalized pinhole area as shown in Figure 7.5. This is because the pinhole defect introduces a competition between the current going through the undamaged part $A_0(1 - A_{\text{ph}})$ of the barrier and the current going through the pinhole area, and only the former accounts for the TMR effect [216].

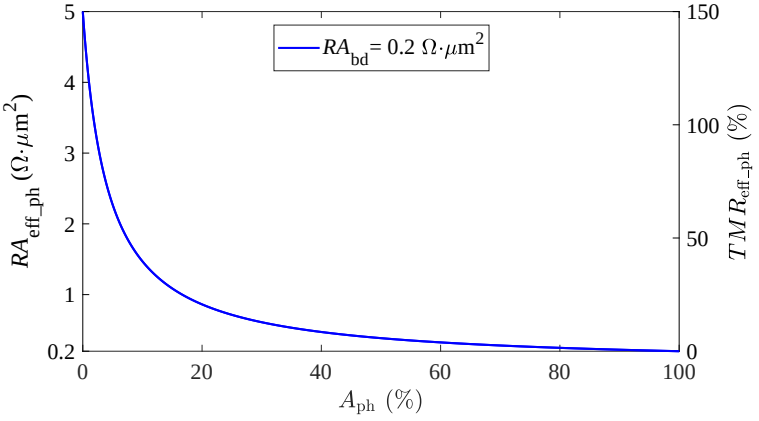


Figure 7.5: Effective *RA* and *TMR* with respect to the normalized pinhole area A_{ph}.

2) Electrical modeling of the defective MTJ device. The mapping from technology parameters to electrical parameters (i.e., $R_{\text{P}}, R_{\text{AP}}, I_{\text{c}}, t_{\text{w}}$) is realized by a number of physical models, which are mainly described by Equations (4.1–4.8). For the defective MTJ model, we replaced the original *RA* and *TMR* parameters with the effective ones in Equations (7.1–7.2). Thus, we obtained a pinhole-adjustable defective Verilog-A PMA-MTJ model with an input argument A_{ph}. With this model, we are able to evaluate how the pinhole defect impacts the MTJ's electrical behavior.

Figure 7.6 shows that the pinhole defect leads to a shrunk R-V hysteresis loop, indicating that both write and read operations are affected. As the hysteresis loop shrinks below a certain threshold (depending on the pinhole area A_{ph}), it becomes impossible to distinguish between the two states, leading to a stuck-at-fault (SAF). Figure 7.7 illustrates that the critical switching current I_{c} gradually increases with A_{ph} when less than ~80%. When larger than ~80%, I_{c} increases exponentially. The increase in I_{c} results from the degradation of spin polarization P due to the pinhole defect. This means more current is

required in order to switch the MTJ state. However, it is worth noting that for A_{ph} larger than 10% I_c is not that important any more, since the MTJ behaves as a SAF as we discussed previously. Figure 7.8 shows the effect of pinhole defects on the STT switching time t_w. It can be seen that t_w decreases with A_{ph} and stabilizes around A_{ph} = 10%. The decrease in t_w is due to an increase in the write current margin (see Equation (4.8). Note that although I_c increases with the pinhole defect as shown in Figure 7.7, the write current increases faster, as the MTJ resistance declines significantly with the pinhole defect (see Figure 7.6). This indicates that the writability (write latency) of MTJ is enhanced by pinhole defects. However, it is worth noting that the increased programming current also makes the MTJ device more vulnerable to a permanent breakdown.

3) Fitting and model optimization. In this step, we use the measurement data of

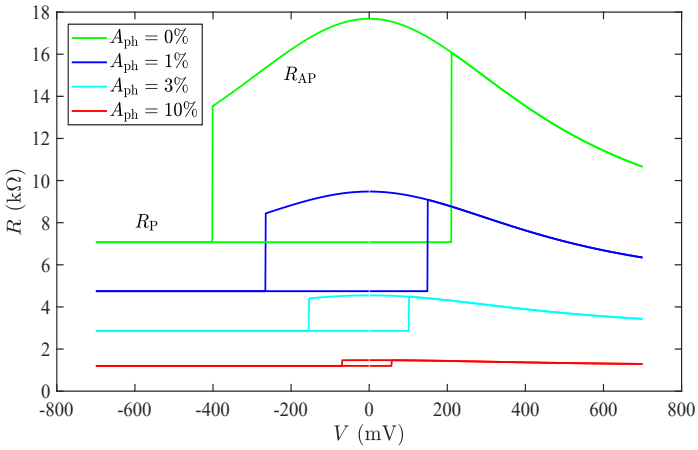


Figure 7.6: Impact of pinhole defects on the R-V hysteresis loop.

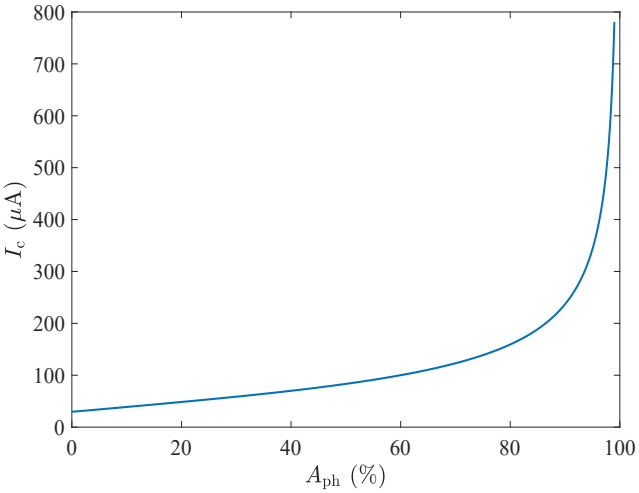


Figure 7.7: Impact of pinhole defects on the critical switching current I_c.

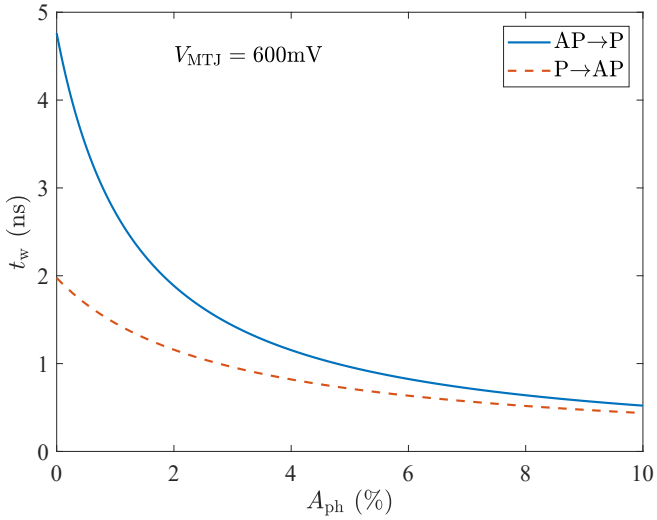


Figure 7.8: Impact of pinhole defects on the switching time t_w.

MTJ devices with and without pinhole defects to calibrate our model. To this end, we performed comprehensive electrical and magnetic characterizations of MTJ devices at both t=0 and t>0 (i.e., stress test), as presented in the previous section. Based on the measurement results for defect-free MTJs, we take A_0 = 2827.4 nm^2, RA_{df}=4.52 Ω·μm^2, and TMR_{df}=139%. By constantly stressing the devices with a small pinhole while tracking its RA and TMR values, we obtained RA_{bd}=0.41 Ω·μm^2 after extrapolating the fitting curve to the point where TMR=0, as shown in Figure 7.4. The output of device-aware defect modeling is an optimized defective MTJ model, as show in Figure 7.9. After verifying the defective MTJ model in MATLAB, we moved this model to Verilog-A so as to

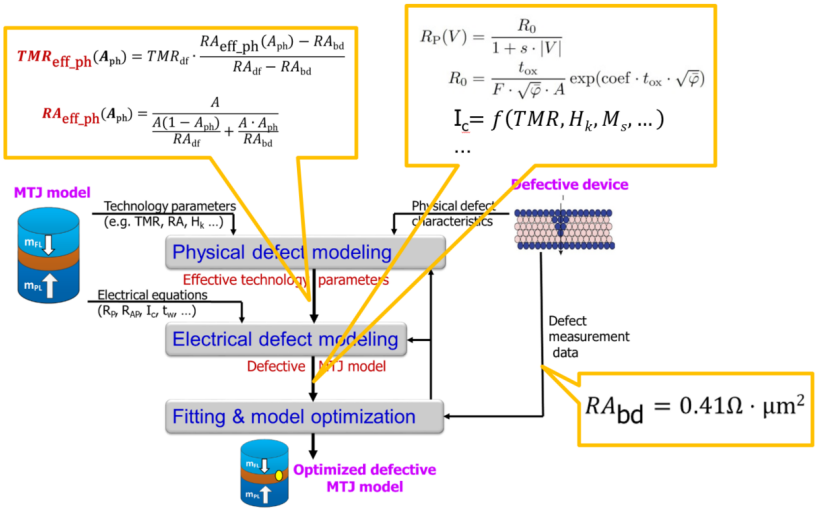


Figure 7.9: Device-aware defect modeling process for pinhole defects.

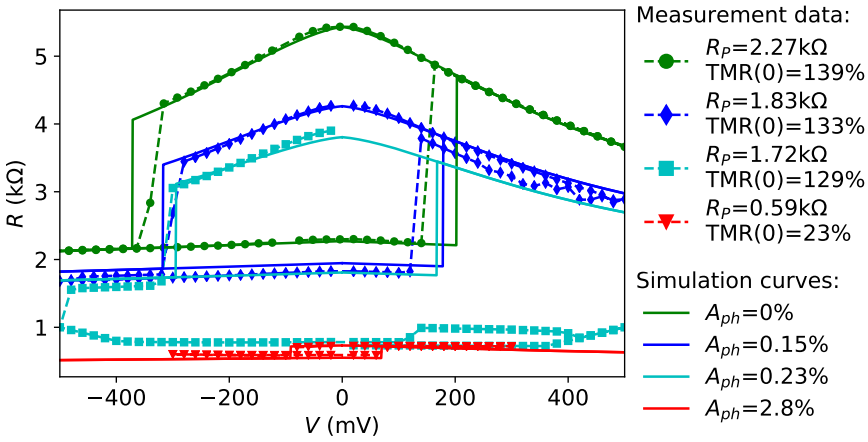


Figure 7.10: Spectre simulation results vs. measurement data.

make it compatible with circuit simulators such as Cadence Spectre for subsequent fault modeling.

Figure 7.10 shows the Spectre simulation results (solid curves) of R-V hysteresis loops with various A_{ph} values. It can be seen that the simulation results with our proposed defective MTJ model match the measured silicon data in terms of resistance and switching voltage. Note that our simulation results represent the green R-V loop with an injection of pinhole defects. However, the other three measured R-V hysteresis loops belong to three distinct defective devices, which may have different RA_{df} and TMR_{df} due to process variation. Based on the proposed defective MTJ model, accurate fault modeling of pinhole defects and subsequent test development can be performed.

7.5. DEVICE-AWARE FAULT MODELING FOR PINHOLES

Device-aware fault modeling is the second step in our proposed DAT approach. We have worked out this step for pinhole defects based on circuit simulations of STT-MRAM circuits presented in Chapter 3. Note that we replaced the defect-free MTJ model with the calibrated defective MTJ model for the defect injection, instead of adding a series or parallel resistor in the conventional test approach (see Figure 6.1). The pinhole size is represented by an input parameter A_{ph} (the pinhole area normalized the cross-sectional area of the MTJ device) of the defective MTJ model. In our simulations, we swept A_{ph} from 0% to 100%. Next, we present the fault modeling results and compare them with those using the conventional approach.

The upper part of Table 7.2 shows the fault modeling results of pinhole defects in MTJ devices using our proposed DAT approach; the fault detection condition for each pinhole size range are listed in the last column. It can be seen that sufficiently large pinholes (A_{ph}>0.61%) make the MTJ device fall into the resistance range of '0' state or even of 'L' state, sensitizing easy-to-detect faults; the corresponding fault primitives are listed in the table. Among those FPs, S1F0=⟨1/0/−⟩ and S1FL=⟨1/L/−⟩ (marked with bold font) are easy to detect with a read '1' (r1) operation. As the pinhole gets smaller

Table 7.2: Single-cell static fault modeling results of pinhole defects.

	Defect Model	Defect Strength	Sensitized Fault Primitive	Detection Condition
DAT	Pinhole area A_{ph}	(0.04, 0.07]	S1FU, W1DFU, W1TFU, dR1DFU	Stress tests/ DfT designs
		(0.07, 0.32]	S0FL, S1FU, W0DFL, W1DFU, W1TFU, W0TFL, dR0DFL, dR1DFU	
		(0.32, 0.35]	S0FL, S1FU, W0DFL, W1DFU, W1TFU, W0TFL, dR0DFL, rR1DFU	
		(0.35, 0.61]	S0FL, S1FU, W0DFL, W1DFU, W1TFU, W0TFL, dR0DFL, **iR1DFU**	⇳ (...1, r1, ...)
		(0.61, 0.78]	S0FL, **S1F0**, W0DFL, W1DF0, W1TF0, W0TFL, dR0DFL, iR1DF0	⇳ (...1, r1, ...)
		(0.78, 100]	S0FL, **S1FL**, W0DFL, W1DFL, W1TFL, W0TFL, dR0DFL, iR1DFL	⇳ (...1, r1, ...)
Conventional	Parallel resistor R_{pd}	[0,1.1k)	**iR1NF1**, W1TF0, W0TF1	⇳ (...1, r1, ...)
		[1.1k,3.1k)	**iR1NF1**, W0TF1	
	Series resistor R_{sd}	(466, 870]	**iR0NF0**	⇳ (...0, r0, ...)
		(870, 1.6K]	**iR0NF0**, W0TF1	
		(1.6k, +∞]	**iR0NF0**, W0TF1, W1TF0	

(A_{ph}∈(0.07%,0.61%]), it makes R_P fall into 'L' state and R_{AP} into 'U' state. Depending on the exact MTJ resistance in the AP state, the readout value can be one of the following three cases: (a) '0', (b) random ('?'), and (c) '1'. In Case (a) where R_{AP} is significantly smaller than the resistance of the reference cell (i.e., A_{ph}∈(0.35%,0.61%]), the readout value of the device in AP state is '0'. In this case, a r1 operation can detect the sensitized FP iR1DFU=⟨1r1/U/0⟩ (marked with bold font). In Case (b) where R_{AP} is close to the resistance of the reference cell (i.e., A_{ph}∈(0.32%,0.35%]), the readout value is random, leading to strong hard-to-detect faults. In other words, the read operation is unstable, and therefore both '0' and '1' are possible readout values. Thus, a r1 operation cannot guarantee the detection. In Case (c) where R_{AP} is much larger than the resistance of the reference cell while it is still out of the spec. of the logic '1' (i.e., A_{ph}∈(0.07%,0.32%]), the readout is '1'. In this case, strong hard-to-detect faults are sensitized which cannot be detected by March tests. As the pinhole area becomes smaller between 0.04% to 0.07%, R_{AP} falls into a 'U' state, while R_P remains in the correct range. Similarly, the sensitized strong hard-to-detect faults cannot be detected by March tests. If the pinhole size is smaller than 0.04%, it leads to a weak fault, while the cell still behaves logically correct.

Conventionally, MTJ-internal defects irrespective of their physical natures are modeled as linear resistors either in parallel to (R_{pd}) or in series with (R_{sd}) to an *idea defect-free* MTJ model, as mentioned previously. The fault modeling results using R_{pd} and R_{sd} as the pinhole defect model are shown in the lower part of Table 7.2. Comparing the fault modeling results of our DAT approach and the conventional approach reveals the following.

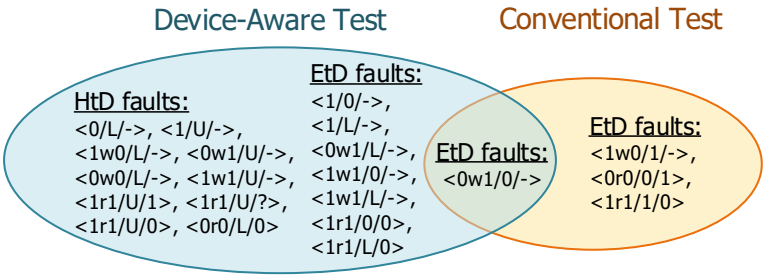


Figure 7.11: Comparison of sensitized FPs due to pinhole defects: device-aware test vs. conventional test approach based on linear resistors.

- The faulty behavior of the memory due to a pinhole defect *cannot* be covered by the conventional resistor-based defect models. Figure 7.11 shows our DAT approach results in 18 FPs. Among these FPs, 17 FPs are not observed with resistor models R_{pd} and R_{sd} while only a single EtD fault (W1TF0=⟨0w1/0/–⟩) is in overlap. Among the unique 17 FPs generated by our DAT approach, 10 FPs are HtD faults and the rest 7 FPs are EtD faults. With the resistor-based defect models, only '0' and '1' states were observed in the simulations, leading to 4 EtD faults. This is because the MTJ device is considered as a **black box** and **ideal**. However, our simulations and measurement data clearly show that pinhole defects can lead the device to 'U' or even 'L' state.

- Conventional resistor-based defect models may result in *wrong* fault models. Figure 7.11 shows that R_{pd} and R_{sd} result in 3 FPs which are not applicable to pinhole defects (i.e., not observed with our device-aware pinhole defect model).

The above observations clearly indicate that test algorithms developed with the conventional resistor-based defect modeling approach not only cannot guarantee the detection of pinhole defects leading to test escapes, but also may waste test time and resources as they target non-existing faults. Hence, more attention needs to be paid to the analysis and modeling of defects in MTJ devices, since those defects cannot be simply modeled as linear resistors but they have significant impacts on the data-storing MTJ devices in STT-MRAMs.

7.6. DEVICE-AWARE TEST DEVELOPMENT FOR PINHOLES

Based on the previous fault analysis results, appropriate test solutions can be developed to detect pinhole defects with different sizes. It is clear that the larger the pinhole, the larger its fault effect; hence, the easier it is to be detected. Combining the last three rows in the DAT part of Table 7.2, we can see that any March algorithm including the element ⇕(w1,r1) can guarantee the detection of a pinhole defect with A_{ph}>0.35% as it sensitizes only easy-to-detect faults. Large pinhole results in the faulty effect equivalent to the conventional *stuck-at-0* (SA0) fault.

However, for smaller pinhole defects (A_{ph}≤0.35%), HtD faults are sensitized. They are typically related to the cell being in a forbidden state (i.e., H, L, or U) or to random

readout values. Obviously, March tests cannot guarantee the detection of such faults, although they may detect some of them. For example, iR1DFU=⟨1r1/U/0⟩ may be detected by a March test {⇕(w1), ⇕(r1)}. Applying March tests multiple times with different data background and address sequences [36, 188] will increase the detection probability of such faults. As small pinhole defects grow in area over time due to the accumulated Joule heating, they would cause an early breakdown in the field if not detected during manufacturing tests [62]. Hence, guaranteeing their detection is a must.

Using DfT or stress tests are common practices to further increase the change of detecting HtD faults. One possible solution is to subject the STT-MRAM to a hammering write '1' operation sequence with elevated voltage or prolonged pulse width to deliberately speedup the growth of pinhole defects, so as to transform hard-to-detect faults to easy-to-detect faults. Figure 7.3 shows the measurement data of four selected MTJ devices under a stress test. In this test, we constantly applied hammering write '1' operations (P→AP switching) to hundreds of MTJ devices for 400k cycles; the pulse amplitude and width are −0.8 V and 50 ns, respectively. As can be seen in the figure, device A (green wide line on the top) which represents the majority of devices under test survived this stress test. In contrast, three devices broke down within the first 40 cycles (denoted as B, C, D). The resistance (R_{AP}) of device C (blue) in AP state was already below the nominal R_P value (~2 kΩ) of good devices before this stress test. Thus, this pinhole defect can be easily detected by March tests. However, detecting pinhole defects in devices B and D cannot be guaranteed by March tests at t=0, since these two devices have small pinholes and their initial R_{AP} values are close to the nominal R_{AP} of defect-free devices (e.g., device A). Under pulse stress, the pinhole defects quickly grow up into larger ones leading to a reduction in the resistance of the MTJ devices. Hence, stress test is an effective way to detect devices with small pinhole defects.

It is worth noting that stress tests are prohibitively expensive for high-volume testing. In addition, the amplitude and duration of the hammering write pulse need to be carefully tuned to avoid any inadvertent destruction of good devices while maintaining an acceptable test effectiveness and efficiency.

8

DAT FOR SYNTHETIC ANTI-FERROMAGNET FLIP (SAFF) DEFECTS

Understanding the manufacturing defects in MTJs and their resultant faulty behaviors are crucial for developing high-quality test solutions. This chapter introduces a new type of MTJ defect: synthetic anti-ferromagnet flip (SAFF) defect, wherein the magnetization in both the hard layer and reference layer of MTJ devices undergoes an unintended flip to the opposite direction. Both magnetic and electrical measurement data of SAFF defect in fabricated MTJ devices is presented; it shows that such a defect reverses the polarity of stray field at the free layer of MTJ, while it has no electrical impact on the single isolated device. We demonstrate that using the conventional fault modeling and test approach fails to appropriately model and test such a defect. Therefore device-aware fault modeling and test approach is used. It first physically models the defect and incorporate it into a Verilog-A MTJ compact model, which is afterwards calibrated with silicon data. The model is thereafter used for fault analysis and modeling within an STT-MRAM array; simulation results show that a SAFF defect may lead to an intermittent Passive Neighborhood Pattern Sensitive Fault (PNPSF1$_i$) when all neighboring cells are in logic '1' state. Finally, test solutions for such fault are discussed.

The contents of this chapter have been published in ITC'20 as a distinguished paper [63].

8.1. SAFF DEFECT CHARACTERIZATION

We did comprehensive magnetic and electrical characterization on MTJs with diameters ranging from 35 nm to 175 nm on four wafers. We observed a small fraction of devices across different sizes with *horizontally* flipped R-H loops and *normal* R-V loops. We attribute the root cause to the flip of magnetization in both HL and RL, which we name as *Synthetic Anti-Ferromagnet Flip* (SAFF) defect in this thesis. Next, we will present both magnetic and electrical measurement data of a representative SAFF-defective device as well as a defect-free device for the purpose of comparison. Thereafter, we briefly review the SAFF defect and its potential causes.

8.1.1. MAGNETIC CHARACTERIZATION

Measurement of the R-H hysteresis loop of MTJ device is a useful and fast technique to characterize the device's magnetic properties such as the *coercivity* H_c (defined as the reverse field needed to drive the magnetization of a ferromagnet to zero [189]) and $H^z_{s\_intra}$. In this measurement, a perpendicular magnetic field is applied to the device and swept in the range of +/-3 kOe. We monitor the resistance of the MTJ device at every value of the applied field using a small sense current. At certain threshold field, the magnetization of the FL reverses from its initial direction resulting in an abrupt shift in the resistance of the MTJ (i.e., $R_P \rightarrow R_{AP}$ or $R_{AP} \rightarrow R_P$).

Figure 8.1a shows the measured R-H hysteresis loops (averaged over ten cycles) of a defect-free device (upper) and a defective device (lower), with the same size eCD=55 nm; eCD stands for *electrical Critical Diameter* which is used to describe the MTJ size as a common practice in the MRAM community [32, 199]. Due to the existence of $H^z_{s\_intra}$ at the FL, the *positive* switching field H_{sw_p} and the *negative* switching field H_{sw_n} are asymmetric. The R-H loop of the defect-free device shifts to the right side, which is reflected by the offset field $H_{offset} = \frac{1}{2}(H_{sw\_p} + H_{sw\_n})$ marked in the figure. Therefore, $H^z_{s\_intra} = -H_{offset}$ and $H_c = \frac{1}{2}(H_{sw\_p} - H_{sw\_n})$. In contrast, the defective device shows a

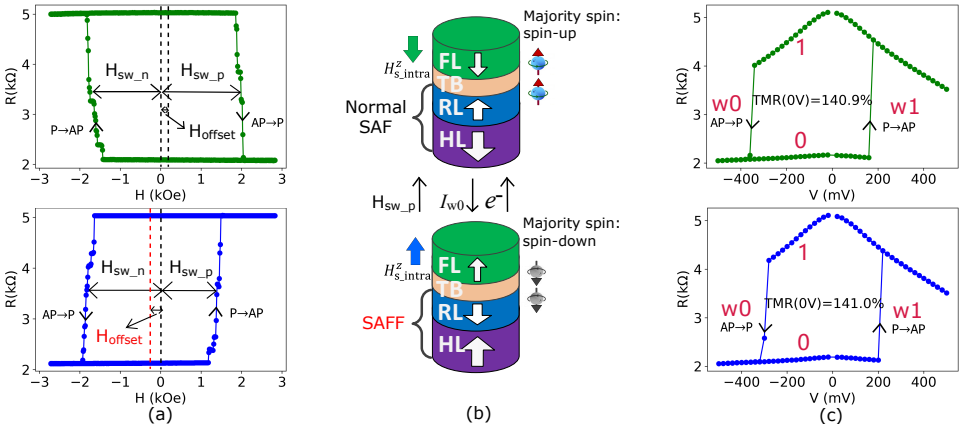


Figure 8.1: Comparison between a defect-free MTJ (upper) and a SAFF-defective MTJ (lower) with the same nominal size eCD=55 nm: (a) R-H loop, (b) schematic of AP state, and (c) R-V loop.

horizontally flipped R-H loop which shifts to the left side rather than the right side. This indicates that the *polarity* of $H^z_{s\_intra}$ *reverses* for the defective device while its *coercivity* H_c is *not influenced*. In addition, the switching direction (i.e., AP→P or P→AP) also flips for a given switching field. For example, a positive field H_{sw_p} induces a P→AP transition for the defective device while it leads to an AP→P transition for the defect-free device, as illustrated in the figure. Based on these observations, it is clear that the magnetization in the RL of the defective device flips to the opposite direction in comparison to the defect-free device, as illustrated with the device schematics in Figure 8.1b. Due to the AFC relation between the RL and HL, the latter also flips to the opposite direction.

8.1.2. ELECTRICAL CHARACTERIZATION

Apart from the R-H loops, we also measured the R-V loops of the same devices; the results are shown in Figure 8.1c. During the measurements, a ramped DC current was applied flowing through the device under test to switch its state. It can be seen that the R-V loop of the SAFF-defective device has the same shape and follows the same switching directions as the defect-free device; their R_P, R_{AP}, and TMR values at 0 V are almost the same. However, one can clearly see there is a marginal shift in the switching voltage, which could be attributed to the intrinsic switching stochasticity and process variations.

The STT-switching mechanisms in both cases can be explained theoretically as follows. Figure 8.1b shows the schematics of a defect-free device (upper) and a SAFF-defective device (lower) with both in AP state. In case of AP→P switching, a write current I_{w0} is applied from the FL to HL; note that electrons flow in the opposite direction from the HL to FL as illustrated in the figure. For the defect-free MTJ device, the RL polarizes the incoming electrons to align with its magnetization direction, making spin-up the majority spin. Once the spin-up electrons tunnel through the MgO-based TB, they exert a torque on the FL, thereby switching its magnetization to the opposite direction. For the SAFF-defective device, spin-down becomes the majority spin, as the magnetization in the RL (spin polarizer) flips with the HL. Therefore, it is the majority spin which switches the magnetization of the FL in both cases. This indicates that the critical switching current I_c would not change if the magnetizations in the RL and HL flipped for a single MTJ device. More details about the STT-switching principle can be found in [26].

8.1.3. SAFF DEFECT MECHANISM AND POTENTIAL CAUSES

Given the strong anti-ferromagnetic coupling strength between the HL and RL (>10 kOe measured by vibrating sample magnetometer (VSM) [26]) for our devices and the RL flip observed in Figure 8.1a, a probable cause of the SAFF defect is an initial HL reversal. Due to inhomogeneities arising during device fabrication steps, HL with significantly reduced H_c may exist in certain outlier devices. Based on the measurement results presented previously, the SAFF defect has no impact on the switching current direction. However, the polarity of $H^z_{s\_intra}$ is reversed by the defect, compared to defect-free devices. This may affect the way the SAFF-defective MTJ manifests itself at the functional level in an STT-MRAM array. Hence, modeling the SAFF defect and analyzing its impact at the behavior level is a must in order to develop appropriate test solutions if needed.

8.2. LIMITATIONS OF THE CONVENTIONAL TEST APPROACH

As already mentioned in the previous chapters, the conventional test approach models any defect in a MTJ device as a linear resistor either in parallel to or in series with a defect-free MTJ model. Applying the conventional fault modeling approach to the SAFF defect results in four FPs: iR1NF1 = ⟨1r1/1/0⟩, iR0NF0 = ⟨0r0/0/1⟩, W1TF0 = ⟨0w1/0/–⟩, W0TF1 = ⟨1w0/1/–⟩, as shown in Table 8.1. These four FPs can be used to generate test solutions such as March algorithms. First, each sensitized FP is assigned its own detection condition. For instance, iR0NF0 requires a read operation on the faulty cell at state '0' to guarantee its detection, denoted as ⇳(...0,r0,...), where ⇳ means that the detection condition does not depend on the addressing direction. The detection condition for W0TF1 is ⇳(...1,w0,r0,...), meaning that a down-transition write followed by a read is enough to detect this fault, regardless of the addressing direction. The detection conditions of all sensitized FPs are compiled into the following optimal March test with three march elements:

$$\{⇳(w0); ⇑(w1,r1); ⇓(w0,r0)\}.$$

Note that different versions of March tests can be generated (e.g., with two march elements) as long as the test satisfies all the detection conditions.

Table 8.1: Static fault modeling results of SAFF defect using resistive models.

Defect model	Resistance (Ω)	Sensitized FP	Fault Model &FP Name	Detection Condition
Series resistor R_{sd}	(466, 870]	⟨0r0/0/1⟩	incorrect Read Non-destructive Fault: iR0NF0	⇳(...0,r0,...)
	(870, 1.6k]	⟨0r0/0/1⟩	incorrect Read Non-destructive Fault: iR0NF0	⇳(...0,r0,...)
		⟨1w0/1/–⟩	Write Transition Fault: W0TF1	⇳(...1,w0,r0,...)
	(1.6k, +∞]	⟨0r0/0/1⟩	incorrect Read Non-destructive Fault: iR0NF0	⇳(...0,r0,...)
		⟨1w0/1/–⟩	Write Transition Fault: W0TF1	⇳(...1,w0,r0,...)
		⟨0w1/0/–⟩	Write Transition Fault: W1TF0	⇳(...0,w1,r1,...)
Parallel resistor R_{pd}	[0, 1.1k)	⟨1r1/1/0⟩	incorrect Read Non-destructive Fault: iR1NF1	⇳(...1,r1,...)
		⟨1w0/1/–⟩	Write Transition Fault: W0TF1	⇳(...1,w0,r0,...)
		⟨0w1/0/–⟩	Write Transition Fault: W1TF0	⇳(...0,w1,r1,...)
	[1.1k, 3.1k)	⟨1r1/1/0⟩	incorrect Read Non-destructive Fault: iR1NF1	⇳(...1,r1,...)
		⟨1w0/1/–⟩	Write Transition Fault: W0TF1	⇳(...1,w0,r0,...)

We verified the effectiveness of the generated March algorithm on our fabricated devices. However, we observed that the test is not able to distinguish the SAFF-defective MTJs from defect-free ones. This conclusion can also be drawn by comparing the two R-V loops in Figure 8.1c. In both defect-free and defective cases, the MTJ devices were initialized to state '0' with w0 operations. The loop starts with an up-transition (w1) operation followed by a down-transition (w0) operation. All the points in the two R-V loops are readout resistance (r0 or r1) under a voltage of 20 mV. It is clear that these two loops have the same shape and switching directions.

The above suggests that the generated FPs using the conventional fault modeling approach (and covered by our test) are *not qualified* to describe the actual faulty behavior of an STT-MRAM cell with the SAFF defect. As these FPs are derived by circuit simulations with the injection of resistive models, we can infer that the SAFF defect cannot be sim-

ply modeled as a linear resistor. As explained in Section 8.1.3, the main change induced by the SAFF defect is that the polarity of the stray field at the FL is reversed. To capture the changes in the MTJ's magnetic properties, we need a more sophisticated defect modeling approach in replacement of the conventional resistor-based defect modeling approach.

8.3. DEVICE-AWARE DEFECT MODELING FOR SAFF

As an alternative to the conventional test approach, we will apply our Device-Aware Test (DAT) approach to the SAFF defect in the remainder of this paper. The DAT approach consists of three steps as follows.

- ***Device-aware defect modeling.*** Instead of modeling manufacturing defects in MTJs as linear resistors, the DAT approach integrates the defect effects into MTJ device model. This is achieved by first identifying and modifying the affected technology parameters of MTJ; thereafter, the impact is mapped into device's electrical parameters.
- ***Device-aware fault modeling.*** This step defines the complete fault space for STT-MRAMs by expanding the conventional fault primitive notation. Subsequently, a systematic fault analysis based on circuit simulations is performed to validate realistic faults in the space in the presence of a device defect.
- ***Device-aware test development.*** The obtained accurate and realistic faults from the previous step are utilized to develop high-quality test solutions.

In this section, we will work out the first step for the SAFF defect. It consists of three sub-steps: 1) physical defect modeling, 2) electrical modeling of defective device, 3) fitting and model optimization. Next, we will follow these three sub-steps to develop a physics-based model for the SAFF defect. To this end, we first model the impact of SAFF defect on the overall stray field H^z_{stray} (including both intra- and inter-cell stray fields) at the FL of the defective cell within a memory array; the rest of technology parameters in Table 3.1 are not impacted as suggested in Section 8.1. Thereafter, its impact is mapped to MTJ's electrical parameters I_c and t_w; the MTJ resistance is not influenced by this defect. Finally, we calibrate the SAFF-defective MTJ compact model with silicon data.

8.3.1. PHYSICAL DEFECT ANALYSIS AND MODELING

INTRA-CELL STRAY FIELD MODELING

$H^z_{\text{s\_intra}}$ in a single MTJ device can be physically modeled based on the bound current theory and Biot-Savart law [198, 217]. For a thin ferromagnet (i.e., HL, RL, or FL), the generated field is identical to the field that would be produced by the bound current I_b [198], under the assumption that it is uniformly magnetized. I_b is a macroscopic current flowing around the boundary of the ferromagnet, as all internal molecular current loops cancel each other out while those at the edge are left uncanceled, as illustrated in Figure 5.4a. The magnetic moment m of the ferromagnet is $m{=}I_b{\cdot}A$, where A is the cross-sectional area. In addition, m can also be expressed as in [198]: $m{=}M_s{\cdot}A{\cdot}t$, where M_s is the saturation magnetization and t is the thickness of this ferromagnet. Therefore,

one can easily derive $I_b = M_s \cdot t$. Here, the $M_s \cdot t$ product can be measured at blanket film level by VSM measurements [26].

With the derived bound current I_b for each ferromagnet in the MTJ stack, the generated stray field at any point in the space can be calculated by the Biot-Savart law [217], as shown in Figure 5.4b. In this way, we can calculate the out-of-plane component of the stray field at the FL from both HL ($H^z_{s\_HL}$) and RL ($H^z_{s\_RL}$). Thus, the net intra-cell stray field: $H^z_{s\_intra} = H^z_{s\_HL} + H^z_{s\_RL}$.

INTER-CELL STRAY FIELD MODELING

In addition to the intra-cell stray field from the device itself, all neighboring cells also produce stray fields acting on each other in a memory array. The magnitude of the inter-cell stray field $H^z_{s\_inter}$ depends on device size as well as array pitch [197, 217]. Therefore, it is crucial to model and take into account $H^z_{s\_inter}$ especially when it comes to high-density STT-MRAM arrays at advanced technology nodes. To the end, we built up two 3×3 memory arrays (defect-free vs. defective) in Cartesian Coordinates to calculate $H^z_{s\_inter}$ at the FL of the central cell from all the eight neighboring cells. Figure 8.2a shows a memory array consisting of nine defect-free MTJ devices, while Figure 8.2c shows an array composed of eight defect-free devices (C0-C7) and a SAFF-defective device (C8) in the center. In both cases, C0-C3 are considered as *direct* neighbors with the same distance to C8; each of them produces an inter-cell stray field H_{dir} acting on C8 as illustrated in the figure. Similarly, C4-C7 are in symmetric *diagonal* positions; each of them exerts a

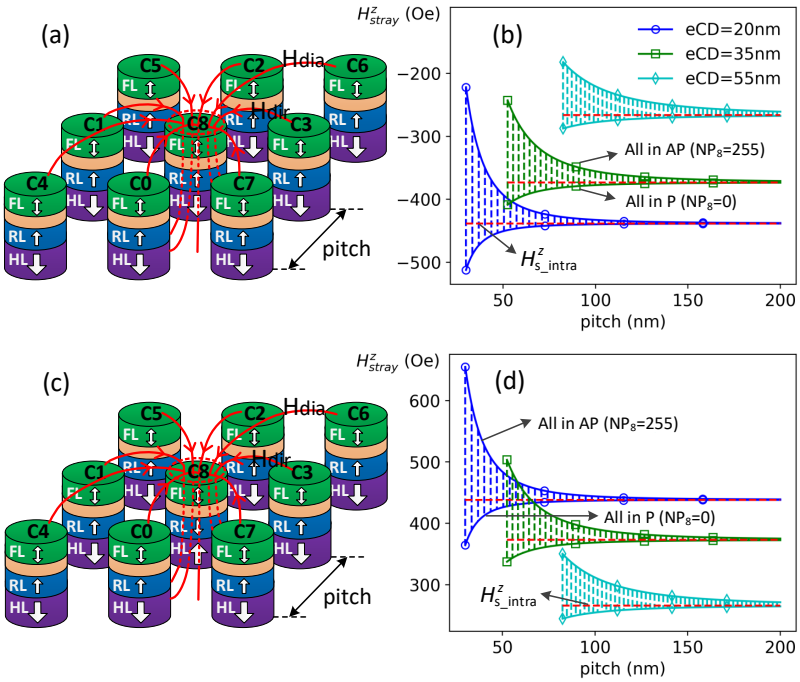


Figure 8.2: (a) 3×3 array of all defect-free MTJs, (b) the overall out-of-plane stray field H^z_{stray} at the FL of the defect-free cell C8, (c) 3×3 array of eight defect-free MTJs (C0-C7) and a SAFF-defective cell C8 in the center, and (d) H^z_{stray} at the FL of the SAFF-defective cell C8.

field H_{dia} on C8. With the previously introduced stray field modeling approach, we can also calculate $H^z_{s\_inter}$ at the FL of victim cell C8 from C0-C7 as follows:

$$H^z_{s\_inter} = \sum_{i=0}^{7} (H^z_{s\_HL}(Ci) + H^z_{s\_RL}(Ci) + H^z_{s\_FL}(Ci)). \tag{8.1}$$

For each cell, both the polarity and magnitude of $H^z_{s\_HL}$ and $H^z_{s\_RL}$ are fixed for a given design (i.e., device size and array pitch). However, the polarity of $H^z_{s\_FL}$ changes dynamically depending on the data stored in the MTJ device although its magnitude remains the same. As a result, $H^z_{s\_inter}$ depends on the *Neighborhood Pattern* (NP) in the eight neighboring cells C0-C7, denoted as NP_8. In the binary form, NP_8 can be expressed as: $[d_0,...,d_7]_2$, where $d_i \in \{0,1\}$ represents the data stored in cell Ci. NP_8 can also be denoted in the decimal form: $[n]_{10}$, where $n \in [0,255]$.

OVERALL STRAY FIELD

Figure 8.2b shows the overall stray field ($H^z_{stray} = H^z_{s\_intra} + H^z_{s\_inter}$) at the FL of the defect-free cell C8 for the configuration of Figure 8.2a at varying pitches with respect to three different eCD values representing device sizes. In our simulations, we set the minimum pitch to 1.5×eCD according to [156] for high-density STT-MRAMs and the maximum pitch to 200 nm which is adopted by Intel [30]. The shaded areas indicate all possible H^z_{stray} values depending on the NP_8 in C0-C7; the uppermost curve of each shaded area represents NP_8=255 (all in AP state), while the lowermost curve represents NP_8=0 (all in P state). It can also be seen that the magnitude of H^z_{stray} increases as eCD decreases (i.e., smaller MTJs) and the variation range of H^z_{stray} increases as the pitch goes down (i.e., MTJs become closer to each other). The red dotted lines mark $H^z_{s\_intra}$ for isolated devices.

In contrast, Figure 8.2d shows H^z_{stray} at the FL of the SAFF-defective cell C8 in the configuration of Figure 8.2c. It can be seen that the SAFF-defective cell experiences a *positive stray field* rather than a negative one in the defect-free case. In absolute number, H^z_{stray} in the presence of SAFF defect is much larger than that of the defect-free case, especially for smaller pitches; e.g., for eCD=20 nm at pitch=30 nm, H^z_{stray} increases by up to 70%. Furthermore, the magnitude of H^z_{stray} reaches the peak when NP_8=255 in the defective case, whereas the maximum H^z_{stray} occurs when NP_8=0 in defect-free case.

8.3.2. ELECTRICAL MODELING OF SAFF-DEFECTIVE MTJ DEVICES

With the obtained physics-based model of H^z_{stray}, we can map the SAFF-induced change in H^z_{stray} to the two key electrical parameters: I_c and t_w. Under the influence of stray field H^z_{stray}, I_c can be expressed as follows [75]:

$$I_c(H^z_{stray}) = \frac{1}{\eta}\frac{2\alpha e}{\hbar} M_s \cdot V \cdot H_k \cdot (1 + T \cdot \frac{H^z_{stray}}{H_k}), \tag{8.2}$$

$$T = (-1)^{j+l}, \quad j,l \in \{0,1\}. \tag{8.3}$$

In Equation (8.2), η is the STT efficiency, α the magnetic damping constant, e the elementary charge, $\hbar$ the reduced Planck constant, M_s the saturation magnetization, V the

volume of the FL, H_k the magnetic anisotropy field. We added the term T (see Equation (8.3)) to identify the switching direction for both defect-free and defective devices; j=1(0) indicates a defective (defect-free) MTJ device. In addition, l=1(0) represents an AP→P (P→AP) switching direction. Consequently, one can derive I_c(AP→P)>I_c(P→AP) in both defect-free and defective cases, which is consistent with the experimental results and theoretical analysis in Section 8.1.2. Note that the magnitude of I_c(AP→P) (or I_c(P→AP)) in the defective case differs from that in the defect-free case, since the H^z_{stray} magnitudes in the two cases are not same for a given eCD, pitch, and NP_8, as shown in Figure 8.2b and Figure 8.2d.

Furthermore, the switching time t_w in the precessional regime (namely, switched by the STT-effect) can be estimated using the Sun's model as follows [62]:

$$\mu(t_w) = \left(\frac{2}{C+\ln(\frac{\pi^2\Delta}{4})} \cdot \frac{\mu_B P}{e\cdot m\cdot(1+P^2)} \cdot I_d\right)^{-1}, \tag{8.4}$$

$$I_d = \frac{V_p}{R(V_p)} - I_c(H^z_{stray}), \tag{8.5}$$

$$t_w \sim \mathcal{N}(\mu(t_w), \sigma(t_w)^2). \tag{8.6}$$

Here, C≈0.577 is Euler's constant, Δ the thermal stability factor, μ_B the Bohr magneton, P the spin polarization, and m the FL magnetic moment. V_p is the voltage applied on the MTJ device to switch its state. $R(V_p)$ is the resistance of the MTJ device; it shows a non-linear dependence on V_p [62]. In addition, we assume that t_w obeys a normal distribution for a given V_p (Equation 8.6).

8.3.3. FITTING AND MODEL OPTIMIZATION

Finally, the obtained electrical model of SAFF-defective MTJ device (Equations 8.2–8.6) has to be calibrated with silicon data. To this end, we performed comprehensive pulsed-switching characterization on the identified SAFF-defective MTJ devices at IMEC. In the measurements, the pulse width t_p was swept from 5 ns to 40 ns; these t_p values represent the typical write speed for STT-MRAM designs in practice. The interval of pulse amplitude V_p at each t_p point was carefully tuned to cover the entire switching spectrum, namely *switching probability* P_{sw} from 0% to 100%, as the switching events are intrinsically stochastic. To obtain a statistical result of the stochastic switching characteristics with an acceptable accuracy, we applied 1k-cycle pulses for each combination of V_p and t_p. For instance, we observed that the number of successful P→AP switching events is 63 out of 1k pulses at V_p = 0.4 V, t_p = 10 ns, leading to P_{sw} = 6.3%. As V_p increases to 0.5 V at the same t_p, 885 successful switching events were observed, resulting in P_{sw} = 88.5%. In this way, we obtained the three-dimensional statistics of P_{sw} vs. V_p vs. t_p for both P→AP and AP→P switching directions.

Figure 8.3 shows the measured V_p vs. t_p at switching probability P_{sw}=0.16, 0.50 and 0.84 for a SAFF-defective MTJ with eCD=35 nm. These three P_{sw} values are the outputs of the cumulative distribution function $F(\mu-\sigma)$, $F(\mu)$, and $F(\mu+\sigma)$ of normal distribution, respectively. The two curves with P_{sw}=$F(\mu)$=0.50 in Figure 8.3 are used to calibrate $\mu(t_w)$ (see Equations 8.4–8.5). By carefully tuning some physical parameters such M_s and H_k, we are able to fit our device model to the measurement data. Figure 8.4a shows the final

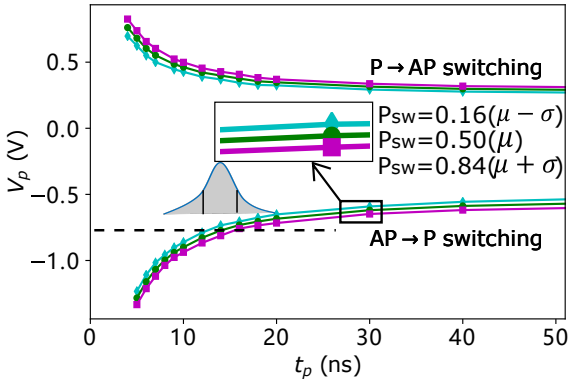


Figure 8.3: Measured pulse amplitude V_p vs. pulse width t_p in determining the switching behavior at switching probability P_{sw}=0.16, 0.50, 0.84 for a SAFF-defective MTJ with eCD=35 nm

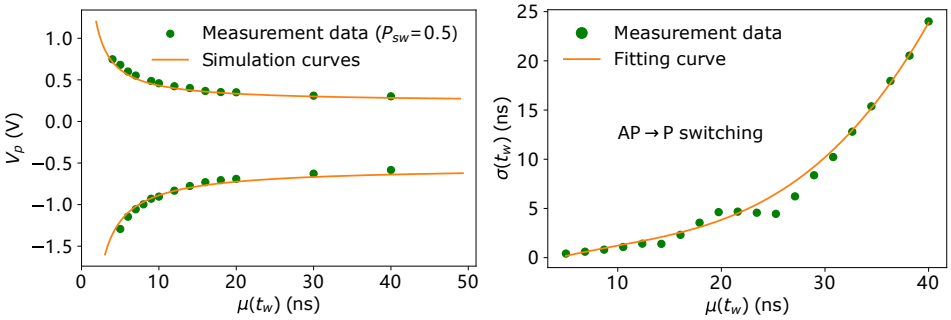


Figure 8.4: (a) Simulation results vs. measurement data at P_{sw}=0.50, (b) the extracted mean $\mu(t_w)$ and standard deviation $\sigma(t_w)$ of the AP→P switching time at different V_p.

fitting results; it can be seen that our simulation results match the silicon data very well. In addition, the measurement data in Figure 8.3 also allows us to extract the standard deviation $\sigma(t_w)$ for a given V_p, which is marked with the dashed line in the figure. Figure 8.4b shows the extracted data for $\sigma(t_w)$ vs. $\mu(t_w)$ as well as the fitting curve with a three-degree polynomial for the AP→P switching direction. The data corresponding to the other switching direction is similar, thus not presented due to space limitation.

The output of device-aware defect modeling is a calibrated Verilog-A SAFF-defective MTJ compact model. After verifying and calibrating the MTJ model in Python as presented previously, we moved this model to Verilog-A so as to make it compatible with analog circuit simulations for subsequent fault modeling. To integrate the inter-cell magnetic coupling effect, we added four ports to the Verilog-A MTJ model: H_{dir_in}[0:3], H_{dia_in}[0:3], H_{dir_out}, and H_{dia_out}; H_{dir_in}[0:3] are input inter-cell stray fields from the four direct neighbors C0-C3 while H_{dia_in}[0:3] are input inter-cell stray fields from the other four diagonal neighbors C4-C7 (see Figure 8.4b). H_{dir_out} and H_{dia_out} are the output stray fields from C8 itself; they go to direct neighbors and diagonal neighbors of C8, respectively. This enables us to simulate the SAFF-defective MTJ device in the presence of magnetic coupling effect in a circuit simulator.

8.4. DEVICE-AWARE FAULT MODELING FOR SAFF

In this section, we apply the device-aware fault modeling to obtain realistic and accurate fault models for the SAFF defect. It consists of two steps: 1) fault space definition, 2) fault analysis. The fault space has already been defined with the fault primitive $\langle S/F_n/R\rangle$, presented in Section 6.4. For the fault analysis we used the same experimental set-up as that in the previous chapters but with some modifications as follows. First, for defect injection we replaced the defect-free MTJ model in the victim cell C8 with our SAFF-defective MTJ compact model, as shown in Figure 8.5. As the SAFF defect does not affect the magnitude of the magnetizations of the RL and HL (only their directions are flipped), the SAFF defect size or strength plays no role here.

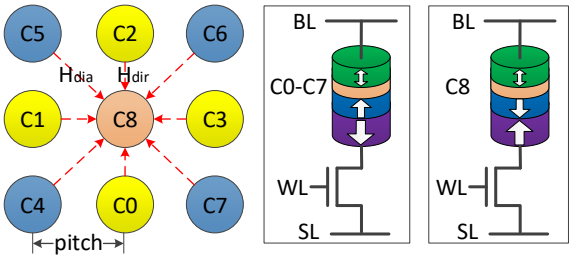


Figure 8.5: SAFF defect injection in the DAT approach.

Second, two array pitches were selected: 200 nm [30], 52.5 nm (=1.5×eCD) [156], representing high-performance and high-density STT-MRAM designs, respectively. Third, each sensitizing sequence S was simulated 10k cycles using Monte Carlo simulations, as the MTJ model has the stochastic switching property (see Figure 8.3c and Equations 8.4–8.6).

Simulation results reveal interesting observations. For pitch=200 nm, no faults were observed in the presence of the SAFF defect; no single-cell, no two-cell, neither nine-cell faults. This clearly indicates that the inter-cell magnetic coupling and SAFF defect effects are negligible at this pitch.

For pitch=52.5 nm the results show some interesting fault behaviors in some cases. No single-cell and two-cell faults were observed at all. However, C8 failed to undergo a 0w1 transition in 1150 cycles out of the simulated 10k cycles, when all neighborhood cells were in state '1' (i.e., NP_8=255). This corresponds to an occurrence rate of 11.5%. Although the observed fault looks like the known fault model: *Passive Neighborhood Pattern Sensitive Fault* (PNPSF) for DRAMs [36], its nature is different; the fault is intermittent rather than permanent, due to the STT-switching stochasticity. Thus, we refer to the observed fault as *intermittent* PNPSF, denoted as PNPSF1_i=$\langle 1;1;1;1;1;1;1;1;1w0/1_\text{i}/-\rangle$. As this fault is a type of hard-to-detect fault [67], testing it is not quite easy!

Figure 8.6a compares the switching time t_{1w0} histograms for 10k-cycle 1w0 operations in defect-free (blue) and defective (yellow) cases; the write pulse width is set long enough to cover the 3σ corner, as demarcated with the vertical dotted line in the figure. However, due to the SAFF defect, the t_{1w0} histogram shifts towards the right side. This means that PNPSF1_i takes place in those cycles where the required switching time

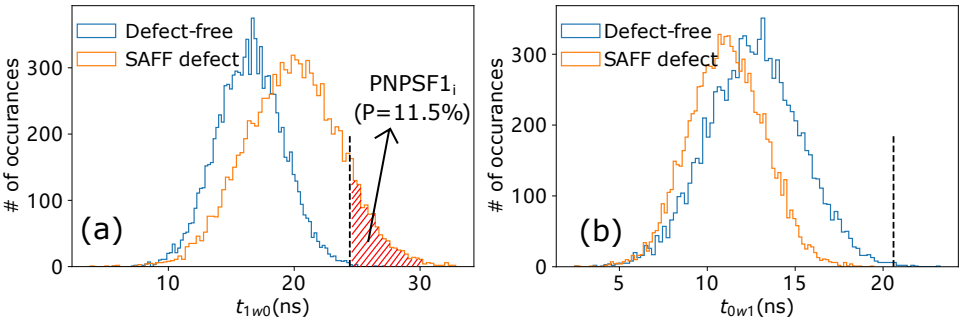


Figure 8.6: PNPSF1$_i$ with an occurrence rate of 11.5% for write '0' operations.

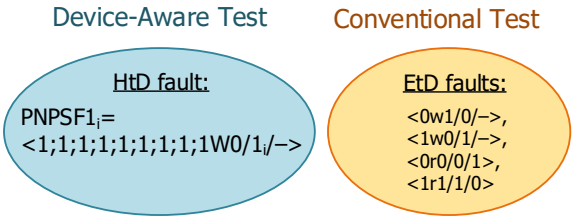


Figure 8.7: Comparison of sensitized FPs due to SAFF defect: device-aware test vs. conventional test approach based on linear resistors.

is larger than the applied write pulse width. It is worth noting that a higher write voltage reduces the shift, thus slightly alleviating the faulty effect of this defect. But it comes with the cost of more power consumption, smaller endurance, and more susceptible to back-hopping effect. In addition, this fault becomes worse as the MTJ device scales down, due to the increased stray fields in smaller MTJs at narrower pitches [65]. Figure 8.6b shows the results for 0w1 operations; the t_{0w1} histogram of SAFF-defective device shifts towards the left side, indicating a faster transition on average in comparison to the defect-free device. Therefore, no faults were observed for 0w1 operations.

Figure 8.7 compares the fault modeling results using our DAT approach and the conventional test approach based on linear resistor injection. It can be seen that the SAFF defect results in a HtD fault (PNPSF1$_i$) using our DAT approach. This cannot be obtained by the conventional fault modeling approach where a linear resistor is injected in parallel with or in series with an ideal defect-free MTJ device model. In contrast, the conventional approach results in four EtD faults, as shown in the figure. This indicates that these four faults are not qualified to cover the SAFF defect in STT-MRAMs. Accordingly, the March tests targeting these four faults obviously cannot guarantee the defection of the SAFF defect.

8.5. DEVICE-AWARE TEST DEVELOPMENT FOR SAFF

The last step of DAT is to develop appropriate test solutions for the derived fault PNPSF1$_i$. Next, two test solutions will be discussed. One straightforward test solution could be a March algorithm such as:

$$\{\Updownarrow (w1); \Updownarrow (w0, r0, w1)^n\}.$$

In the above algorithm, n ($n \in \mathbb{Z}^+$) denotes the number of times that the second march element should be repeated. The first march element ⇕(w1) initializes all memory cells to state '1', while the second applies three operations: w0 to sensitize the fault, r0 to *probabilistically* detect it, and w1 to reset the cell back to state '1'. As our experiments showed that $PNPSF1_i$ occurs with a probability of 11.5%, when NP_8 is 255; it is a random process, indepenent on the previous operations. With a repetition of n times, the detection probability $P_{dt}=1-(1-11.5\%)^n$; hence the higher n, the higher P_{dt}. E.g., P_{dt}=90% requires n=19, while P_{dt}=99.99% requires n=76. Clearly, getting high confidence in the detection comes at the cost of long test time (large n); 100% detection is hard to guarantee.

The second test solution aims at guaranteeing the detection by incorporating *magnetic* write operations in the March test:

$$\{⇕(w0_H); ⇕(r0)\} \text{ or } \{⇕(w1_H); ⇕(r1)\}.$$

Here, the first element $w0_H$ ($w1_H$) indicates a magnetic write '0' ('1') operation; i.e., an *external* field H_{ext} is applied to switch the MTJ state rather than driving an electric current through the MTJ device. Note that H_{ext} should be set higher than the coercivity of the FL but smaller than that of the RL and HL (i.e., $H_c(FL) < H_{ext} < H_c(RL) < H_c(HL)$) to avoid switching of the RL and HL. This can be realized by adding a perpendicular magnetic generator to a test chamber, similar to the wafer-level magnetic characterization tool developed by Hprobe [218]. As an entire STT-MRAM chip or even multiple chips can be reset to certain state by an external field in one shot, the additional cost due to this handling is limited. Figure 8.8 illustrates the test process with a $w0_H$ operation to guarantee the detection of SAFF defect. Irrespective of the initial state, a $w0_H$ operation sets the magnetization of the FL to the same direction as the field H_{ext}. This makes the defect-free MTJ stay in P(0) state, while the SAFF-defective MTJ goes to AP(1) state, as shown in the figure. Thereafter, a r0 operation can easily distinguish defective devices from defect-free ones.

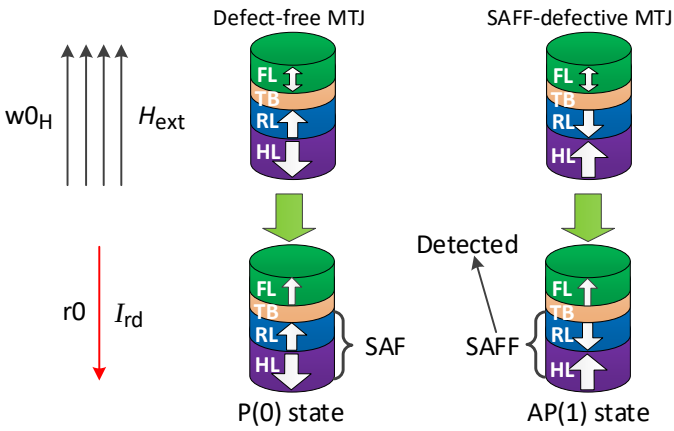


Figure 8.8: Testing SAFF defects using a magnetic write '0' operation ($w0_H$).

9

DAT FOR INTERMEDIATE (IM) STATE DEFECTS

Understanding the manufacturing defects in MTJs and their resultant faulty behaviors are paramount for developing high-quality test solutions. This chapter characterizes, models, and tests intermediate (IM) state defect in MTJ devices based on silicon measurements and circuit simulations. Base on the comprehensive characterization of MTJ devices with diameter ranging from 60 nm *to* 120 nm*, we observe that this defect manifests itself as a third resistive state with a certain occurrence probability depending on the switching direction, device size, and bias voltage, in addition to the normal bi-stable states. We demonstrates that using the conventional fault modeling approach fails to derive appropriate fault models for this defect. Therefore, device-aware test (DAT) is used. We first physically models the defect and incorporate it into a Verilog-A MTJ compact model, which is calibrated with measured silicon data. Thereafter, this model is used for a systematic fault analysis based on circuit simulations to validate accurate and realistic fault models in a pre-defined fault space. Our simulation results show that the IM state defect leads to intermittent write transition faults: W1TFU$_i$ and W0TFU$_i$. Finally, we present a device-aware test solution based on weak write operations specifically targeting this defect.*

Parts of this chapter have been accepted by DATE'21 as a best paper award candidate [64].

9.1. IM STATE DEFECT MECHANISM

Normally, MTJ devices only have two bi-stable magnetic states: parallel (P) and anti-parallel (AP) states. P state exhibits relatively low resistance (R_P) and AP state exhibits high resistance (R_{AP}). These two distinct states represent one bit of data in an STT-MRAM cell. However, the fabrication and integration process of MTJ devices is vulnerable to several defects, as introduced in Section 3.4. Among these manufacturing defects, *intermediate* (IM) state defect is considered as a critical type. An IM state manifests itself as a third resistive state between R_P and R_{AP}, leading to unintended memory faulty behaviors.

There are several prior works on studying IM states in MTJ devices based on experiments and/or simulations, as listed in Table 9.1. Yao *et al.* [219] observed stable IM states in both P→AP and AP→P switching directions after the removal of write pulses; the read pulse width is 200 ms, indicating that the retention time of IM state (RT_{IM}) is at least 200 ms. They attributed the physical causes of IM state to the multi-structure of the FL induced by the dipole field and large device size. Aoki *et al.* [220] also observed IM states during STT-switching with sub-10ns pulses and claimed that those IM states are metastable meaning that they disappear after the removal of write pulses; the claimed physical cause is similar to the above one. Subsequently, more research works [79, 132, 221] were conducted and reported that the observed IM states are metastable due to the inhomogeneous distribution of stray field at the FL and unreversed magnetic bubbles, as elaborated in the table. In recent two years, studies in [133, 222] on IM states reveal that IM states in MTJ devices take place due to Skymion formation and their retention time can be as long as the bi-stable P and AP states.

Table 9.1: Related work on IM state in MTJ devices in the literature.

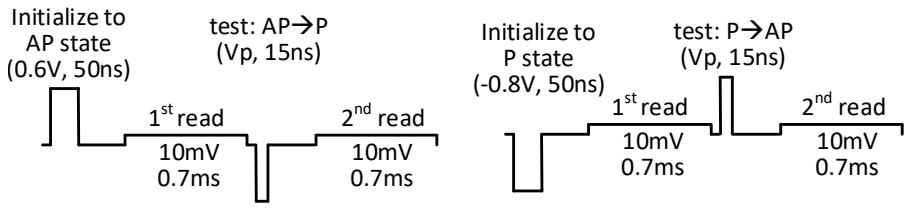
Institute	Method	Stability & Retention	Claimed Physical Cause
Minnesota Unv. (2008) [219]	Experiments	Stable, RT_{IM}>200ms	Multi-domain structure of the FL induced by the dipole field and large device size
Tohoku Univ. (2010) [220]	Experiments	Metastable, RT_{IM} =?	Inhomogeneous magnetization behavior induced by multi-domain and/or vortex creation
NYU&STT Inc. (2016) [221]	Experiments	Metastable, RT_{IM}= 1us	Inhomogeneous distribution of stray field at the FL from SAF layers
CNRS (2016) [132]	Experiments	Metastable, RT_{IM} =?	Unreversed magnetic bubble forms during the switching process
Intel Corp. (2018) [79]	Experiments	Metastable, RT_{IM} =?	Inhomogeneous distribution of stray field at the FL from SAF layers
Beihang Univ. (2018) [133]	Simulations	Stable, RT_{IM}= RT_P/RT_{AP}	Skyrmion formation due to non-uniformity of stray field and DMI effect
UCLA (2019) [222]	Experiments +Simulations	Stable, RT_{IM}= RT_P/RT_{AP}	Skyrmions formed in MTJs without the DMIs

9.2. IM STATE DEFECT CHARACTERIZATION

Electrical characterization with pulses is a common practice to evaluate the write performance of STT-MRAM devices. When we performed comprehensive characterization on devices with CD ranging from 60 nm to 120 nm, some devices showed IM states with resistance values between R_P and R_{AP}. In this section, we first introduce the experimental set-up for measuring the IM state. Thereafter, the measured results of an MTJ device without IM state and an MTJ device with IM state are presented and compared. Then, we elaborate the dependence of IM state occurrence probability on bias voltage, device size, and switching direction. Finally, we briefly review the related work in the literature and discuss the potential causes of IM state.

9.2.1. MEASUREMENT SET-UP

Figure 9.1a and 9.1b show the pulse configurations in each cycle for AP→P and P→AP switching characterization, respectively. For AP→P switching characterization, a positive voltage pulse (V_p=0.6 V, t_p=50 ns) was applied to the MTJ device under test to initialize it to AP state, as illustrated in Figure 9.1a. The pulse was followed by a read operation using a relatively long but small voltage pulse (V_p=10 mV, t_p=0.7 ms) to check whether the device has been initialized to AP state successfully. After the read, a negative pulse with t_p=15 ns was applied to the device to study AP→P switching. Similarly, a second read was applied to read out the resistive state of the device. As the switching behavior is intrinsically stochastic, we repeated these four operations for 10k cycles to obtain a statistical result. To cover the switching probability P_{sw} from 0% to 100%, we swept the pulse amplitude V_p of the second pulse in a carefully-tuned range. For P→AP switching characterization, a similar measurement was conducted with the polarity of both write pulses reversed, as shown in Figure 9.1b.



(a) AP→P switching characterization.

(b) P→AP switching characterization.

Figure 9.1: Pulse configuration in each cycle.

9.2.2. IDENTIFICATION OF IM STATE DEFECTS

Figure 9.2a and 9.2b show the measured results of a representative normal MTJ A (nominal CD=100 nm) for AP→P switching and P→AP switching, respectively; each point represents a readout resistance of the second read pulse in Figure 9.1. It can be seen that when V_p=−0.74 V, AP→P switching probability is 100% in the measured 10k cycles. When V_p=0.45 V, P→AP switching probability is 99.2%, meaning that 0.8% of the 10k cycles experience failed transitions (marked with red triangles), due to the STT-switching stochasticity. In both cases, there is no third resistive state observed. In contrast, Figure 9.2c and 9.2d show the measurement data of a typical device with IM state (MTJ B) with the same

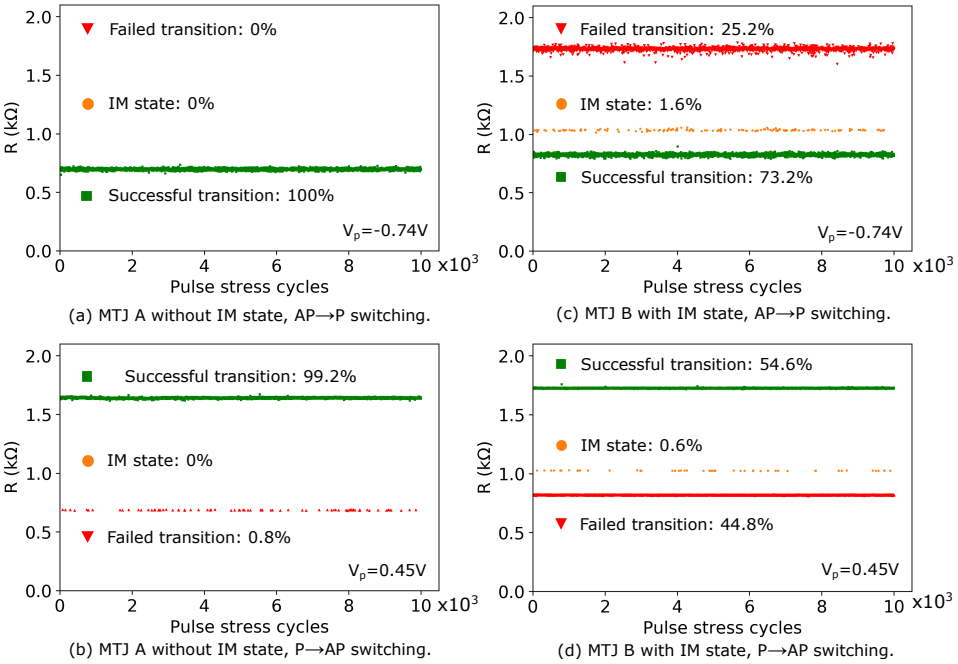


Figure 9.2: Measurement results: MTJ A without IM state (left) vs. MTJ B with IM state (right).

size and experimental conditions. It is clear that a line of unexpected orange points (i.e., IM state) show up between the two lines representing AP and P states. The occurrence probability of IM state in AP→P switching direction is 1.6% when V_p=−0.74 V while it is 0.6% in the opposite switching direction when V_p=0.45 V. It is also worth noting that the probability of failed transition of MTJ B is much higher than that of MTJ A under the same applied pulses. The disparity of R_P (red lines) and R_{AP} (green lines) between these two devices is attributed to process variations; the slight TMR drop in this defective MTJ was not a common rule in all measured defective MTJs, compared to good MTJs.

9

9.2.3. DEPENDENCE OF IM STATE DEFECTS

We observed that the occurrence of IM state significantly depends on the applied bias voltage, switching direction (i.e., AP→P or P→AP), and device size in our experiments. Figure 9.3a and Figure 9.3b show the bias voltage dependence of IM state of four different MTJ devices in AP→P and P→AP switching directions, respectively; the nominal CD of MTJ C and D is 100 nm while it is 120 nm for MTJ E and F. It can be seen that the *successful transition probability* (P_{ST}) between P and AP states (marked with green square points corresponding to the left y-axis) increases from 0% to 100%, as the amplitude of V_p increases in both switching directions. The orange circle points represent the *occurrence probability of IM state* (P_{IM}) corresponding to the right y-axis at various V_p points; the V_p measurement points for AP→P switching are from −0.8 V to −0.6 V in a step of 0.02 V, whereas the V_p points for P→AP switching are from 0.35 V to 0.55 V in a step of 0.02 V. One can observe that P_{IM} increases with the amplitude of V_p until reach-

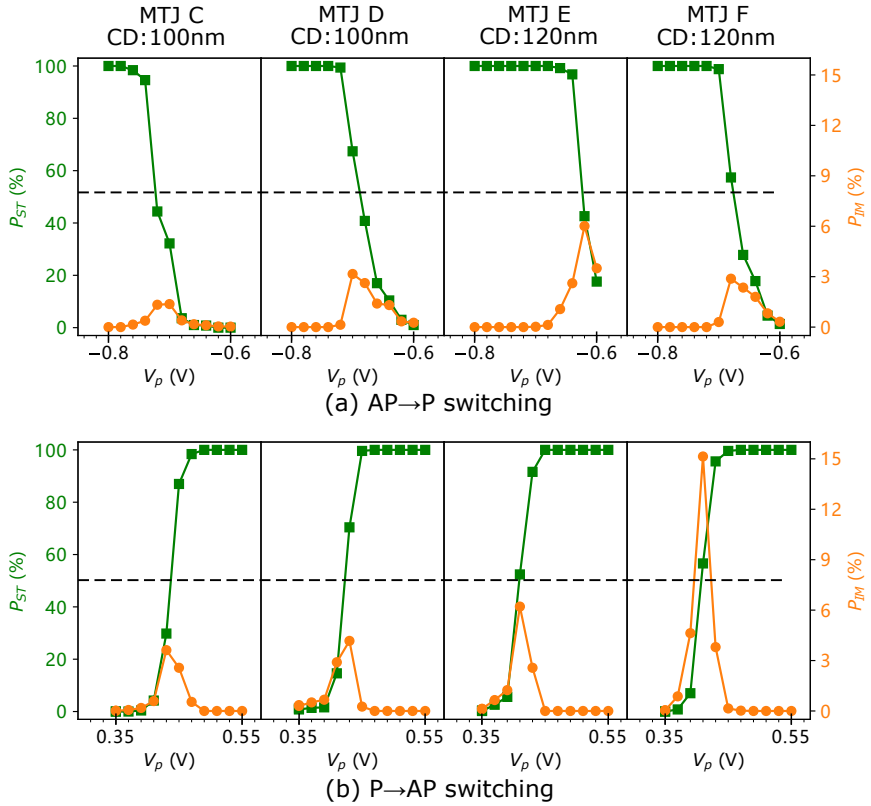


Figure 9.3: Bias voltage and switching direction dependence of IM state.

ing a peak at $P_{ST} \approx 50\%$ (marked with the horizontal dash line), then it decreases as V_p further increases; this rule applies for all four devices in both switching directions despite the peak height of P_{IM} varies from one device to another. Furthermore, even for the same device, there is a large difference in the peak height of P_{IM} for the AP→P and P→AP switching directions. This indicates that the occurrence probability of IM state also depends on the switching direction.

To investigate whether the MTJ size plays a role in determining the occurrence probability of IM state, we repeated the same measurements on MTJ devices with four different sizes, i.e., CD=60 nm, 75 nm, 100 nm, and 120 nm. For each size, we measured 60 devices; the number of devices with IM state is shown with the blue histogram (left y-axis) in Figure 9.4. It is clear that the smaller the MTJ device (i.e., smaller CD), the less likely to see IM states in our devices. More specifically, 57 devices out of the measured 60 devices with CD=120 nm exhibit IM states in the measurement, whereas the number is 5 and 0 for MTJs with CD=75 nm and 60 nm respectively. Among those devices with observed IM states, the median of the maximum occurrence probability of IM state (i.e., the peak height of P_{IM} in Figure 9.3) becomes smaller when CD decreases, as shown with the two orange curves corresponding the right y-axis in Figure 9.4. It is also worth noting that the median of the maximum P_{IM} in AP→P switching direction is slightly smaller than that in

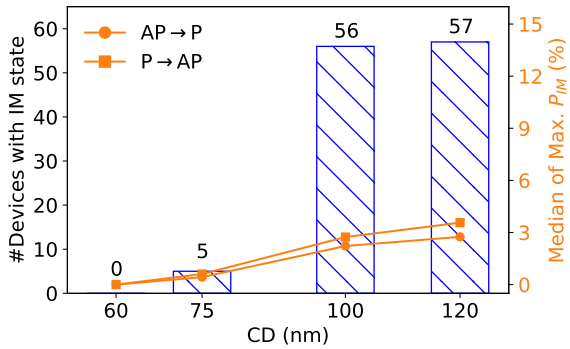


Figure 9.4: MTJ size dependence of IM state.

P→AP switching direction for a given MTJ size. This is probably because AP→P switching generates more Joule heating than the opposite switching direction, which reduces the retention time of IM state; thus, the captured number of IM states on average is smaller in AP→P switching direction under the same measurement set-up. Interestingly, Intel also presented similar measurement results in [79]. Based on the above observations, it can be inferred that STT-MRAM technology down-scaling is helpful in reducing the probability of having IM states in MTJ devices, thus leading to a more deterministic and uniform transition between the bi-stable AP and P states.

In summary, the above measurement data clearly demonstrates the existence of IM states in MTJ devices especially for large sizes above 75 nm. The occurrence of IM state is probabilistic depending on the switching direction, applied bias voltage, and device size. In addition, we swept the read pulse width from 50 μs to 10 ms in our measurements; the results show that the IM states occur in all these configurations indicating that RT_{IM} is larger than 10 ms after the removal of write pulses. The root causes can be attributed to some physical imperfections such as unreversed magnetic bubbles [132], inhomogeneous distribution of stray field [79] or even skyrmion generation [133].

9.3. LIMITATIONS OF THE CONVENTIONAL TEST APPROACH

As already mentioned in the previous chapter, the conventional test approach models any defect in an MTJ device as a linear resistor either in parallel to (R_{pd}) or in series with (R_{sd}) a defect-free MTJ model. The physical mechanism of defect is never taken into account and manifested as a difference in the defect model. This can be found in the prior works on STT-MRAM testing [49, 50, 52, 54–56, 61]. Applying the conventional fault modeling approach to the IM state defect results in four FPs: iR1NF1 = ⟨1r1/1/0⟩, iR0NF0 = ⟨0r0/0/1⟩, W1TF0 = ⟨0w1/0/–⟩, W0TF1 = ⟨1w0/1/–⟩, as shown in Table 9.2. These four FPs can be used to generate test solutions such as March algorithms. First, each sensitized FP is assigned its own detection condition. For instance, iR0NF0 requires a read operation on the faulty cell at state '0' to guarantee its detection, denoted as ⇕(...0,r0,...), where ⇕ means that the detection condition does not depend on the addressing direction. The detection condition for W0TF1 is ⇕(...1,w0,r0,...), meaning that a down-transition write followed by a read is enough to detect this fault, regardless of the addressing direction. The detection conditions of all sensitized FPs are compiled into the following optimal

Table 9.2: Static fault modeling results of IM state defect using resistive models.

Defect model	Resistance (Ω)	Sensitized FP	Fault Model &FP Name	Detection Condition
Series resistor R_{sd}	(466, 870]	⟨0r0/0/1⟩	incorrect Read Non-destructive Fault: iR0NF0	⇳ (...0, r0, ...)
	(870, 1.6k]	⟨0r0/0/1⟩	incorrect Read Non-destructive Fault: iR0NF0	⇳ (...0, r0, ...)
		⟨1w0/1/–⟩	Write Transition Fault: W0TF1	⇳ (...1, w0, r0, ...)
	(1.6k, +∞]	⟨0r0/0/1⟩	incorrect Read Non-destructive Fault: iR0NF0	⇳ (...0, r0, ...)
		⟨1w0/1/–⟩	Write Transition Fault: W0TF1	⇳ (...1, w0, r0, ...)
		⟨0w1/0/–⟩	Write Transition Fault: W1TF0	⇳ (...0, w1, r1, ...)
Parallel resistor R_{pd}	[0, 1.1k)	⟨1r1/1/0⟩	incorrect Read Non-destructive Fault: iR1NF1	⇳ (...1, r1, ...)
		⟨1w0/1/–⟩	Write Transition Fault: W0TF1	⇳ (...1, w0, r0, ...)
		⟨0w1/0/–⟩	Write Transition Fault: W1TF0	⇳ (...0, w1, r1, ...)
	[1.1k, 3.1k)	⟨1r1/1/0⟩	incorrect Read Non-destructive Fault: iR1NF1	⇳ (...1, r1, ...)
		⟨1w0/1/–⟩	Write Transition Fault: W0TF1	⇳ (...1, w0, r0, ...)

March test with three march elements:

$$\{⇳ (w0); ⇑ (w1, r1); ⇓ (w0, r0)\}.$$

Note that different versions of March tests can be generated (e.g., with two march elements) as long as the test satisfies all the detection conditions.

Based on our measurement results in the previous section, one can easily observe that the sensitized four FPs using the conventional fault modeling approach cannot cover the faulty behaviors of IM state defect in MTJ device. This is because the IM state defect manifests itself as a resistive state between R_P and R_{AP} with an occurrence probability. This means that this defect may turn an MTJ device into the undefined state 'U' and this faulty behavior occurs intermittently. The conventional fault modeling and test approach consider the MTJ device as an *ideal black box* (only state '0' and '1'). Therefore it fails to capture the above-mentioned characteristics of IM state defect. As the four FPs are inappropriate in presenting the IM state defect, March tests that target these faults obviously cannot guarantee the detection of such a defect. Therefore, we need to apply DAT to the IM state defect for accurate defect and fault models, which will eventually lead to high-quality test solutions that we desire.

9.4. DEVICE-AWARE DEFECT MODELING FOR IM STATE

In order to investigate the faulty behavior of memory cell in the presence of an IM state defect, first an appropriate physics-based defect model needs to be developed. In this section, we will follow the device-aware defect modeling approach proposed in [47], which consists of three steps: 1) physical defect analysis and modeling, 2) electrical modeling of defective MTJ device, and 3) fitting and model optimization. Next, we will work out these three steps for the IM state defect.

9.4.1. PHYSICAL DEFECT ANALYSIS AND MODELING

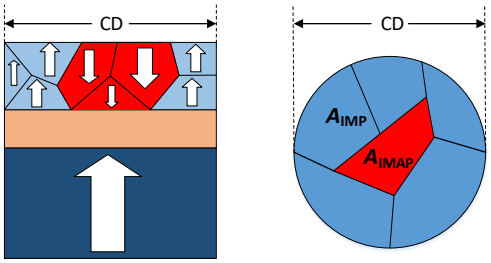
Based on the characteristics and potential forming mechanisms of IM state, as presented with silicon measurements in Section 9.2, we physically model the IM state at three key aspects as follows.

PARTIAL SWITCHING BEHAVIOR OF THE FL

As explained in the previous section, the most probable cause of IM state in MTJ devices is that some parts of the FL switch to the intended state under a write pulse while the rest remain in their initial state due to unreversed magnetic bubbles, inhomogeneous distribution of stray field at the FL, or even skyrmion generations. Therefore, we model this partial switching behavior by splitting the FL into two regions: 1) P-state region and 2) AP-state region with the assumption that these two regions are independent magnetically and electrically. Figure 9.5a and Figure 9.5b show the vertical and horizontal cross-section schematics of an MTJ device with both P-state and AP-state regions, respectively. As a result, we can derive:

$$1 = \frac{A_{\mathrm{P}}}{A_0} + \frac{A_{\mathrm{AP}}}{A_0} = A_{\mathrm{IMP}} + A_{\mathrm{IMAP}}, \tag{9.1}$$

where A_{P} and A_{AP} are the cross-sectional area of the P-state and AP-state regions, respectively. A_{IMP} and A_{IMAP} are the normalized area with respect to the entire area A_0; they can be any value in $[0,1]$. Note that this model also covers the defect-free case where the P and AP states exist exclusively; i.e., A_{IMP}=0 represents AP state and A_{IMP}=1 means P state.



(a) Vertical cross-section (b) Horizontal cross-section

Figure 9.5: MTJ schematics with both P-state and AP-state regions in the FL.

PROBABILISTIC OCCURRENCE OF IM STATE

As introduced previously, the IM state does not show up in all write cycles. Instead, we observed experimentally that it has a certain occurrence probability depending on the applied bias voltage V_{p}, MTJ size CD, and the switching direction. Apart from that, it is expected that the FL thickness (t_{FL}) also plays a role in determining the IM occurrence probability, as it significantly influences the thermal stability of the device [75].

We define a discrete random variable X as whether or not the IM state occurs. For a given V_{p}, CD, and t_{FL}, X obeys a Bernoulli distribution. Its probability mass function Pr(X) is:

$$Pr(X) = \begin{cases} 1 - P_{\mathrm{IM}}(V_{\mathrm{p}}, CD, t_{\mathrm{FL}}), & X = 0; \\ P_{\mathrm{IM}}(V_{\mathrm{p}}, CD, t_{\mathrm{FL}}), & X = 1. \end{cases} \tag{9.2}$$

As shown in Figure 9.3, the correlation between P_{IM} and V_{p} exhibits a curve which is quite similar to Gaussian function (Bell curve). Thus, we model the V_{p} dependence of P_{IM} as:

$$P_{\mathrm{IM}} = H_{\mathrm{IM}} \cdot \exp\left(\frac{-(V_{\mathrm{p}} - V_{\mathrm{pk}})^2}{2V_{\mathrm{wd}}^2}\right), \tag{9.3}$$

where V_{pk} is the applied bias voltage when P_{IM} reaches its peak H_{IM}, and V_{wd} is a parameter controlling the width of the Bell curve. Note that the polarity of V_p determines the switching direction; a negative V_p results in an AP→P transition while a positive V_p leads to a reversed transition. Since H_{IM} shows a linear scaling trend with CD, as shown in Figure 9.4, it can be modeled as a linear piecewise function:

$$H_{IM} = \begin{cases} S_{lp} \cdot (CD - 60), & CD \geq 60; \\ 0, & CD < 60. \end{cases} \tag{9.4}$$

S_{lp} is the slope of the curve. Since all the measurements we performed were on MTJ devices with the same t_{FL}, it is assumed that t_{FL} has no impact on P_{IM}. However, for a generic model for devices with different P_{IM}, such impact should be incorporated. Combing Equations (9.2–9.4), S_{lp}, V_{pk}, and V_{wd} are three fitting parameters which can be tuned and fitted to measurement data, which will be covered later.

RETENTION TIME ESTIMATION OF IM STATE

The retention time of IM state (RT_{IM}) indicates how long the IM state remains after the removal of write pulses; it determines the time period where the memory fault behavior appears in the presence of the IM state. Thus, it is important to estimate RT_{IM} of our devices and integrate it into the defect model if necessary. Conventionally, the following static model is used to roughly estimate the retention time of AP or P state for a given Δ [81]:

$$RT = \tau_0 \exp(\Delta), \tag{9.5}$$

where τ_0 is the inverse of the attempt frequency (~1 ns). However, the retention time for STT-MRAMs has intrinsic stochasticity, as the magnetization flip induced by thermal fluctuation is unpredictable. This static model fails to capture the stochastic property. Actually, the calculated retention time using Equation (9.5) corresponds to the time after which the MTJ state flips at a probability of 63%, as pointed out in [82]. As an alternative, a statistic model derived from the switching model in thermal-activation regime is widely used, as can found in [75, 82, 83]:

$$RT = \tau_0 \exp(\Delta) \cdot (\frac{1}{1 - P_{RT}}), \tag{9.6}$$

where P_{RT} is the switching probability of a certain MTJ state due to thermal fluctuation after time RT (i.e., the confidence in the estimation of RT). Next, we will model the retention time of IM state RT_{IM} based on this statistic model.

As illustrated in Figure 9.5, the IM state takes place when some parts of the FL switch while the rest remain in their initial state. Thus, the retention time of IM state RT_{IM} is the time period before the magnetization of the P-state or AP-state region spontaneously flips to the opposite direction under the influence of thermal perturbation such that the two regions merge again into an entire one. In other words, RT_{IM} is the smaller one in

the retention time of the P-state region and AP-state region.

$$RT_{\mathrm{IM}} = min\{RT_{\mathrm{IMP}}, RT_{\mathrm{IMAP}}\}, \tag{9.7}$$

$$RT_{\mathrm{IMP}} = \tau_0 \exp(\Delta_{\mathrm{P}} \cdot \sqrt{A_{\mathrm{IMP}}}) \cdot (\frac{1}{1-P_{\mathrm{RT}}}), \tag{9.8}$$

$$RT_{\mathrm{IMAP}} = \tau_0 \exp(\Delta_{\mathrm{AP}} \cdot \sqrt{A_{\mathrm{IMP}}}) \cdot (\frac{1}{1-P_{\mathrm{RT}}}). \tag{9.9}$$

In the above equations, Δ_{P} and Δ_{AP} are the thermal stability factor of the normal P and AP states of MTJ, respectively. RT_{IMP} and RT_{IMAP} are the retention time of the P-state and AP-state regions in IM state, respectively. The modeling principle for RT_{IMP} and RT_{IMAP} is based on the observation with device-level silicon measurements that Δ scales linearly with CD (i.e., $\sqrt{A}$) when CD>40 nm [223].

Figure 9.6 shows the estimated retention time in IM state RT_{IM} as a function of A_{IMP}. It can be seen that RT_{IM} increases with A_{IMP} until reaching a peak at $A_{\mathrm{IMP}} = 0.64$, after which it goes down. The maximum RT_{IM} can be up to 1 day for both P_{RT}=63.0% and 99.9%. However, it is still more than three orders of magnitude smaller than RT_{P}; note that RT_{P} is smaller than RT_{AP} due to the existence of stray field at the FL. Furthermore, the large amount of Joule heating generated under switching pulses may increase the junction temperature by more than 50 °C [224]. This will further reduce RT_{IM} in practice.

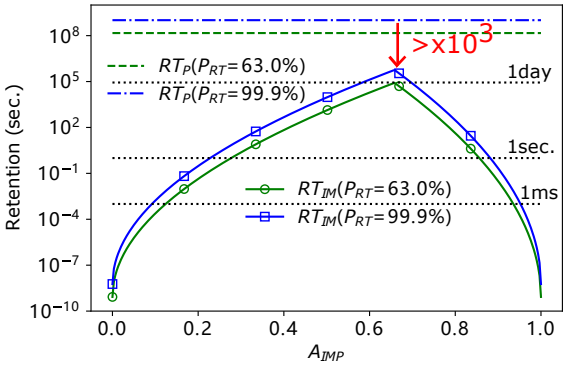


Figure 9.6: IM state retention time estimation.

9

9.4.2. ELECTRICAL MODELING OF MTJ DEVICES WITH A SINGLE IM STATE

With the obtained physical model of IM state, we can map it to the three key electrical parameters: R, I_{c}, and t_{w} as a reflection of the impact on the device's electrical behavior.

As we model the IM state by splitting the FL into AP-state and P-state regions (see Figure 9.5), electrons can go through via either the P-state region or the AP-state region under an electric field. Therefore, the overall conductance of IM state is the sum of the conductance of these two parallel regions.

$$G_{\mathrm{IM}}(A_{\mathrm{IMP}}) = G_{\mathrm{P}} \cdot A_{\mathrm{IMP}} + G_{\mathrm{AP}} \cdot (1 - A_{\mathrm{IMP}}), \tag{9.10}$$

where G_{P} and G_{AP} are the conductance when the entire FL is in P and AP states, respectively. A_{IMP} is the normalized area of P-state region in IM state with respect to the entire

cross-sectional area of the FL. By replacing conduction with resistance (G=1/R) in the above equation, we can derive:

$$R_{\mathrm{IM}}(A_{\mathrm{IMP}}) = \frac{R_{\mathrm{P}} \cdot R_{\mathrm{AP}}}{R_{\mathrm{P}} \cdot (1 - A_{\mathrm{IMP}}) + R_{\mathrm{AP}} \cdot A_{\mathrm{IMP}}}. \tag{9.11}$$

R_{P} and R_{AP} are both dependent on the bias voltage V_{MTJ} applied across the MTJ device. Figure 9.7a shows the measured R-V loop of MTJ C, the same one shown in Figure 9.3; the red solid curves are fitting curves used to extract the exact resistance at a given bias voltage with the physical model in [62]. With R_{P} and R_{AP} extracted from measurement data at different bias voltages, we can calculate R_{IM} for different A_{IMP} values using Equation (9.11); the results are shown in Figure 9.7b for V_{p}= 10 mV, 300 mV, and 700 mV.

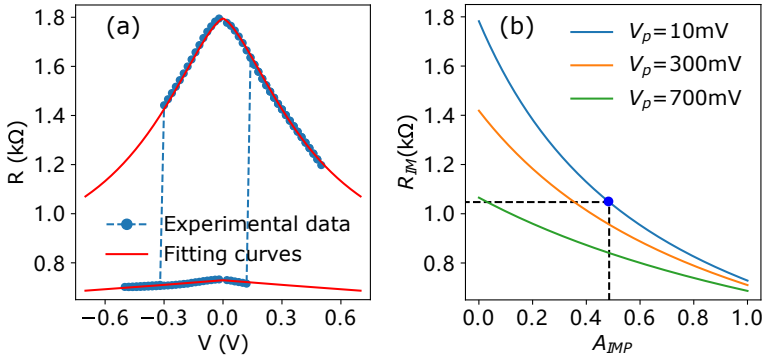


Figure 9.7: (a) R-V loop experimental data vs. fitting curves to extract R_{P} and R_{AP} at varying bias voltage, (b) R_{IM} vs. A_{IMP} with respect to three bias voltages.

Conventionally, the switching spectrum between P and AP states in STT-MRAMs can be divided into two regimes: 1) precessional regime for short pulses (<~40 ns for our devices), 2) thermal activation regime for long pulses [62, 75]. The switching behavior in the precessional regime is dominated by the STT effect while the thermal effect plays a major role in determining the switching behavior in the thermal activation regime. To model the switching behavior between P, AP, and a third IM state, we modify the equation of the critical switching current I_{c} in the STT-switching model as follows [75].

$$I_{\mathrm{c}}(A_{\mathrm{IMP}}) = \begin{cases} \frac{1}{\eta}\frac{2\alpha e}{\hbar} M_{\mathrm{s}} H_{\mathrm{k}} t_{\mathrm{FL}} A_0 A_{\mathrm{IMP}}, & \mathrm{IM(P)} \rightarrow \mathrm{AP}; \\ \frac{1}{\eta}\frac{2\alpha e}{\hbar} M_{\mathrm{s}} H_{\mathrm{k}} t_{\mathrm{FL}} A_0 (1 - A_{\mathrm{IMP}}), & \mathrm{IM(AP)} \rightarrow \mathrm{P}. \end{cases} \tag{9.12}$$

In this equation, η is the STT efficiency, α the magnetic damping constant, e the elementary charge, $\hbar$ the reduced Planck constant. The rest of parameters have already been introduced previously. When $A_{\mathrm{IMP}} = 1$ (indicating P state), the above equation collapses to the original equation for I_{c}(P→AP). When $A_{\mathrm{IMP}} \in (0, 1)$ (indicating IM state), I_{c}(IM→AP) is smaller than I_{c}(P→AP) as only the P-state region in the FL necessitates a flip. Similar interpretation can be inferred for IM(AP)→P switching. Note that the switching from P or AP state to IM state is governed by the aforementioned statistical model in Equation (9.2–9.4).

Furthermore, the switching time t_w in the precessional regime (namely, switched by the STT-effect) can be estimated using the Sun's model as follows [62]:

$$\mu(t_w) = \left(\frac{2}{C+\ln(\frac{\pi^2\Delta}{4})} \cdot \frac{\mu_B P}{e \cdot m \cdot (1+P^2)} \cdot I_d\right)^{-1}, \tag{9.13}$$

$$I_d = \frac{V_p}{R(V_p)} - I_c(A_{IMP}), \tag{9.14}$$

$$t_w \sim \mathcal{N}(\mu(t_w), \sigma(t_w)^2). \tag{9.15}$$

Here, C≈0.577 is Euler's constant, Δ the thermal stability in P or AP or IM depending on the switching direction, μ_B the Bohr magneton, P the spin polarization, and m the FL magnetic moment. V_p is the bias voltage across the MTJ device to switch its state. $R(V_p)$ is the resistance of the MTJ device; it shows a non-linear dependence on V_p (see Figure 9.7a). In addition, we assume that t_w obeys a normal distribution for a given V_p (Equation 9.15) as model for the switching stochasticity [96].

9.4.3. FITTING AND MODEL OPTIMIZATION

In the third step of our device-aware defect modeling approach, fitting and model optimization can be conducted if silicon data is available. With the measured data presented in the Section 9.2, next we will illustrate this step by fitting the obtained model to a specific device MTJ B as an example. Note that our MTJ compact model is generic and device-to-device variations due to process variations can be modeled by assigning a Gaussian distribution to the key technology parameters of MTJ.

First, R_P and R_{AP} of MTJ C can be extracted from its R-V loop, as shown in Figure 9.7a. As the measured R_{IM}=1050 Ω (see Figure 9.2c and 9.2d) and the read bias is 10 mV, we can calculate the A_{IMP} value based on our model. The result is marked with the blue point (A_{IMP}=0.48) in Figure 9.7b. Second, the fitting results of P_{ST} and P_{IM} are shown in Figure 9.8. On the positive side V_p>0 for P→AP switching, S_{lp}=1e-3, V_{pk}=0.4369, and V_{wd}=0.0145. On the negative side V_p<0 for AP→P switching, S_{lp}=3.9e-4, V_{pk}=-0.7096, and V_{wd}=0.0182. Third, the critical switching current I_c is not directly measurable. Thus, I_c fitting is not applicable here. In addition, the switching time t_w changes with V_p as well. The fitting process and results are presented in [62], thus will not be repeated here.

The output of device-aware defect modeling is a calibrated Verilog-A MTJ compact model. After verifying and calibrating the MTJ model in Python as presented previously, we moved this model to Verilog-A so as to make it compatible with circuit simulators such as Cadence Spectre for subsequent fault modeling. Figure 9.9 shows the verification results of the MTJ compact model integrating the following three variation sources affecting the switching behavior for P→AP switching under pulses with t_p=15 ns as an example.

- **Switching stochasticity (STO)**: In Figure 9.9a, only the switching stochasticity (cycle-to-cycle variation) is enabled while process and temperature variations are disabled. We swept the bias voltage V_p from 0.3 V to 0.5 V in 50 steps, each of which involved a 5k-cycle Mente Carlo simulation to obtain statistical switching results. It can be seen that the circuit simulation results accurately emulate the measurement and fitting results shown in the positive part in Figure 9.8.

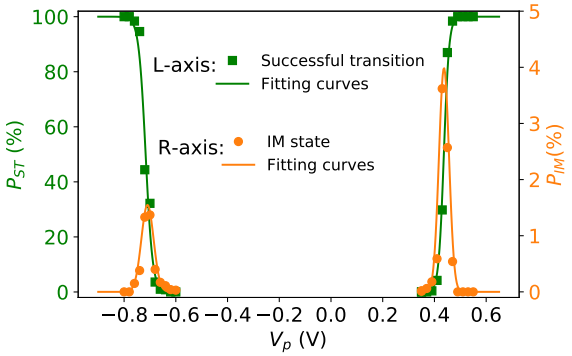


Figure 9.8: Curve fitting of P_{ST} and P_{IM} to measurement data.

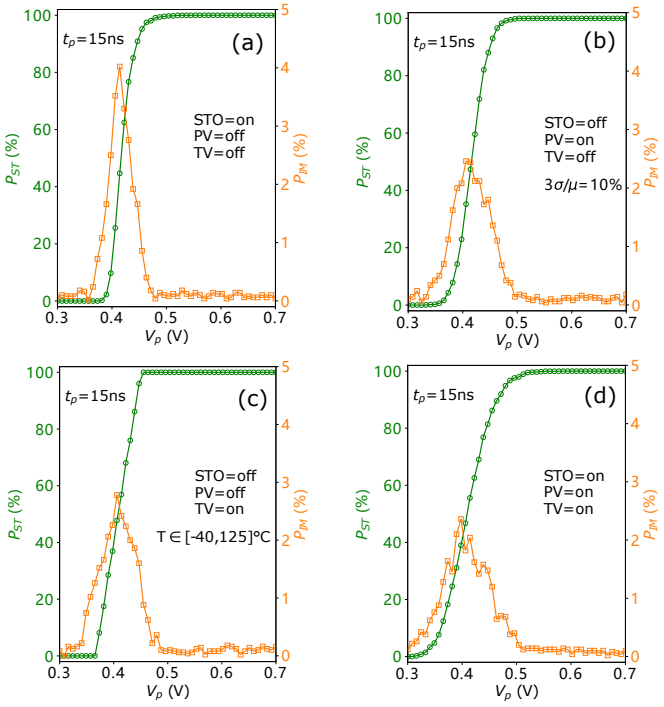


Figure 9.9: Verification of Verilog-A MTJ compact model with Cadence Spectre: (a) Switching stochasticity (STO) enabled only, (b) process variation (PV) enabled only, (c) temperature variation (TV) enable only, and (d) all the three sources of variation enabled simultaneously.

- **Process variation (PV)**: Process variations in MTJ's geometrical parameters (e.g., CD, t_{FL}, t_{TB}) and magnetic properties (e.g., H_{k} and M_{s}) greatly contribute to the device-to-device variation in the switching behavior on top of the intrinsic switching stochasticity, as shown with silicon data in [15, 225]. Our MTJ model takes into account process variation by introducing a Gaussian distribution to each of the

above parameters. Figure 9.9b shows the switching statistics with PV enabled only; we set the 3σ corner at 10% away from the average (i.e., $3\sigma = 0.1\mu$) in our simulations. One can observe that PV on this scale introduces a slightly wider distribution in both P_{ST} and P_{IM} than STO in Figure 9.9a

- **Temperature variation (TV)**: The operating temperature also has a large impact on the switching behavior in STT-MRAM as demonstrated in [4, 15]. In our simulations, we took into account temperature variation by assigning a uniform distribution to the operating temperature from −40 °C to 125 °C (typical industrial standard). Figure 9.9c shows the switching statistics with TV enabled only; it is clear that TV has a contribution as large as STO and PV in the switching variation of STT-MRAM.

Figure 9.9d shows the switching statistics combining all the above three sources of variation. It shows that the switching voltage V_p may span more than 0.2 V from 0% to 100% switching probability; across the entire switching curve, the IM state appears with varying probability as shown in the figure. Due to the large variation in the switching behavior, it is unwise to adopt fixed overdrive pulse amplitude and duration in order to obtain 100% switching in all cells, all cycles, and all operating temperature for write operations in practice.

9.5. DEVICE-AWARE FAULT MODELING FOR IM STATE

In this section, we apply the device-aware fault modeling to obtain realistic and accurate fault models for the SAFF defect. It consists of two sub-steps: 1) fault space definition, 2) fault analysis. The former defines all possible faults theoretically, as already presented in Section 6.4.1. The latter validates realistic faults in the presence of the defect under investigation in the pre-defined fault space using SPICE-based circuit simulations. Next, we will work out these two sub-steps for IM state defects in MTJ devices and compare the fault modeling results with that of the conventional resistive model. Finally, we study the distribution of observed memory faults on write voltage and duration for the purpose of test development.

The simulation circuits are from [46] with a 3×3 1T-1MTJ array and peripheral circuits (e.g., write driver and sense amplifier). All transistors in the netlist are built with the 90 nm predictive technology model (PTM) [98]. Process variations in transistors are lumped into the variation in the threshold voltage V_{th} with 10% away from its nominal value at 3σ corners. For the nine MTJ devices in the memory array, our Verilog-A MTJ compact model with CD=100 nm is adopted; Variations in MTJ performance are covered by enabling STO, PV, and TV options in the MTJ model, as detailed in Section 9.4.3.

The defect injection was executed by replacing the defect-free MTJ model (with only P and AP states) located in the center of the array with a defective one (with P, AP, and IM states) presented in the previous section. The defect strength was configured by assigning a float number to $A_{IMP} \in (0, 1)$ as an input parameter of the Verilog-A MTJ model; it was swept from 0 to 1 in 100 steps in the simulations. The remaining eight MTJs surrounding the central one were always defect-free.

In terms of stimuli, we simulated $S \in \{0, 1, 0w0, 1w1, 0w1, 1w0, 0r0, 1r1\}$, i.e., all static operations. V_{DD} was set to 1.6 V and V_{WL} at 1.8 V. Note that boosting the voltage on the

WL is a common practice in the MRAM community due to the source degeneration (i.e., $V_{GS}<V_{DD}$) of NMOS selectors [30, 95]. The write pulse width was set to 20 ns and read pulse width at 5 ns. Due to the large variation in the switching behavior induced by STO, PV, and TV, we conducted 2k Monte Carlo simulations for each sensitizing sequence S.

Since the simulation overhead is immense due to Monte Carlo simulations (2k cycles), we performed the circuit simulations in a cluster with eight compute nodes to speedup the simulation by exploiting job-level parallelism. We first ran the simulation with a defect-free netlist. Thereafter, the whole simulation process was repeated after injecting an IM state defect with certain A_{IMP} value into the netlist. Finally, fault analysis and FP identification can be conducted by comparing the simulation results of the above defect-free and defective cases.

Table 9.3 lists the fault modeling results due to IM state defects. When $A_{IMP}\in[0.30, 0.61]$, two FPs were observed: $\langle 0w1/U_i/-\rangle$ and $\langle 1w0/U_i/-\rangle$. The intermittent write transition fault $W1TFU_i=\langle 0w1/U_i/-\rangle$ means that an up-transition operation on a memory cell with inital state '0' transforms the memory cell into a 'U' state with a certain probability (i.e., intermittently). Similarly, the intermittent write transition fault $W0TFU_i=\langle 1w0/U_i/-\rangle$ was also observed. Since these two FPs both involve the 'U' state and are intermittent, they belong to hard-to-detect faults [46]. Their detection cannot be guaranteed by March tests and thus requires DfT solutions. Note that transition failures due to switching stochasticity are typically not considered as memory faults induced by defects [56]; thus, they are excluded here.

Figure 9.10 shows a Venn diagram which compares the fault modeling results using our device-aware (DA) defect model and the conventional resistive model. Clearly, the DA model leads to two hard-to-detect faults while the resistive model results in four

Table 9.3: Fault modeling results of IM state defects using our device-aware (DA) defect model.

Defect model	**A_{IMP}**	**Sensitized FP**	**FP name and abbreviation**	**Detection condition**
DA model	[0.30, 0.61]	$\langle 0w1/U_i/-\rangle$	Intermittent write transition fault: $W1TFU_i$	DfT
		$\langle 1w0/U_i/-\rangle$	intermittent write transition fault: $W0TFU_i$	

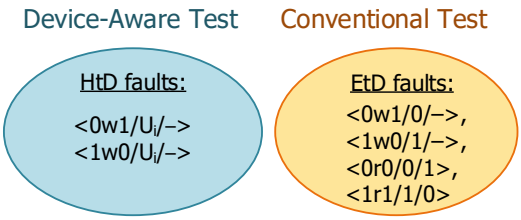


Figure 9.10: Comparison of sensitized fault primitives between our DA model (left) and the conventional resistive model (right).

easy-to-detect faults. There is no overlap between the two circles. This means that IM state defects in MTJ devices exhibits unique faulty behaviors which cannot be covered by the resistor-based defect models. The two FPs sensitized using our DA model are intermittent and involve the 'U' state, which make them hard to be detected by March tests. In contrast, the resistive models resulted in only easy-to-detect faults, since the MTJ device was considered as an *ideal black box* and thus only '0' and '1' states were observed in the simulations.

To investigate the dependence of the observed write transition faults on write voltage and duration, we swept V_{WL} from 1.4 V to 2.2 V and t_{p} from 10 ns to 40 ns in our circuit simulations. Figure 9.11 shows the simulation result statistics of S=0w1 at varying V_{WL} and t_{p} in the defect-free case. The successful transition probability P_{ST} rises from 0% (red area) to 100% (blue area) as V_{WL} and t_{p} increase. However, one can observe that the transition area occupies a large area in the contour map, which poses a big design challenge for reliable and deterministic write operations in STT-MRAMs. This clearly indicates that write schemes with a fixed configuration of write voltage and duration are unwise in practice with four drawbacks: 1) large energy consumption, 2) long write latency (performance loss), 3) more susceptible to back-hopping effect [134, 148], and 4) reduced endurance or even early breakdown induced by aggressively wearing out the untra-thin MgO tunnel barrier under a large switching current. This has led to the introduction of more flexible write schemes such as write-verify-write scheme by Intel [30] and self-write-termination scheme by TSMC [102].

Figure 9.12 shows the IM state statistics in S=0w1 operations at varying V_{WL} and t_{p} in the defective case (A_{IMP}=0.48 as an example). It can be seen that the IM state shows up with different probability P_{IM} in a large area of the contour plot, especially in the area where P_{ST} is near 50%. Obviously, the closer to the top-right corner, the less likely to see an IM state and more likely to have a successful transition. However, large V_{WL} and t_{p} incur the aforementioned four drawbacks. Hence, in practice, a trade-off has to be made and a flexible and self-adaptive write scheme is more desirable. The simulation results for S=1w0 are similar, thus they are excluded here.

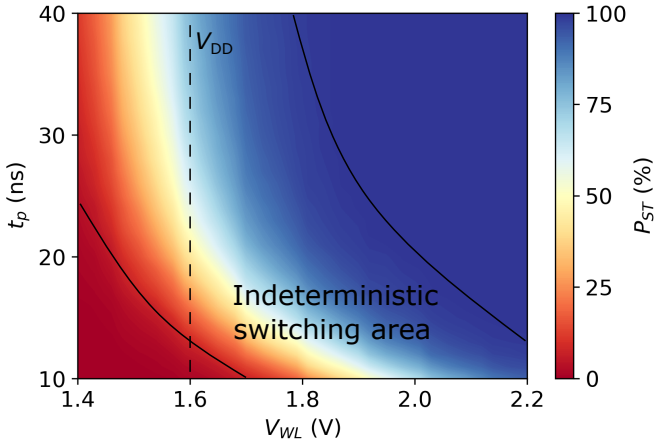


Figure 9.11: Successful transition probability P_{ST} statistics in 0w1 operations at varying word line voltage V_{WL} and pulse width t_{p} in the defect-free case.

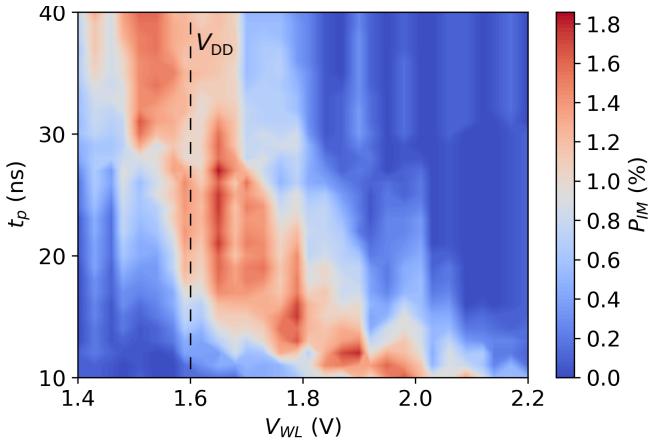


Figure 9.12: Occurrence probability of IM state P_{IM} statistics in 0w1 operations at varying word line voltage V_{WL} and pulse width t_{p} in the defective case.

In summary, our device-aware defect model for IM state defects results in two intermittent write transition faults, whereas the conventional resistive defect models lead to four different memory faults. Hence, test solutions targeting for memory faults based on resistive defect models will lead to not only test escapes but also a waste of test time and resources. As an alternative, device-aware test can be a complimentary approach which specifically targets device-internal defects. Next, the fault modeling results obtained in this section will be used to develop a dedicated test solution for IM state defects.

9.6. DEVICE-AWARE TEST DEVELOPMENT FOR IM STATE

The last step of DAT is to develop appropriate test solutions for the derived faults: W1TFU$_{\mathrm{i}}$ and W0TFU$_{\mathrm{i}}$. In this section, we first explain the test philosophy. Thereafter, a test solution with weak write operations is introduced. Its circuit implementation will also be presented and discussed.

9.6.1. TEST PHILOSOPHY

To detect IM state defects, the following two key steps are crucial: 1) fault sensitization, 2) fault detection. The former forces a defective MTJ into the IM state so that it exhibits faulty behavior, whereas the latter distinguishes it from the normal memory behavior. Figure 9.13a illustrates the energy barrier diagram of a defect-free MTJ with bi-stable AP and P states. The energy barrier in AP→P switching is larger than that of the opposite switching direction, due to the existence of stray field which is in favor of AP state. Figure 9.13b illustrates the energy barrier diagram of a defective MTJ with AP, P, and IM states. As already discussed in previous sections, the IM state can be set with write operations with certain occurrence probability P_{IM}; the peak of P_{IM} occurs at the bias voltage where P_{ST}=~0.5 (see Figure 9.3). Once the IM state is set, the device may stay in IM state without external interference for certain period of time (i.e., retention time of the IM state) or fall back to AP or P state in an accelerated process under external interference. This is

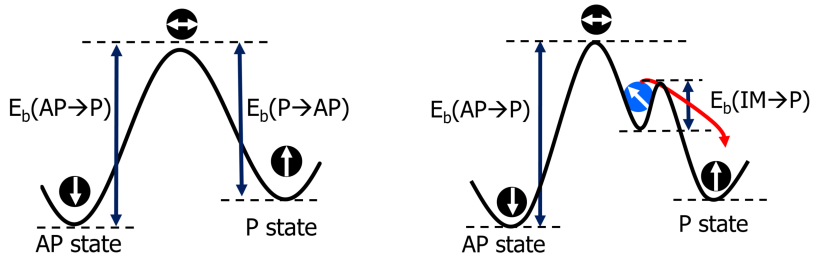


Figure 9.13: Comparison of energy barriers between: (a) a defect-free MTJ with bi-stable AP and P states and (b) a defective MTJ with AP, P, and IM states.

because the energy barrier from IM to P (or AP) is much smaller than that between P and AP states, as illustrated with the height of the two-way arrows in Figure 9.13b. Hence, to distinguish the IM state from P and AP states, a feasible solution is to provide sufficient external energy to push the device in IM state back to P (or AP) state while avoiding disturbing devices in AP (or P) state.

Typically, there are mainly three sources of external energy which can be provided to affect the thermal stability factor Δ of MTJ. They are thermal energy reflected as temperature (T), electric current (I), and magnetic field (H). The quantitative correlation between these three variables and Δ can be approximately expressed as follows [75, 226]:

$$\Delta(T, I, H) = \frac{E_B}{k_B T} \cdot (1 - \frac{I}{I_c}) \cdot (1 \pm \frac{H}{H_k})^2. \tag{9.16}$$

First, the above equation indicates that Δ can be reduced by heating up the MTJ devices (i.e., burn-in test). The elevated temperature leads to an increase in thermal perturbation, which in turn increases the chance of spontaneous flip of one state to the others. Although this approach is effective in kicking an MTJ device out of the IM state, the switching direction (i.e., IM→P or IM→AP) is not controllable. Thus, burn-in test is an unsuitable approach to detect IM state defects. Second, applying an electric current I going through the MTJ is also an approach to reduce Δ due to its Joule heating effect. After being spin-polarized, it is also used to switch the magnetization in the FL. More importantly, current-induced switching is bipolar, meaning that the switching direction is controlled by the current direction. Third, external magnetic field H has a large influence on Δ. It is widely used in the characterization test of MRAM and serves as the write method in the first generation of MRAM technology, also referred to as Toggle MRAM. Field-induced switching is also bipolar, as the direction of H determines the switching direction of magnetization in the FL.

In summary, the detection of IM state defects can be achieved by applying a weak write current/field, which provides a moderate energy to push a defective MTJ out of its IM state without disturbing the bi-stable P and AP states of defect-free MTJs. Next, we will elaborate the test process with weak write operations.

9.6.2. TEST SOLUTION WITH WEAK WRITE OPERATIONS

To detect IM state defects, the following March algorithm can be used, as illustrated in Figure 9.14.

$$\{\Updownarrow (w0); \Uparrow (w1, r1); \Downarrow (\hat{w}0/\hat{w}0_H, r1)\}.$$

The first march element $\Updownarrow$(w0) initializes all memory cells to state '0' in normal mode. The second march element is composed of two operations in normal mode; the first one is an up-transition write and the second one is a read. For a defect-free MTJ, the MTJ state switches from '0' to '1' as intended and the readout is logic '1'. Note that we do not take into account failed transitions caused by the switching stochasticity, since they can be mitigated by circuit-level designs such as write-verify-write as mentioned previously. For a defective MTJ with IM state, the w1 operation may result in a transition to '1' (AP) or 'U' (IM) state. If the device ends up in the 'U' state, the readout value can be *random* ('?'); i.e., sometimes '0', sometimes '1', unpredictably. The third march element consists of a weak down-transition operation in DfT mode and a read operation in normal mode. The weak write operation can be implemented as a relatively weak current ($\hat{w}0$) or field ($\hat{w}0_H$) with reduced amplitude or duration in comparison to normal write operations. The weak write operation induces an IM→P transition while it is not strong enough to change AP state. As a result, the readout is expected to be logic '1' for the MTJs which are in AP state before the weak write. However, the readout of those MTJs which are in IM state before the weak write is logic '0'. Since the occurrence of IM state is probabilistic, this test cannot 100% guarantee the detection of IM state defects with a single shot. To increase to detection confidence, repeating the above march test for a certain number of times can be considered.

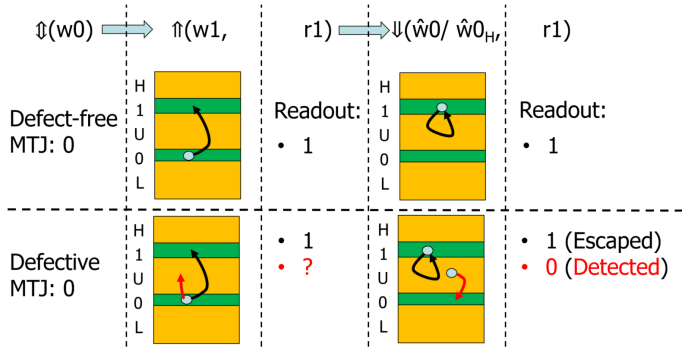


Figure 9.14: Proposed March algorithm with a weak write operation $\hat{w}0/\hat{w}0_H$.

The implementation of weak write operations requires dedicated DfT. Since STT-MRAM exploits an electric current for w0 and w1 operations in normal mode, adding a DfT circuit to write drivers to tune the write voltage or duration will provide a feasible solution with minimal area overhead. For example, if a weak write voltage on the WL ($\hat{V}_{WL}$) is utilized for the DfT circuit, it has to meet the following requirement: $V_{WL}(P_{SIM}{=}1) < \hat{V}_{WL} < V_{WL}(P_{ST}{=}0)$, where P_{SIM} is the switching probability of IM state to either P or AP state and P_{ST} is the switching probability between P and AP states. This ensures that defective memory cells are detected while defect-free ones are not over killed.

Given this consideration, $\widehat{V}_{WL}$ can be set to a point in the black curve in the bottom-left corner of Figure 9.11; it marks the boundary of the area where P_{ST}=0. Hamdioui *et al.* [227] proposed a programmable DfT scheme for weak write operations to detect open defects in RRAMs; this DfT scheme can also be adopted here to configure the weak write operations for STT-MRAMs. In addition, Naik *et al.* [4] proposed an internal bias control design for setting optimal write bias voltages in STT-MRAM in order to adapt to different operating temperature. This bias control design for normal write operations can also be reused to select $\widehat{V}_{WL}$ in DfT mode.

We implemented the above March test and verified the design based on circuit simulations. Figure 9.15 shows the waveforms of five key signals in both defect-free and defective cases. First, both the defect-free and defective MTJs are initialized to state '0' (P), as shown with the MTJ resistance (R_{MTJ}) waveform. The normal w1 operation turns the defect-free MTJ into AP state as intended and the defective MTJ into IM state (sensitizing the W1TFU$_i$ fault). Note that V_{DD}=1.6V whereas V_{WL_en} and V_{WL} are both boosted to 1.8V. Next, the r1 operation reads out the MTJ state on the signal V_{out}. The readout of IM state is unpredictable; on the waveform, it outputs a fake '1'. The third operation is a weak write 0 operation $\widehat{w}0$ with V_{WL} degraded to 1.4V and t_p unchanged at 20ns in DfT mode. It switches the defective MTJ from IM state to P state, while the defect-free MTJ remains in AP state as the provided energy is not high enough to invoke a full transition from AP state to P state. The last r1 operation detects the IM state defect, since the defective MTJ outputs a '0' while the defect-free case is '1', as illustrated in the figure.

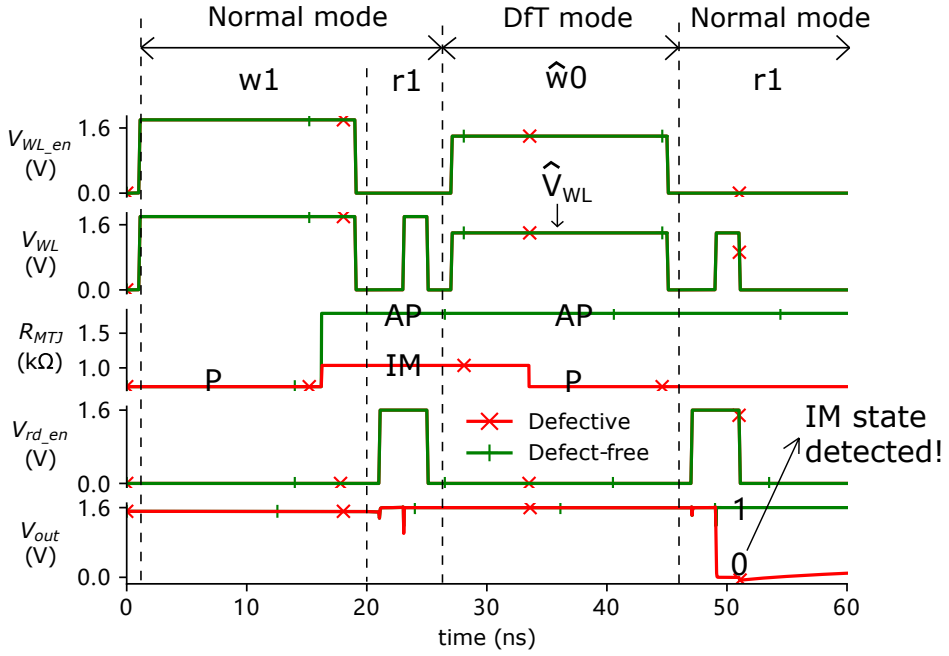


Figure 9.15: Test implementation and verification.

10

CONCLUSION

10.1 Summary
10.2 Future Research Directions

This dissertation demonstrates a paradigm shift in memory testing for emerging memory technologies such as STT-MRAM, RRAM, and PCM. In the conventional memory testing approach, all physical defects are modeled as linear resistors (i.e., opens, shorts, and bridges) which have been widely used for developing tests for existing memories such as SRAMs and DRAMs. This test approach has also been inherited to develop tests for emerging memory technologies, as can be found in the prior art presented in Section 1.3. However, we have demonstrated with both experiments and simulations that this conventional approach is unsuitable for defects in STT-MRAM devices, where magnetic properties are as important as electrical ones. Therefore, we propose a new test approach: device-aware test (DAT), which specifically targets device-internal defects. DAT goes beyond the present cell-aware test and enables future test programs for DPPB level. This dissertation is dedicated to introducing DAT and applying it to STT-MRAM for the purpose of developing high-quality yet cost-efficient tests for STT-MRAM, which are required for high-volume production. This chapter summarizes the overall achievements of this dissertation and highlights some future research directions in STT-MRAM testing.

10.1. SUMMARY

Chapter 1: Introduction

This chapter emphasizes the importance of VLSI test and explains some fundamental concepts in this field. It briefly introduces three types of emerging NVM technologies: PCM, RRAM, and MRAM (including Toggle MRAM, STT-MRAM, and SOT-MRAM). By comparing them to existing memory technologies with various metrics, STT-MRAM stands out due to its advantageous features such as fast speed, high endurance, high density, low-power consumption, radiation immunity, and CMOS compatibility. The promising prospect of STT-MRAM has attracted large amounts of attention in the semiconductor industry to commercialize this technology. However, prior to high-volume production, high-quality test solutions are crucial. This chapter presents the state of the art in STT-MRAM testing, which suggests that STT-MRAM testing is still in the infant stage and many challenges remain at different abstraction levels. At the physical defect level, STT-MRAM defects especially those related to MTJ devices have not been well understood yet. Accurate defect models which appropriately represent all possible physical defects at the electrical level are required. At the memroy fault level, accurate and realistic fault models which describe the faulty behaviors in the presence of a specific defect need to be developed. At the highest abstraction level, optimal test solutions have to be established to effectively weed out defective STT-MRAM chips and pass all good chips (preferably at DPPB level) in a cost-efficient way.

Chapter 2: STT-MRAM Behavior and Architecture

This chapter introduces the behavior and architecture of STT-MRAM using Everspin's 1Gb ST-DDR4 STT-MRAM chip as an example. Depending on different abstraction levels, STT-MRAM can be described using behavioral, functional, electrical, and layout models. The behavioral STT-MRAM model shows that the STT-MRAM chip behaves like a standalone DRAM device with DDR4 interface. The advantages of STT-MRAM include data persistence and low power, compared to existing DRAM chips. This disadvantages of this specific STT-MRAM product include its access speed, density, and cost per bit. The functional STT-MRAM model reveals the internal functional blocks of the STT-MRAM chip and how they behave individually and collectively. A key difference in the internal behavior between STT-MRAM and DRAM lies on the execution of refresh command. DRAM needs refresh operations to retain data stored in cell capacitors, whereas STT-MRAM keeps the refresh command for the purpose of compatibility but redefines its function to move data from open page buffers to persistent STT-MRAM arrays.

Chapter 3: STT-MRAM Technology and Implementation

This chapter elaborates MTJ device organization and working principles for data-storing (thermal stability), data-retrieving (TMR effect), and data-recording (STT effect). In addition, it presents our STT-MRAM circuit design including 1T-1MTJ bit cell and peripheral circuits; all the research works carried out in this dissertation are based on SPICE circuit simulations with this STT-MRAM design. This chapter also surveys the manufacturing process and defect space for STT-MRAMs. STT-MRAM manufacturing defects are classified into two groups: FEOL defects and BEOL defects. The latter is further divided

into interconnect defects and MTJ defects. We give special attention to MTJ defects such as pinhole in the MgO barrier and synthetic anti-ferromagnetic flip, as they are unique to STT-MRAMs. Moreover, this chapter reviews the evolution course of MTJ and commercialization attempts of four generations of MRAM in the past decades. Thanks to its attractive features and tunability between speed, endurance, and retention, STT-MRAM products can be SRAM-like, or DRAM-like, or flash-like. This tunability enables STT-MRAM to be fitted into the present memory hierarchy spanning from last level cache to solid-state storage. From an application point of view, STT-MRAM is a perfect memroy solution in both stand-alone and embedded forms for a variety of applications such as enterprise SSD, AIoT, automotive, and aerospace. Finally, the remaining challenges facing STT-MRAM are also discussed in this chapter.

Chapter 4: Testing STT-MRAM with Conventional Approach

This chapter presents a Verilog-A MTJ compact model for defect-free MTJ devices. It describes the electrical behaviors of physical PMA-MTJ devices by mapping device technology parameters to electrical ones. We have optimized and calibrated this model with silicon data measured on fabricated MTJ devices at imec, thus allowing fast and accurate SPICE circuit simulations along with other circuit elements such as transistors, resistors, and capacitors. In addition, this chapter explores STT-MRAM testing based on the conventional fault modeling and test approach. Specifically, all STT-MRAM manufacturing defects irrespective of their physical natures are modeled as linear resistors (i.e., opens, shorts, and bridges). These resistors are injected separately into our STT-MRAM circuit netlist and subsequently simulated for fault analysis. The fault modeling results suggest that resistive defects only lead to two types of fault: transition faults and incorrect read faults. Therefore, March tests such March C- can be used to detect them.

Chapter 5: Magnetic-Field-Aware Compact Model of pMTJ

This chapter presents a magnetic-field-aware compact model of pMTJ for magnetic/electrical co-simulation of STT-MRAM circuits, based on the fact that magnetic fields including internal intra- and inter-cell stray fields and external disturbance fields have a large impact on STT-MRAM performance. Our measurement results of MTJ devices reveal that the intra-cell stray field becomes larger as MTJ size scales down. We physically modeled the internal stray fields and calibrated the model with the measured silicon data. To quantitatively evaluate the coupling strength, we proposed the inter-cell magnetic coupling factor Ψ ($\Psi \in [0,1]$) as a indicator. The larger the Ψ value, the higher the inter-cell magnetic coupling strength. This magnetic coupling model has been implemented in Verilog-A and integrated into our compact MTJ model. Using this model, we demonstrated magnetic/electrical co-simulation of STT-MRAM full circuits under PVT variations and various magnetic configurations. Hence, our magnetic-field-aware compact MTJ model can be used for fast and robust device/circuit co-design of STT-MRAM under the simulation environment of existing commercial CAD tools.

Chapter 6: Device-Aware Test Approach

This chapter demonstrates a paradigm shift in defect and fault modeling for STT-MRAMs. It has been shown based on device measurements and circuit simulations using cali-

brated MTJ models that modeling MTJ-internal defects as linear resistors between or at the device terminals is inaccurate. Inaccurate defect models may result in unrealistic fault models, which in turn lead to test escapes and a waste of test time and resources. This motivated us to propose a new test approach: Device-Aware Test (DAT), which goes beyond cell-aware test and sets up a step towards high-quality ICs at DPPB level. Our proposed DAT approach consists of three steps: device-aware defect modeling, device-aware fault modeling, and device-aware test development. The defect modeling does not assume that a defect in a device can be modeled electrically as a linear resistor, but it rather incorporates the impact of the physical defect on the technology parameters of the device and thereafter on its electrical parameters. Once the defective electrical model is obtained, a systematic fault analysis based on SPICE circuit simulations is performed to derive accurate fault models within a pre-defined complete fault space. Finally, the derived fault models corresponding to this specific defect are used to develop test solutions.

Chapter 7: DAT for Pinhole Defects

This chapter applies DAT to pinhole defects which are seen as a key type of STT-MRAM manufacturing defects. Pinhole defects take place in the ultra-thin MgO tunnel barrier in MTJ devices due to imperfect deposition processes. This chapter presents comprehensive characterization on MTJ devices with pinhole defects both during manufacturing test (t=0) and in the field (t>0). A defective MTJ compact model with parameterized pinhole defect is developed and calibrated with the measured silicon data. This model is then used to derive accurate faults in the presence of a pinhole defect with different sizes. Our simulation results show that a large pinhole defect results in easy-to-detect faults (together equivalent to the traditional stuck-at-0 fault), while a small pinhole defect leads to hard-to-detect faults. The easy-to-detect faults can be detected by applying March tests. However, detecting the hard-to-detect faults require stress tests with hammering write 1 operations under elevated voltage and/or prolonged pulses. This test intentionally enlarges the pinhole under the effect of Joule heating and/or the electric field across the pinhole circumference, thus transforming the hard-to-detect faults to easy-to-detect faults.

Chapter 8: DAT for Synthetic Anti-Ferromagnetic Flip (SAFF) Defects

This chapter applies DAT to a new type of MTJ defect: Synthetic Anti-Ferromagnetic Flip (SAFF) defect. SAFF means that the magnetization in both the hard layer and reference layer of MTJ devices undergoes an unintended flip to the opposite direction. It takes place due to an initial reversal of hard layer with significantly reduced coercivity, which is attributed to inhomogeneities arising during device fabrication steps. Both magnetic and electrical measurement data of SAFF defect in fabricated MTJ devices is presented; it shows that such a defect reverses the polarity of stray field at the free layer of MTJ, while it has no electrical impact on the single isolated device. We demonstrate that using the conventional fault modeling and test approach fails to appropriately model and test such a defect. Therefore, our proposed DAT is applied. The fault modeling results show that a SAFF defect may lead to an intermittent Passive Neighborhood Pattern Sensitive Fault ($PNPSF1_i$) when all neighboring cells are in logic '1' state. Finally, two test

solutions are discussed. The first one is a March algorithm consisting of normal write and read operations, which cannot guarantee the detection of SAFF defect. In contrast, the second one is a new March algorithm, incorporating a magnetic write operation; it can guarantee the detection of SAFF defect with an affordable cost.

Chapter 9: DAT for Intermediate (IM) State Defects

This chapter applies DAT to intermediate (IM) state defect. The IM state defect manifest itself as an abnormal third resistive state apart from the two bi-stable states of MTJ. It takes place due to due to some physical imperfections such as unreversed magnetic bubbles, inhomogeneous distribution of stray field or even skyrmion generation. We performed silicon measurements on MTJ devices with diameter ranging from 60 nm to 120 nm. The results reveal that the occurrence probability of IM state strongly depends on the switching direction, device size, and applied voltage bias. We also demonstrate that the traditional fault modeling and test approach fails to accurately model this defect at the functional behavior; hence it fails in detecting such a defect during manufacturing tests. Therefore, our proposed DAT is applied. Our simulation results show that an IM state defect leads to intermittent write transition faults. To detect such faults and the underlying IM state defect, we propose and implement a DfT with weak write operations.

10.2. FUTURE RESEARCH DIRECTIONS

In this dissertation, we have investigated STT-MRAM defect space with special attention to MTJ-internal defects. Some of these defects have been characterized and modeled accurately. The obtained defect models are then used to develop accurate and realistic faults on our STT-MRAM circuit simulation platform. Test development is also covered for each type of STT-MRAM defect. Nevertheless, limited by time and resources, we have only covered several topics in STT-MRAM design and test. There exists many topics that need to be explored. These topics include, but not limited to, the following.

1) Characterization and modeling of MTJ-internal defects

As MTJ devices are the data-storing elements in STT-MRAMs, detecting MTJ-internal defects are the key to a high-quality manufacturing test. Therefore, accurate defect models are required for developing accurate and realistic fault models, which are typically the targets of test solutions. In this dissertation, we have addressed several MTJ-internal defects. But there still exists many defects that need to be characterized and modeled using the DAT approach as we did in the dissertation. Some examples are back-hopping, atom inter-diffusion, and interface roughness. Furthermore, the formation mechanism of defect is very much dependent on fabrication processes and tools which may differ between fabs. With the evolution of STT-MRAM technology as well as processing technology over time, the afore-mentioned STT-MRAM defects may disappear and new defects may arise. To generate an optimal test solution, understanding defects in MTJ devices in a specific manufacturing process plays a extremely important role.

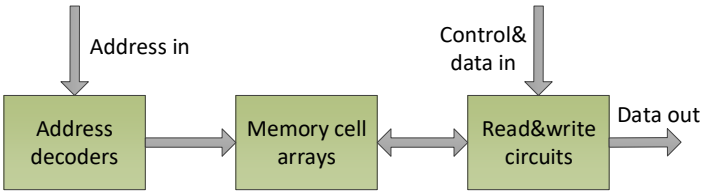


Figure 10.1: Reduced functional memory model.

2) Experimental analysis of peripheral circuits

A memory chip can be functionally reduced to three blocks: address decoders, memory cell arrays, and read&write circuits, as illustrated in Figure 10.1. Depending on the location where the fault take place, faults can be divided into three groups with each of which corresponding to a block in Figure 10.1. For STT-MRAM testing, all theoretical and experimental research including this dissertation has been concentrated on faults in STT-MRAM arrays. Nothing has been published on faults that occur in address decoders and read&write circuits. Therefore, investigating the faulty behavior in the presence of defects in these two circuit blocks, developing appropriate fault models, and eventually covering them in test programs are crucial.

3) DfT, mangetic test, stress/burn-in test, and BIST techniques

With the increase in STT-MRAM design complexity and capacity (1Gb standalone STT-MRAM chips were commercialized in 2019), making the chip test manageable will become an important aspect in STT-MRAM testing. In addition, March tests sometimes cannot detect some types of defect; an example is the IM state defect which leads to intermittent write transition faults. Therefore, developing and implementing DfT techniques to facilitate STT-MRAM testing is required.

For conventional memory technologies such as SRAM and DRAM, only electrical properties matter. When it comes to STT-MRAMs, magnetic properties are as important as electrical ones. Furthermore, magnetic fields can be utilized exclusively or in conjunction with electric current to switch the state of MTJ. Therefore, magnetic test techniques are of great interest to facilitate STT-MRAM testing. For example, we proposed a March algorithm combining the conventional electrical write operations and magnetic write operations to detect SAFF defect in MTJ devices. Whether magnetic tests are a must to detect other STT-MRAM defects is still an open question. Another interesting topic is to explore the possibility of using magnetic tests as a helper or accelerator for conventional manufacturing tests. A key motivation behind this idea is that applying a perpendicular magnetic field can significantly change the thermal stability of MTJ devices, thus changing write current, write time, and retention time.

Stress/burn-in tests with elevated voltage/temperature are widely used for reliability tests. For STT-MRAM testing, burn-in test with elevated temperature (i.e., baking in a oven) is a well known technique to characterize the retention time. In this dissertation, we have proposed a stress test with hammering write '1' operation sequence with elevated voltage or prolonged pulse width to detect small pinhole defects in the tunnel barrier of MTJ. For other STT-MRAM defects, stress/burn-in test techniques can also be

explored to examine its effectiveness.

BIST is an indispensable technique for memroy testing. STT-MRAM has been proven to be a competitive memory technology for both stand-alone and embedded applications. In embedded applications, neither the address, read/write and data input liens are controllable nor are the data output lines observable. BIST is a favorable solution for this problem. Furthermore, BIST also has the advantages of at-speed test, reduced test time, and elimination of external test equipment. All of these highlight the importance of developing BIST solutions for STT-MRAMs.

4) DAT applications in a broad test scope

This thesis mainly focuses on applying DAT to STT-MRAMs. DAT has also been proven to be an indispensable test approach for developing high-quality tests for RRAMs, as can be found in [67, 207, 228]. Apart from STT-MRAM and RRAM, it is expected that DAT can be applied to other memory technologies including advanced volatile memories (e.g., SRAM and DRAM) and non-volatile ones (e.g., PCM and FeRAM). Moreover, it can also be applied to logic circuits especially for technology nodes below 10 nm, where it has been shown that many failure mechanisms cannot be modeled with linear resistors [212].

REFERENCES

[1] Huawei Inc., "Huawei Kirin 990 SoC features," https://consumer.huawei.com/en/campaign/kirin-990-series/, accessed in Jun. 2020.

[2] M. Nicolaidis *et al.*, "On-line testing for VLSI—a compendium of approaches," *Journal of Electronic Testing*, vol. 12, no. 1-2, pp. 7–20, Feb. 1998, doi:10.1023/A:1008244815697.

[3] M. Bushnell *et al.*, *Essentials of electronic testing for digital, memory and mixed-signal VLSI circuits*. Springer Science & Business Media, 2004, vol. 17, https://link.springer.com/book/10.1007/b117406.

[4] V.B. Naik *et al.*, "Manufacturable 22nm FD-SOI embedded MRAM technology for industrial-grade MCU and IOT applications," in *IEEE Int. Electron Devices Meeting*, Dec. 2019, pp. 2.3.1–2.3.4, doi:10.1109/IEDM19573.2019.8993454.

[5] K.K. Chang *et al.*, "Understanding latency variation in modern dram chips: Experimental characterization, analysis, and optimization," *SIGMETRICS Perform. Eval. Rev.*, vol. 44, no. 1, p. 323–336, Jun. 2016, doi:10.1145/2964791.2901453.

[6] J.L. Hennessy *et al.*, *Computer architecture: a quantitative approach (Fifth Edition)*. Elsevier, 2012, http://acs.pub.ro/~cpop/SMPA/Computer%20Architecture%20A%20Quantitative%20Approach%20(5th%20edition).pdf, accessed in Jul., 2020.

[7] C. Matsui *et al.*, "Design of hybrid ssds with storage class memory and nand flash memory," *Proceedings of the IEEE*, vol. 105, no. 9, pp. 1812–1821, Sep. 2017, doi:10.1109/JPROC.2017.2716958.

[8] G.W. Burr *et al.*, "Overview of candidate device technologies for storage-class memory," *IBM Journal of Research and Development*, vol. 52, no. 4.5, pp. 449–464, Jul. 2008, doi:10.1147/rd.524.0449.

[9] Y. Chen *et al.*, "Recent technology advances of emerging memories," *IEEE Design & Test*, vol. 34, no. 3, pp. 8–22, 2017, doi:10.1109/MDAT.2017.2685381.

[10] A. Chen, "A review of emerging non-volatile memory (NVM) technologies and applications," *Solid-State Electronics*, vol. 125, pp. 25–38, Nov. 2016, doi:10.1016/j.sse.2016.07.006.

[11] H. Noguchi *et al.*, "Variable nonvolatile memory arrays for adaptive computing systems," *IEEE Int. Electron Devices Meeting*, pp. 617–620, Dec. 2013, doi:10.1109/IEDM.2013.6724690.

[12] Y. Choi *et al.*, "A 20nm 1.8V 8Gb PRAM with 40MB/s program bandwidth," in *IEEE Int. Solid-State Circuits Conf.*, Feb. 2012, pp. 46–48, doi:10.1109/ISSCC.2012.6176872.

[13] S. Sills *et al.*, "High-density reRAM for storage class memory," in *15th Non-Volatile Memory Tech. Symp.*, Oct. 2015, pp. 1–4, doi:10.1109/NVMTS.2015.7499973.

[14] K. Lee *et al.*, "1Gbit high density embedded STT-MRAM in 28nm FDSOI technology," in *IEEE Int. Electron Devices Meeting*, Dec. 2019, pp. 2.2.1–2.2.4, doi:10.1109/IEDM19573.2019.8993551.

[15] S. Aggarwal *et al.*, "Demonstration of a reliable 1 Gb standalone spin-transfer torque MRAM for industrial applications," in *IEEE Int. Electron Devices Meeting*, Dec. 2019, pp. 2.1.1–2.1.4, doi:10.1109/IEDM19573.2019.8993516.

[16] Y. Huai *et al.*, "High density 3D cross-point STT-MRAM," in *IEEE Int. Memory Workshop*, May 2018, pp. 1–4, doi:10.1109/IMW.2018.8388833.

[17] G.W. Burr *et al.*, "Recent progress in phase-change memory technology," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 6, no. 2, pp. 146–162, Apr. 2016, doi:10.1109/JETCAS.2016.2547718.

[18] S.W. Fong *et al.*, "Phase-change memory—towards a storage-class memory," *IEEE Transactions on Electron Devices*, vol. 64, no. 11, pp. 4374–4385, Sep. 2017, doi:10.1109/TED.2017.2746342.

[19] C. Villa *et al.*, "A 45nm 1Gb 1.8V phase-change memory," in *IEEE Int. Solid-State Circuits Conf.*, Feb. 2010, pp. 270–271, doi:10.1109/ISSCC.2010.5433916.

[20] Intel, "Intel Optane technology and products," https://www.intel.com/content/www/us/en/architecture-and-technology/optane-dc-persistent-memory.html, accessed in Jun. 2020.

[21] H.S.P. Wong *et al.*, "Metal–oxide RRAM," *Proceedings of the IEEE*, vol. 100, no. 6, pp. 1951–1970, May 2012, doi:10.1109/JPROC.2012.2190369.

[22] H. Wu *et al.*, "Resistive random access memory for future information processing system," *Proceedings of the IEEE*, vol. 105, no. 9, pp. 1770–1789, May 2017, doi:10.1109/JPROC.2017.2684830.

[23] Panasonic, "Panasonic starts world's first mass production of ReRAM mounted microcomputers," https://news.panasonic.com/global/press/data/2013/07/en130730-2/en130730-2.html, accessed in Jun. 2020.

[24] R. Fackenthal *et al.*, "19.7 A 16Gb ReRAM with 200MB/s write and 1GB/s read in 27nm technology," in *IEEE Int. Solid-State Circuits Conf.*, Feb. 2014, pp. 338–339, doi:10.1109/ISSCC.2014.6757460.

[25] T. Liu *et al.*, "A 130.7-mm^2 2-layer 32-Gb ReRAM memory device in 24-nm technology," *IEEE Journal of Solid-State Circuits*, vol. 49, no. 1, pp. 140–153, Sep. 2014, doi:10.1109/JSSC.2013.2280296.

[26] D. Apalkov *et al.*, "Magnetoresistive random access memory," *Proceedings of the IEEE*, vol. 104, no. 10, pp. 1796–1830, Aug. 2016, doi:10.1109/JPROC.2016.2590142. http://ieeexplore.ieee.org/document/7555318/

[27] Everspin Technologies, "Toggle MRAM commerical products from everspin technologies," https://www.everspin.com/toggle-mram-technology, accessed in Jun. 2020.

[28] S. Ikegawa *et al.*, "Magnetoresistive random access memory: Present and future," *IEEE Trans. Electron Devices*, vol. 67, no. 4, pp. 1407–1419, Jan. 2020, doi:10.1109/TED.2020.2965403.

[29] Avalanche Technology, "Avalanche STT-MRAM products," http://www.avalanche-technology.com/products/, accessed in Jun. 2020.

[30] L. Wei *et al.*, "A 7Mb STT-MRAM in 22FFL FinFET technology with 4ns read sensing time at 0.9V using write-verify-write scheme and offset-cancellation sensing technique," in *IEEE Int. Solid-State Circuits Conf.*, Feb. 2019, pp. 214–216, doi:10.1109/ISSCC.2019.8662444.

[31] N. Sato, "CMOS compatible process integration of SOT-MRAM with heavy-metal bi-layer bottom electrode and 10ns field-free SOT switching with STT assist," in *IEEE Symp. VLSI Technol.*, Jun. 2020, pp. 1–2, doi:10.1109/VLSITechnology18217.2020.9265028.

[32] W.J. Gallagher *et al.*, "22nm STT-MRAM for reflow and automotive uses with high yield, reliability, and magnetic immunity and with performance and shielding options," in *IEEE Int. Electron Devices Meeting*, Dec. 2019, pp. 2.7.1–2.7.4, doi:10.1109/IEDM19573.2019.8993469.

[33] T. Coughlin, "The growing market for MRAMs," http://electronicdesign.com/products/industry-first-1gb-ddr4-perpendicular-st-mram-device-introduced, accessed in Jun. 2020.

[34] K. Garello *et al.*, "Manufacturable 300mm platform solution for field-free switching SOT-MRAM," in *IEEE Symp. VLSI Circuits*, Jun. 2019, pp. T194–T195, doi:10.23919/VLSIC.2019.8778100.

[35] M.A. Breuer *et al.*, *Diagnosis & reliable design of digital syst.*, ser. Digital syst. design series. Computer Science Press, 1976.

[36] A.J. Van de Goor, *Testing semiconductor memories: theory and practice*. John Wiley & Sons, Inc., 1991.

[37] I. Schanstra *et al.*, "Industrial evaluation of stress combinations for march tests applied to SRAMs," in *IEEE Int. Test Conf.*, Aug. 1999, pp. 983–992, doi:10.1109/test.1999.805831.

[38] A.J. van de Goor *et al.*, "Industrial evaluation of DRAM tests," in *Design, Autom. & Test in Europe Conf. & Exhib.*, Mar. 1999, pp. 623–630, doi:10.1109/DATE.1999.761194.

[39] S. Hamdioui *et al.*, "An experimental analysis of spot defects in SRAMs: realistic fault models and tests," in *IEEE Asian Test Symp.*, Dec. 2000, pp. 131–138, doi:10.1109/ATS.2000.893615.

[40] E.I. Vatajelu *et al.*, "Analyzing resistive-open defects in SRAM core-cell under the effect of process variability," in *IEEE Eur. Test. Symp.*, May 2013, pp. 1–6, doi:10.1109/ETS.2013.6569373.

[41] F. Hapke *et al.*, "Defect-oriented cell-internal testing," in *IEEE Int. Test Conf.*, Nov. 2010, pp. 1–10, doi:10.1109/TEST.2010.5699229.

[42] Z. Gao *et al.*, "Application of cell-aware test on an advanced 3nm CMOS technology library," in *IEEE Int. Test Conf.*, Nov. 2019, pp. 1–6, doi:10.1109/ITC44170.2019.9000164.

[43] TSMC, "TSMC 5nm process technology in volume production," https://www.tsmc.com/english/dedicatedFoundry/technology/logic/l_5nm, accessed in Jun. 2020.

[44] S. Borkar, "Microarchitecture and design challenges for gigascale integration," in *37th Int. Symp. Microarchitecture*, Dec. 2005, pp. 3–3, doi:10.1109/micro.2004.24.

[45] A.N. Bhoj *et al.*, "Fault models for logic circuits in the multigate era," *IEEE Trans. Nanotech.*, vol. 11, no. 1, pp. 182–193, Jan. 2012, doi:10.1109/TNANO.2011.2169807.

[46] L. Wu *et al.*, "Defect and fault modeling framework for STT-MRAM testing," *IEEE Trans. Emerg. Topics Comput.*, pp. 1–15, Dec. 2019, doi:10.1109/TETC.2019.2960375.

[47] L. Wu *et al.*, "Electrical modeling of STT-MRAM defects," in *IEEE Int. Test Conf.*, Oct. 2018, pp. 1–10, doi:10.1109/TEST.2018.8624749.

[48] P. Girard *et al.*, "A survey of test and reliability solutions for magnetic random access memories," *Proc. IEEE*, pp. 1–21, Oct. 2020, doi:10.1109/JPROC.2020.3029600.

[49] C.L. Su *et al.*, "MRAM defect analysis and fault modeling," in *IEEE Int. Test Conf.*, Oct. 2004, pp. 124–133, doi:10.1109/TEST.2004.1386944.

[50] C. Su *et al.*, "Testing MRAM for write disturbance fault," in *IEEE Int. Test Conf.*, Oct. 2006, pp. 1–9, doi:10.1109/TEST.2006.297702.

[51] J. Azevedo *et al.*, "Impact of resistive-bridge defects in TAS-MRAM architectures," *IEEE Asian Test Symp.*, pp. 125–130, 2012, doi:10.1109/ATS.2012.37.

[52] J. Azevedo *et al.*, "A complete resistive-open defect analysis for thermally assisted switching MRAMs," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 22, no. 11, pp. 2326–2335, Nov. 2014, doi:10.1109/TVLSI.2013.2294080.

[53] A. Chintaluri *et al.*, "A model study of defects and faults in embedded spin transfer torque (STT) MRAM arrays," in *IEEE Asian Test Symp.*, Nov. 2015, pp. 187–192, doi:10.1109/ATS.2015.39.

[54] A. Chintaluri *et al.*, "Analysis of defects and variations in embedded spin transfer torque (STT) MRAM arrays," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 6, no. 3, pp. 319–329, Sep. 2016, doi:10.1109/JETCAS.2016.2547779.

[55] I. Yoon *et al.*, "EMACS: efficient MBIST architecture for test and characterization of STT-MRAM arrays," in *IEEE Int. Test Conf.*, Nov. 2016, pp. 1–10, doi:10.1109/TEST.2016.7805834.

[56] S.M. Nair *et al.*, "Defect injection, fault modeling and test algorithm generation methodology for STT-MRAM," in *IEEE Int. Test Conf.*, Oct. 2018, pp. 1–10, doi:10.1109/TEST.2018.8624725.

[57] G. Radhakrishnan *et al.*, "A parametric DFT scheme for STT-MRAMs," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 27, no. 7, pp. 1685–1696, Jul. 2019, doi:10.1109/TVLSI.2019.2907549.

[58] G. Radhakrishnan *et al.*, "Monitoring aging defects in STT-MRAMs," *IEEE Trans. Comput.-Aided Design of Integr. Circuits and Syst.*, pp. 1–1, Mar. 2020, doi:10.1109/TCAD.2020.2982145.

[59] E.I. Vatajelu *et al.*, "Challenges and solutions in emerging memory testing," *IEEE Trans. Emerg. Topics Comput.*, vol. 7, no. 3, pp. 493–506, Jul.–Sep. 2019, doi:10.1109/TETC.2017.2691263.

[60] R. Bishnoi *et al.*, "Read disturb fault detection in STT-MRAM," in *IEEE Int. Test Conf.*, Oct. 2014, pp. 1–7, doi:10.1109/TEST.2014.7035342.

[61] L. Wu *et al.*, "Survey on STT-MRAM testing: Failure mechanisms, fault models, and tests," *arXiv preprint*, pp. 1–24, Jan. 2020, arXiv:2001.05463.

[62] L. Wu *et al.*, "Pinhole defect characterization and fault modeling for STT-MRAM testing," in *IEEE Eur. Test Symp.*, May 2019, pp. 1–6, doi:10.1109/ETS.2019.8791518.

[63] L. Wu *et al.*, "Characterization, modeling and test of synthetic anti-ferromagnet flip defect in STT-MRAMs," in *IEEE Int. Test Conf.*, Nov. 2020, pp. 1–10, doi:10.1109/ITC44778.2020.9325258.

[64] L. Wu *et al.*, "Characterization and fault modeling of intermediate state defects in STT-MRAM," in *Design, Autom. & Test in Europe Conf. & Exhib.*, Feb. 2021, pp. 1–6.

[65] L. Wu *et al.*, "Impact of magnetic coupling and density on STT-MRAM performance," in *Design, Autom. & Test in Europe Conf. & Exhib.*, Mar. 2020, pp. 1211–1216, doi:10.23919/DATE48585.2020.9116444.

[66] L. Wu *et al.*, "Device-aware test for emerging memories: Enabling your test program for DPPB level," in *IEEE Eur. Test Symp.*, May 2020, pp. 1–2, doi:10.1109/ETS48528.2020.9131559.

[67] M. Fieback *et al.*, "Device-aware test: A new test approach towards DPPB level," in *IEEE Int. Test Conf.*, Nov. 2019, pp. 1–10, doi:10.1109/ITC44170.2019.9000134.

[68] L. Wu *et al.*, "Characterization and fault modeling of intermediate state defects in STT-MRAM," *IEEE Trans. Comput.*, pp. 1–14, 2020, under review.

[69] Z. Al-Ars, "DRAM fault analysis and test generation," Ph.D. dissertation, Delft University of Technology, Jun. 2005, pp. 26–27, http://ce-publications.et.tudelft.nl/publications/842_dram_fault_analysis_and_test_generation.pdf.

[70] Everspin Technologies, "EMD4E001G– 1Gb spin-transfer torque MRAM datasheet (Rev. 1.2)," Sep. 2020, https://www.everspin.com/family/emd4e001g?npath=3557, accessed in Nov. 2020.

[71] JEDEC, "DDR4 synchronous DRAM standard: JESD79-4C," https://www.jedec.org/standards-documents/docs/jesd79-4a, accessed in Jun. 2020.

[72] B. Keeth *et al.*, *DRAM circuit design: fundamental and high-speed topics*. John Wiley & Sons, 2007, vol. 13, http://ce-publications.et.tudelft.nl/publications/842_dram_fault_analysis_and_test_generation.pdf.

[73] O. Mutlu, "Computer architecture: Main memory," https://course.ece.cmu.edu/~ece740/f11/lib/exe/fetch.php?media=wiki:lectures:onur-740-fall11-lecture25-mainmemory.pdf, accessed in Nov. 2020.

[74] Micron, "Technical note: DDR3 ZQ calibration introduction," https://www.micron.com/support#SupportDocumentationandDownloads, accessed in Nov. 2020.

[75] A.V. Khvalkovskiy *et al.*, "Basic principles of STT-MRAM cell operation in memory arrays," *J. Phys. D: Appl. Phys*, vol. 46, no. 13, p. 139601, Feb. 2013, doi:10.1088/0022-3727/46/13/139601.

[76] G.S. Kar *et al.*, "Co/Ni based p-MTJ stack for sub-20nm high density stand alone and high performance embedded memory application," in *IEEE Int. Electron Devices Meeting*, Dec. 2014, pp. 19.1.1–19.1.4, doi:10.1109/IEDM.2014.7047080.

[77] D.W. Abraham *et al.*, "Rapid-turnaround characterization methods for MRAM development," *IBM Journal of Research and Development*, vol. 50, no. 1, pp. 55–67, 2006, doi:10.1147/rd.501.0055.

[78] S. Van Beek *et al.*, "Impact of processing and stack optimization on the reliability of perpendicular STT-MRAM," in *IEEE Int. Reliability Phys. Symp.*, Apr. 2017, pp. 5A–1.1–5A–1.5, doi:10.1109/IRPS.2017.7936318.

[79] O. Golonzka *et al.*, "MRAM as embedded non-volatile memory solution for 22FFL FinFET technology," in *IEEE Int. Electron Devices Meeting*, Dec. 2018, pp. 18.1.1–18.1.4, doi:10.1109/IEDM.2018.8614620.

[80] Y.C. Wu *et al.*, "Impact of operating temperature on the electrical and magnetic properties of the bottom-pinned perpendicular magnetic tunnel junctions," *Applied Physics Letters*, vol. 113, no. 14, p. 142405, Oct. 2018, doi:10.1063/1.5042028.

[81] X. Fong *et al.*, "Spin-transfer torque memories: devices, circuits, and systems," *Proc. IEEE*, vol. 104, no. 7, pp. 1449–1488, 2016, doi:10.1109/JPROC.2016.2521712.

[82] N. Strikos *et al.*, "Low-current probabilistic writes for power-efficient STT-RAM caches," in *IEEE Int. Conf. Comput. Design*, Oct. 2013, pp. 511–514, doi:10.1109/ICCD.2013.6657095.

[83] R. Bowles *et al.*, *STTRAM Scaling and Retention Failure*, May 2013, vol. 17, no. 1. ISBN 9781934053560. https://www.intel.com/content/dam/www/public/us/en/documents/research/2013-vol17-iss-1-intel-technology-journal.pdf#page=54

[84] A. Jog *et al.*, "Cache revive: architecting volatile STT-RAM caches for enhanced performance in CMPs," in *Proc. Annual Design Autom. Conf.*, Jun. 2012, pp. 243–252, doi:10.1145/2228360.2228406.

[85] L. Tillie *et al.*, "Data retention extraction methodology for perpendicular STT-MRAM," in *IEEE Int. Electron Devices Meeting*, Dec. 2016, pp. 27.3.1–27.3.4, doi:10.1109/IEDM.2016.7838492.

[86] K. Lee *et al.*, "22-nm FD-SOI embedded MRAM with full solder reflow compatibility and enhanced magnetic immunity," in *IEEE Symp. VLSI Technol.*, Jun. 2018, pp. 183–184, doi:10.1109/VLSIT.2018.8510655.

[87] W.J. Gallagher *et al.*, "Recent progress and next directions for embedded MRAM technology," in *IEEE Symp VLSI Technol.*, Jun. 2019, pp. T190–T191, doi:10.23919/VLSIT.2019.8776547.

[88] H. Jin *et al.*, "Tunnel magnetoresistance effect," in *The Physics of Ferromagnetism*. Springer Series in Materials Science, 2012, vol. 158, pp. 403–432, doi:10.1007/978-3-642-25583-0.

[89] D. Wang *et al.*, "70% TMR at room temperature for SDT sandwich junctions with CoFeB as free and reference layers," *IEEE Trans. Magnetics*, vol. 40, no. 4, pp. 2269–2271, Aug. 2004, doi:10.1109/TMAG.2004.830219.

[90] S. Ikeda *et al.*, "Tunnel magnetoresistance of 604% at 300k by suppression of ta diffusion in CoFeB\MgO\CoFeB pseudo-spin-valves annealed at high temperature," *Appl. Phys. Lett.*, vol. 93, no. 8, p. 082508, Aug. 2008, doi:10.1063/1.2976435.

[91] Y. Wang *et al.*, "Reliability analysis of spintronic device based logic and memory circuits," Ph.D. dissertation, Télécom ParisTech, Apr. 2017, https://pastel.archives-ouvertes.fr/tel-01743849/.

[92] H. Noguchi *et al.*, "A 3.3ns-access-time 71.2μW/MHz 1Mb embedded STT-MRAM using physically eliminated read-disturb scheme and normally-off memory architecture," *IEEE Int. Solid-State Circuits Conf.*, vol. 58, pp. 1–3, Feb. 2015, doi:10.1109/ISSCC.2015.7062963.

[93] C.J. Lin *et al.*, "45nm low power CMOS logic compatible embedded STT MRAM utilizing a reverse-connection 1T/1MTJ cell," in *IEEE Int. Electron Devices Meeting*, Dec. 2009, pp. 1–4, doi:10.1109/IEDM.2009.5424368.

[94] Y.M. Lee *et al.*, "Highly scalable STT-MRAM with MTJs of top-pinned structure in 1T/1MTJ Cell," in *IEEE Symp. VLSI Technol.*, Jun. 2010, pp. 49–50, doi:10.1109/VLSIT.2010.5556123.

[95] D. Lee *et al.*, "High-performance low-energy STT MRAM based on balanced write scheme," in *ACM Int. Symp. on Low Power Electron. and Design*, Jul. 2012, pp. 9–14, doi:10.1145/2333660.2333665.

[96] A.K. Jones *et al.*, "Asymmetry of MTJ switching and its implication to STT-RAM designs," in *Design, Autom. & Test in Europe Conf. & Exhib.*, Mar. 2012, pp. 1313–1318, doi:10.1109/DATE.2012.6176695.

[97] W. Zhao *et al.*, "Design considerations and strategies for high-reliable STT-MRAM," *Microelectronics Rel.*, vol. 51, no. 9-11, pp. 1454–1458, Sep.-Nov. 2011, doi:10.1016/j.microrel.2011.07.001.

[98] Nanoscale Integration and Modeling (NIMO) Group at ASU, "Predictive technology model," http://ptm.asu.edu/, accessed in Jun. 2019.

[99] L. Zhang *et al.*, "A 16 kb spin-transfer torque random access memory with self-enable switching and precharge sensing schemes," *IEEE Trans. Magnetics*, vol. 50, no. 4, Apr. 2014, doi:10.1109/TMAG.2013.2291222.

[100] D. Suzuki *et al.*, "Cost-efficient self-terminated write driver for spin-transfer-torque RAM and logic," *IEEE Trans. Magnetics*, vol. 50, no. 11, pp. 1–4, Nov. 2014, doi:10.1109/TMAG.2014.2322387.

[101] R. Bishnoi *et al.*, "Self-timed read and write operations in STT-MRAM," *IEEE Trans. Very Large Scale Integ. (VLSI) Systems*, vol. 24, no. 5, pp. 1783–1793, May 2016, doi:10.1109/TVLSI.2015.2496363.

[102] Q. Dong *et al.*, "A 1Mb 28nm STT-MRAM with 2.8ns read access time at 1.2V VDD using single-cap offset-cancelled sense amplifier and in-situ self-write-termination," in *IEEE Int. Solid-State Circuits Conf.*, Feb. 2018, pp. 480–482, doi:10.1109/ISSCC.2018.8310393.

[103] W. Zhao *et al.*, "High speed, high stability and low power sensing amplifier for mtj/cmos hybrid logic circuits," *IEEE Trans. Magnetics*, vol. 45, no. 10, pp. 3784–3787, Oct. 2009, doi:10.1109/TMAG.2009.2024325.

[104] M. Jefremow *et al.*, "Time-differential sense amplifier for sub-80mV bitline voltage embedded STT-MRAM in 40nm CMOS," *IEEE Int. Solid-State Circuits Conf.*, vol. 56, pp. 216–217, Feb. 2013, doi:10.1109/ISSCC.2013.6487706.

[105] H. Lee *et al.*, "Design of a fast and low-power sense amplifier and writing circuit for high-speed MRAM," *IEEE Trans. Magnetics*, vol. 51, no. 5, May 2015, doi:10.1109/TMAG.2014.2367130.

[106] C. Kim *et al.*, "A covalent-bonded cross-coupled current-mode sense amplifier for STT-MRAM with 1T1MTJ common source-line structure array," *IEEE Int. Solid-State Circuits Conf.*, vol. 58, pp. 134–135, Feb. 2015, doi:10.1109/ISSCC.2015.7062962.

[107] D. Zhang *et al.*, "Reliability-enhanced separated pre-charge sensing amplifier for hybrid CMOS/MTJ logic circuits," *IEEE Trans. Magnetics*, vol. 53, no. 9, pp. 1–5, Sep. 2017, doi:10.1109/TMAG.2017.2702743.

[108] WikiChip, "Intel expands 22FFL with production-ready RRAM and MRAM on FinFET," https://fuse.wikichip.org/news/2801/intel-expands-22ffl-with-production-ready-rram-and-mram-on-finfet/, accessed in Jun. 2020.

[109] Y.J. Song *et al.*, "Highly functional and reliable 8Mb STT-MRAM embedded in 28nm logic," in *IEEE Int. Electron Devices Meeting*, Dec. 2016, doi:10.1109/IEDM.2016.7838491.

[110] D. Shum *et al.*, "CMOS-embedded STT-MRAM arrays in 2x nm nodes for GP-MCU applications," in *Symp. VLSI Technol.*, Jun. 2017, pp. T208–T209, doi:10.23919/VLSIT.2017.7998174.

[111] Y.G. Fedorenko, "Ion-beam-induced defects in CMOS technology: methods of study," *arXiv preprint*, pp. 1–32, Apr. 2017, arXiv:1701.07087.

[112] M. Sachdev *et al.*, *Defect-oriented testing for nano-metric CMOS VLSI circuits*. Springer Science & Business Media, 2007.

[113] J.C.M. Li *et al.*, "Diagnosis of resistive-open and stuck-open defects in digital CMOS ICs," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 24, no. 11, pp. 1748–1759, Nov. 2005, doi:10.1109/TCAD.2005.852457.

[114] N.Z. Haron *et al.*, "On defect oriented testing for hybrid CMOS/Memristor memory," in *IEEE Asian Test Symp.*, Nov. 2011, pp. 353–358, doi:10.1109/ATS.2011.66.

[115] H. Sato *et al.*, "Properties of magnetic tunnel junctions with a MgO/CoFeB/Ta/CoFeB/MgO recording structure down to junction diameter of 11 nm," *Appl. Phys. Lett.*, vol. 105, no. 6, p. 062403, Jul. 2014, doi:10.1063/1.4892924.

[116] W. Zhao *et al.*, "Failure analysis in magnetic tunnel junction nanopillar with interfacial perpendicular magnetic anisotropy," *Materials*, vol. 9, no. 1, pp. 1–17, Jan. 2016, doi:10.3390/ma9010041.

[117] B. Oliver *et al.*, "Two breakdown mechanisms in ultrathin alumina barrier magnetic tunnel junctions," *J. Appl. Phys.*, vol. 95, no. 3, pp. 1315–1322, Jan. 2004, doi:10.1063/1.1636255.

[118] Y. Wang *et al.*, "Compact model of dielectric breakdown in spin-transfer torque magnetic tunnel junction," *IEEE Trans. Electron Devices*, vol. 63, no. 4, pp. 1762–1767, Apr. 2016, doi:10.1109/TED.2016.2533438.

[119] H. Meng *et al.*, "Annealing effects on CoFeB-MgO magnetic tunnel junctions with perpendicular anisotropy," *J. Appl. Phys.*, vol. 110, no. 3, p. 033904, Aug. 2011, doi:10.1063/1.3611426.

[120] H. Maehara *et al.*, "Tunnel Magnetoresistance above 170% and resistance–area product of 1 Ω (μm^2) attained by in situ annealing of ultra-thin MgO tunnel barrier," *Appl. Phys. Express*, vol. 4, no. 3, p. 033002, Mar. 2011, doi:10.1143/apex.4.033002.

[121] B. Bhusan Singh *et al.*, "Effect of MgO spacer and annealing on interface and magnetic properties of ion beam sputtered NiFe/Mg/MgO/CoFe layer structures," *J. Appl. Phys.*, vol. 112, no. 6, p. 063906, Sep. 2012, doi:10.1063/1.4752264.

[122] J.G. Park *et al.*, "Challenging issues for terra-bit-level perpendicular STT-MRAM," in *IEEE Int. Electron Devices Meeting*, Dec. 2014, pp. 19.2.1–19.2.4, doi:10.1109/IEDM.2014.7047081.

[123] W. Boullart *et al.*, "STT MRAM patterning challenges," in *Proc. SPIE 8685, Advanced Etch Technology for Nanopatterning II*, vol. 8685, Mar. 2013, p. 86850F, doi:10.1117/12.2013602.

[124] K. Sugiura *et al.*, "Ion beam etching technology for high-density spin transfer torque magnetic random access memory," *Jpn. J. Appl. Phys.*, vol. 48, no. 8 Part 2, Aug. 2009, doi:10.1143/JJAP.48.08HD02.

[125] K. Nagahara *et al.*, "Ion-beam-etched profile control of MTJ cells for improving the switching characteristics of high-density MRAM," *IEEE Trans. Magnetics*, vol. 42, no. 10, pp. 2745–2747, Oct. 2006, doi:10.1109/TMAG.2006.878862.

[126] K. Nagahara *et al.*, "Magnetic tunnel junction (MTJ) patterning for magnetic random access memory (MRAM) process applications," *Jpn. J. Appl. Phys.*, vol. 42, no. Part 2, No. 5B, pp. L499–L501, May 2003, doi:10.1143/jjap.42.l499.

[127] E.H. Kim *et al.*, "Evolution of etch profile of magnetic tunnel junction stacks etched in a CH3OH/Ar plasma," *Journal of The Electrochemical Society*, vol. 159, no. 3, pp. H230–H234, Jan. 2012, doi:10.1149/2.012203jes.

[128] A.A. Garay *et al.*, "Inductively coupled plasma reactive ion etching of magnetic tunnel junction stacks in a CH3COOH/Ar gas," *ECS Solid State Letters*, vol. 4, no. 10, pp. P77–P79, Aug. 2015, doi:10.1149/2.0071510ssl.

[129] Y.H. Wang *et al.*, "Impact of stray field on the switching properties of perpendicular MTJ for scaled MRAM," in *IEEE Int. Electron Devices Meeting*, Dec. 2012, pp. 29.2.1–29.2.4, doi:10.1109/IEDM.2012.6479127.

[130] H. Jiancheng *et al.*, "Effect of the stray field profile on the switching characteristics of the free layer in a perpendicular magnetic tunnel junction," *J. Appl. Phys.*, vol. 117, no. 17, p. 17B721, Mar. 2015, doi:10.1063/1.4916037.

[131] J.H. Jeong *et al.*, "Novel oxygen showering process (OSP) for extreme damage suppression of sub-20nm high density p-MTJ array without IBE treatment," in *Symp. VLSI Technol.*, Jun. 2015, pp. T158–T159, doi:10.1109/VLSIT.2015.7223660.

[132] T. Devolder *et al.*, "Size dependence of nanosecond-scale spin-torque switching in perpendicularly magnetized tunnel junctions," *Phys. Rev. B*, vol. 93, p. 224432, Jun. 2016, doi:10.1103/PhysRevB.93.224432.

[133] X. Zhang *et al.*, "Skyrmions in magnetic tunnel junctions," *ACS Appl. Mater. Interfaces*, vol. 10, no. 19, pp. 16 887–16 892, May 2018, doi:10.1021/acsami.8b03812.

[134] W. Kim *et al.*, "Experimental observation of back-hopping with reference layer flipping by high-voltage pulse in perpendicular magnetic tunnel junctions," *IEEE Trans. Magnetics*, vol. 52, no. 7, pp. 1–4, Feb. 2016, doi:10.1109/TMAG.2016.2536158.

[135] G. Hu *et al.*, "Reliable five-nanosecond writing of spin-transfer torque magnetic random-access memory," *IEEE Magnetics Letters*, vol. 10, pp. 1–4, Jul. 2019, doi:10.1109/LMAG.2019.2928243.

[136] M. Julliere, "Tunneling between ferromagnetic films," *Physics Letters A*, vol. 54, no. 3, pp. 225–226, Sep. 1975, doi:10.1016/0375-9601(75)90174-7.

[137] J.S. Moodera *et al.*, "Large magnetoresistance at room temperature in ferromagnetic thin film tunnel junctions," *Phys. Rev. Lett.*, vol. 74, pp. 3273–3276, Apr. 1995, doi:10.1103/PhysRevLett.74.3273.

[138] W.H. Butler *et al.*, "Spin-dependent tunneling conductance of Fe|MgO|Fe sandwiches," *Physical Review B*, vol. 63, no. 5, p. 054416, Jan. 2001, doi:10.1103/PhysRevB.63.054416.

[139] S. Yuasa *et al.*, "Giant room-temperature magnetoresistance in single-crystal Fe/MgO/Fe magnetic tunnel junctions," *Nature Materials*, vol. 3, no. 12, pp. 868–871, Oct. 2004, doi:10.1038/nmat1257.

[140] R. Scheuerlein *et al.*, "A 10 ns read and write non-volatile memory array using a magnetic tunnel junction and FET switch in each cell," in *IEEE Int. Solid-State Circuits Conf.*, Aug. 2000, pp. 128–129, doi:10.1109/ISSCC.2000.839717.

[141] L. Savtchenko *et al.*, "Method of writing to scalable magnetoresistance random access memory element," Apr. 8 2003, uS Patent 6,545,906.

[142] I.L. Prejbeanu *et al.*, "Thermally assisted switching in exchange-biased storage layer magnetic tunnel junctions," *IEEE Trans. Magnetics*, vol. 40, no. 4, pp. 2625–2627, Aug. 2004, doi:10.1109/TMAG.2004.830395.

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

[144] J. Slonczewski, "Current-driven excitation of magnetic multilayers," *Journal of Magnetism and Magnetic Materials*, vol. 159, no. 1-2, pp. L1–L7, 1996, doi:10.1016/0304-8853(96)00062-5.

[145] J.A. Katine *et al.*, "Current-driven magnetization reversal and spin-wave excitations in Co/Cu/Co pillars," *Phys. Rev. Lett.*, vol. 84, pp. 3149–3152, Apr. 2000, doi:10.1103/PhysRevLett.84.3149.

[146] F.J. Albert *et al.*, "Spin-polarized current switching of a Co thin film nanomagnet," *Appl. Phys. Lett.*, vol. 77, no. 23, pp. 3809–3811, Oct. 2000, doi:10.1063/1.1330562.

[147] Z. Diao *et al.*, "Spin transfer switching and spin polarization in magnetic tunnel junctions with MgO and AlO_x barriers," *Appl. Phys. Lett.*, vol. 87, no. 23, p. 232502, Oct. 2005, doi:10.1063/1.2139849.

[148] G. Hu *et al.*, "Spin-transfer torque MRAM with reliable 2 ns writing for last level cache applications," in *IEEE Int. Electron Devices Meeting*, Dec. 2019, pp. 2.6.1–2.6.4, doi:10.1109/IEDM19573.2019.8993604.

[149] A. Chernyshov *et al.*, "Evidence for reversible control of magnetization in a ferromagnetic material by means of spin–orbit magnetic field," *Nature Physics*, vol. 5, no. 9, pp. 656–659, Aug. 2009, doi:10.1038/nphys1362.

[150] I.M. Miron *et al.*, "Perpendicular switching of a single ferromagnetic layer induced by in-plane current injection," *Nature*, vol. 476, no. 7359, pp. 189–193, Aug. 2011, doi:10.1038/nature10309.

[151] T. Nozaki *et al.*, "Voltage-induced perpendicular magnetic anisotropy change in magnetic tunnel junctions," *Appl. Phys. Lett.*, vol. 96, no. 2, p. 022506, Jan. 2010, doi:10.1063/1.3279157.

[152] W.G. Wang *et al.*, "Voltage-induced switching in magnetic tunnel junctions with perpendicular magnetic anisotropy," *J. Phys. D: Appl. Phys.*, vol. 46, no. 7, p. 074004, Feb. 2013, doi:10.1088/0022-3727/46/7/074004.

[153] J. Park *et al.*, "Enhancement of data retention and write current scaling for sub-20nm STT-MRAM by utilizing dual interfaces for perpendicular magnetic anisotropy," in *Symp. VLSI Technol.*, Jun. 2012, pp. 57–58, doi:10.1109/VLSIT.2012.6242459.

[154] N. Nishimura *et al.*, "Magnetic tunnel junction device with perpendicular magnetization films for high-density magnetic random access memory," *J. Appl. Phys.*, vol. 91, no. 8, pp. 5246–5249, Jan. 2002, doi:10.1063/1.1459605.

[155] D. Wang *et al.*, "Magnetostatic coupling in spin dependent tunnel junctions," *IEEE Trans. Magnetics*, vol. 36, no. 5, pp. 2802–2805, Sep. 2000, doi:10.1109/20.908594.

[156] V.D. Nguyen *et al.*, "Towards high density STT-MRAM at sub-20nm nodes," in *Int. Symp. VLSI Technol., Syst. Appl.*, Apr. 2018, pp. 1–2, doi:10.1109/VLSI-TSA.2018.8403867.

[157] M. Durlam *et al.*, "A 0.18 /spl mu/m 4Mb toggling MRAM," in *IEEE Int. Electron Devices Meeting*, Dec. 2003, pp. 34.6.1–34.6.3, doi:10.1109/IEDM.2003.1269448.

[158] B.N. Engel *et al.*, "A 4-Mb toggle MRAM based on a novel bit and switching method," *IEEE Trans. Magnetics*, vol. 41, no. 1, pp. 132–136, Jan. 2005, doi:10.1109/TMAG.2004.840847.

[159] M. Hosomi *et al.*, "A novel nonvolatile memory with spin torque transfer magnetization switching: spin-RAM," in *IEEE Int. Electron Devices Meeting*, Dec. 2005, pp. 459–462, doi:10.1109/IEDM.2005.1609379.

[160] T. Kawahara *et al.*, "2Mb spin-transfer torque RAM (SPRAM) with bit-by-bit bidirectional current write and parallelizing-direction current read," in *IEEE Int. Solid-State Circuits Conf.*, Feb. 2007, pp. 480–617, doi:10.1109/ISSCC.2007.373503.

[161] R. Beach *et al.*, "A statistical study of magnetic tunnel junctions for high-density spin torque transfer-MRAM (STT-MRAM)," in *IEEE Int. Electron Devices Meeting*, Dec. 2008, pp. 1–4, doi:10.1109/IEDM.2008.4796679.

[162] S. Chung *et al.*, "Fully integrated 54nm STT-RAM with the smallest bit cell dimension for high density memory application," in *IEEE Int. Electron Devices Meeting*, Dec. 2010, pp. 12.7.1–12.7.4, doi:10.1109/IEDM.2010.5703351.

[163] N.D. Rizzo *et al.*, "A fully functional 64 Mb DDR3 ST-MRAM built on 90 nm CMOS technology," *IEEE Trans. Magnetics*, vol. 49, no. 7, pp. 4441–4446, Jul. 2013, doi:10.1109/TMAG.2013.2243133.

[164] D.C. Worledge *et al.*, "Switching distributions and write reliability of perpendicular spin torque MRAM," in *IEEE Int. Electron Devices Meeting*, Dec. 2010, pp. 12.5.1–12.5.4, doi:10.1109/IEDM.2010.5703349.

[165] Y.J. Lee *et al.*, "Demonstration of chip level writability, endurance and data retention of an entire 8mb stt-mram array," in *Int. Symp. VLSI Technol., Syst. and Appl.*, Apr. 2013, pp. 1–2, doi:10.1109/VLSI-TSA.2013.6545595.

[166] C. Park *et al.*, "Systematic optimization of 1 Gbit perpendicular magnetic tunnel junction arrays for 28 nm embedded STT-MRAM and beyond," in *IEEE Int. Electron Devices Meeting*, Dec. 2015, pp. 26.2.1–26.2.4, doi:10.1109/IEDM.2015.7409771.

[167] S.W. Chung *et al.*, "4Gbit density STT-MRAM using perpendicular MTJ realized with compact cell structure," in *IEEE Int. Electron Devices Meeting*, Dec. 2016, pp. 27.1.1–27.1.4, doi:10.1109/IEDM.2016.7838490.

[168] Y. Shih *et al.*, "Logic process compatible 40nm 16Mb, embedded perpendicular-MRAM with hybrid-resistance reference, sub-μA sensing resolution, and 17.5ns read access time," in *IEEE Symp. VLSI Circuits*, Jun. 2018, doi:10.1109/VLSIC.2018.8502260.

[169] C. Bi *et al.*, "6 - spin-orbit torque magnetoresistive random-access memory (SOT-MRAM)," in *Advances in Non-Volatile Memory and Storage Technology (Second Edition)*, ser. Woodhead Publishing Series in Electronic and Optical Materials, B. Magyari-Köpe *et al.*, Eds. Woodhead Publishing, Jun. 2019, pp. 203 – 235, doi:10.1016/B978-0-08-102584-0.00007-3.

[170] P.K. Amiri *et al.*, "Voltage-controlled magnetic anisotropy in spintronic devices," *SPIN*, vol. 02, no. 03, p. 1240002, 2012, doi:10.1142/S2010324712400024.

[171] Y.C. Wu *et al.*, "Deterministic and field-free voltage-controlled MRAM for high performance and low power applications," in *IEEE Symp. VLSI Technol.*, Jun. 2020, pp. 1–2, doi:10.1109/VLSITechnology18217.2020.9265057.

[172] Everspin, "DDR3 STT-MRAM in enterprise SSD," https://www.everspin.com/storage-solutions, accessed in Jun. 2020.

[173] IBM, "Using MRAM in high-speed enterprise storage caches," https://www.flashmemorysummit.com/Proceedings2019/08-05-Monday/20190805_MRAMDD_App_Briefs_Thangaraj.pdf, accessed in Jun. 2020.

[174] B. Sun *et al.*, "System demonstration of MRAM co-designed processing-in-memory CNN accelerator for mobile and IoT applications." https://www.flashmemorysummit.com/Proceedings2019/08-05-Monday/20190805_MRAMDD_App_Briefs_Thangaraj.pdf, accessed in Sept. 2020.

[175] Everspin, "MRAM for aerospace applications," https://www.everspin.com/aerospace, accessed in Sept. 2020.

[176] Y. Wang *et al.*, "Compact model for perpendicular magnetic anisotropy magnetic tunnel junction. (version 5)," *nanoHUB*, 2017.

[177] J. Yun *et al.*, "MBIST support for reliable eMRAM sensing," in *IEEE Eur. Test Symp.*, May 2020, pp. 1–6, doi:10.1109/ETS48528.2020.9131564.

[178] H. Lim *et al.*, "A survey on the modeling of magnetic tunnel junctions for circuit simulation," *Active and Passive Electronic Components*, vol. 2016, pp. 1–32, May 2016, doi:10.1155/2016/3858621.

[179] NIST, "OOMMF micromagnetic simulation tool," https://math.nist.gov/oommf/, accessed in Sept. 2020.

[180] G.D. Panagopoulos *et al.*, "Physics-based SPICE-compatible compact model for simulating hybrid MTJ/CMOS circuits," *IEEE Trans. Electron Devices*, vol. 60, no. 9, pp. 2808–2814, Sep. 2013, doi:10.1109/TED.2013.2275082.

[181] Y. Zhang *et al.*, "Compact modeling of perpendicular-anisotropy CoFeB/MgO magnetiac tunnel junctions," *IEEE Trans. Electron Devices*, vol. 59, no. 3, pp. 819–826, Mar. 2012, doi:10.1109/TED.2011.2178416.

[182] W.F. Brinkman *et al.*, "Tunneling conductance of asymmetrical barriers," *J. Appl. Phys.*, vol. 41, no. 5, pp. 1915–1921, 1970, doi:10.1063/1.1659141.

[183] D.C. Worledge *et al.*, "Spin torque switching of perpendicular Ta/CoFeB/MgO-based magnetic tunnel junctions," *Appl. Phys. Lett*, vol. 98, no. 2, pp. 93–96, Jan. 2011, doi:10.1063/1.3536482.

[184] R. Heindl *et al.*, "Validity of the thermal activation model for spin-transfer torque switching in magnetic tunnel junctions," *J. Appl. Phys.*, vol. 109, no. 7, Apr. 2011, doi:10.1063/1.3562136.

[185] S. Hamdioui, "Testing multi-port memories: theory and practice," Ph.D. dissertation, Delft University of Technology, Oct. 2001, https://repository.tudelft.nl/islandora/object/uuid%3Accced323-46dd-4557-aaf5-dee4840b31fe.

[186] S. Hamdioui *et al.*, "Memory fault modeling trends: a case study," *J. Electronic Testing*, vol. 20, no. 3, pp. 245–255, Jun. 2004, doi:10.1023/B:JETT.0000029458.57095.bb.

[187] A.J. Van De Goor, "Using march tests to test SRAMs," *IEEE Design Test of Computers*, vol. 10, no. 1, pp. 8–14, Mar. 1993, doi:10.1109/54.199799.

[188] S. Hamdioui *et al.*, "Memory test experiment: industrial results and data," *IEE Proc. Computers and Digital Techniques*, vol. 153, no. 1, pp. 1–8, Jan. 2006, https://ieeexplore.ieee.org/abstract/document/1576336.

[189] G. Han *et al.*, "Control of offset field and pinning stability in perpendicular magnetic tunnelling junctions with synthetic antiferromagnetic coupling multilayer," *J. Appl. Phys.*, vol. 117, no. 17, p. 17B515, Mar. 2015, doi:10.1063/1.4913942.

[190] C. Chappert *et al.*, "The emergence of spin electronics in data storage," in *Nanoscience and Technology*. World Scientific, 2009, pp. 147–157, doi:10.1142/9789814287005_0015.

[191] C. Wang *et al.*, "Impact of external magnetic field on embedded perpendicular STT-MRAM technology qualified for solder reflow," in *IEEE Int. Electron Devices Meeting*, Dec. 2017, pp. 21.1.1–21.1.4, doi:10.1109/IEDM.2017.8268432.

[192] M. Frankowski *et al.*, "Micromagnetic model for studies on magnetic tunnel junction switching dynamics, including local current density," *Physica B: Condensed Matter*, vol. 435, pp. 105 – 108, 2014, doi:10.1016/j.physb.2013.08.051.

[193] F.O. Heinz *et al.*, "Fast simulation of spin transfer torque devices in a general purpose tcad device simulator," in *Int. Conf. Simulation of Semiconductor Processes and Devices (SISPAD)*, Sep. 2013, pp. 127–130, doi:10.1109/SISPAD.2013.6650591.

[194] Y. Wang *et al.*, "Compact thermal modeling of spin transfer torque magnetic tunnel junction," *Microelectronics Reliability*, vol. 55, no. 9, pp. 1649 – 1653, Aug.–Sep. 2015, doi:https://doi.org/10.1016/j.microrel.2015.06.029.

[195] H. Lim *et al.*, "Advanced circuit-level model for temperature-sensitive read/write operation of a magnetic tunnel junction," *IEEE Trans. Electron Devices*, vol. 62, no. 2, pp. 666–672, Feb. 2015, doi:10.1109/TED.2014.2380819.

[196] R. De Rose *et al.*, "A compact model with spin-polarization asymmetry for nanoscaled perpendicular MTJs," *IEEE Trans. Electron Devices*, vol. 64, no. 10, pp. 4346–4353, 2017, doi:10.1109/TED.2017.2734967.

[197] C. Augustine *et al.*, "Numerical analysis of typical STT-MTJ stacks for 1T-1R memory arrays," in *IEEE Int. Electron Devices Meeting*, Dec. 2010, pp. 22.7.1–22.7.4, doi:10.1109/IEDM.2010.5703416.

[198] D.J. Griffiths, *Introduction to electrodynamics*. Pearson, 2013.

[199] L. Thomas *et al.*, "Perpendicular spin transfer torque magnetic random access memories with high spin torque efficiency and thermal stability for embedded applications," *J. Appl. Phys.*, vol. 115, no. 17, p. 172615, Apr. 2014, doi:10.1063/1.4870917.

[200] V. Drewello *et al.*, "Evidence for strong magnon contribution to the TMR temperature dependence in MgO based tunnel junctions," *Phys. Rev. B*, vol. 77, p. 014440, Jan. 2008, doi:10.1103/PhysRevB.77.014440.

[201] J.G. Alzate *et al.*, "Temperature dependence of the voltage-controlled perpendicular anisotropy in nanoscale MgO|CoFeB|Ta magnetic tunnel junctions," *Appl. Phys. Lett.*, vol. 104, no. 11, p. 112410, Mar. 2014, doi:10.1063/1.4869152.

[202] T. Devolder *et al.*, "Single-shot time-resolved measurements of nanosecond-scale spin-transfer induced switching: stochastic versus deterministic aspects," *Physical Review Letters*, vol. 100, no. 5, p. 057206, 2008, doi:10.1103/PhysRevLett.100.057206.

[203] L. Faber *et al.*, "Dynamic compact model of spin-transfer torque based magnetic tunnel junction (MTJ)," in *Int. Conf. Design Tech. Integrated Syst. Nanoscale Era*, Apr. 2009, pp. 130–135, doi:10.1109/DTIS.2009.4938040.

[204] C. Park *et al.*, "Low RA magnetic tunnel junction arrays in conjunction with low switching current and high breakdown voltage for STT-MRAM at 10 nm and beyond," in *IEEE Symp. VLSI Technol.*, Jun. 2018, pp. 185–186, doi:10.1109/VLSIT.2018.8510653.

[205] S. Hamdioui, "Device-aware test," Patent P6086853NL, 2020, (pending).

[206] M. Hu *et al.*, "Tightening the mesh size of the cell-aware ATPG net for catching all detectable weakest faults," in *IEEE Eur. Test Symp.*, May 2020, pp. 1–6, doi:10.1109/ETS48528.2020.9131567.

[207] M. Fieback *et al.*, "Testing resistive memories: Where are we and what is missing?" in *IEEE Int. Test Conf.*, Oct. 2018, pp. 1–9, doi:10.1109/TEST.2018.8624895.

[208] S. Hamdioui *et al.*, "Test and reliability of emerging non-volatile memories," in *IEEE Asian Test Symp.*, 2017, pp. 175–183, doi:10.1109/ATS.2017.42.

[209] N.Z. Haron *et al.*, "DfT schemes for resistive open defects in RRAMs," in *Design, Autom. & Test in Europe Conf. & Exhib.*, Mar. 2012, pp. 386–391, doi:10.1109/DATE.2012.6176603.

[210] S. Kannan *et al.*, "Sneak-path testing of memristor-based memories," in *Int. Conf. VLSI Design*. IEEE, Jan. 2013, pp. 386–391, doi:10.1109/VLSID.2013.219.

[211] G.C. Medeiros *et al.*, "DFT scheme for hard-to-detect faults in FinFET SRAMs," in *IEEE Eur. Test Symp.*, May 2019, pp. 1–2, doi:10.1109/ETS.2019.8791517.

[212] M. Shah *et al.*, "Special session: A quality and reliability driven DFT and DFR strategy for automotive and industrial markets," in *IEEE VLSI Test Symp.*, Apr. 2019, pp. 1–1, doi:10.1109/VTS.2019.8758640.

[213] S. Mukherjee *et al.*, "Role of boron diffusion in CoFeB/MgO magnetic tunnel junctions," *Phys. Rev. B*, vol. 91, no. 8, p. 085311, 2015, doi:10.1103/PhysRevB.91.085311.

[214] D.V. Dimitrov *et al.*, "Dielectric breakdown of MgO magnetic tunnel junctions," *Appl. Phys. Lett.*, vol. 94, no. 12, pp. 1–4, Mar. 2009, doi:10.1063/1.3109792.

[215] J.H. Lim *et al.*, "Investigating the statistical-physical nature of MgO dielectric breakdown in STT-MRAM at different operating conditions," in *IEEE Int. Electron Devices Meeting*, Dec. 2018, pp. 25.3.1–25.3.4, doi:10.1109/IEDM.2018.8614515.

[216] B. Oliver *et al.*, "Tunneling criteria and breakdown for low resistive magnetic tunnel junctions," *J. Appl. Phys.*, vol. 94, no. 3, pp. 1783–1786, May 2003, doi:10.1063/1.1590064.

[217] I. Yoon *et al.*, "Modeling and analysis of magnetic field induced coupling on embedded STT-MRAM arrays," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 37, no. 2, pp. 337–349, Feb. 2018, doi:10.1109/TCAD.2017.2697963.

[218] M. Shulaker *et al.*, "Special session (new topic): emerging computing and testing techniques," in *VLSI Test Symp.*, Apr. 2019, pp. 1–2, doi:10.1109/VTS.2019.8758598.

[219] X. Yao *et al.*, "Observation of intermediate states in magnetic tunnel junctions with composite free layer," *IEEE Trans. Magnetics*, vol. 44, no. 11, pp. 2496–2499, Nov. 2008, doi:10.1109/TMAG.2008.2003072.

[220] T. Aoki *et al.*, "Dynamic magnetic intermediate state during nanosecond spin transfer switching for MgO-based magnetic tunnel junctions," *Appl. Phys. Express*, vol. 3, no. 5, p. 053002, Apr. 2010, doi:10.1143/apex.3.053002.

[221] C. Hahn *et al.*, "Time-resolved studies of the spin-transfer reversal mechanism in perpendicularly magnetized magnetic tunnel junctions," *Phys. Rev. B*, vol. 94, p. 214432, Dec. 2016, doi:10.1103/PhysRevB.94.214432.

[222] N. Penthorn *et al.*, "Experimental observation of single skyrmion signatures in a magnetic tunnel junction," *Phys. Rev. Lett.*, vol. 122, no. 25, p. 257201, Jun. 2019, doi:10.1103/PhysRevLett.122.257201.

[223] L. Thomas *et al.*, "Solving the paradox of the inconsistent size dependence of thermal stability at device and chip-level in perpendicular STT-MRAM," in *IEEE Int. Electron Devices Meeting*, Dec. 2015, pp. 26.4.1–26.4.4, doi:10.1109/IEDM.2015.7409773.

[224] S. Van Beek *et al.*, "Impact of self-heating on reliability predictions in STT-MRAM," in *IEEE Int. Electron Devices Meeting*, Dec. 2018, pp. 25.2.1–25.2.4, doi:10.1109/IEDM.2018.8614617.

[225] J.G. Alzate *et al.*, "2 MB array-level demonstration of STT-MRAM process and performance towards L4 cache applications," in *IEEE Int. Electron Devices Meeting*, Dec. 2019, pp. 2.4.1–2.4.4, doi:10.1109/IEDM19573.2019.8993474.

[226] A. Iyengar *et al.*, "Retention testing methodology for STTRAM," *IEEE Design & Test*, vol. 33, no. 5, pp. 7–15, Oct. 2016, doi:10.1109/MDAT.2016.2591554.

[227] S. Hamdioui *et al.*, "Testing open defects in memristor-based memories," *IEEE Trans. Comput.*, vol. 64, no. 1, pp. 247–259, Oct. 2015, doi:10.1109/TC.2013.206.

[228] M. Fieback *et al.*, "Testing scouting logic-based computation-in-memory architectures," in *IEEE Eur. Test Symp.*, May 2020, pp. 1–6, doi:10.1109/ETS48528.2020.9131604.

NOMENCLATURE

SYMBOLS

α	magnetic damping
γ	gyromagnetic ratio
e	elementary charge
$\hbar$	reduced Planck constant
$\bar{\varphi}$	potential barrier height of the tunnel barrier
η	STT efficiency
C	Euler's constant ($\approx$0.577)
Ψ	inter-cell magnetic coupling factor
μ_0	vacuum permeability
τ_0	inverse of the attempt frequency ($\sim$1 ns)
k_B	Boltzmann constant
μ_B	Bohr magneton
μ_0	vacuum permeability
w0	electrical write 0 operation
$\mathrm{w0_H}$	magnetic write 0 operation
w1	electrical write 1 operation
$\mathrm{w1_H}$	magnetic write 1 operation
r0	read 0 operation
r1	read 1 operation
$\Uparrow$	up addressing
$\Downarrow$	down addressing
$\Updownarrow$	addressing order irrelevant
$\widehat{\mathrm{w}}0$	weak write operation using an electric current
$\widehat{\mathrm{w}}0_\mathrm{H}$	weak write operation using a magnetic field
S	sensitization sequence
F	faulty effect
F_n	faulty effect with n denoting its nature: p, or i, or t
p	permanent fault
i	intermittent fault
t	transient fault
L	extreme low resistance state
U	undefined resistance state
H	extreme high resistance state
R	readout value
?	random readout
$n\uparrow$	the number of spin-ups
$n\downarrow$	the number of spin-downs

Δ	thermal stability factor
Δ_0	intrinsic thermal stability factor
Δ_P	thermal stability factor in P state
Δ_AP	thermal stability factor in AP state
A_0	cross-sectional area of MTJ
A_ph	normalized pinhole area with respect to A_0
A_P	cross-sectional area of the P-state region in an intermediate state
A_AP	cross-sectional area of the AP-state region in an intermediate state
A_IMP	normalized cross-sectional area of the P-state region in an intermediate state with respect to the entire cross-sectional area of MTJ
A_IMAP	normalized cross-sectional area of the AP-state region in an intermediate state with respect to the entire cross-sectional area of MTJ
A_ds	quantitative design space
E_B	energy barrier between P and AP states
$E_\mathrm{B}(\mathrm{P}\rightarrow\mathrm{AP})$	energy barrier for P→AP switching
$E_\mathrm{B}(\mathrm{AP}\rightarrow\mathrm{P})$	energy barrier for AP→P switching
$F(x)$	cumulative distribution function
G_P	MTJ conductance in P state
G_AP	MTJ conductance in AP state
G_IM	MTJ conductance in IM state
H_k	magnetic anisotropy field
H_c	coercivity
$\boldsymbol{H}_\mathrm{stray}$	stray field at the free layer
$\boldsymbol{H}_\mathrm{ext}$	external magnetic field
$\boldsymbol{H}_\mathrm{s\_intra}$	intra-cell stray field
$H^z_\mathrm{s\_intra}$	the out-of-plane component of intra-cell stray field
$H^{x-y}_\mathrm{s\_intra}$	the in-plane component of intra-cell stray field
$\boldsymbol{H}_\mathrm{s\_inter}$	inter-cell stray field
$H^z_\mathrm{s\_inter}$	the out-of-plane component of inter-cell stray field
$H^{x-y}_\mathrm{s\_inter}$	the in-plane component of inter-cell stray field
H_offset	offset field in R-H loops
$H_\mathrm{sw\_n}$	Negative switching field from P state to AP state
$H_\mathrm{sw\_p}$	Positive switching field from AP state to P state
H_dir	inter-cell stray field to direct neighbors
H_dia	inter-cell stray field to diagonal neighbors
I_b	bound current
I_c	critical switching current
I_rd	current flowing through an STT-MRAM cell in a read operation
I_ref	current flowing through a reference STT-MRAM cell
I_w0	current flowing through an STT-MRAM cell in a w0 operation
I_w1	current flowing through an STT-MRAM cell in a w1 operation
I_DDQ	supply current I_DD in the quiescent state
$\boldsymbol{m}_\mathrm{FL}$	magnetization of the free layer
$\boldsymbol{m}_\mathrm{RL}$	magnetization of the reference layer
$\boldsymbol{m}_\mathrm{HL}$	magnetization of the hard layer
M_s	saturation magnetization

NP_8	neighborhood pattern in the eight neighboring cells surrounding the central cell in a 3×3 memory array
P	spin polarization
P_{FL}	spin polarization of the free layer
P_{PL}	spin polarization of the pinned layer
P_{sw}	switching probability
P_{dt}	detection probability
P_{ST}	successful transition probability
P_{RT}	switching probability due to thermal fluctuation after time RT
P_{IM}	occurrence probability of intermediate state
RA	resistance-area product
RA_{bd}	resistance-area product after hard breakdown
RA_{df}	defect-free resistance-area product
RA_{eff}	effective resistance-area product
R_P	resistance in parallel state
R_{AP}	resistance in anti-parallel state
R_{IM}	resistance in intermediate state
R_{sd}	resistance of series resistor as a defect model
R_{pd}	resistance of parallel resistor as a defect model
RT	retention time
RT_{IM}	retention time of IM state
TMR	tunneling magneto-resistance ratio
TMR_{df}	defect-free tunneling magneto-resistance ratio
TMR_{eff}	effective tunneling magneto-resistance ratio
t_w	Switching time
t_w^{pr}	Switching time in the precessional regime
t_w^T	Switching time in the thermal activation regime
T	Temperature
t_{FL}	thickness of the free layer
t_{TB}	thickness of the tunnel barrier
t_{PL}	thickness of the pinned layer
t_p	pulse width
V	volume of the free layer
V_c	switching voltage
V_p	pulse amplitude
V_{BL}	voltage on the bit line
V_{SL}	voltage on the source line
V_{WL}	voltage on the word line
V_{set}	set voltage
V_{rst}	reset voltage
V_{fm}	forming voltage
V_{DD}	supply voltage

ACRONYMS

AC	Alternating current
AFC	Anti-Ferromagnetic Coupling
AIoT	Artificial Intelligence + Internet of Things
AP	Anti-Parallel
BA	Bank Address within a given bank group
BEC	Bottom Electrode Contact
BEOL	Back-End-Of-Line
BER	Bit Error Rate
BG	Bank Group address
BIST	Built-In-Self-Test
BL	Bit Line
BPPM	Defective Part Per Billion
Ca	aggressor cell
CA	Column Address
CAFM	Conducting Atomic Force Microscopy
CAT	Cell-Aware Test
CD	Critical Diameter
CF	Conductive Filament or Coupling Fault
CIPT	Current-In-Plane Tunneling
CMOS	Complementary Metal–Oxide–Semiconductor
CMP	Chemical Mechanical Polishing
CNN	convolutional neural network
CPU	Central Processing Unit
CRC	Cyclic Redundancy Check
Cv	victim cell
DAT	Device-Aware Test
DC	Direct Current
DDR	Double Data Rate
DDR4	the 4th generation of high-bandwidth DDR interface
DfT	Design for Testability
DLL	Delay-locked loop
DPPB	Defective Part Per Billion
DPPM	Defective Part Per Million
DQ	data bus
DQS	DQ Strobe
DRAM	Dynamic Random Access Memory
ECC	Error Correction Code
eCD	electrical Critical Diameter
EtD	Easy-to-Detect
FBGA	Fine Ball Grid Array
FeRAM	Ferroelectric Random Access Memory
FEOL	Front-End-Of-Line
FIFO	First-In-First-Out data buffer
FinFET	Fin Field-Effect Transistor

FL	Free Layer
FP	Fault Primitive
FTL	Flash Translation Layer
GPU	Graphics Processing Unit
HDD	Hard Disk Drive
HL	Hard Layer
HRS	High Resistance State
HtD	Hard-to-Detect
IBE	Ion Beam Etching
IC	Integrated Circuits
IM	InterMediate
IMA	In-Plane Magnetic Anisotropy
IoT	Internet-of-Things
iSAF	inner Synthetic Anti-Ferromagnet
JEDEC	Joint Electron Tube Engineering Council, a standards organization for the Microelectronics industry
LLG	Landau-Lifshitz-Gilbert equation
LRS	Low Resistance State
MBIST	Memory Built-In-Self-Test
MCU	Micro-Controller Unit
MLC	Multi-Level Cell
MOSFET	Metal–Oxide–Semiconductor Field-Effect Transistor
MRAM	Magnetic Random Access Memory
MTJ	Magnetic Tunnel Junction
NPSF	Neighborhood Pattern Sensitive Fault
NPU	Neural Processing Unit
NVM	Non-Volatile Memory
ODT	on-die termination
OOMMF	The Object Oriented MicroMagnetic Framework
OSP	Oxygen Showering Post-treatment
P	Parallel
PCM	Phase-Change Memory
PL	Pinned Layer
PMA	Perpendicular Magnetic Anisotropy
pMTJ	Perpendicular (Magnetic Anisotropy) Magnetic Tunnel Junction
PNPSF	Passive Neighborhood Pattern Sensitive Fault
$PNPSF_i$	Intermittent Passive Neighborhood Pattern Sensitive Fault
PTM	Predictive Technology Model
PV	Process Variation
RDF	Read Destructive Fault
RIE	Reactive Ion Etching
RL	Reference Layer

RNF	Read Non-destructive Fault
RRAM	Resistive Random Access Memory
SAF	Synthetic Anti-Ferromagnet or Stuck-At Fault
SA0	Stuck-at 0
SA1	Stuck-at 1
SAFF	Synthetic Anti-Ferromagnet Flip
SCM	Storage-Class Memory
SEM	Scanning Electron Microscope
SF	State Fault
SL	Source line
SLC	Single-Level Cell
SoC	System on Chip
SOT	Spin-Orbit Torque
SPI	Serial Peripheral Interface
SPICE	Simulation Program with Integrated Circuit Emphasis
SRAM	Static Random Access Memory
SSD	Solid State Drive
ST-DDR4	Everspin's DDR4 interface tailored for STT-MRAM
STT	Spin-Transfer Torque
STT-MRAM	Spin-Transfer Torque Magnetic Random Access Memory
STO	Switching stochasticity
TB	Tunnel Barrier
TCAD	Technology Computer-Aided Design
TEC	Top Electrode Contact
TEM	Transmission Electron Microscope
TMR	Tunneling Magneto-Resistance
TOPS/W	Tera Operations Per Second per Watt
TSMC	Taiwan Semiconductor Manufacturing Company
TV	Temperature Variation
VCMA	Voltage-Controlled Magnetic Anisotropy
VLSI	Very Large Scale Integration
VSM	Vibrating Sample Magnetometry
WDF	Write Destructive Fault
WER	Write Error Rate
WL	World Line
WTF	Write Transition Fault

CURRICULUM VITÆ

Lizhou WU (吴利舟)

PERSONAL INFO.

Emails: njuwulizhou@163.com or @gmail.com
10-01-1991 Born in Quzhou, Zhejiang, China

RESEARCH INTERESTS

STT-MRAM design, test, and reliability
NVM design and test
Emerging computing paradigm based on NVM devices

EDUCATION

2016–2020 Ph.D. degree in Computer Engineering
Delft University of Technology (TU Delft), the Netherlands
Thesis title: Testing STT-MRAM: Manufacturing Defects, Fault Models, and Test Solutions
Supervisors: S. Hamdioui (TU Delft), M. Taouil (TU Delft), E.J. Marinissen (imec), S. Rao (imec)

2013–2016 M.Eng. degree in Computer Science & Engineering
National University of Defense Technology (NUDT), China
Thesis title: NAND Flash Controller Design and Optimization
Supervisors: F. Liu, N. Xiao

2009–2013 B.Sc. degree in Electronic Science and Engineering
Nanjing University (NJU), China

AWARDS & HONORS

2021 Best Paper Award nomination at DATE'21
2020 Distinguished Paper & Best Paper Award candidate at ITC'20
2020 Best Paper Award at DATE'20
2019 Best Paper Award nomination at ETS'19
2014 Outstanding Student Research Funding of NUDT
2013 The First Prize of Technology Innovation of NJU
2012 The First Prize of Int. Embedded System Design Contest (Intel Cup)
2011 Freshman Tutor Award, EE, NJU
2011 Vice Chairman of Student Union, EE, NJU

List of Publications

International Conferences

9. **L. Wu**, S. Rao, M. Taouil, E.J. Marinissen, G.S. Kar, S. Hamdioui, "Characterization and fault modeling of intermediate state defects in STT-MRAM," *IEEE/ACM Design, Automation & Test in Europe Conference (**DATE**)*, Grenoble, France, Feb. 2021, pp. 1–6. (**Best Paper Award nomination**)

8. **L. Wu**, S. Rao, M. Taouil, E.J. Marinissen, G.S. Kar, S. Hamdioui, "Characterization, modeling and test of synthetic anti-ferromagnet flip defect in STT-MRAMs," in *IEEE International Test Conference (**ITC**)*, Washington, DC, USA, Nov. 2020, pp. 1–10, doi:10.1109/ITC44778.2020.9325258. (**Distinguished paper & BPA candidate, BPA to be announced at ITC'21**)

7. **L. Wu**, M. Fieback, M. Taouil, S. Hamdioui, "Device-aware test for emerging memories: enabling your test program for DPPB level," in *IEEE European Test Symposium (**ETS**)*, Tallinn, Estonia, May 2020, pp. 1–2, doi: 10.1109/ETS48528.2020.9131559.

6. R. Bishnoi, **L. Wu**, M. Fieback, C. Münch, S.M. Nair, M. Tahoori, Y. Wang, H. Li, S. Hamdioui, "Special session–emerging memristor based memory and CIM architecture: test, repair and yield analysis," in *IEEE VLSI Test Symposium (**VTS**)*, San Diego, CA, USA, Apr. 2020, pp. 1–10, doi: 10.1109/VTS48691.2020.9107595.

5. **L. Wu**, S. Rao, M. Taouil, E.J. Marinissen, G.S. Kar, S. Hamdioui, "Impact of magnetic coupling and density on STT-MRAM performance," in *IEEE/ACM Design, Automation & Test in Europe Conference (**DATE**)*, Grenoble, France, Mar. 2020, pp. 1211–1216. (**Best Paper Award**) doi: 10.23919/DATE48585.2020.9116444.

4. G.C. Medeiros, C.C. Gürsoy, **L. Wu**, M. Fieback, M. Jenihhin, M. Taouil, S. Hamdioui, "A DFT scheme to improve coverage of hard-to-detect faults in FinFET SRAMs," in *IEEE/ACM Design, Automation & Test in Europe Conference (**DATE**)*, Grenoble, France, Mar. 2020, pp. 792–797, doi: 10.23919/DATE48585.2020.9116278.

3. M. Fieback, **L. Wu**, G.C. Medeiros, H. Aziza, S. Rao, E.J. Marinissen, M. Taouil, S. Hamdioui, "Device-aware test: a new test approach towards DPPB level," in *IEEE International Test Conference (**ITC**)*, Washington, DC, USA, 2019, pp. 1–10, doi: 10.1109/ITC44170.2019.9000134.

2. **L. Wu**, S. Rao, G.C. Medeiros, M. Taouil, E.J. Marinissen, F. Yasin, S. Couet, S. Hamdioui, G.S. Kar, "Pinhole defect characterization and fault modeling for STT-MRAM testing," in *IEEE European Test Symposium (**ETS**)*, Baden-Baden, Germany, May 2019, pp. 1–6, doi: 10.1109/ETS.2019.8791518. (**Best Paper Award Nomination**)

1. **L. Wu**, M. Taouil, S. Rao, E.J. Marinissen, S. Hamdioui, "Electrical modeling of STT-MRAM defects," in *IEEE International Test Conference (**ITC**)*, Phoenix, AZ, USA, Oct. 2018, pp. 1–10, doi: 10.1109/TEST.2018.8624749.

INTERNATIONAL JOURNALS

5. **L. Wu**, S. Rao, M. Taouil, E.J. Marinissen, G.S. Kar, S. Hamdioui, "A Field-Aware Compact Model of Perpendicular Magnetic Tunnel Junction for STT-MRAM," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (**TCAD**)*, pp. 1–14, 2021. (Under review, 2019 impact factor= 2.402)

4. **L. Wu**, S. Rao, M. Taouil, E.J. Marinissen, G.S. Kar, S. Hamdioui, "Characterization, Fault Modeling, and Test of Intermediate State Defect in STT-MRAM," *IEEE Transactions on Computers (**TC**)*, pp. 1–14, 2021. (Under review, 2019 impact factor= 3.131)

3. G.C. Medeiros, M. Fieback, **L. Wu**, M. Taouil, L.B. Poehls, S. Hamdioui, "Hard-to-Detect Faults in FinFET SRAMs: Analyses and Test Solutions," *IEEE Transactions on Very Large Scale Integration Systems (**TVLSI**)*, 2020. (Under review, 2019 impact factor= 2.360)

2. M. Fieback, G.C. Medeiros, **L. Wu**, H. Aziza, R. Bishnoi, M. Taouil, S. Hamdioui, "Defect and fault modeling, and test development framework for RRAM," *ACM Journal on Emerging Technologies in Computing Systems (**JETC**)*, 2020. (Under review, 2019 impact factor= 1.652)

1. **L. Wu**, S. Rao, M. Taouil, G.C. Medeiros, M. Fieback, E.J. Marinissen, G.S. Kar, S. Hamdioui, "Defect and fault modeling framework for STT-MRAM testing," *IEEE Transactions on Emerging Topics in Computing (**TETC**)*, pp. 1–15, Dec. 2019, doi:10.1109/TETC.2019.2960375. (2019 impact factor=6.043)

OTHER PUBLICATIONS

1. **L. Wu**, M. Taouil, S. Rao, E.J. Marinissen, S. Hamdioui, "Survey on STT-MRAM testing: failure mechanisms, fault models, and tests," *arXiv preprint*, pp. 1–24, Jan. 2020, arXiv:2001.05463.

2016.10-2021.3

Amazing time in TU

TU Delft

Quantum & Computer Engineering

ISBN 9789463841993

Invitation

You are cordially invited to attend the public defense of my PhD dissertation, entitled

Testing STT-MRAM:
Manufacturing Defects, Fault Models, and Test Solutions

on Monday
Feb. 22, 2021
at 3:00 pm

To comply with measures against the spread of COVID-19, the defense will be live streamed to you via the link below

collegerama.tudelft.nl/Mediasite/Showcase/phd-defence/

Or scan me