Variation Analysis, Fault Modeling and Yield Improvement of Emerging Spintronic Memories

Zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften

von der KIT-Fakultät für Informatik
des Karlsruher Instituts für Technologie (KIT)

genehmigte

Dissertation

von

Sarath Mohanachandran Nair

aus Cherthala, Kerala, India

Tag der mündlichen Prüfung: 19.05.2020

1. Referent/Referentin: Prof. Dr.-Ing. Mehdi B. Tahoori
Karlsruher Institut für Technologie (KIT); Germany

2. Referent/Referentin: Prof. Dr.-Ing. Lorena Anghel
Grenoble Institute of Technology; France

Sarath Mohanachandran Nair
Pfinzstr. 66
76227 Karlsruhe

Hiermit erkläre ich an Eides statt, dass ich die von mir vorgelegte Arbeit selbstständig verfasst habe, dass ich die verwendeten Quellen, Internet-Quellen und Hilfsmittel vollständig angegeben haben und dass ich die Stellen der Arbeit - einschlielich Tabellen, Karten und Abbildungen - die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Karlsruhe, 05.2020
Sarath Mohanachandran Nair

Acknowledgement

It would not have been possible to write this doctoral thesis without the help and support of the kind people around me, to only some of whom it is possible to give particular mention here.

My deep gratitude goes first to my advisor Prof. Mehdi Tahoori who expertly guided me through my research, not to mention his unsurpassed knowledge. His unwavering enthusiasm kept me constantly engaged with my research. In addition, I want to thank Prof. Lorena Anghel for co-supervising my thesis.

My appreciation extends to all my colleagues in the Chair of Dependable Nano Computing (CDNC) at KIT for their academic and technical support. I take this opportunity to thank Dr. Rajendra Bishnoi, Dr. Mohammad Saber Golanbari, Dr. Nour Sayed, Dr. Anteneh Gebregiorgis, Dennis Gnad, Dennis Weller, Jonas Krautter and Christopher Münch for their collaboration and valuable insights.

I would also like to thank Ms. Iris Schröder-Piepka for her support and for handling all the required documentation during the course of my PhD. Along with that, I thank Ms. Audrey Bohlinger and Mr. Thomas Griesbaum for the help in the preparation of my defense.

The good advice, support and friendship of my wonderful colleagues and friends Dr. Arunkumar Vijayan, Dr. Arun Chandrasekharan, Samir Ben Dodo, Farhan Rasheed, Ahmet Erozan and Manu Sreedharan have been invaluable on both an academic and a personal level, for which I am extremely grateful.

I am forever indebted to my parents Mohanachandran Nair and Rethiamma and my sister Saritha for their unparalleled love and support and for giving me the opportunities and experiences that have made me who I am.

Most of all, I am fully indebted to my wife Karthi Vijayan for her love, understanding, enthusiasm and encouragement. There has been many challenging times and the unequivocal support of my wife throughout, as always, to see me achieve this goal has pushed me farther than I thought I could go.

List of Publications

List of first author publications included in this thesis

1. **S. M. Nair**, R. Bishnoi, M. S. Golanbari, F. Oboril and M. B. Tahoori, "VAET-STT: A variation aware estimator tool for STT-MRAM based memories," Design, Automation & Test in Europe Conference & Exhibition (DATE), pp. 1456-1461, 2017.

2. **S. M. Nair**, R. Bishnoi, M. S. Golanbari, F. Oboril, F. Hameed, and M. B. Tahoori, "VAET-STT: Variation aware STT-MRAM analysis and design space exploration tool," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 37, no. 7, pp. 1396-1407, 2017.

3. S. M. Nair, R. Bishnoi and M. B. Tahoori, "Parametric failure modeling and yield analysis for STT-MRAM," Design, Automation & Test in Europe Conference & Exhibition (DATE), pp. 265-268, 2018.

4. **S. M. Nair**, R. Bishnoi, and M. B. Tahoori, "A Comprehensive Framework for Parametric Failure Modeling and Yield Analysis of STT-MRAM ," IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 27, no. 7, pp. 1697-1710, 2019.

5. **S. M. Nair**, R. Bishnoi, M. B. Tahoori, G. Tshagharyan, H. Grigoryan, G. Harutyunyan, and Y. Zorian, "Defect Injection, Fault Modeling and Test Algorithm Generation Methodology for STT-MRAM," IEEE International Test Conference (ITC), pp. 1-10, 2018.

6. **S. M. Nair**, R. Bishnoi, M. Tahoori, H. Grigoryan, and G. Tshagharyan, "Variation-aware Fault Modeling and Test Generation for STT-MRAM," IEEE 25th International Symposium on On-Line Testing and Robust System Design (IOLTS), pp. 80-83, 2019

List of first author publications not included in this thesis

7. **S. M. Nair**, R. Bishnoi, M.B. Tahoori, H. Zahedmanesh, K. Croes, K. Garello, G. Sankar Kar, and F. Catthoor, "Variation-aware physics based electromigration modeling and experimental calibration for VLSI interconnects," IEEE International Reliability Physics Symposium (IRPS), pp. 1-6, 2019.

8. **S. M. Nair**, R. Bishnoi, A. Vijayan and M. B. Tahoori, "Dynamic Faults based Hardware Trojan Design in STT-MRAM," Design, Automation & Test in Europe Conference & Exhibition (DATE), 2020.

9. **S. M. Nair**, R. Bishnoi and M. B. Tahoori, "Mitigating Read Failures in STT-MRAM," VLSI Test Symposium (VTS), 2020.

10. **S. M. Nair**, R. Bishnoi, M.B. Tahoori, H. Zahedmanesh, K. Croes, K. Garello, G. Sankar Kar, and F. Catthoor, "Physics based modeling of bimodal electromigration failure distributions and variation analysis for VLSI interconnects," IEEE International Reliability Physics Symposium (IRPS), 2020.

11. **S. M. Nair**, C. Muench and M. B. Tahoori, "Defect Characterization and Test Generation for Spintronic-based Compute-In-Memory," European Test Symposium (ETS), 2020.

List of co-author publications not included in this thesis

12. M.B. Tahoori, **S.M. Nair**, R. Bishnoi, S. Senni, J. Mohdad, F. Mailly, L. Torres, P. Benoit, P. Nouet, R. Ma, M. Kreiig, F. Ellinger, K. Jabeur, P. Vanhauwaert, G. Di Pendina and G. Prenat, "GREAT: heteroGeneous integRated magnetic tEchnology using multifunctional standardized sTack," IEEE Computer Society Annual Symposium on VLSI (ISVLSI), pp. 344-349, 2017.

13. N. Sayed, **S. M. Nair**, R. Bishnoi and M. B. Tahoori, "Process variation and temperature aware adaptive scrubbing for retention failures in STT-MRAM," 23rd Asia and South Pacific Design Automation Conference (ASP-DAC), pp. 203-208, 2018.

14. M.B. Tahoori, **S. M. Nair**, R. Bishnoi, S. SENNI, J. Mohdad, F. Mailly, L. Torres, P. Benoit, A. Gamatie, P. Nouet, K. Jabeur, P. Vanhauwaert, A. Atitoaie, I. Firastrau, G. Di Pendina, and G. Prenat, "Using Multifunctional Standardized Stack as Universal Spintronic Technology for IoT," Design, Automation & Test in Europe Conference & Exhibition (DATE), pp. 931-936, 2018.

15. S. Ben Dodo, R. Bishnoi, **S. M. Nair** and M. B. Tahoori, "A Spintronics Memory PUF for Resilience Against Cloning Counterfeit," IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 27, no. 11, pp. 2511-2522, 2019.

16. M.B. Tahoori, **S. M. Nair**, R. Bishnoi, L. Torres, G. Partigeon, G. DiPendina, and G. Prenat, "A Universal Spintronic Technology Based on Multifunctional Standardized Stack," Design, Automation & Test in Europe Conference & Exhibition (DATE), 2020.

17. R. Bishnoi, L. Wu, M. Fieback, C. Muench, **S. M. Nair**, M. B. Tahoori, Y. Wang, H. Li and S. Hamdioui, "Emerging Memristor Based Memory and CIM Architecture: Test, Repair and Yield Analysis," VLSI Test Symposium (VTS), 2020.

Abstract

The demand for high performance and low power consumption for modern computing devices have resulted in aggressive technology down-scaling, driven by Moore's law. However, the scaling challenges limit the use of conventional Complementary Metal Oxide Semiconductor (CMOS) memories such as Static Random Access Memory (SRAM) and Dynamic Random Access Memory (DRAM) in advanced technology nodes, primarily due to the increased leakage power. Hence, several new novel non-volatile memory (NVM) technologies are being researched to replace conventional CMOS memories. Among these, the Spin Transfer Torque Magnetic Random Access Memory (STT-MRAM) is the most promising candidate, as shown by several recent industrial demonstrations. It has several advantageous features such as zero standby leakage, high density, nearly unlimited endurance, and process and voltage compatibility with CMOS technology. Therefore, with all its advantages, STT-MRAM has the capacity to become a universal memory since it can potentially fit into every level of the memory hierarchy.

STT-MRAM exhibits stochastic switching behavior and has higher sensitivity to process variation as compared to CMOS memories, significantly affecting its performance, energy and reliability. In addition, the parametric variations in STT-MRAM exacerbates its stochastic switching behavior, leading to both test time fails and reliability failures in the field. Since an STT-MRAM memory array consists of both CMOS and magnetic components, the system level failures in STT-MRAM depends on variations in both these components. As this memory technology gains rapid industrial adoption, analyzing the impact of manufacturing variations and defects becomes extremely important to predict the yield and implement effective yield improvement techniques.

In this thesis, we investigate the impact of variations and defects on this novel memory technology. Several tools and frameworks are developed for variability, yield analysis and fault modeling, which can help in the design of reliable STT-MRAM memories.

First, we developed a system-level variation-aware framework which can quantify the effect of stochasticity and process variations from the bit-cell level to the overall memory system and estimate the latencies and energies of STT-MRAM based memories. This tool can perform a variation-aware design space exploration and memory configuration optimization for energy or performance while meeting reliability constraints. It also reports various failure rates for read, write and retention and can evaluate the effectiveness of different Error Correcting Code (ECC) schemes to reduce the failure rates. The results show that our framework provides more realistic margins and the optimized variation-aware memory configuration could be significantly different from the conventional analysis based on nominal values.

Next, we have modeled the system level parametric failures of STT-MRAM, considering the spatial correlation among bit-cells as well as the impact of peripheral components. The proposed approach provides realistic fault distribution maps and equips the designer to investigate the efficacy of different combinations of defect tolerance techniques for an effective design-for-yield exploration. The results show that the fault distribution and yield depend on the correlation coefficient and the temperature, which in turn determine the correct choice of defect tolerance schemes such as ECC or redundancy to be adopted to mitigate them to improve the yield.

Finally, a defect characterization and fault modeling methodology is developed for STT-MRAM, based on extensive defect injection campaign, by considering the netlist and the layout. Both spot defects, manifesting as resistive opens and shorts, as well as STT-MRAM specific defects have been evaluated. Furthermore, the impacts of test environment, namely the temperature and voltage, as well as process variations on the manifestation of defects are analyzed. Based on the results of fault analysis, efficient test algorithms have been developed to cover the unique faults of STT-MRAM.

Zusammenfassung

Die Nachfrage für hochperformante und energiesparende moderne Computersysteme hat aufgrund des Mooreschen Gesetzes zu einer aggressiven Technologieskalierung geführt. Dennoch beschränken die auftretenden Skalierungsprobleme die Nutzbarkeit von konventionellen komplementären Metall-Oxid Halbleitern (CMOS) Speichern wie beispielsweise statischer Random-Access Speicher (SRAM) und dynamischer Random-Access Speicher (DRAM) bei fortschrittlichen Fertigungstechniken, hauptsächlich bedingt durch den Anstieg der Leckströme. Daher wird an verschiedenen neuen, innovativen nicht-flüchtigen Speichertechnologien (NVM) geforscht, um die herkömmlichen CMOS-basierten Speicher zu ersetzen. Unter diesen ist der auf Spin-Transfer-Torque basierende magnetische Random-Access Speicher (STT-MRAM) der vielversprechendste Kandidat, wie verschiedene industrielle Demonstrationen erst kürzlich zeigten. Er hat eine Reihe von vorteilhaften Eigenschaften, wie die Zero-Standby-Leakage, eine hohe Integrationsdichte, nahezu unbegrenzte Haltbarkeit und ist prozess- und spannungskompatibel mit CMOS-Technologie. Deshalb bietet sich STT-MRAM als universeller Speicher an und hat potentielle Anwendungsgebiete auf allen Ebenen der Speicherhierarchie.

STT-MRAM zeigt stochastisches Schaltverhalten und ist anfälliger gegenüber Prozessvariationen im Vergleich zu CMOS-Speichern, was sowohl die Performance als auch die Zuverlässigkeit beeinträchtigt. Zusätzlich verschlechtert sich das stochastische Schaltverhalten aufgrund von parametrischen Variationen im STT-MRAM, was wiederum sowohl zu Testfehlern als auch zu Zuverlässigkeitsfehlern im Einsatz führt. Da STT-MRAM sowohl aus CMOS als auch magnetischen Komponenten besteht, sind Fehler auf Systemebene abhängig von der Variationen aus beiden Komponentengruppen. Mit der wachsenden industriellen Verbreitung der Speichertechnik wird die Analyse des Einflusses von Herstellungsvariationen und -defekten zur Prognose der Fertigungsausbeute und zu deren Verbesserung äuerst wichtig.

In dieser Arbeit untersuchen wir den Einfluss von Prozessvariationen und -defekten auf diese neue Speichertechnologie. Zur Unterstützung im Entwurf von zuverlässigen

STT-MRAM Speichern werden verschiedene Werkzeuge und Frameworks entwickelt, welche Modelle zur Variabilität, zur Prozessvariationsanalyse und zur Fehleranalyse bereitstellen.

Zuerst entwickelten wir ein variationssensibles Framework auf Systemebene zur Quantifizierung des Effekts der Stochastik und der Prozessvariationen von der Zellebene bis zum Gesamtsystem und schätzen so die Verzögerungen und den Energieverbrauch des STT-MRAM-basierenden Speichers. Dieses Werkzeug kann zum Durchsuchen des Entwurfsraumes genutzt werden, um Speicherkonfigurationen hinsichtlich ihres Energieverbrauches und ihrer Performance zu optimieren, während die Auflagen bezüglich der Zuverlässigkeit eingehalten werden. Es meldet ebenfalls Fehlerraten für Lese- und Schreibzugriffe, sowie für Retentionsfehler und ist in der Lage die Effektivität von fehlerkoriegerenden Codes (ECC) zur Reduktion von Fehlerraten zu evaluieren. Die Ergebnisse zeigen, dass unser Framework im Vergleich zu konventionellen Analysen mit nominalen Werten realistischere Fehlermargen berechnen kann und die optimierten variationssensiblen Speicherkonfigurationen sich deutlich von den konventionellen unterscheiden.

Als nächstes modellierten wir, sowohl unter Beachtung des räumlichen Zusammenhangs der Bitzellen als auch des Einflusses der peripheren Komponenten, die auftretenden parametrischen Fehler des STT-MRAM auf Systemebene. Der vorgeschlagene Ansatz ermöglicht das Erstellen von realistischen Fehlerverteilungsplänen und stattet den Entwickler mit der Möglichkeit zum Untersuchen von verschiedenen Kombinationen von fehlertolerierenden Techniken aus, um so einen effektiven Fertigungsausbeute-orientierten Entwurf zu erstellen. Die Ergebnisse zeigen, dass die Fehlerverteilung und die Ausbeute zum Korrelationskoeffizient und zur Temperatur in Abhängigkeit stehen, welche wiederum die korrekte Wahl des Schemas zur Tolerierung von Defekten, wie beispielsweise ECC oder Redundanzen, bestimmt, um diese zu vermeiden und die Ausbeute zu erhöhen.

Zum Abschluss wurde eine Defektcharakterisierungs- und Fehlermodelierungsmethode für STT-MRAM entwickelt. Diese basieren auf ausgiebig durchgeführten Defektinjektionen unter Berücksichtigung der Netzliste und des Layouts. Sowohl Punktdefekte, welche sich durch fehlenden Kontakt oder einen Kurzschluss auszeichnen, also auch STT-MRAM spezifische Defekte wurden untersucht. Des Weiteren wurde der Einfluss der Testumgebung, im Speziellen der Temperatur und der Spannung, wie auch die Prozessvariationen auf die Manifestation von Defekten analysiert. Auf Basis dieser Fehleranalyse wurden Testalgorithmen entwickelt, welche die speziellen Fehler von STT-MRAM abdecken.

Contents

List of Publications **ix**

Abstract **xii**

Zusammenfassung **xiv**

Contents **xix**

List of Figures **xix**

List of Tables **xxi**

1 Introduction **1**
1.1 Dissertation Contributions 5
1.2 Dissertation Outline 6

2 Background **7**
2.1 STT-MRAM Technology 7
2.2 Read and Write Operations in STT-MRAM 10
2.2.1 Read Operation 10
2.2.2 Write Operation 12
2.3 Parametric Variations in STT-MRAM 12
2.3.1 Random Variations 12
2.3.1.1 Variations in MTJ 12
2.3.1.2 Variations in Access Transistor 13
2.3.1.3 Variations in Peripheral Circuitry 14
2.3.2 Systematic Variations 14
2.4 Defects in MTJs and STT-MRAM array 14
2.5 Defect Tolerance Techniques 15
2.6 Summary 16

3 Variation-aware STT-MRAM Analysis and Design Space Exploration **17**
3.1 Overview 17
3.1.1 Related Work 17
3.1.2 Contributions 19

3.2 Variation-aware Analysis Framework 20
3.2.1 Overview 20
3.2.2 Component-level Variation Analysis 21
3.2.2.1 Bit-cell Latency 21
3.2.2.2 Bit-cell Energy 22
3.2.2.3 Peripheral Latency and Energy 23
3.2.3 System-level Variation Analysis 23
3.2.3.1 Latency Calculation 23
3.2.3.2 Energy Calculation 25
3.2.4 Reliability Analysis 25
3.2.4.1 Errors in STT-MRAM 25
3.2.4.2 Error Mitigation using ECCs 26
3.2.5 Design Space Exploration 27
3.3 Results 27
3.3.1 Experimental Setup 27
3.3.2 Results 28
3.3.2.1 Overall Latency and Energy Distributions 28
3.3.2.2 Latency vs Error rates 28
3.3.2.3 Read Disturb 29
3.3.2.4 Retention Failures 30
3.3.2.5 Scaling Effects 31
3.3.2.6 Effect of ECCs 32
3.3.2.7 Design Space Exploration 32
3.3.3 Validation 33
3.4 Summary 34

4 Parametric Failure Modeling and Yield Analysis 35
4.1 Overview 35
4.1.1 Related Work 36
4.1.2 Contributions 36
4.2 Yield Analysis Framework 37
4.2.1 Obtaining Correlation maps 38
4.2.2 Permanent Fault Analysis 38
4.2.3 Reliability Fault Analysis 41
4.2.4 Analysis Complexity Reduction Techniques 44
4.2.5 Yield Exploration 45
4.3 Results 47
4.3.1 Experimental Setup 47
4.3.2 Results 47
4.4 Summary 55

5 Defect Injection, Fault Modeling and Test Algorithm Generation 57
5.1 Overview 57
5.1.1 Related Work 58

5.1.2 Contributions . . . 58
5.2 Fault Analysis Framework . . . 59
5.2.1 Defect Injection Methodology . . . 59
5.2.2 Simulation Framework . . . 60
5.2.3 Impact of Variability . . . 62
5.3 Results . . . 62
5.4 Test Pattern Generation . . . 67
5.5 Summary . . . 68

6 Conclusions and Outlook **71**
6.1 Conclusions . . . 71
6.2 Outlook . . . 72

Bibliography **73**

Glossary **81**

List of Figures

1.1 50 years of microprocessor trend data obtained from [1] 1
1.2 Conventional memory (SRAM and DRAM) capacity trend [2, 3] . . . 2
1.3 Leakage and static power at various technology nodes [4] 3
1.4 Comparison of bit-cell architecture of SRAM, DRAM and STT-MRAM 3
1.5 A typical memory hierarchy using conventional memories 4

2.1 STT-MRAM storing device and its bit-cell architecture 8
2.2 Layout and cross-sectional view of STT-MRAM bit-cell 8
2.3 Bit-cell storing '0' and '1' 9
2.4 STT-MRAM memory array organization 9
2.5 Typical read and write circuits for STT-MRAM 11
2.6 Waveform showing a read '1' operation in STT-MRAM 11

3.1 Memory array organization as proposed in NVSim [5]. 18
3.2 High-level composition of a subarray with individual components and their latencies. 19
3.3 Proposed Variation Aware STT-MRAM Analysis and Design Space Exploration Tool (VAET-STT) flow. 20
3.4 Q-Q plots for bit-cell read/write latency and energy. For setup please refer to Section 3.3.1. 22
3.5 Overall latency and energy distributions for a subarray size of 1024 × 1024 at a temperature of 300K for 45 nm node. 29
3.6 Variations of overall write latency with temperature for a subarray size of 1024 × 1024. 30
3.7 Overall read and write latencies for various error rates. 30
3.8 Read disturb probabilities for different read periods. 31
3.9 Histogram of mean retention time for 10,000 samples. 31
3.10 Effect of ECCs on write latency for WER of 1×10^{-18}. 32
3.11 Validation of our hybrid method with full Monte-Carlo for a subarray size of 256 × 32. 33

4.1 Proposed yield analysis flow. 37
4.2 Correlation maps for radius (μ = 20 nm, σ = 6%) and V_{th} (μ = 397.9 mV, σ = 3.76%) for a 32×32 array for $\Phi = 0$ and $\Phi = 0.25$. 39
4.3 Correlation maps for radius (μ = 20 nm, σ = 6%) and V_{th} (μ = 397.9 mV, σ = 3.76%) for a 32×32 array for $\Phi = 0.5$ and $\Phi = 0.75$. . 40

4.4 Read, write, retention and read-disturb failure distribution map for a 32×32 memory array for radius ($\mu = 20$ nm, $\sigma = 6\%$), V_{th} ($\mu = 397.9$ mV, $\sigma = 3.76\%$) for $\Phi = 0$ and $\Phi = 0.25$. 42
4.5 Read, write, retention and read-disturb failure distribution map for a 32×32 memory array for radius ($\mu = 20$ nm, $\sigma = 6\%$), V_{th} ($\mu = 397.9$ mV, $\sigma = 3.76\%$) for $\Phi = 0.50$ and $\Phi = 0.75$. 43
4.6 Analysis Complexity Reduction Technique. [Region 1 - Definite 'Fail'; Region 2 - Definite 'Pass'; Region 3 and Region 4 - Simulation required] 44
4.7 Write latency reduction using current boosting technique. 46
4.8 Circuit for write current boosting. 46
4.9 Line fault distribution for a 512×512 memory array at 25 °C or various values of correlation coefficient (Φ) 48
4.10 Percentage of chips with their fault types for a 512×512 memory array at 25 °C for various values of correlation coefficient (Φ) 49
4.11 Yield improvement using different defect tolerance techniques versus their area overhead for write and read faults for $\Phi = 0.50$ (RR - Redundancy Repair; CB - Current Boost) 50
4.12 Yield improvement using different defect tolerance techniques versus their area overhead for retention and read-disturb faults for $\Phi = 0.50$ (RR - Redundancy Repair) 51
4.13 Temperature dependence of yield for various faults ($\Phi = 0.25$) 53
4.14 Yield analysis for various faults and combined yield considering all faults (Temperature = 25 °C) 54
4.15 Combined yield and its boundaries (Temperature = 25 °C) [Y_c - Combined Yield; Y_i - Yield due to fault type i] 54
4.16 Comparison of defect tolerance techniques for yield improvement considering all faults for various correlation coefficients (Φ) (Temperature = 25 °C) 55
4.17 Yield analysis for different Δ values assuming same percentage variation ($\Phi = 0.5$, Temperature = -25 °C). For setup, please refer to Section 4.3.1. 55

5.1 Advanced Inductive Failure Analysis (AIFA) flow 59
5.2 Layout based fault injection. The injected resistors are shown in red. 61
5.3 Waveform showing dIRF1-3 dynamic fault 64
5.4 Dynamic fault with process variation at LVHT corner 65
5.5 Waveform showing the signals of a victim cell affected by dCFir1-2 . 66
5.6 Coupling fault with process variation at LVHT corner 66
5.7 Test algorithm generation flow 68
5.8 STAR Memory System 68

List of Tables

1.1 Comparison of STT-MRAM with existing memory technologies ($F-$Feature size) . . . 4

3.1 Parameters of PMA MTJ at 45 nm and 65 nm . . . 28
3.2 Overall latency and energy values at 45 nm and 65 nm . . . 32
3.3 Design space exploration of a 128KB RAM for an input WER of 1×10^{-18} 33
3.4 Comparison of run-time of the proposed method with full Monte-Carlo 33
3.5 Comparison of WER for a memory array with 512 columns . . . 34

4.1 Operation of current boosting circuitry . . . 45
4.2 Parameters of the MTJ . . . 47

5.1 Defects and Fault Models in STT-MRAM . . . 60
5.2 Defect Injection and Write Fault Models at different operating conditions . . . 62
5.3 Defect Injection and Read Fault Models . . . 63
5.4 Corner Analysis for Write Fault Models . . . 63
5.5 Dynamic Faults . . . 64
5.6 Corner Analysis for dIRF1 dynamic fault (SL $-$ IN short) . . . 65
5.7 Corner Analysis for dCFir1 (BL $-$ IN short) . . . 65

Chapter 1

Introduction

Aggressive technology scaling, driven by Moore's law [6, 7], has resulted in tremendous increase in the complexity (in terms of number of transistors) and performance (in terms of operating frequency) for various microprocessor generations in the past five decades as illustrated in Fig. 1.1 [1]. Moore's law predicts that the number of transistors in a semiconductor chip doubles approximately every 18 months. The advantages of technology scaling have also benefited conventional Complementary Metal-Oxide Semiconductor (CMOS) memory technologies, such as Static Random Access Memory (SRAM) and Dynamic Random Access Memory (DRAM), as shown in Fig. 1.2, where the memory capacity increases exponentially year on year. The increase in the transistor density per chip has also been driven by Dennard's scaling [8], which states that, as the transistors get smaller, the power density remains a constant, which means that the voltage and current also scale down with the length.

However, several factors affect the scalability of SRAM and DRAM in advanced technology nodes [9–11]. In particular, as the technology scales down to sub 100 nm,

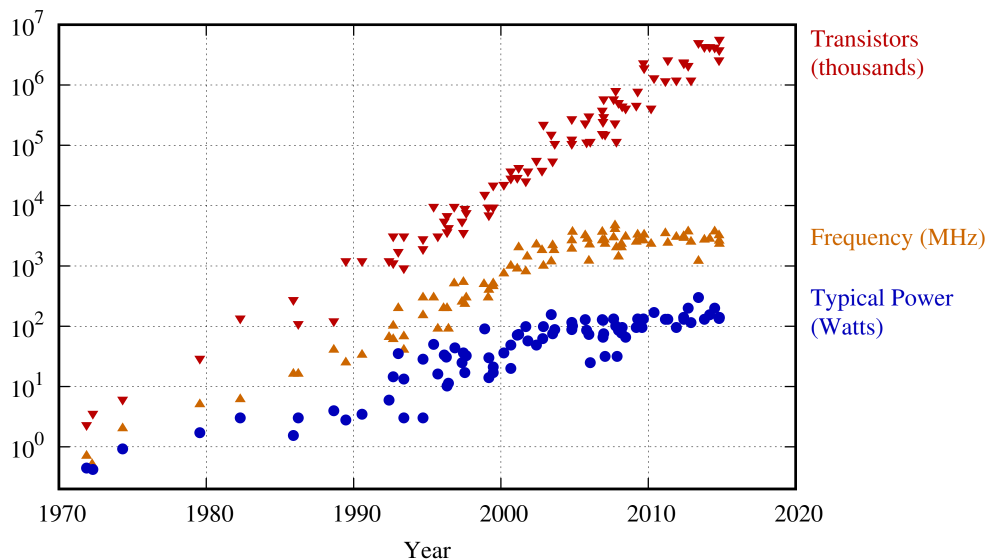
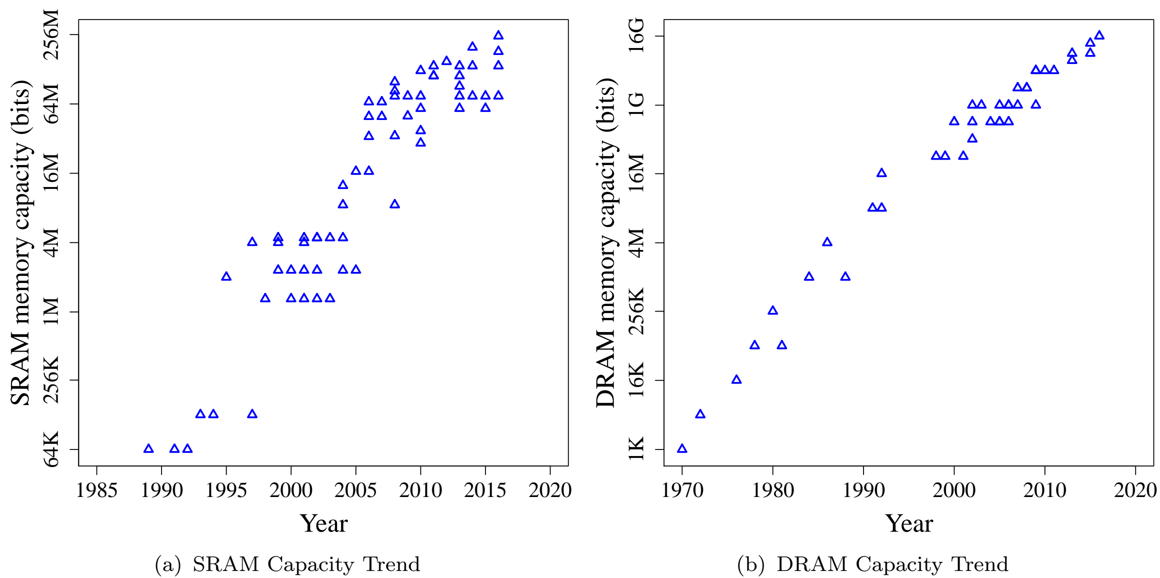


Figure 1.1: 50 years of microprocessor trend data obtained from [1]



(a) SRAM Capacity Trend

(b) DRAM Capacity Trend

Figure 1.2: Conventional memory (SRAM and DRAM) capacity trend [2, 3]

Dennard's scaling starts to break down. This is illustrated in Fig. 1.3, where an increase in power density is observed as the technology scales down from 90 nm to 20 nm [4]. Furthermore, as shown in this figure, in advanced technology nodes (28 nm and below), the leakage or static power is equal to or greater than the dynamic power, which means that the total power consumption is dominated by the leakage power. This is due to the fact that as transistor sizes shrink, the leakage current increases, thus increasing the leakage power. Since conventional CMOS-based memories (SRAM and DRAM) are volatile, they need to be constantly powered on to retain the stored data, even when they are not being accessed for memory read and write operations. Hence, the increased leakage power has become a major factor affecting the scalability of these memories. The alternative is to use non volatile memory (NVM) technologies, which have zero leakage power and can retain their data in the power off state. The NVMs can provide normally-off/instant-on capabilities, where the memory can be powered down in every idle cycle, thus helping in drastically reducing the power consumption.

Several NVM technologies are being considered to replace conventional CMOS memories. These include the Ferroelectric Random Access Memory (FeRAM), the Phase Change Random Access Memory (PCRAM), the Resistive Random Access Memory (ReRAM), the Spin Transfer Torque Magnetic Random Access Memory (STT-MRAM), among others. Among these, the spintronic-based STT-MRAM is the most promising candidate due to its several advantageous features such as scalability, high endurance, long retention, fast accesses and immunity to radiation-induced soft errors [12–17]. Moreover, this technology has process and voltage compatibility with the existing CMOS technology. These features give it an edge over other competing emerging NVM technologies.

Moreover, STT-MRAM has many favorable qualities of existing memory technologies. Fig. 1.4 compares the bit-cell architectures of SRAM, DRAM and STT-MRAM. As

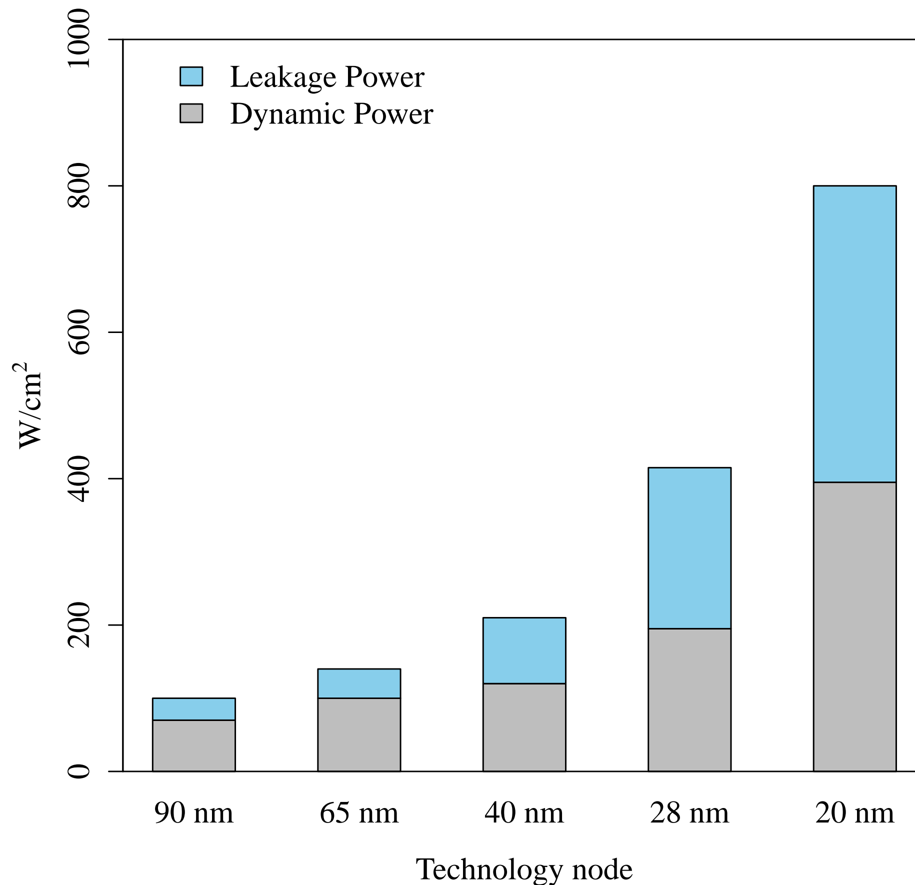
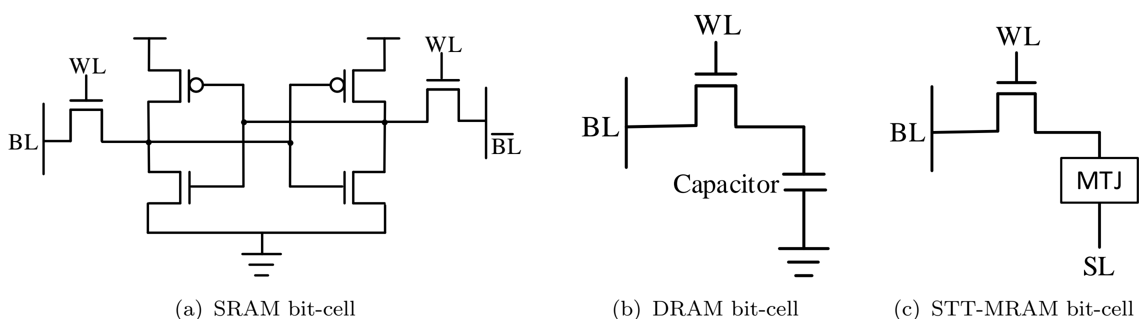


Figure 1.3: Leakage and static power at various technology nodes [4]

seen in the figure, the density of STT-MRAM is similar to that of DRAM, since both these technologies need only one transistor in the bit-cell. However, unlike DRAM, which uses a capacitor to store the data, STT-MRAM uses a Magnetic Tunnel Junction (MTJ) as the storage element. The MTJ is fabricated in the back-end-of-line (BEOL), above several metal layers and hence do not cause much area overhead to the bit-cell. In addition, since STT-MRAM is non-volatile just like the Flash memory, it can retain the stored data even during power-off. This is in contrast with the DRAM technology, which requires periodic refresh to retain the stored data. Furthermore, STT-MRAM has access speeds similar to those of SRAM.



(a) SRAM bit-cell (b) DRAM bit-cell (c) STT-MRAM bit-cell

Figure 1.4: Comparison of bit-cell architecture of SRAM, DRAM and STT-MRAM

	SRAM	DRAM	NOR-FLASH	NAND-FLASH	STT-MRAM
Cell Size (F^2)	**80 – 120**	**6 – 10**	**10**	**5**	**6-20**
Read Latency	**Low**	**Medium**	**High**	**High**	**Low**
Write Latency	**Low**	**Medium**	**High**	**High**	**Medium**
Write Energy	**Low**	**Low**	**High**	**High**	**Medium**
Other Power Consumption	**Leakage**	**Refresh**	**None**	**None**	**None**
Supply Voltage	**Low**	**Medium**	**High**	**High**	**Low**
Non-Volatility	**No**	**No**	**Yes**	**Yes**	**Yes**
Scalability	**Problem**	**Problem**	**Problem**	**Problem**	**Scalable**
Endurance	**High**	**High**	**Low**	**Low**	**High**
Soft-Error Impact	**Yes**	**Yes**	**No**	**No**	**No**

Table 1.1: Comparison of STT-MRAM with existing memory technologies (F–Feature size)

The parameters of STT-MRAM is compared with those of existing CMOS-based memory technologies in Table 1.1. As can be seen from the table, STT-MRAM triumphs conventional memory technologies in almost every aspect.

A typical memory hierarchy consists of mass storage devices (hard disk and Flash), main memory (DRAM), various levels of caches (SRAM) and registers as shown in Fig. 1.5. Such hierarchical memory organization is required to hide the access latency of the much slower memory from the fast processor. As we have seen, STT-MRAM can be used as SRAM, DRAM and Flash replacement, paving the way to becoming a universal memory since it can potentially fit into every level of the memory hierarchy [18, 19]. The viability of using STT-MRAM as DRAM, SRAM and Flash replacement is further illustrated by various recent industrial demonstrations [20–32].

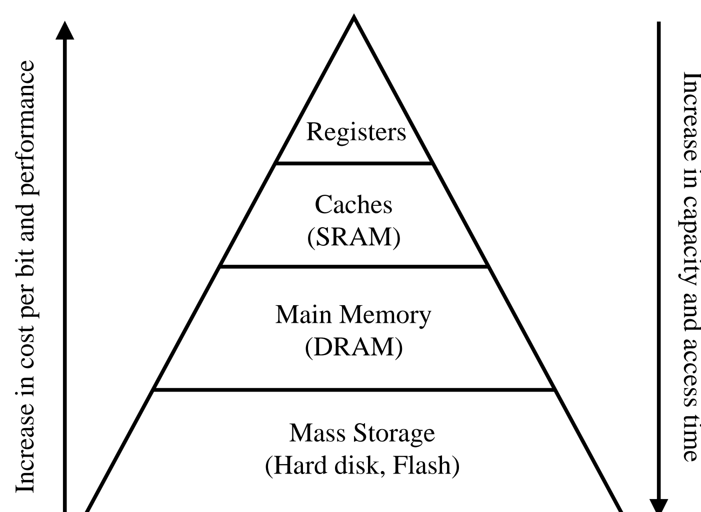


Figure 1.5: A typical memory hierarchy using conventional memories

1.1 Dissertation Contributions

As STT-MRAM technology gains in popularity, especially for low power applications such as the "Internet of Things" (IoT), the reliability challenges of adopting this technology need to be investigated. Since this technology is targeted for advanced nodes, manufacturing variations and operating conditions of voltage and temperature have a higher impact on its performance, energy and reliability. Furthermore, STT-MRAM exhibits stochastic switching behavior caused by thermal fluctuations, which is exacerbated by the impact of process variations.

The fabrication of STT-MRAM requires two processes, a magnetic process for the MTJ and a CMOS process for the access transistor and other peripheral components. Although the MTJ manufacturing process is compatible with the CMOS process, the defects and the associated fault models of STT-MRAM are different from those of conventional CMOS memories, due to the introduction of new materials and manufacturing steps.

The objective of this thesis is to analyze the impact of process variation, voltage, temperature, extreme parametric variations and manufacturing defects on the performance, energy and reliability of STT-MRAM. We have developed several tools and frameworks which can help in this analysis and enable the implementation of effective reliability and yield improvement techniques. Although this thesis focuses on the Spin Transfer Torque (STT) technology, the tools and methodologies developed can easily be extended to other spintronic-based memories such as the Spin Orbit Torque (SOT) memories [33].

Overall, the contributions of this thesis are listed below.

- **Variation-aware analysis and design space exploration** [34, 35]: This work analyses the impact of stochastic switching of the MTJ and process variations in the MTJ, access transistor and peripheral circuitry on the latency and energy of the entire memory system. For this purpose, we have developed a variation-aware design space exploration and memory configuration optimization tool, which can also report the failure rates for read, write and retention based on the required latency margins. Furthermore, this tool can evaluate the effectiveness of error correction techniques (such as ECCs) to reduce the failure rates. This variation-aware analysis can provide more realistic latency and energy values as compared to the conventional analysis based on nominal values.
- **Parametric Failure Modeling and Yield Analysis** [36, 37]: In this work, we model the impact of extreme parametric variations on the reliability failures and permanent faults. We consider the bit-cell and peripheral components as well as the spatial correlation among the bit-cells to get a fault distribution map of the entire memory array. Analysis of these fault maps can help in an effective design-for-yield exploration by evaluating the effectiveness of traditional error mitigation techniques (such as ECC or redundancy) or techniques specific to STT-MRAM, such as current boosting.

- **Defect Injection, Fault Modeling and Test Algorithm Generation** [38, 39]: This work deals with analyzing the impact of defects on STT-MRAM by considering the netlist as well as the layout. Both spot defects (resistive opens and shorts) and MTJ specific defects are considered. Moreover, the impact of temperature, voltage and process variation on the manifestation of defects are evaluated. The various fault models resulting from the injected defects are then analyzed to develop efficient test algorithms to cover the unique faults of STT-MRAM.

1.2 Dissertation Outline

This chapter presented the motivation and contributions of this thesis. The remainder of the thesis is organized into five chapters:

- Chapter 2 provides the preliminaries of the STT-MRAM technology. The reliability challenges associated with this technology are presented and the existing reliability improvement solutions are discussed.
- Chapter 3 presents a variation-aware analysis and design space exploration methodology for STT-MRAM. The resulting access latency, energy and failure rates are reported.
- Chapter 4 deals with parametric failure modeling and yield analysis. The fault distribution map of the memory array is presented along with the yield estimate. The efficacy of different yield improvement techniques is also evaluated.
- Chapter 5 discusses a defect injection flow for STT-MRAM by considering the netlist and layout. Based on the injected spot defects and MTJ specific defects, various fault models are identified and efficient test algorithms to cover them are developed.
- Chapter 6 concludes the thesis and provides some perspectives about future research directions on reliability aspects of emerging spintronic memories.

Chapter 2

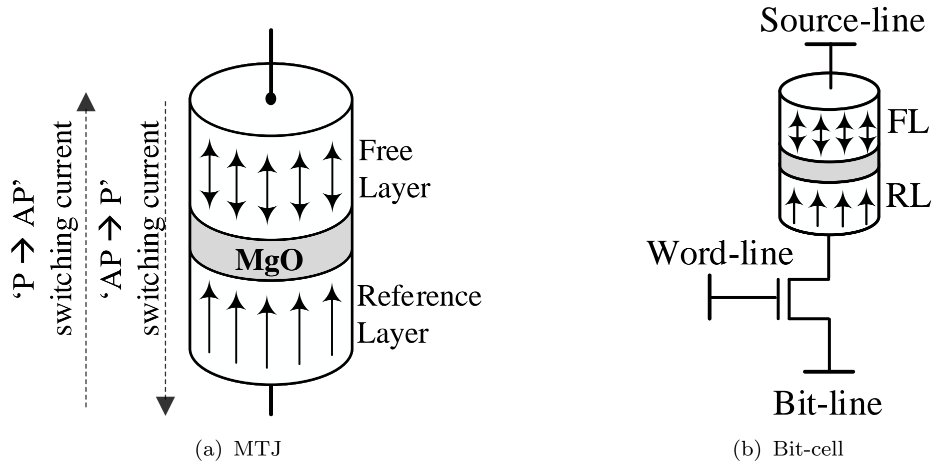
Background

This chapter presents an overview of the STT-MRAM technology and explains the reliability challenges associated with this technology. In particular, the impact of stochastic switching, process variation, and defects on STT-MRAM are explained and some of the existing solutions to mitigate these reliability challenges are discussed.

2.1 STT-MRAM Technology

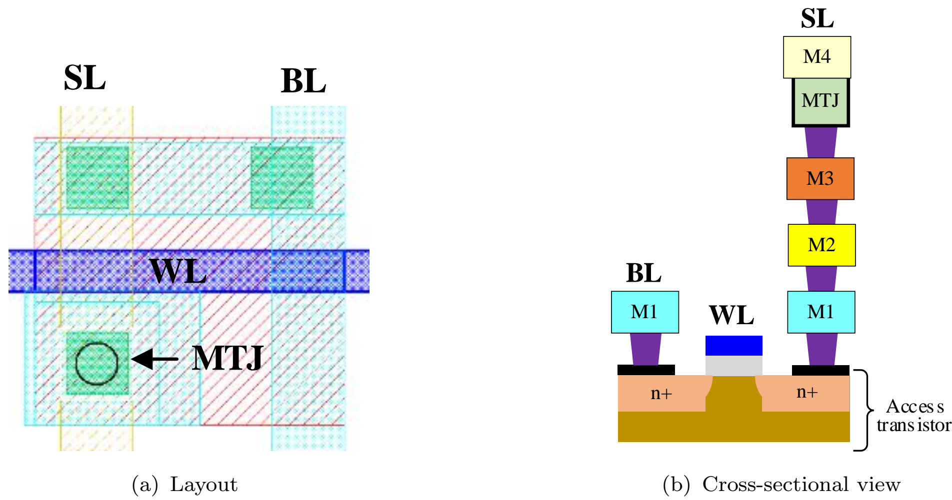
The *Magnetic Tunnel Junction* or *MTJ* is the basic storage element in STT-MRAM memory technology. It comprises of two ferromagnetic layers separated by a thin oxide layer (e.g., MgO) as shown in Fig. 2.1(a). The magnetic orientation of one of the layers is fixed and is known as the *Reference Layer* (RL) or the Pinned Layer. The magnetic orientation of the other layer, which is known as the *Free Layer* (FL), can be freely rotated. The resistance of the MTJ depends on the magnetic orientations of the FL with respect to the RL. When the magnetic orientation of the FL is parallel to that of the RL, the MTJ is in low resistance state ('P' state). On the other hand, when the magnetic orientation of FL is anti-parallel to that of RL, the MTJ is in high resistance state ('AP' state). The low resistance state is typically denoted by R_P and the high resistance state by R_{AP}. The magnetic orientations can be switched from one state to the other by passing a spin polarized current in the proper direction. To switch the MTJ from the anti-parallel (parallel) to the parallel (anti-parallel) state, the current has to flow from the FL (RL) to RL (FL). The switching in MTJ is asymmetric, which means that the switching latency from the parallel ('P') to the anti-parallel ('AP') state is different from (and higher than) that of the latency for 'AP' to 'P' switching.

To read the value stored in the MTJ, a low unidirectional current is passed through the MTJ, which is then sensed using a sense amplifier. The sensed current is then compared with a reference current to determine the state of the MTJ. When the sensed current is more that the reference current, the MTJ is in low resistance (R_P) state. Similarly, when the sensed current is less than the reference current, the MTJ is in high resistance (R_{AP}) state. The accuracy of sensing depends on the



(a) MTJ (b) Bit-cell

Figure 2.1: STT-MRAM storing device and its bit-cell architecture



(a) Layout (b) Cross-sectional view

Figure 2.2: Layout and cross-sectional view of STT-MRAM bit-cell

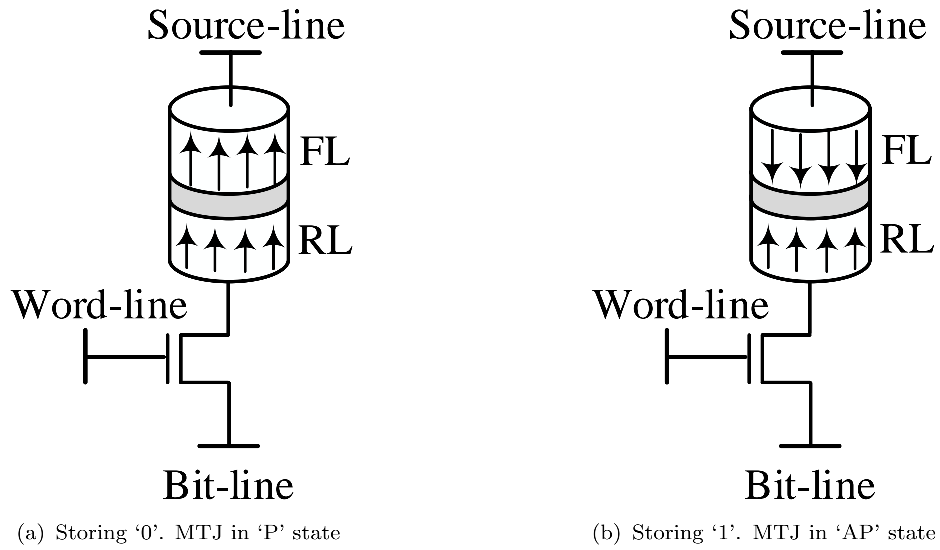
resistance difference between the R_P and R_{AP} resistance quantified by the Tunneling Magnetoresistance (TMR) ratio, given by

$$TMR = \frac{R_{\text{AP}} - R_{\text{P}}}{R_{\text{P}}}. \tag{2.1}$$

The *Thermal Stability Factor* (Δ) is an important parameter of the MTJ which is modeled as:

$$\Delta = \frac{V \cdot H_k \cdot M_s}{2 \cdot K_B \cdot T}, \tag{2.2}$$

where V, M_s, K_B, T and H_k are the volume of the free layer, the saturation magnetization, the Boltzmann constant, the temperature in Kelvin and the effective field anisotropy respectively.



(a) Storing '0'. MTJ in 'P' state

(b) Storing '1'. MTJ in 'AP' state

Figure 2.3: Bit-cell storing '0' and '1'

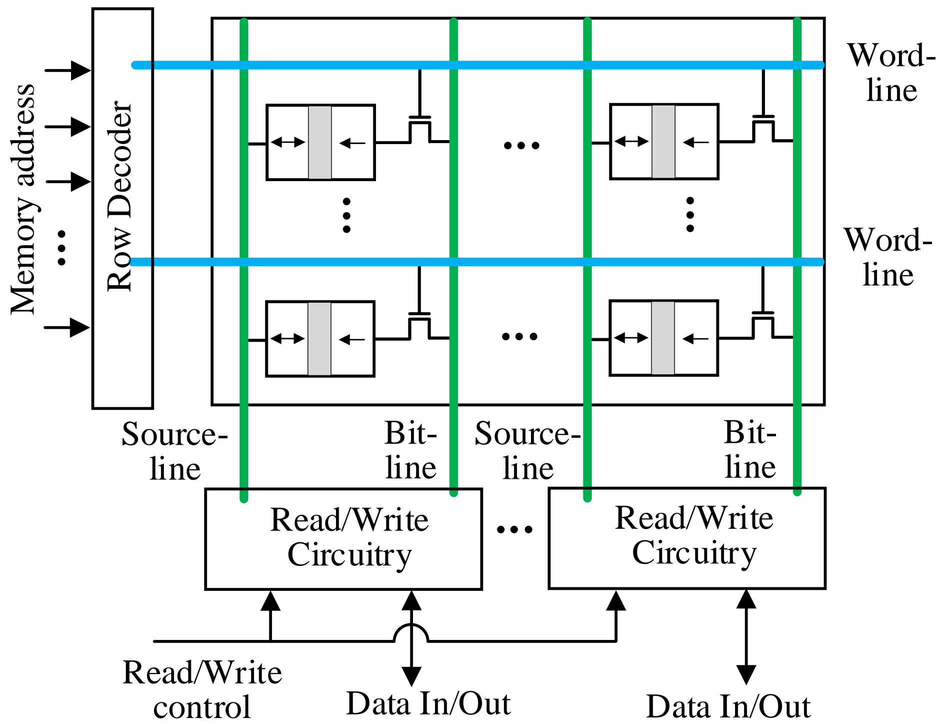


Figure 2.4: STT-MRAM memory array organization

A typical STT-MRAM bit-cell architecture is shown in Fig. 2.1(b). It consists of one access transistor (1T) and one MTJ (1MTJ) and is typically referred to as the 1T-1MTJ configuration. The NMOS access transistor is connected to the word-line and is used to select the MTJ for memory operations. In Fig. 2.2, the layout and the cross-sectional view of the 1T-1MTJ bit-cell is shown. The access transistor is fabricated in the front-end-of-line (FEOL), whereas the MTJ is fabricated in the back-end-of-line (BEOL) above several metal layers, as illustrated in Fig. 2.2(b).

Fig. 2.3 shows how a bit-cell can be used to store the '0' state and the '1' state. Here, the low resistance ('P') state is used to store '0' and the high resistance ('AP') state is used to store '1'.

A typical STT-MRAM memory array organization is illustrated in Fig. 2.4. It consists of the 1T-1MTJ bit-cell as well as the peripheral circuitry. The peripheral components are CMOS-based and include the row decoder and other read/write circuitry such as column decoders, sense amplifiers, column multiplexers, word-line drivers, and output drivers. The row decoder is used to select the correct memory row for read/write operation based on the incoming address. The read/write circuitry is used to read the value stored in the bit-cell or write a value to the bit-cell based on the selected memory address.

2.2 Read and Write Operations in STT-MRAM

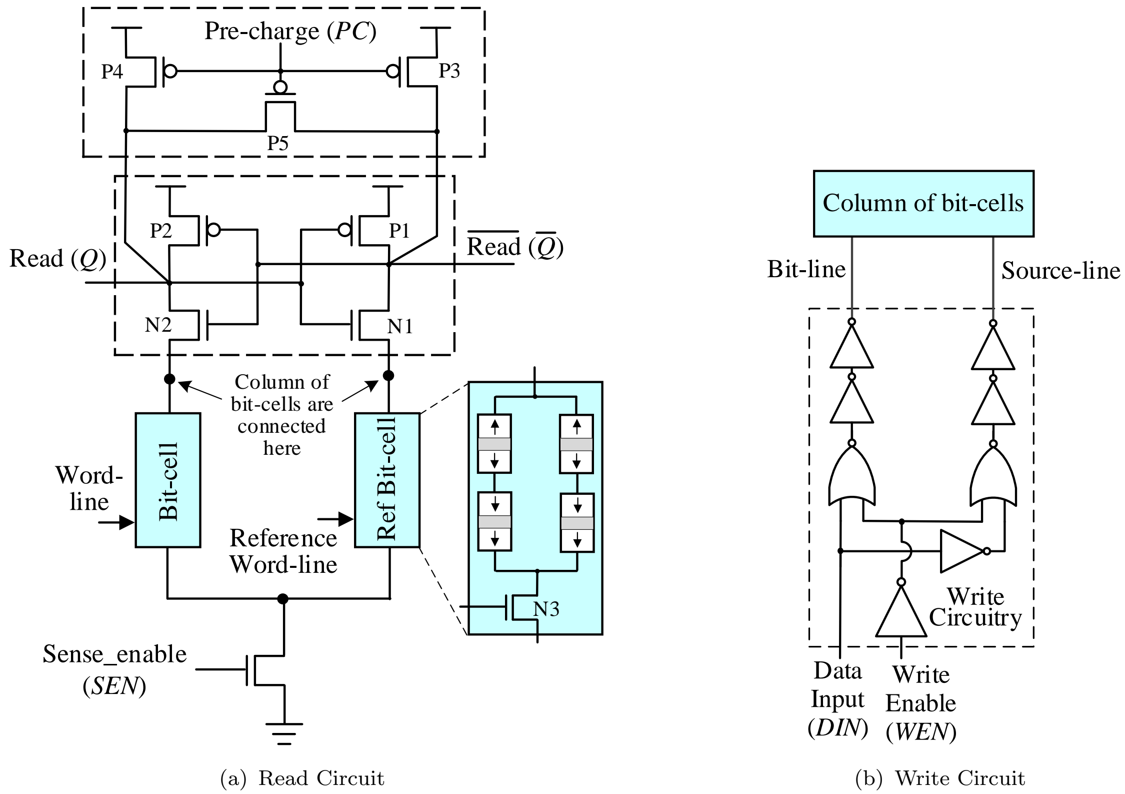
2.2.1 Read Operation

The TMR effect is used to read the value stored in an MTJ. Due to this effect, the resistance of the MTJ in the 'P' state is lower than that in the 'AP' state. The value stored in the bit-cell can be read using a sense amplifier. Various sensing schemes have been proposed for STT-MRAM. Among these, the conventional and the pre-charge based sensing schemes are the most popular [40].

In our work, we use the pre-charge sense amplifier (PCSA) as shown in Fig. 2.5(a). The PCSA consists of a pre-charge circuit which charges the output nodes (Q and $\overline{Q}$) to a high voltage (VDD) before the read operation starts. To start the read operation, the SEN signal is set to '1' and the output nodes Q and $\overline{Q}$ start discharging through the bit-cell and reference bit-cell, respectively. The resistance of the reference bit-cell is set as the average of the $\mathrm{R_{AP}}$ and $\mathrm{R_P}$ values, given by

$$R_{\text{ref}} = \frac{R_{\text{AP}} + R_{\text{P}}}{2}. \tag{2.3}$$

This means that $\mathrm{R_P} < \mathrm{R_{ref}}$ and $\mathrm{R_{AP}} > \mathrm{R_{ref}}$. Now, if the MTJ is in $\mathrm{R_P}$ state (storing a '0'), then the current through the bit-cell ($\mathrm{I_{cell}}$) will be higher that the current through the reference cell ($\mathrm{I_{ref}}$). In this case, the output node Q discharges faster than node $\overline{Q}$. This action is positively reinforced by the back-to-back connected transistor chain $N1\,P1\,P2\,N2$, resulting in Q to output '0' and $\overline{Q}$ to output '1'. Similarly, when the



(a) Read Circuit (b) Write Circuit

Figure 2.5: Typical read and write circuits for STT-MRAM

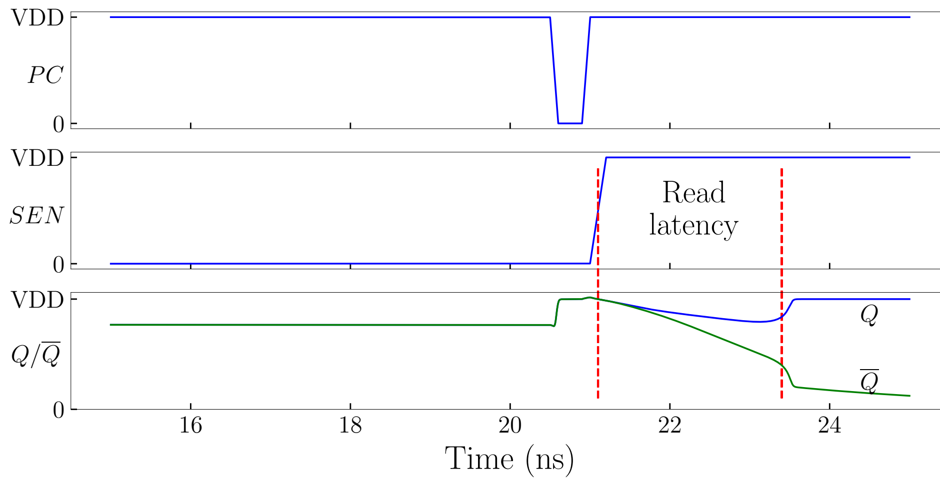


Figure 2.6: Waveform showing a read '1' operation in STT-MRAM

MTJ is in R_{AP} state (storing a '1'), $I_{cell} < I_{ref}$ and hence Q output will be '1' and $\overline{Q}$ output will be '0'. The read latency is the time required for the output nodes to output a stable value of '0' or '1'. For instance, the waveforms of a read '1' operation is shown in Fig. 2.6. Here the read latency is the time required for the output nodes Q and $\overline{Q}$ reach a stable value of '1' and '0', respectively.

2.2.2 Write Operation

The write operation in STT-MRAM is achieved by the Spin-Transfer Torque (STT) effect. Due to this effect, a spin polarized current can be used to switch the magnetic orientation of the free layer (FL) by transferring the angular momentum of the electrons. To switch the state of the FL, a write current which is greater than a threshold current (also called *critical current*, I_c) has to be passed for a sufficient duration. The MTJ can be switched from the 'AP' ('P') state to the 'P' ('AP') state by passing a current from the FL (RL) to RL (FL). Thus, the write current in an MTJ has a bidirectional path. The magnetic switching in MTJ is asymmetrical in nature due to its inherent properties. The switching from 'P'→'AP' takes considerably more time than that from 'AP'→'P' [12, 41].

A typical circuit for STT-MRAM write operation is shown in Fig. 2.5(b). To start the write operation, the WEN signal is first set to '1'. Then, the data to be written is made available at the data input DIN. When DIN is '1', the current flows from the bit-line to the source-line (i.e., from RL to FL, see Fig. 2.1(b)), thus switching the MTJ state to the 'AP' ('1') configuration. Similarly, when DIN is '0', current flows from the source-line to the bit-line (i.e., from FL to RL), causing the MTJ to switch to the 'P' ('0') state.

2.3 Parametric Variations in STT-MRAM

The fabrication of STT-MRAM requires two different fabrication processes, a magnetic process for the MTJs which is done in the BEOL and a CMOS process for the access transistors and peripherals, done in the FEOL. Variations in either process can affect the characteristics of the memory. These variations are a combination of both random effects and systematic effects. The random effects are caused due to issues such as dopant density fluctuations [42], whereas the systematic effects are caused by issues such as lithographic lens aberrations [43–45]. The systematic variations exhibit spatial correlation, which means that the neighboring cells have similar parameters values. This contrasts with random variations, where the parameters of each and every cell are random and have no correlation with those of the neighboring cells.

2.3.1 Random Variations

2.3.1.1 Variations in MTJ

Imperfections in the magnetic manufacturing process cause variations mainly in tunneling oxide thickness and the radius of the MTJ. These variations in turn affect the thermal stability factor (Δ) and the critical current (I_c), which are important parameters of the MTJ. These parameters primarily depend on the volume of the

free layer (V), which in turn depends on the radius (r) of the MTJ [46]:

$$I_c, \Delta \propto V \propto r^2 \tag{2.4}$$

The variations in r alters the switching threshold current (critical current), resistance values and the TMR [47, 48], significantly affecting the bit-cell latency for both read and write operations.

Moreover, the magnetic switching of MTJ is stochastic in nature due to random thermal fluctuations. The write probability of a bit-cell can be modeled by the following equation [49]:

$$WP_{\text{bit}}(t) = exp\left[\frac{-\pi^2(I-1)\triangle}{4(Ie^{C(I-1)t}-1)}\right],\ I = \frac{I_w}{I_c}, \tag{2.5}$$

where t is the write period, I_w is the write current, I_c is the critical current and C is a constant determined by the material and technology parameters. The Write Error Rate (WER) is given by:

$$WER_{bit}(t) = 1 - WP_{bit}(t). \tag{2.6}$$

For the read operation, a read decision failure occurs when a wrong decision is made by the sensing circuitry. In addition, STT-MRAM is also affected by read-disturb, where the read current accidentally flips the data stored in the MTJ. The read disturb probability (P_{RD}) is given by [50]:

$$P_{RD} = 1 - exp[-\frac{t_r}{\tau \cdot e^{\Delta(1-\frac{I_r}{I_c})}}], \tag{2.7}$$

where t_r is the read period, I_r is the read current, I_c is the critical current and τ is a constant with value 1 ns.

A retention failure in STT-MRAM happens when the magnetic orientation of the MTJ spontaneously flips due to thermal noise. The retention failure probability (P_{RF}) for a given time period (t) can be computed as [51]:

$$P_{RF} = 1 - exp[-\frac{t}{\tau \cdot e^{\Delta}}] \tag{2.8}$$

The retention failures are heavily dependent on temperature (T) since $\Delta \propto \frac{1}{T}$ (see Eq. 2.2). Hence an increase in temperature reduces Δ, thereby causing an increased number of retention failures.

2.3.1.2 Variations in Access Transistor

The STT-MRAM bit-cell is also influenced by variations in the CMOS access transistor caused by variations in the CMOS fabrication process, mainly due Random

Dopant Fluctuation (RDF), Line-Edge Roughness (LER) and Shallow-Trench Isolation (STI) stress [52, 53]. The standard deviation of the threshold voltage (σV_{th}) due to these random variations is given by the Pelgrom law [54, 55]:

$$\sigma V_{th} \propto \frac{1}{\sqrt{WL}}, \tag{2.9}$$

where L and W are the effective length and width of the transistors respectively.

2.3.1.3 Variations in Peripheral Circuitry

The peripheral components of an STT-MRAM memory consists of row decoders, column decoders, sense amplifiers, column multiplexers, word-line drivers and output drivers. The variations in the threshold voltage as per Eq. 2.9 affect the on-current of these CMOS based peripheral components causing variations in the read/write current of the bit-cell. Similarly, for the read operation, the variations in the sensing circuitry can cause read decision failures, especially for cells with low TMR values. These variations significantly affect the read/write and read-disturb error rates. To keep the error rates within acceptable limits, sufficient read/write pulse margins have to be provided. Hence the combined effect of bit-cell and peripheral variations significantly impact the overall access latency of the memory system. Furthermore, extreme parametric variations can cause the latencies to extend beyond the design margins, resulting in permanent faults.

2.3.2 Systematic Variations

In addition to the random variations described in Section 2.3.1, STT-MRAM bit-cell is also affected by systematic variations. These variations show strong spatial correlations, which means that the variations among neighboring cells are much smaller compared to cells that are far apart from each other.

2.4 Defects in MTJs and STT-MRAM array

The manufacturing process of STT-MRAM can lead to several defects and faults in MTJ. The defects in MTJ can be classified into four categories [56].

- Short fault: In this fault, the two ferromagnetic layers (i.e., RL and FL) are shorted together due to the sputtering effects while performing the ion beam etching. As a result, such MTJs can have very low resistance values depending on the damage in the barrier oxide layer.

- Open fault: MTJs can have an open connection because of the damage in via or metal connections that can lead to very high resistance value.
- Stuck-at-P fault: In this case, the MTJ is permanently or temporarily stuck to the 'P' magnetization state. In other words, the magnetization of the FL is either fixed to this state or requires very high current to change its state which is not possible in the given design environment.
- Stuck-at-AP fault: Similar to the previous fault, in this fault, the magnetic orientation of the FL of the MTJ is fixed permanently or temporarily to 'AP' state.

These MTJ defects can lead to write and read faults. The defects that can impact the switching of the MTJ will result in write faults. However, some MTJ faults, such as those affecting the oxide barrier thickness, affect the resistances of MTJ states, which in turn result in read faults. These defects can be modeled as resistance values. The defects during the fabrication process of other components of the STT-MRAM can be modeled as resistive opens and shorts.

2.5 Defect Tolerance Techniques

Error Correcting Code (ECC), Redundancy Repair (RR) and Fault Masking (FM) techniques are typically employed to mitigate faults in logic and memory chips [57]. ECCs are typically used to detect and correct reliability faults whereas RR and FM are typically used to repair permanent faults.

Process variations and stochastic switching leads to a long tail in the write latency distribution of STT-MRAM memories [58]. This necessitates large write margins to keep error rates within acceptable limits. ECC can be employed to reduce the write margin and correct some of the resulting errors, so as to meet the target reliability requirements. Various advanced error correction techniques have also been proposed for STT-MRAM [57, 59–61], which provide improvements over the conventional ECC technique.

An ECC scheme for correcting e errors in k data bits is represented as ECC (n, k, e) where n is the word size. Here the number of check bits is $n - k$. The storage overhead of ECC ($\frac{n-k}{n}$) increases as the number of errors e in the data increases.

If WER_{bit} is the WER for a bit, then the WER for a word WER_{word} is given by:

$$WER_{word} = 1 - \sum_{i=0}^{\mathrm{e}} \binom{k}{i} \cdot WER_{bit}^{i} \cdot (1 - WER_{bit})^{k-i} \tag{2.10}$$

If no ECC is implemented (i.e. $e = 0$), then from Eq. (2.10):

$$WER_{word} = 1 - (1 - WER_{bit})^{k} \tag{2.11}$$

Clearly, the WER_{word} in Eq. (2.10) is much less than that in Eq. (2.11). An increase in e reduces the WER_{word}.

The ECC scheme is responsible for encoding of the original data bits to be stored in the memory, and decoding of the stored data bits in the memory into the original data bits. This results in additional encoding and decoding circuity which in turn increases the access latency, chip area and power consumption. Hence an appropriate combination of the write margin and ECC should be selected based on the conflicting requirements of ECC overhead and acceptable WER.

In RR technique, a faulty row or column is replaced with a spare one. Hence this technique results in a large overhead, since an entire row/column is required to repair even a single fault. This can be overcome by FM, but at the cost of more complex addressing and accessing schemes. There are also some solutions proposed to improve the redundancy efficiency by combining the ECC and RR techniques [62].

2.6 Summary

In this chapter, an overview of the STT-MRAM memory technology was presented. The various reliability challenges associated with this technology were explained, followed by a discussion on some of the existing mitigation techniques to overcome the reliability issues.

Chapter 3

Variation-aware STT-MRAM Analysis and Design Space Exploration

3.1 Overview

In order to support early-stage memory design space exploration, an estimator tool for STT-MRAM technology is needed for a system-level evaluation before (or without) fabricating a real chip. NVSim [5] is the most popular among existing estimator tools for non-volatile memories. It estimates access latency, access energy and silicon area of non-volatile memories for an optimal memory configuration satisfying input design options. However, it has been shown that the latency and energy of STT-MRAM are severely affected by process variations and stochastic nature of the MTJ switching [58, 63, 64]. A realistic estimation of latency/energy of STT-MRAM memories requires taking into account the effects of these variations in the analysis. NVSim falls short in this regard since it does not consider variability in its analysis. A comprehensive variation-aware estimation and optimization tool for energy, performance and reliability is missing for system-level evaluation of STT-MRAM based memories.

3.1.1 Related Work

There are some works considering variability in STT-MRAM [47, 58, 63–67]. An adaptive write current boosting technique is proposed in [58] to fix the worst-case write latency due to process variations. The work in [47] considers the variations in the bit-cell parameters such as the tunneling oxide thickness and cross-sectional area to come up with an optimal memory configuration, access transistor size and read drive circuitry to maximize the yield. In [65], the authors propose a methodology for an STT-MRAM cell reliability prediction under the joint impact of fabrication and aging-induced process variability, supply voltage and temperature variations. A

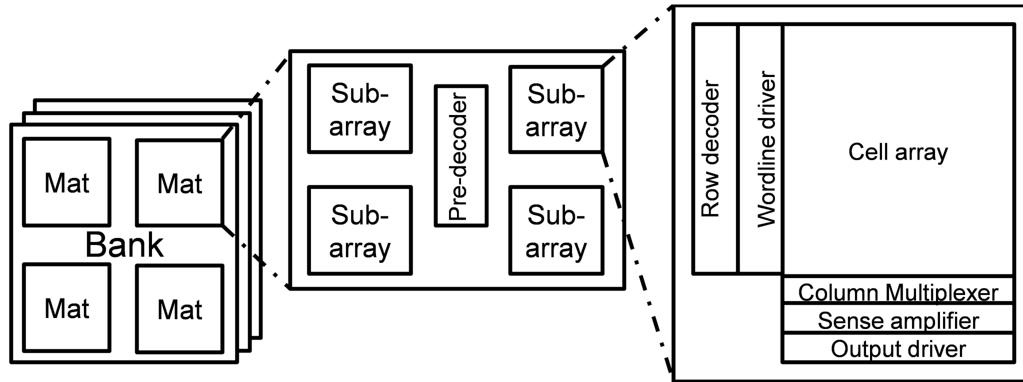


Figure 3.1: Memory array organization as proposed in NVSim [5].

unified device-circuit-architecture co-design strategy to optimize the yield of STT-MRAM, in which the memory array is optimized by tuning the bit-cell parameters together with different ECC schemes, is proposed in [63]. The work in [66] considers variations only in the read circuitry and analyzes the impact of variability on sensing schemes. The impact of process variations on the stochastic switching behavior of MTJ is investigated in [64] and probabilistic design techniques are proposed to enhance the performance while maintaining a low write failure probability. In [67], the authors have modeled the combined effect of process variation and stochastic write behavior on the write latency of STT-MRAM bit-cell. However, none of the above works provide an estimation of the overall read/write latency and energy of an entire memory system consisting of STT-MRAM bit-cells and CMOS peripherals.

The most popular tool for a memory system-level analysis for emerging non-volatile memories is NVSim [5], which is based on the well-known CACTI [68]. A typical memory system as modeled in NVSim [5] consists of banks, mats and subarrays as shown in Fig. 3.1. Such hierarchical organization helps in reducing the wire routing delay and memory array access delay, thereby reducing the overall access latencies. The subarray is the basic building block of the memory system. It consists of the bit-cell array and the peripheral components as shown in Fig. 3.2. The peripheral components are CMOS based and include row decoders, column decoders, sense amplifiers, column multiplexers, word-line drivers and output drivers. Several subarrays and a pre-decoder block form a mat. Multiple mats connected in either H-tree or bus-like manner form a bank. Each bank can be operated independently and is a fully functional memory unit. A non-volatile memory chip can have several banks.

NVSim finds an optimal memory configuration based on input design options and reports the latencies and energies of the overall memory system. The analysis in NVSim is based on nominal values and does to take variations into account. NVSim-VX^{S} [69] is an improved variation-aware version of NVSim. However, the variation-aware statistical model in NVSim-VX^{S} is limited to only the bit-cell level for the write operation. It does not consider the read operation in its analysis. In NVSim-VX^{S}, the bit-cell WER is reported for ‘0’→‘1’ and ‘1’→‘0’ transitions. It also includes a simple extension of the cell-level WER to the block level based on a user input switching pattern. However, the user input switching pattern does not reflect the switching

Figure 3.2: High-level composition of a subarray with individual components and their latencies.

pattern of the optimal memory configuration and hence the block-level extension in NVSim-VX^{S} does not accurately reflect the energies/error rates.

3.1.2 Contributions

None of the existing estimator tools for STT-MRAM considers the combined impact of variability and stochasticity in the bit-cell and process variations in the peripherals for a system-level analysis. To bridge this gap, we have developed a *Variation Aware STT-MRAM Analysis and Design Space Exploration Tool* (VAET-STT) [34, 35], an early stage design exploration tool for STT-MRAM, which considers process variation, stochasticity and reliability requirements in its analysis and memory configuration optimization. In this tool, we employ a hierarchical and hybrid approach of analytical curve fitting and Monte-Carlo simulations for estimating STT-MRAM access latencies, energies and error rates. We also allow the user the ability to evaluate the effectiveness of various ECC schemes on the performance of STT-MRAM memories. The tool can also report various failure rates considering the impact of process variations.

Overall, our contributions are as follows:

- We develop a statistical model for the entire memory architecture including the bit-cell array and the peripheral components.
- The latency and energy values are estimated based on the required margins to satisfy the reliability (overall system-level failure rate) requirements provided by the user.
- We analyze the impact of variations on various failure rates, particularly read disturb and retention failures, for the entire memory.

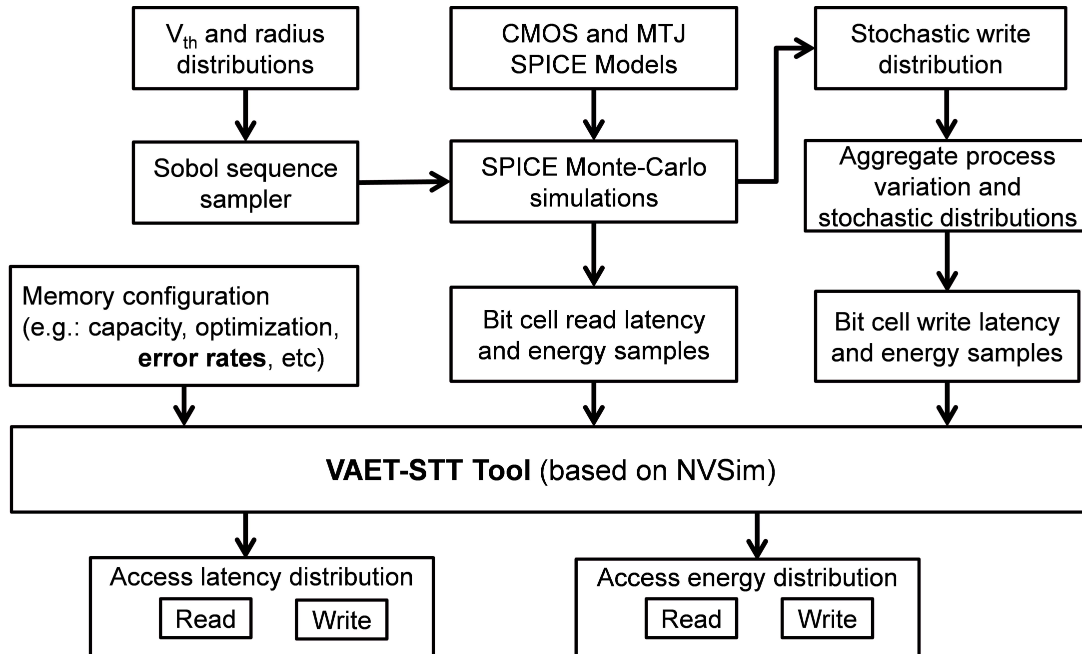


Figure 3.3: Proposed Variation Aware STT-MRAM Analysis and Design Space Exploration Tool (VAET-STT) flow.

- We evaluate various ECC schemes and analyze their impact on the latency and energy of the entire memory.
- We perform a design space exploration that generates an optimal memory configuration satisfying the input reliability requirements.

3.2 Variation-aware Analysis Framework

3.2.1 Overview

Fig. 3.3 presents the *Variation Aware STT-MRAM Analysis and Design Space Exploration Tool* (VAET-STT) flow. We have used a hierarchical and hybrid approach of analytical curve fitting and Monte-Carlo to obtain the overall memory latency and energy distribution samples. This method involves fitting each of the constituent latency/energy samples to known distributions and then combining these distributions using Monte-Carlo. This approach allows us to have a good tradeoff of fast run-time, since component-level samples are taken from their distributions, while achieving high accuracy.

It has been shown that for the bit-cell and the peripherals, the latency and energy follow known distributions for reasonable accuracy [69, 70]. Accesses to a memory array does not read/write a single cell but a line which has N bit-cells. The effective cell latency is then the maximum of these N latencies which is shown to follow a Generalized Extreme Value (GEV) distribution [71]. Hence, at the system-level, when parameters

following diverse and different distributions (e.g., normal, log-normal, GEV) are combined, they cannot simply be fitted to known distributions. For instance, in [70], the cell latency GEV distributions have been approximated with a normal distribution to facilitate an equation based analytical approach. However, such approximations can compromise the accuracy of the results. Hence, at the system-level, a sampling approach, such as Monte-Carlo can be used to get the overall latency and energy distributions.

By combining the bit-cell and peripheral distributions for different operating temperatures and technology nodes using Monte-Carlo, the system-level results can be obtained to reflect the impact of temperature or technology node on the variation of overall latencies and energies.

3.2.2 Component-level Variation Analysis

3.2.2.1 Bit-cell Latency

The first step in the latency calculation is to get the latency samples at the bit-cell level. The bit-cell latency samples are obtained from Monte-Carlo SPICE simulations of the standard 1T-1MTJ bit-cell. For simplicity, bit-cell loading effect is not considered in this work. The variations in the CMOS components are lumped into normally distributed threshold voltage (V_{th}) variations according to the Pelgrom law [54, 55]. For the MTJ, we have assumed a normally distributed radius variation. We then generate samples for V_{th} and radius using low-discrepancy Sobol sequence [72]. As compared to the conventional Monte-Carlo approach, this method requires 2-8$\times$ less samples [73]. The V_{th} and radius samples thus obtained are then used for performing Monte-Carlo SPICE simulations on the bit-cell to obtain the read/write latency and energy samples.

The read latency samples obtained using the above method are the final samples considering the impact of process variations. However, the write latency is impacted by stochastic switching in addition to process variations. The radius variations impact the parameters Δ and I_c of the stochastic switching in Eq. (2.5) according to Eq. (2.4). To consider the effect of these variations, we aggregate the stochastic distribution in Eq. (2.5) for each process variation sample using the approach described in [67] to get the bit-cell write latency samples.

The Quantile-Quantile (Q-Q) plot of the bit-cell read latency and write latency is shown in Fig. 3.4. It can be seen that the read latency follows an extreme value distribution whereas the write latency follows a log-normal distribution. If m is the mean and d is the standard deviation of the samples, then the parameters of the latency distributions can be calculated as follows.

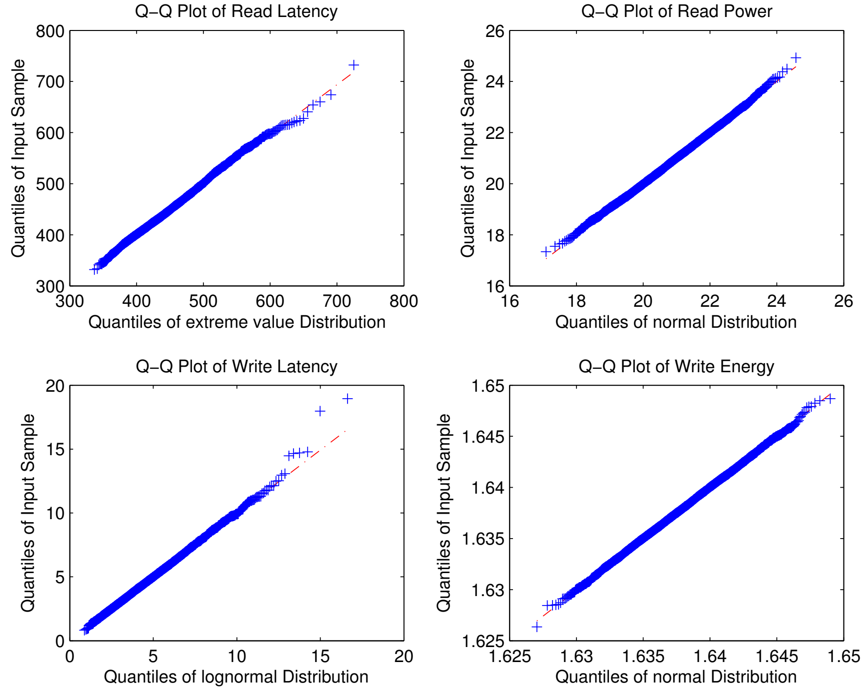


Figure 3.4: Q-Q plots for bit-cell read/write latency and energy. For setup please refer to Section 3.3.1.

The parameters of the extreme value distribution (μ and β) are given as:

$$\beta = \frac{d \times \sqrt{6}}{\pi},$$
$$\mu = m - \gamma\beta,$$

where $\gamma = 0.5772$ is the Euler-Mascheroni constant.

Similarly, parameters of the log-normal distribution (μ and σ) are calculated as:

$$\mu = \log_e\left(\frac{m}{\sqrt{1 + \frac{d^2}{m^2}}}\right),$$
$$\sigma = \sqrt{\log_e\left(1 + \frac{d^2}{m^2}\right)}.$$

3.2.2.2 Bit-cell Energy

For the read operation, the bit-cell level read power samples are obtained from Monte-Carlo simulations in the presence of variations. The read power follows a normal distribution as shown by the Q-Q plot in Fig. 3.4. These generated samples are used to calculate the parameters (μ and σ) of the normal distribution. For a normal distribution, the mean and standard deviation of the samples is equal to the μ and σ of the distribution respectively. If P_r is the read power and t_r is the read period,

then the read energy is $P_r \times t_r$. Please note that t_r is fixed according to the required error rate.

Similar to the read operation, the write period t_w is fixed based on the target WER. The write energy E_w is then calculated as given below:

$$E_w = I_1^2 \times R_1 \times t_s + I_2^2 \times R_2 \times (t_w - t_s)$$

where I_1 and R_1 are the write current and MTJ resistance before switching and I_2 and R_2 are the write current and MTJ resistance after switching respectively and t_s is the switching latency. The write energy also follows a normal distribution whose parameters are calculated in a similar manner as that of the read energy distribution.

3.2.2.3 Peripheral Latency and Energy

The process variations in the CMOS peripheral components can be lumped into variations in the threshold voltage (V_{th}). The V_{th} variations are assumed to be normally distributed with the standard deviation (σV_{th}) as given in Eq. (2.9) [54, 55]. This approach has already been used in process variation aware SRAM tools like VARIUS [70]. The V_{th} variations in turn affect the on-current and mobility which causes variations in the latency and dynamic energy of individual peripheral component. Assuming normally distributed V_{th}, the on-current and mobility variations are obtained from SPICE simulations by performing a sensitivity analysis. We then run Monte-Carlo simulations and obtain samples for the peripheral component latencies and energies. The samples are then used to obtain the distribution parameters (μ and σ) of the normally distributed peripheral latencies and energies.

3.2.3 System-level Variation Analysis

3.2.3.1 Latency Calculation

We follow a two-step process to obtain the latencies at the system-level.

- The latency distributions of the individual components in Fig. 3.2 are extracted by Monte-Carlo and fitted to known distributions such as normal, log-normal and GEV.
- The system-level Monte-Carlo takes samples from the fitted distributions in the previous step and use them to calculate the total read and write latencies.

More details regarding the process is as follows.

The statistical latency distribution of the bit-cell is extracted by running SPICE simulation as explained in Section 3.2.2.1. The latency distributions of CMOS based peripheral components shown in Fig. 3.2 are extracted internally by varying the

transistor parameters based on Pelgrom model [54, 55] for local variations in smaller Monte-Carlo loops. It is worth mentioning that the parameters should be assigned randomly to each and every transistor in the CMOS circuits. The latencies thus extracted from Monte-Carlo are fitted to known distributions which is used in the system-level Monte-Carlo.

In the system-level Monte-Carlo loop, samples are drawn from the extracted distributions in the previous step, and the total latency is calculated by combining the component latencies (the samples) in an iterative manner. The combining process is either by taking the maximum of the latency samples or adding the latencies. When a number of instances, say M, of a component C is accessed in parallel, we take M samples from the corresponding distribution ($\{l_{C,1}, ..., l_{C,M}\}$) and obtain the maximum of these M latencies to generate a single sample of latency. For example, if a memory array has N_C columns (see Fig. 3.2), a single sample of the memory array latency is obtained by taking the maximum of N_C randomly picked samples from the bit-cell latency distribution. Therefore, the latency of memory array can be explained as:

$$\begin{aligned} l_{cell,i} &\sim \text{Distribution}\{\text{Cell Latency}\}, \\ L_{cell} &= \text{Max}\{l_{cell,1}, ..., l_{cell,N_C}\} \sim \text{GEV}, \end{aligned}$$

where $l_{cell,i}$ is a random sample taken from the distribution of memory cell latency. The number of columns in a typical memory array is large, hence, L_{cell} can be explained by a GEV distribution, because the maximum of a number of *independent and identically distributed* (i.i.d.) random variables is taken. Similarly, the random variables L_{RD}, L_{BL}, L_{BLM}, L_{BLMD}, L_{SA}, L_{SAM}, and L_{SAMD} follow GEV distributions, because they are the maximums of i.i.d. random variables from the distributions of Row Decoder, Bitline, Bitline Mux, Bitline Mux Decoder, Sense Amplifier, Sense Amplifier Mux, and Sense Amplifier Mux Decoder, respectively. Furthermore, since all the decoders are invoked in parallel, it is possible to simplify the latencies of all decoders into a single *decoder latency* random variable:

$$L_D = \text{Max}\{L_{RD}, L_{BLMD}, L_{SAMD}\}.$$

When the elements are not accessed in parallel, the latency is the summation of the latencies of individual components. Thus, the total read latency of the memory subarray is the summation of the decoder latency (L_D), cell read latency (L_{cell}), bitline latency (L_{BL}), and sense-amplifier latency (L_{SA}) as given below:

$$L = L_D + L_{cell} + L_{BL} + L_{SA}.$$

In this work, we have assumed current sensing scheme for read, and used the L_{BL} equation similar to that of NVSim. L_{SA} is obtained from SPICE simulations.

The latency calculation flow described above is applicable to read operation. The write latency can also be calculated in a similar manner by appropriately combining the latencies of the individual components in the write path.

3.2.3.2 Energy Calculation

The bit-cell dynamic read/write energy depends on the read/write pulse which is fixed based on a target error rate as explained in 3.2.2.2. The total dynamic energy of the memory subarray is the summation of the dynamic energies of all the components. Similar to the latency calculation, the overall dynamic read and write energy samples are obtained by combining the bit-cell energy distribution and the peripheral energy distributions using Monte-Carlo method. We follow the same flow as the one performed for total latency calculation with the exception that when M instances are accessed, we take summation of the energies instead of taking the maximum. Using the same notation explaining the latencies, we define E_{cell}, E_{RD}, E_{BL}, E_{BLM}, E_{BLMD}, E_{SA}, E_{SAM}, and E_{SAMD} which are the summation of i.i.d. samples from their corresponding energy distributions. For example, the dynamic energy of the memory array is the summation of N_C instances from cell energy distribution:

$$e_{cell,i} \sim \text{Distribution}\{\text{Cell Energy}\},$$
$$E_{cell} = \sum_{i=1}^{N_C} e_{cell,i}.$$

Based on the *Central Limit Theorem*, we can conclude that the energy distribution would be a normal distribution. Additionally, the total dynamic energy is calculated as:

$$\begin{aligned} E = {} & E_{cell} + E_{RD} + E_{BL} + E_{BLM} \\ & + E_{BLMD} + E_{SA} + E_{SAM} + E_{SAMD}. \end{aligned}$$

The total leakage power can also be obtained in a similar manner. Since the leakage power of STT-MRAM bit-cell is zero, the total leakage power is the sum of the leakage powers of the peripheral components.

3.2.4 Reliability Analysis

3.2.4.1 Errors in STT-MRAM

The STT-MRAM bit-cell is subjected to various failure mechanisms as explained in Section 2.3. The error rates for the read and write operations, namely the Read Error Rate (RER) and the Write Error Rate (WER), are calculated based on a given read/write period. For STT-MRAM, the WER for '0'→'1' switching is different from that of '1'→'0' switching. In our analysis, we have assumed worst case switching of the cells, since the switching pattern of the memory array is not known at design time. Since the bit-cell latencies are fitted to known distributions, the Cumulative Distribution Function (CDF) of the respective distributions is used to calculate the RER/WER. If T is a random variable representing the bit-cell read/write latencies

and t_P is the read/write period, then the bit-cell RER/WER e is given by:

$$e = P(T > t_P) = 1 - F_T(t_P), \tag{3.1}$$

where $F_T(t)$ is the CDF of T.

The bit-cell errors can be assumed to be independent, and hence the block-level error rate E for a memory array with N_C columns can then be calculated by the following equation [47]:

$$E = 1 - (1 - e)^{N_C} \tag{3.2}$$

As explained in Section 2.3, STT-MRAM is also affected by read disturb and retention failures. These failures follow the probability distributions given in Eq. (2.7) and Eq. (2.8). We have used a Δ of 40 and $\frac{I_r}{I_c}$ of 0.25 to model these failures.

3.2.4.2 Error Mitigation using ECCs

ECCs can be employed to reduce the read/write margins while maintaining acceptable error rates. The block-level error rate E of an ECC scheme which can correct up to k errors in N_C bits can be computed using the binomial distribution as:

$$E = 1 - \sum_{i=0}^{\mathrm{k}} \binom{N_C}{i} \cdot e^i \cdot (1 - e)^{N_C - i} \tag{3.3}$$

Note that for no ECC (i.e. $k = 0$) Eq. (3.3) can be reduced to Eq. (3.2).

In our tool, the write period t_w is fixed based on a user input target WER. If the target error rate is E for a memory array with N_C columns and no ECC, the error rate of a bit-cell e can be derived from Eq. (3.2) as:

$$e = 1 - (1 - E)^{(\frac{1}{N_C})}.$$

In case of an ECC with k bit error correction, the value of e has to be calculated from Eq. (3.3) by numerical methods. Based on e, the quantile function of the write latency distribution is used to find the write period t_w. The quantile is defined as the inverse of the CDF. From Eq. (3.1), we get

$$\begin{aligned} F_T(t_w) &= (1 - e), \\ t_w &= \text{quantile}(1 - e). \end{aligned} \tag{3.4}$$

It can be deduced that an ECC scheme which can correct more errors (higher value of k) leads to a lower write period t_w and hence lowers the overall latency of the memory system. However, increasing the error correction capabilities comes at the cost of increasing complexity of the encoder and decoder circuitry which may erode

some of the gain in the write latency reduction. Therefore, these trade-offs are very important and is integrated in our analysis tool.

3.2.5 Design Space Exploration

The NVSim tool has an optimization engine which performs a design space exploration and finds an optimal memory configuration based on an input design metric. For instance, if the user wants to find a write latency optimized design, NVSim calculates the overall write latency for different memory configurations and finds the configuration which has the minimal write latency. However, this optimization is based on nominal values of bit-cell and peripheral latencies/energies. A more realistic approach would be to find a *variation-aware* optimal memory configuration considering the impact of process variation.

The bit-cell read/write energy distributions, which depend on the read/write period t_P, is affected by the target error rate. The read period can also be calculated in a similar manner as that of the write period in Eq. (3.4). The read/write period, together with the latency distributions of the peripheral components, are then used to calculate the overall read/write latency. Similarly, the overall read/write energies are obtained by combining the read/write energy distributions of the bit-cell as well as the peripheral components.

To perform a variation-aware memory design space exploration, the overall latency/energy distributions should be obtained for each iteration of the memory configuration. This involves running multiple Monte-Carlo simulations for each iteration which require long run-times. Several techniques can be employed to solve this problem. For instance, in a latency optimized design, based on the trends from the previous iterations, certain memory configurations which tend to increase the latency can be excluded from the exploration. Another approach would be to fix the number of banks and/or mats and perform a limited design space exploration. The optimization engine of NVSim is changed to use 0.9987 quantile (equivalent to $\mu + 3\sigma$ of a normal distribution) of the overall latency/energy distribution as the optimization metric and perform a variation-aware memory configuration optimization. Our results show that the *variation-aware* optimal memory configuration could be different from that of the nominal case.

3.3 Results

3.3.1 Experimental Setup

We have employed the TSMC SPICE models for the CMOS access transistor and the Perpendicular Magnetic Anisotropy (PMA) MTJ model from [74] for the bit-cell simulations. The parameters of our MTJ model are given in Table 3.1. The MTJ radius variation is assumed to be 5% whereas the variations in the CMOS

Table 3.1: Parameters of PMA MTJ at 45 nm and 65 nm

Parameter	Symbol	Value (45 nm)	Value (65 nm)	Unit
Resistance area product	RA	6.12	6.12	$\Omega\mu m^2$
Tunneling magnetoresistance	TMR	1.23	1.23	
Thermal stability factor	Δ	40	40	
Critical current (‘P’→‘AP’)	I_c	33.75	78.52	μA

components are assumed to follow the Pelgrom law [54]. We have not considered the oxide thickness variation in this work, due to the limitations of our MTJ model. However, this can easily be added to the framework if it is supported by the used MTJ model. We have extended NVSim [5] to include process variations in the peripheral components. For a fair comparison with the variation-aware approach, we have used the mean of the bit-cell read/write distribution samples as the read/write period for the nominal case.

We have done our analysis on a 128 KB RAM with word size of 256 bits. We have not considered the global interconnect delay in this work, since our analysis shows that it contributes to only a small percentage of the total delay, especially for memories less than 1 MB in size. We have considered worst case switching of the MTJ (i.e., ‘P’→‘AP’ switching) since reliability requirements have to be met for the worst case conditions. The results have been generated for 45 nm node and 65 nm node at various temperatures (from 300K to 400K), to show the trend of technology scaling on the variability parameters of the entire STT-MRAM memory. We have run Monte-Carlo at the system-level to generate 10000 samples for read/write latencies and energies.

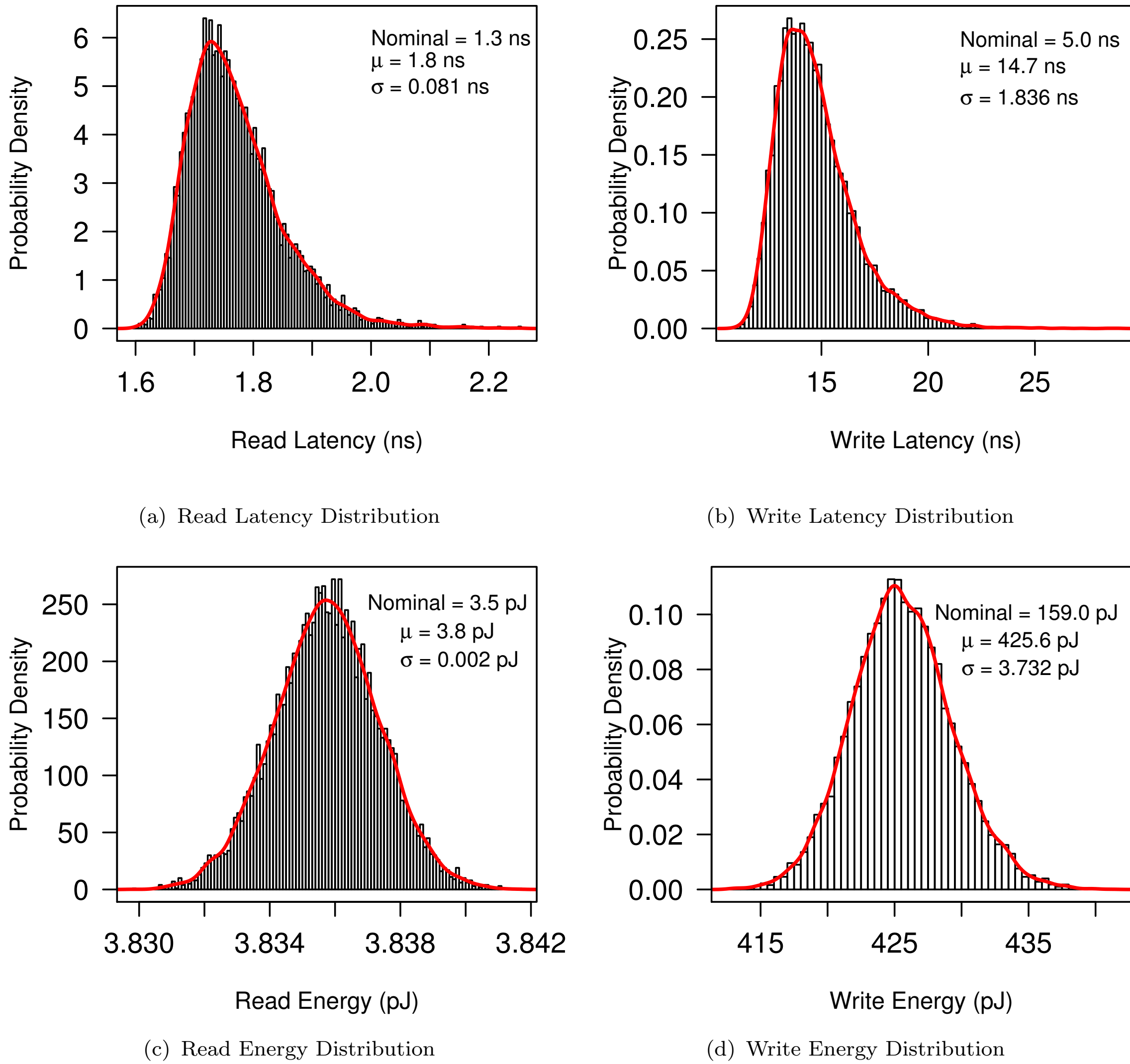
3.3.2 Results

3.3.2.1 Overall Latency and Energy Distributions

The results for overall read/write latencies and energies of the memory are given in Fig. 3.5. The figure also contains the nominal values which are obtained by summing the mean of each of the distributions. The results show that process variations cause a significant increase in the latency and energy values. The variation in energies ($\frac{\sigma}{\mu}$) is less compared to that of latencies. This is because latency calculation involves calculating the maximum of a number of components whereas energy calculation involves calculating the average. The variation of the overall write latency distribution for different temperatures is shown in Fig. 3.6. It can be seen that as the temperature increases from 300K to 375K, its impact on the write latency also increases as seen by the increase in $\frac{\sigma}{\mu}$ of the distribution.

3.3.2.2 Latency vs Error rates

Fig. 3.7 presents the overall latencies for different values of input error rates. As the error rates decrease, the read/write latencies increase. It can be seen from Fig.



(a) Read Latency Distribution

(b) Write Latency Distribution

(c) Read Energy Distribution

(d) Write Energy Distribution

Figure 3.5: Overall latency and energy distributions for a subarray size of 1024×1024 at a temperature of 300K for 45 nm node.

3.7 that the read latency variation is linear whereas the write latency variation is non-linear with respect to the logarithm of the probabilities. This is because the total latency is dominated by the bit-cell level latency which is log-normal for write and follows an extreme value distribution for read. It can be verified that the quantiles of an extreme value distribution is linear whereas the quantiles of a log-normal distribution is non-linear with respect to the logarithm of the probabilities.

3.3.2.3 Read Disturb

The read operation in STT-MRAM is also affected by read disturb. Fig. 3.8 shows the read disturb probabilities for different read periods. Even though a higher read latency leads to a lower RER as per Fig. 3.7, it will lead to increased read disturb probability as shown in Fig. 3.8. Hence the read period should be fixed taking into account the conflicting requirements for RER and read disturb.

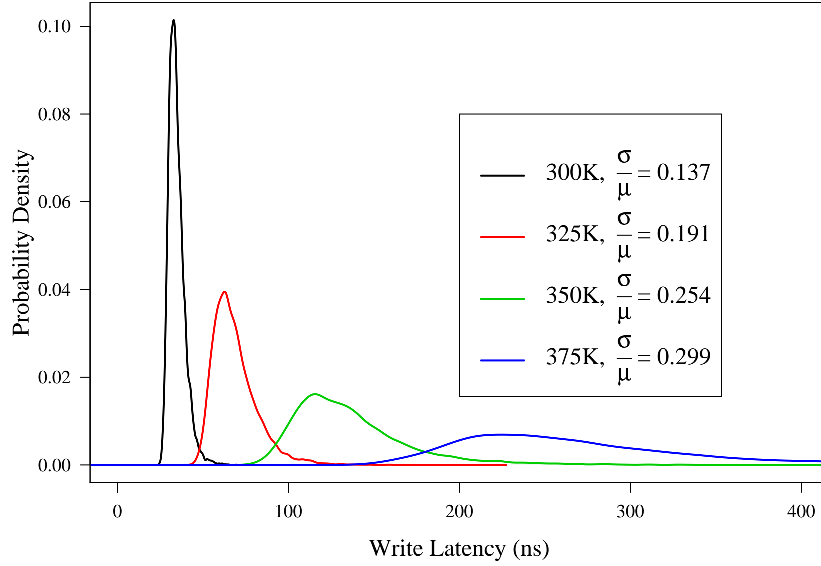
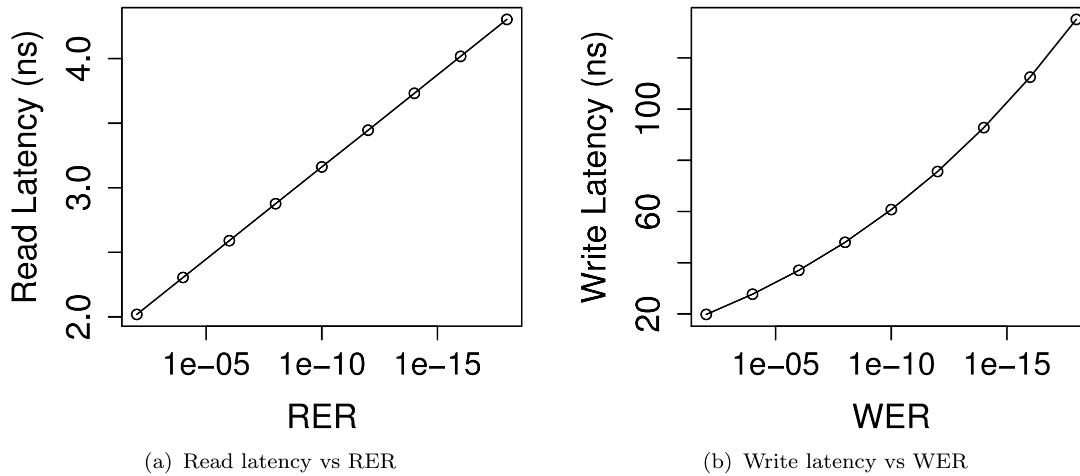


Figure 3.6: Variations of overall write latency with temperature for a subarray size of 1024 × 1024.



(a) Read latency vs RER

(b) Write latency vs WER

Figure 3.7: Overall read and write latencies for various error rates.

3.3.2.4 Retention Failures

In Fig. 3.9, we show the variation in the mean retention time based on Δ variation as per Eq. (2.4). It can be seen that there is a huge variation in the mean retention time with a minimum value of 218 s and a maximum value of around 2×10^{15} s. This is because the mean retention time is exponentially dependent on Δ. Hence, reducing the variations in Δ is critical to maintaining the retention time within acceptable limits. Alternatively, and in addition, appropriate countermeasures should be taken at the architecture level to ensure data integrity and reduce data loss probabilities.

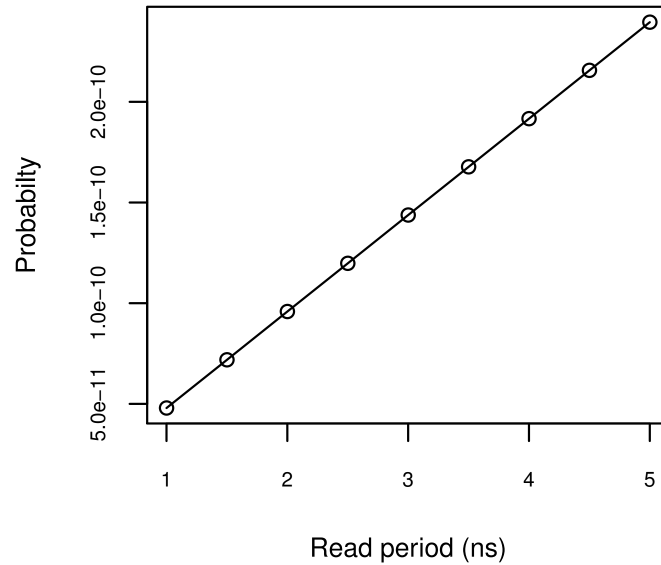


Figure 3.8: Read disturb probabilities for different read periods.

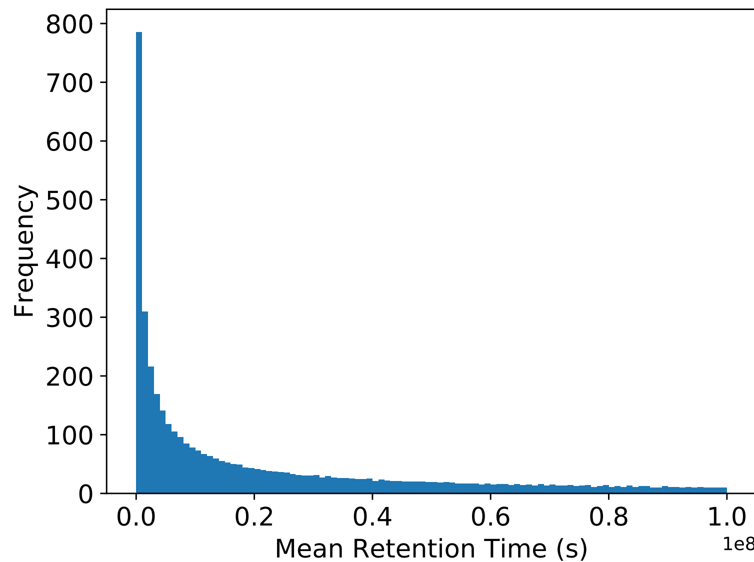


Figure 3.9: Histogram of mean retention time for 10,000 samples.

3.3.2.5 Scaling Effects

In Table 3.2, we compare the latencies and energies at two different technology nodes, 45 nm and 65 nm, to analyze the impact of technology scaling. We see that the nominal values of latency and energy are comparable at both these technology nodes. However, the effect of variations is more pronounced in the smaller technology node (45 nm) as shown by the higher value of $\frac{\sigma}{\mu}$ in this technology node.

Table 3.2: Overall latency and energy values at 45 nm and 65 nm

	45 nm			65 nm		
	Nominal	μ	σ	Nominal	μ	σ
Write Latency	4.99 ns	14.74 ns	1.82 ns	4.47 ns	12.15 ns	1.32 ns
Write Energy	159.02 pJ	425.01 pJ	3.73 pJ	272.83 pJ	512.25 pJ	2.79 pJ
Read Latency	1.27 ns	1.77 ns	0.08 ns	1.22 ns	1.53 ns	0.05 ns
Read Energy	3.48 pJ	4.81 pJ	0.002 pJ	4.82 pJ	5.79 pJ	0.001 pJ

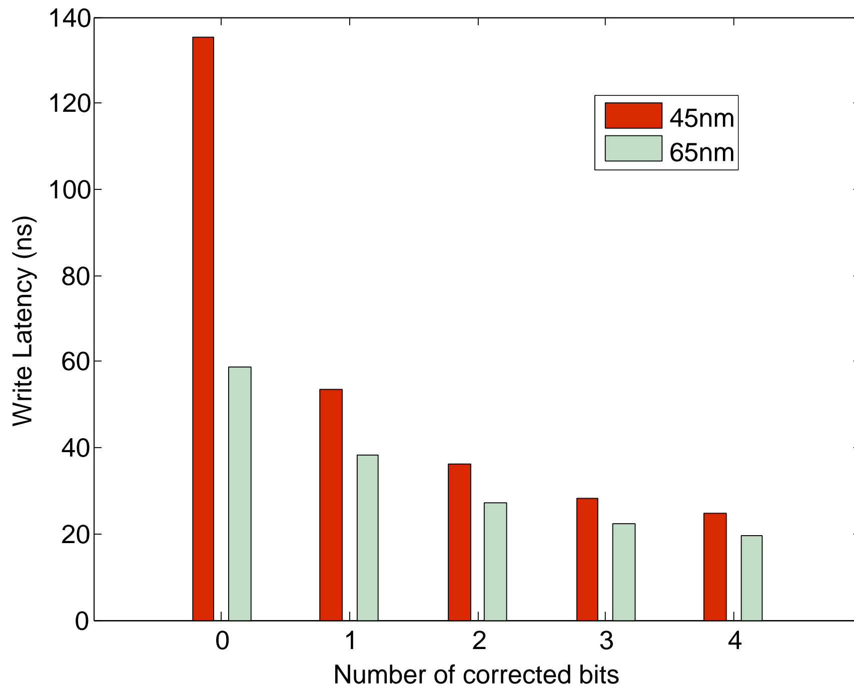


Figure 3.10: Effect of ECCs on write latency for WER of 1×10^{-18}.

3.3.2.6 Effect of ECCs

The effect of various ECC schemes on the write latency is shown in Fig. 3.10. It shows that compared to the case with no ECC (0 bit correction), there is a drastic improvement in latency by using an ECC with one bit error correction. However, the improvement in latency for higher bit error correction is comparatively less.

3.3.2.7 Design Space Exploration

Table 3.3 shows the results of *variation-aware* design space exploration for the write operation for an input WER of 1×10^{-18}. The table also includes the nominal values where the write period is fixed equal to the mean of the bit-cell write latency distribution. The original NVSim optimization, which is not variation-aware, reports an optimal memory configuration of 1024 × 1024, with a lower nominal latency than that of the 2048 × 512 configuration. However, in the *variation-aware* optimization, the optimal memory configuration is 2048 × 512. This configuration has a lower

Table 3.3: Design space exploration of a 128KB RAM for an input WER of 1×10^{-18}

	Rows × Columns 1024 × 1024		Rows × Columns 2048 × 512	
Design Metric	Nominal	Variation-aware	Nominal	Variation-aware
Write Latency	4.99 ns	135.39 ns	5.441 ns	131.98 ns
Write Energy	159.02 pJ	4.25 nJ	208.98 pJ	3.96 nJ

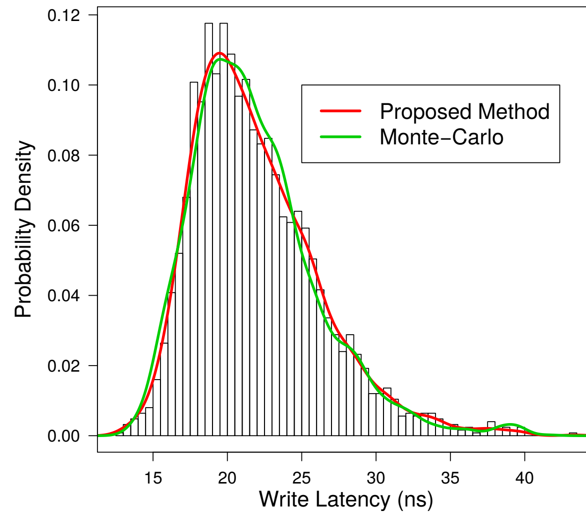


Figure 3.11: Validation of our hybrid method with full Monte-Carlo for a subarray size of 256 × 32.

Table 3.4: Comparison of run-time of the proposed method with full Monte-Carlo

Subarray Size (Rows × Columns)	Run-time (s)		Speed-up
	Our Method	Monte-Carlo	
256 × 32	7	48	6.8×
1024 × 1024	235	10,112	43×

quantile of the overall latency compared to that of the 1024 × 1024 configuration. A similar trend can be seen in the case of write energy as well. These results emphasize the need for an accurate variation-aware optimization instead of a conventional optimization.

3.3.3 Validation

Fig. 3.11 shows the comparison of the proposed hybrid method with a full Monte-Carlo analysis for overall write latency. The results show that our proposed method closely approximates a full Monte-Carlo method. However, a full Monte-Carlo requires large number of samples for the individual components, which increases the run-time. Table 3.4 shows the run-time comparison for two different subarray sizes. Experiments were run on a 64 bit Linux machine having 16 GB of RAM with 16 Intel Xeon cores clocked at 2.53 GHz. It can be seen that our proposed method is

Table 3.5: Comparison of WER for a memory array with 512 columns

Write Period (ns)	WER		
	NVSim-VX^{S} [69]	(VAET-STT)	Monte-Carlo
50	7.882×10^{-03}	4.379×10^{-03}	4.334×10^{-03}
60	1.004×10^{-03}	4.577×10^{-04}	4.463×10^{-04}
70	1.143×10^{-04}	5.708×10^{-05}	5.560×10^{-05}
80	0	8.278×10^{-06}	8.270×10^{-06}
90	0	1.366×10^{-06}	1.160×10^{-06}

much faster than the full Monte-Carlo method for similar accuracy levels. When the number of columns in the subarray increases from 32 to 1024, the run-time for the Monte-Carlo method increases 210×, whereas the run-time for our method increases only 33×. Hence, our method has better scalability compared to the full Monte-Carlo method.

Table 3.5 compares the WER reported by NVSim-VX^{S} with our (VAET-STT) tool for a memory array with 512 columns. It also includes the WER calculated using a detailed Monte-Carlo method. The error rates reported by our tool are more accurate and closer to the detailed Monte-Carlo values. It can also be seen that NVSim-VX^{S} is not able to report the correct error rates beyond a certain value of write period (80 ns as shown in Table 3.5). However, our tool is able to correctly report even low error rates.

3.4 Summary

This chapter presents the details of a system-level *variation-aware* framework to estimate the latencies and energies of STT-MRAM based memories. The tool considers stochastic switching and process variations in the bit-cell as well as process variations in the peripheral components and can perform a system-level analysis at various temperatures and technology nodes. It can report various failure rates and can also analyze the effectiveness of different ECC schemes. Furthermore, the tool can perform a *variation-ware* memory configuration optimization while meeting reliability constraints. The results show that our framework can provide more realistic margins as compared to the conventional framework and that the optimized variation-aware memory configuration could be significantly different from that of the nominal case.

Chapter 4

Parametric Failure Modeling and Yield Analysis

4.1 Overview

Similar to any nano-scale device, STT-MRAM is also affected by manufacturing variations, as the technology scales down. The properties of the magnetic storage device (MTJ) are impacted by the manufacturing imperfections in the magnetic fabrication process [47]. In addition, the CMOS device variations due to RDF, LER and STI stress [52, 53] affect the access transistor in the bit-cell as well as the peripheral components impacting the read and write operations. Hence the combined effect of magnetic and CMOS variations in the bit-cell and the peripheral circuitry result in both reliability failures in the field and permanent faults at the tester in STT-MRAM based memories.

The parametric failures due to extreme process variations result in yield loss during the manufacturing of the chip. A good understanding of the failure behavior and failure map can help the designer to incorporate the right combination of defect tolerance techniques to overcome the yield loss. The existing fault models for conventional CMOS memory technologies cannot be directly applied to STT-MRAM because of the fundamental difference in the operation [75]. In addition, due to non-volatility and stochasticity, some of the failure mechanisms (such as read-disturb, retention failures) [76–78] are unique to STT-MRAM. Hence, unlike conventional SRAM, these failure mechanisms must be considered during the yield analysis of STT-MRAM. Although these failure mechanisms affect run-time reliability of the STT-MRAM and may not show up during manufacturing test, they are originated in parametric variations. Moreover, the memory is no longer considered functional when such run-time failure rates are above a given threshold or retention requirements are not satisfied. The yield analysis framework should also consider the entire memory system including the bit-cell array and the peripherals which can guide the designer to employ appropriate design-for-yield schemes.

4.1.1 Related Work

The defect and fault models for SRAM have been extensively studied in the past such as [79]. However, these fault models are not directly applicable to STT-MRAM because of the fundamental difference in operation of SRAM and STT-MRAM [75]. There are a few works which analyze the transient and permanent faults in STT-MRAM and propose solutions to mitigate these faults. The authors in [57] propose a technique that integrates both the ECC and FM techniques to simultaneously address the transient and permanent faults. This work assumes a fixed number of faults in the memory array which are randomly distributed. However, in reality, faults show strong correlation among neighboring cells, which have to be considered for a realistic failure analysis.

In [80], the authors quantitatively study the persistent and non-persistent errors in STT-MRAM cell operations and propose device level and circuit level solutions. This work, however, does not consider peripherals such as decoders, multiplexers, etc. in the analysis and also does not provide any fault maps, which can help the designer in implementing the right combination of architecture-level solutions such as RR and FM. The work in [75] presents a comprehensive analysis of faults due to both parametric variations as well as defects (opens and shorts) in STT-MRAM. Nonetheless, the array-level analysis in this work is limited to a 2×2 array. This work does not consider peripheral variations or correlation among cells in its analysis.

4.1.2 Contributions

We have developed a framework for yield analysis based on extreme parametric variations [36, 37]. We consider parametric variations in the bit-cells and peripherals as well as the correlation among neighboring cells to get the fault distribution map of the memory array, due to both permanent faults and reliability failures. Based on correlated bit-cell parameters, we perform Monte-Carlo simulations on the memory array to find the number of faulty cells. This process is then repeated for different correlation maps corresponding to different chip instances to obtain the yield. We then explore the yield improvement that can be obtained by incorporating various mitigation techniques.

Overall, our contributions are as follows:

- We model both reliability failures and permanent faults due to parametric variations for the STT-MRAM array considering bit-cell, peripherals and correlation among the cells. The array-level fault map is obtained considering extreme process variations.
- We explore the effect of different temperatures and correlation coefficients on the yield.

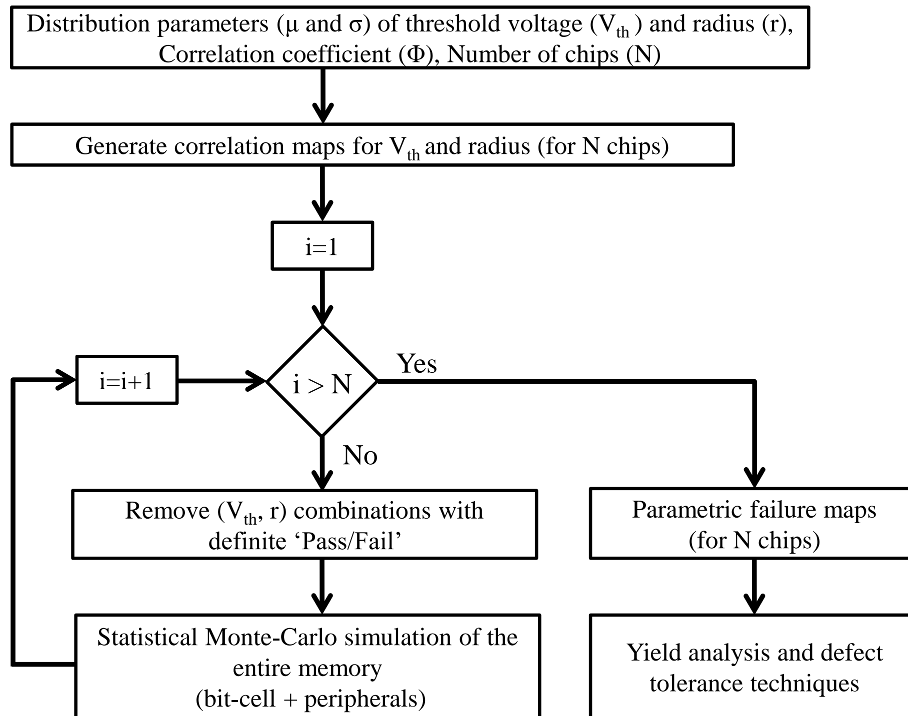


Figure 4.1: Proposed yield analysis flow.

- We use the framework for a design-for-yield exploration combining various defect tolerance techniques (like ECC and RR) to mitigate permanent and reliability failures.

- We observe that unique yield improvement techniques specific to STT-MRAM are far more effective than conventional techniques (such as redundant rows/-columns and ECC). In this work, we show that the current boosting technique can be very effective in mitigating write failures with minimal area overhead.

4.2 Yield Analysis Framework

The overall yield analysis flow is presented in Fig. 4.1. The parameters considered in our analysis are the radius (r) of the MTJ and the threshold voltage (V_{th}) of the CMOS components such as the access transistor and peripheral circuitry. These parameters are assumed to follow a gaussian distribution. The correlation maps for these parameters are then obtained from the VARIUS tool [70]. For each of these correlation maps, we get the parametric failures (both permanent and reliability failures) by performing Monte-Carlo simulations for the entire memory system including the bit-cell and peripheral components. The yield is then obtained by performing Monte-Carlo over multiple maps (corresponding to different chip instances). We then explore the right combination and efficacy of different defect tolerance techniques such as ECC and RR to obtain a target yield.

4.2.1 Obtaining Correlation maps

The correlation map for the bit-cell parameters are obtained from VARIUS tool [70]. In this tool, the systematic variation is modeled using a multivariate normal distribution with a spherical spatial correlation structure as given in Eq. 4.1:

$$\rho(x) = \begin{cases} 1 - \frac{3x}{2\Phi} + \frac{x^3}{3\Phi^3}, & (x \leq \Phi) \\ 0, & \text{otherwise} \end{cases} \tag{4.1}$$

In the above equation, $\rho(x)$ is the correlation function for two points separated by a distance x and Φ is the correlation coefficient, which specifies the range over which two points are correlated, expressed as a fraction of the chip's width. Two cells which are at a distance less than Φ are assumed to be correlated while those with distance greater than Φ have no correlation.

The inputs to the VARIUS tool are the mean (μ) and standard deviation (σ) of the parameter under consideration and also the correlation coefficient (Φ). We run Monte-Carlo analysis using the VARIUS tool [70] to generate independent spatial correlation map of the parameters (V_{th} and r), assuming a gaussian distribution. The correlation map of one Monte-Carlo run for V_{th} and r obtained for a 32×32 array for various values of Φ is shown in Fig. 4.2 and Fig. 4.3. Here, $\Phi = 0$ represents the case where there is no correlation among the parameters of adjacent cells, which means that the parameters are randomly distributed.

4.2.2 Permanent Fault Analysis

The permanent faults are deterministic and can be repeated at the tester. These faults are mainly caused due to extreme process variations or spot defects (opens and shorts). In this work, our focus is only on the permanent faults due to extreme process variations. The parameters considered in our analysis are radius (r) variations of the MTJ and threshold voltage (V_{th}) variations of the CMOS components.

The standard 1T-1MTJ bit-cell structure (see Fig. 2.1(b)) is used for our simulations. The read/write margins are fixed based on the worst-case operating conditions of the cell. For instance, for the write operation, the worst-case operating conditions are minimum supply voltage and temperature, and maximum threshold voltage and radius. Then for extreme variations, we increase the variations in the parameters (V_{th} and r) beyond the nominal variation.

The first step is to get the correlation maps of V_{th} and r as explained in Section 4.2.1. Next, the latency distributions of the peripheral components (see Fig. 3.2) are obtained using a hierarchical and hybrid Monte-Carlo approach as explained in Section 3.2. Then, for each bit-cell, depending on the specific V_{th} and r values for the bit-cell as well as the periphery path, SPICE simulations are performed to determine whether the cell is functional or not (based on the provided margins). The above process is repeated using the Monte-Carlo method for all the bit-cells in a memory array to obtain the fault distribution map.

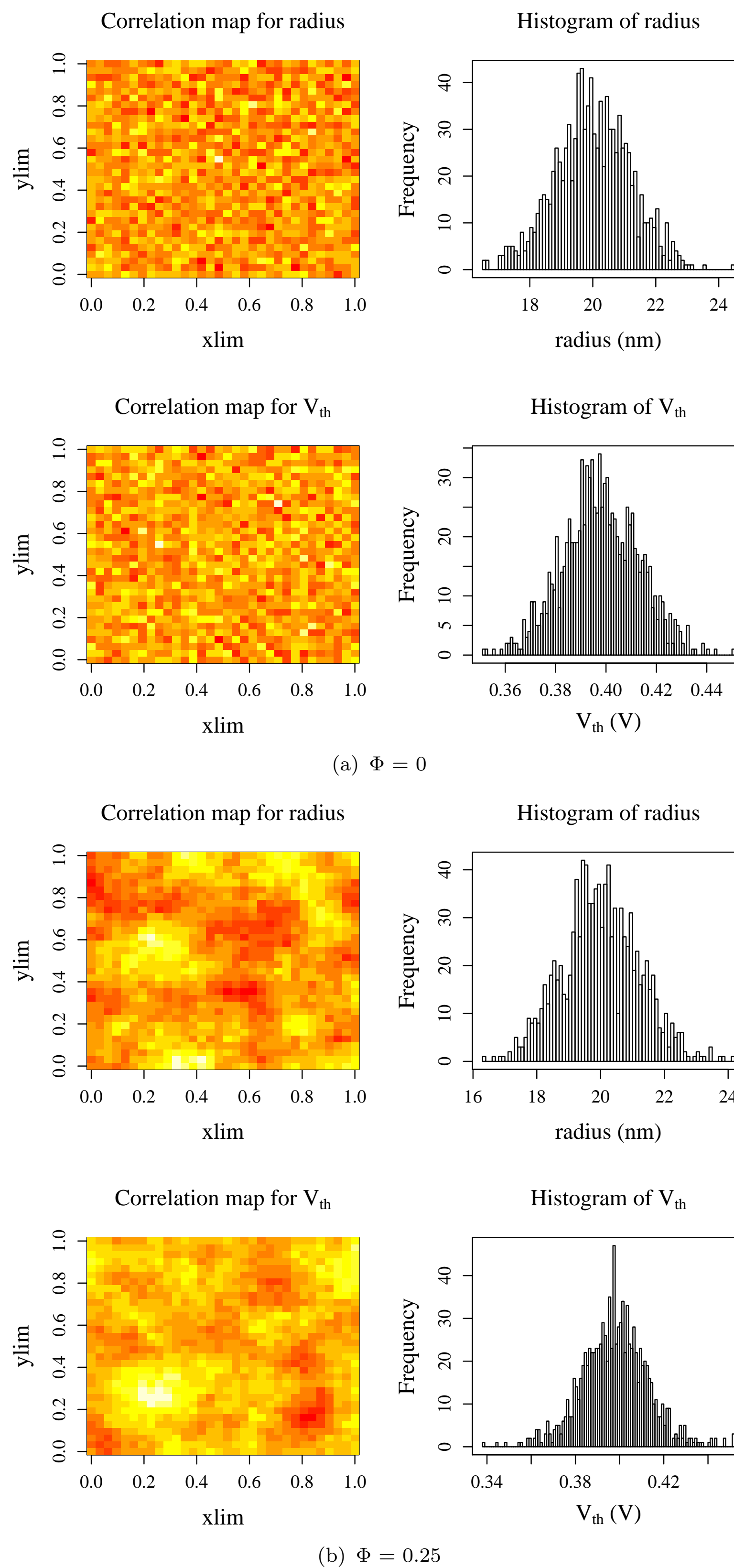
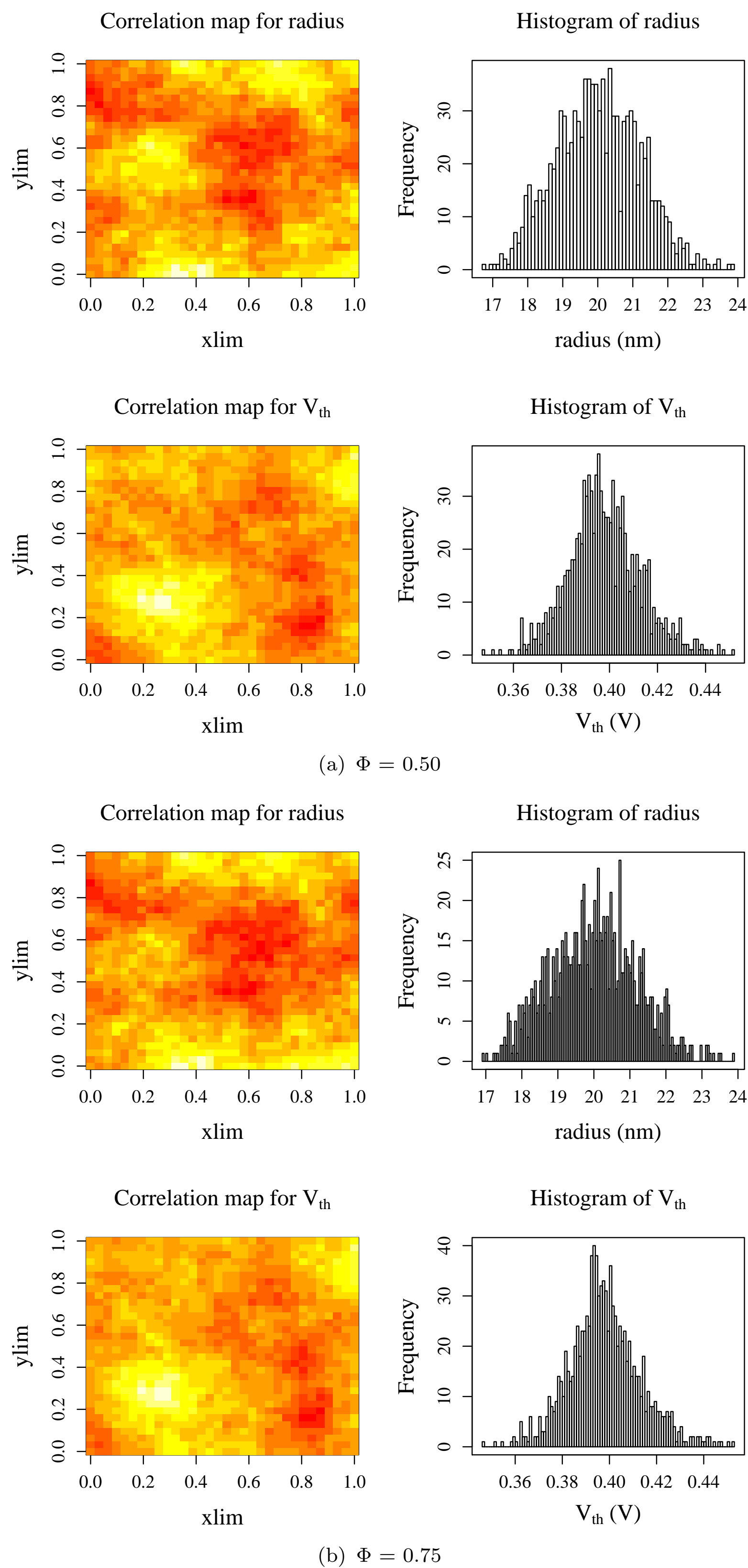


Figure 4.2: Correlation maps for radius (μ = 20 nm, σ = 6%) and V_{th} (μ = 397.9 mV, σ = 3.76%) for a 32×32 array for $\Phi = 0$ and $\Phi = 0.25$.



(a) $\Phi = 0.50$

(b) $\Phi = 0.75$

Figure 4.3: Correlation maps for radius ($\mu = 20\,\text{nm}$, $\sigma = 6\%$) and V_{th} ($\mu = 397.9\,\text{mV}$, $\sigma = 3.76\%$) for a 32×32 array for $\Phi = 0.5$ and $\Phi = 0.75$.

The fault map for read, write and retention and read-disturb failures thus obtained for one of the Monte-Carlo runs for a 32×32 memory array for different values of Φ is shown in Fig. 4.4 and Fig. 4.5. Please note that the exact value of Φ depends on the manufacturing process and represents the amount of variation correlation and fault clustering. The extreme case of $\Phi = 0$ represents random distribution and no clustering, which is often the assumption in typical analysis. Hence we perform our analysis from low to high Φ values to understand the dependence of various fault types on Φ.

Fig. 4.4 and Fig. 4.5 show that when $\Phi = 0$, the faults are randomly distributed, as expected, which leads to large number of line failures. As Φ increases, it can be observed that the faults get more clustered, leading to an increased number of failures per line, however reducing the number of line failures as a consequence. It can also be seen that, as Φ increases, the number of cells that are likely to fail due to retention or read-disturb are much more than those that are likely to fail due to read or write failures. The analysis of the fault maps from different Monte-Carlo runs gives the number of failures and their distribution in the memory array, which provide insights into the defect tolerance techniques required for yield improvement.

4.2.3 Reliability Fault Analysis

The transient faults are non-deterministic faults occurring primarily due the stochastic switching of the MTJ, and are typically expressed by respective error rates. The probability of these failures is modeled in Eq. 2.6, Eq. 2.7 and Eq. 2.8 for write, read-disturb and retention failures, respectively. It can be seen that these failures primarily depend on Δ and the write current, which in turn depends on r and V_{th}. Hence different bit-cells have different failure rates, according to their process points (r, V_{th}).

If e_i is the failure probability of the ith bit-cell and n is the word size, then the failure probability (error rate) of the entire word, E is given by:

$$E = 1 - \prod_{i=1}^{n} (1 - e_i) \tag{4.2}$$

The word error rate E specifies the number of reliability faults per memory access. Since these faults happen in the field, the error rates should be kept to a minimum. For instance, the target values of WER for a memory array should be around 10^{-9} or lower [50].

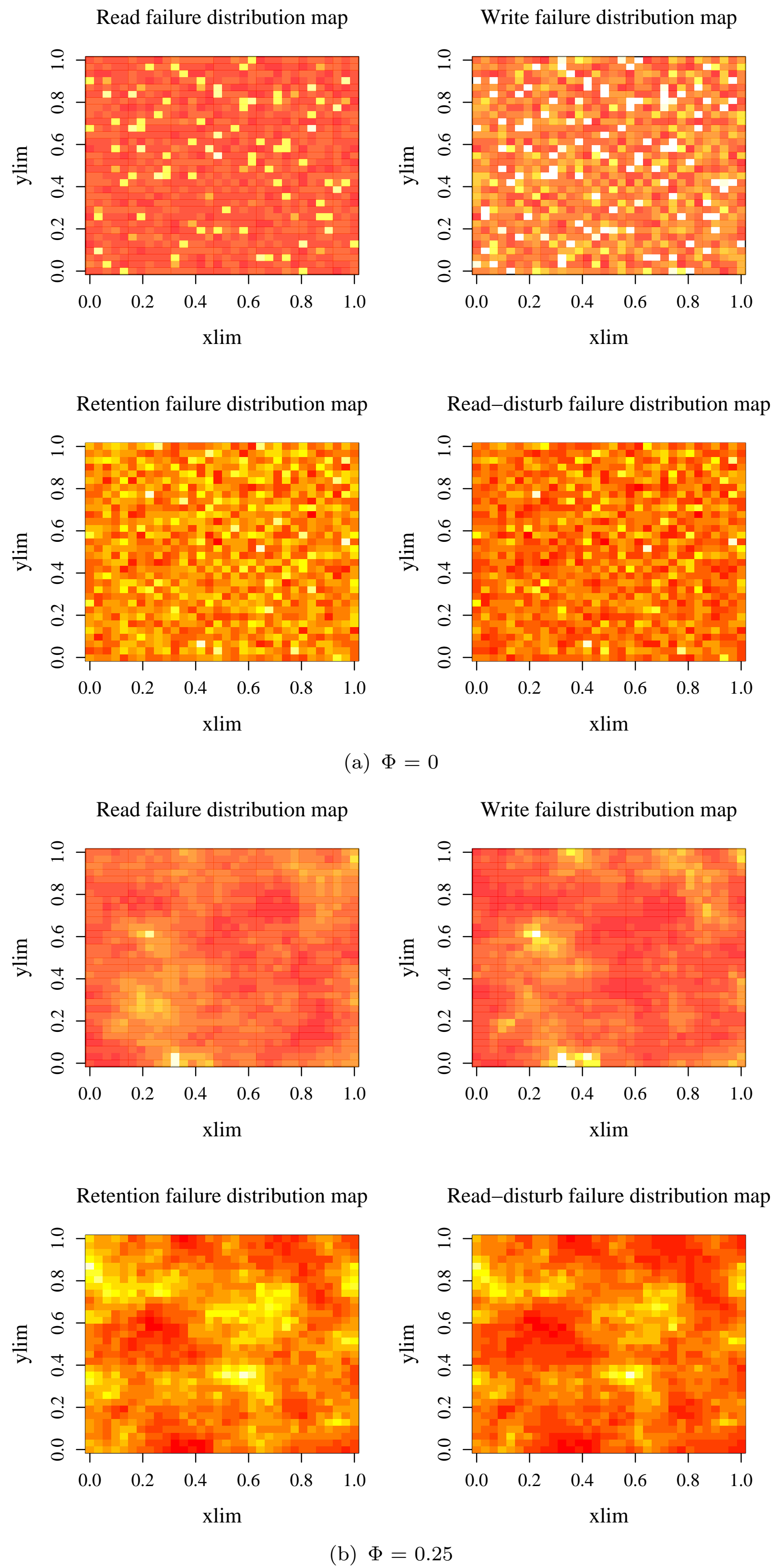


Figure 4.4: Read, write, retention and read-disturb failure distribution map for a 32×32 memory array for radius (μ = 20 nm, σ = 6%), V_{th} (μ = 397.9 mV, σ = 3.76%) for $\Phi = 0$ and $\Phi = 0.25$.

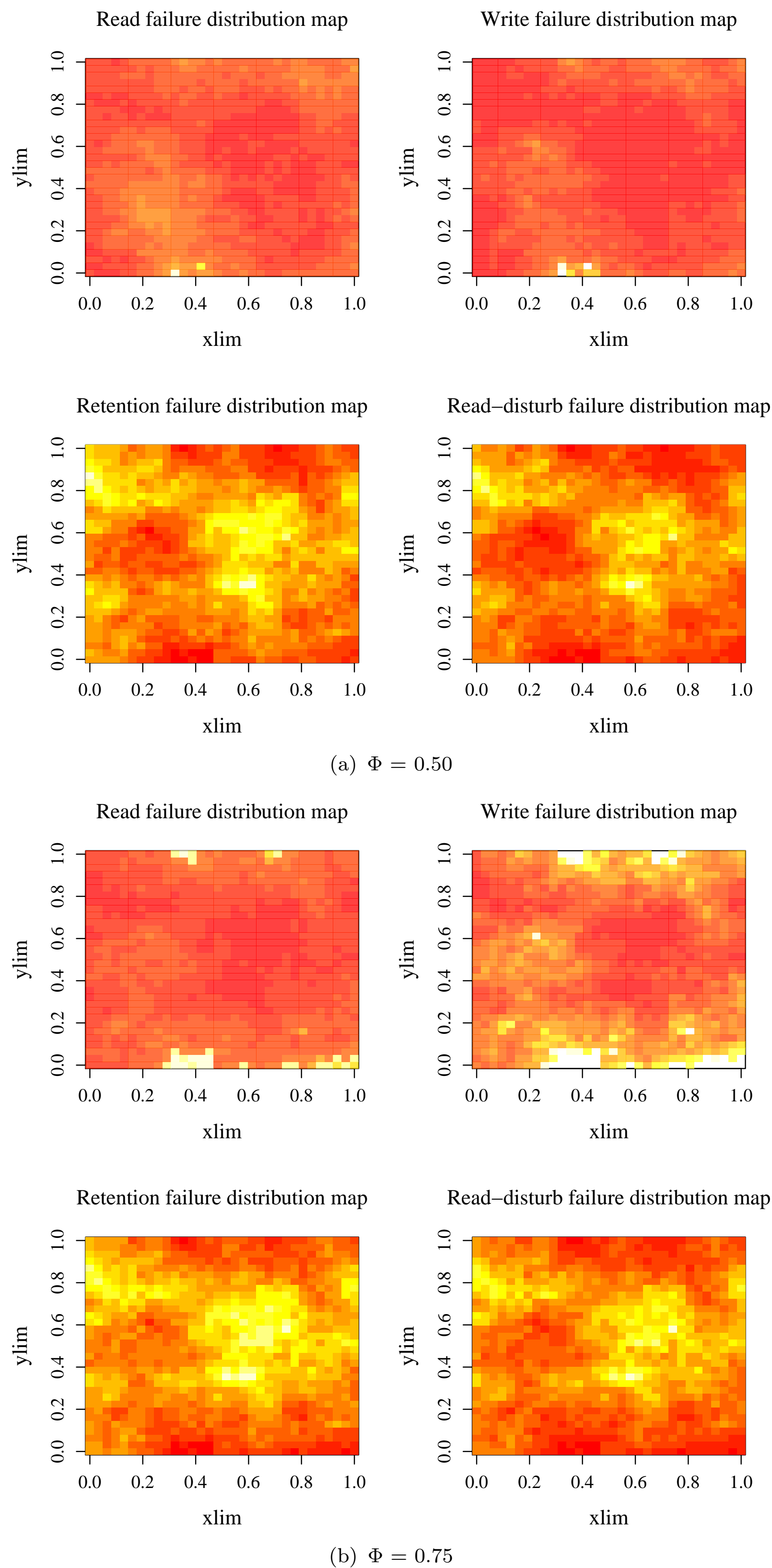


Figure 4.5: Read, write, retention and read-disturb failure distribution map for a 32×32 memory array for radius (μ = 20 nm, σ = 6%), V_{th} (μ = 397.9 mV, σ = 3.76%) for $\Phi = 0.50$ and $\Phi = 0.75$.

Figure 4.6: Analysis Complexity Reduction Technique. [Region 1 - Definite 'Fail'; Region 2 - Definite 'Pass'; Region 3 and Region 4 - Simulation required]

4.2.4 Analysis Complexity Reduction Techniques

Here we discuss some techniques which help to reduce the number of required SPICE simulations for yield analysis. As mentioned earlier, the yield analysis framework requires multiple Monte-Carlo runs each calling several SPICE simulations in the inner loop. This leads to large run-times, especially for analyzing large memory arrays. The run-time can be reduced by reducing the number of SPICE simulations, especially for the analysis of permanent faults.

In permanent fault analysis, we are mainly concerned with 'Pass/Fail' criterion. Hence we can avoid all parameter combinations, which lead to definite 'Pass' or definite 'Fail' conditions and only simulate the remaining parameter combinations. For instance, for the bit-cell simulation, the worst operating conditions are maximum V_{th} (V_{max}) and maximum r (r_{max}). This point is marked as P1 in Fig. 4.6. Then, all combinations with $V_{th} > V_{max}$ and $r > r_{max}$ lead to definite 'Fail' (Region 1) and those with $V_{th} < V_{max}$ and $r < r_{max}$ lead to definite 'Pass' (Region 2). Hence we need to simulate only the rest of the combinations (Region 3 and Region 4). The boundaries for definite 'Pass' and definite 'Fail' conditions are iteratively refined by each subsequent SPICE simulation, if the next simulation results in either a 'Pass' or a 'Fail' scenario. This is illustrated in Fig. 4.6. For instance, consider a point P2 in Region 3. If the simulation yields a 'Pass', it means that any point in Region 5 will also result in 'Pass' condition and hence Region 5 is added to the 'Pass Region'. Similarly, for a point P3 in Region 3, if the simulation yields a 'Fail', the Region 6 will be added to the 'Fail Region'. Similar arguments can be made for points P4 and P5 in Region 4, where Region 7 and Regions 8 are added to the 'Pass Region' and 'Fail Region', respectively. This process is repeated, and more regions are added either to the 'Pass Region' or 'Fail Region'. Our analysis shows that adopting this approach for 100 Monte-Carlo runs of a 512×512 array results in requiring only around 1.22% of the total number of combinations to be simulated, which is approximately

Table 4.1: Operation of current boosting circuitry

Inputs			Outputs				Comments
WBE	BL	SL	D0	D1	D2	D3	
0	X	X	1	0	1	0	M0, M1, M2 & M3 are OFF
1	1	0	0	0	1	1	M0 and M3 are ON, M1 & M2 are OFF
1	0	1	1	1	0	0	M1 and M2 are ON, M0 & M3 are OFF

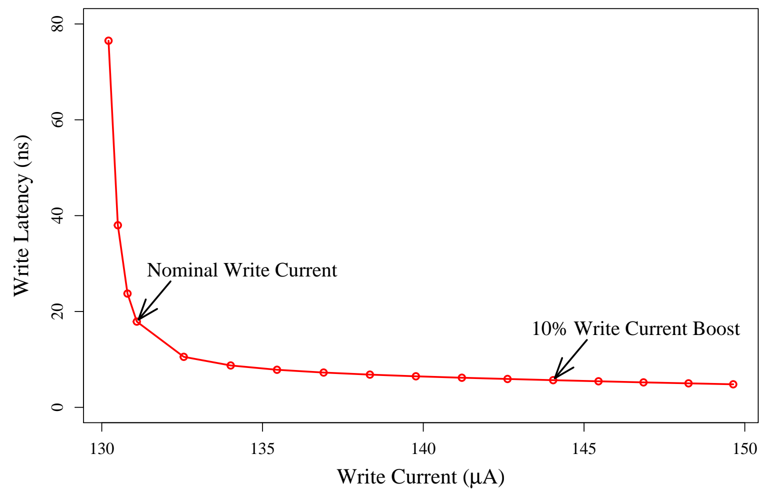
82× speedup. This speedup ratio increases for larger memory size and more chip instances.

4.2.5 Yield Exploration

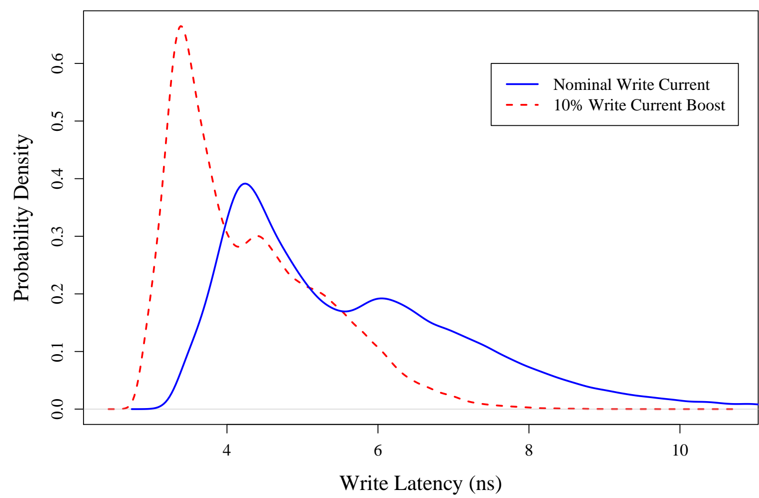
Yield exploration can be done from the failure maps of different Monte-Carlo runs corresponding to different chip instances and by analyzing the number of faults in a row or column. If there are large number of faults per row or column, i.e., the faults are clustered, then RR is a good technique to mitigate these faults. On the other hand, for a small number of faults per row or column, i.e., when faults are more uniformly distributed, ECC would be a good option. In case of more number of single isolated faults, advanced techniques such as those proposed in [57] could be optimal for yield improvement.

Besides the conventional yield improvement techniques, we also explore some of the techniques specific to STT-MRAM. One such technique is the current boosting technique which has been proposed in several previous works for improving the write performance of STT-MRAM [58, 81–83]. In this work, we investigate its effectiveness in improving the yield. The switching probability and latency of STT-MRAM is highly sensitive to the write current as shown in Eq. 2.5. Hence current boosting can significantly decrease the write latency resulting in reduced write failures as shown in Fig. 4.7. From Fig. 4.7(a), it can be seen that a 10% increase in write current can decrease the write latency of a bit-cell to around one-third. This will result in a significant reduction in the number of write failures for the memory array as can be seen from the shift in the write latency distribution to the left in Fig. 4.7(b). The tail of the distribution is much less for the case with current boost as compared to the nominal case. Please note that this current boosting comes at extra energy cost and the amount of current boosting is limited by the oxide breakdown limit (Time Dependent Dielectric Breakdown or TDDB) of the barrier oxide layer. Therefore, it is used only for the columns with low yield, as a specific defect tolerance and yield enhancement scheme.

A typical circuit employed for current boosting is shown in Fig. 4.8. Here, the extra current for boosting is provided by means of four additional transistors (M0 - M3), which are sized according to the amount of current required. The current boosting circuitry is activated by the WBE (Write Boost Enable) signal. The operation of the circuit is shown in Table 4.1. When WBE is low, all the transistors are OFF and the normal operation happens. When WBE is enabled, depending on the values



(a) Write latency versus write current



(b) Write latency distribution

Figure 4.7: Write latency reduction using current boosting technique.

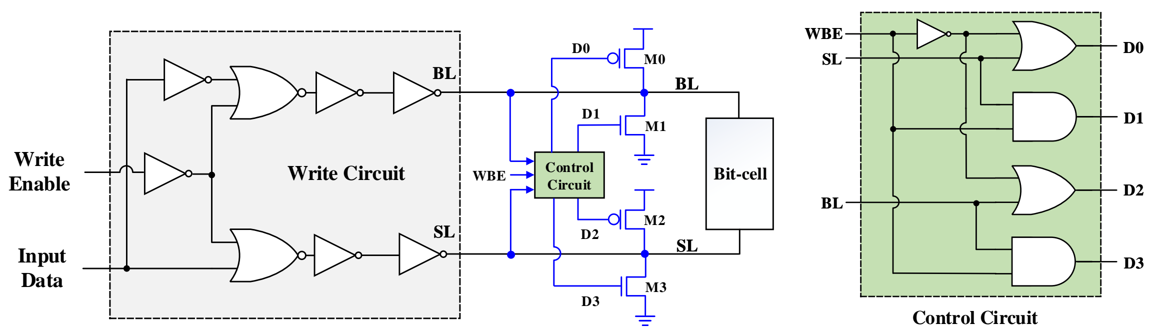


Figure 4.8: Circuit for write current boosting.

Table 4.2: Parameters of the MTJ

Parameter	Value
Radius	20 nm
Free layer thickness	1.3 nm
Oxide thickness	1.48 nm
RA	7.5 $\Omega \mu m^2$
TMR	150%

of BL and SL, either transistors M0 and M3 or transistors M1 and M2 provide the extra write current required for boosting. However, the amount of current boosting is limited to ensure that it does not lead to oxide barrier breakdown of the MTJs. Hence a combination of current boosting and traditional techniques can be the most effective for yield improvement with minimum overheads.

4.3 Results

4.3.1 Experimental Setup

We have used the TSMC SPICE models for the CMOS access transistor and the PMA MTJ model from [84, 85]. The parameters of our MTJ model are given in Table 4.2. The MTJ radius variation is assumed to be 5%, whereas the threshold voltage variations in the CMOS components (bit-cell and periphery) are assumed to follow the Pelgrom law [54]. Based on our MTJ model, we see that a 5% radius variation causes around 8.5% variation in Δ and around 8% variation in the read/write currents for a fixed Δ. We have not considered variations of other MTJ parameters such as free layer and oxide thickness in this work, due to the limitations of our MTJ model. However these can easily be integrated into the framework if supported by the used MTJ model.

For extreme variations, we consider 20% extra variations compared to the nominal variations. We have done our analysis on a 512×512 memory array at 45 nm technology node.

4.3.2 Results

Fig. 4.9 shows the line fault distribution of a 512×512 memory array for different values of Φ. When $\Phi = 0$ (i.e., faults are spatially uncorrelated and randomly distributed), the figure shows that most of the lines have a maximum of 1, 2 or 3 faults. As Φ increases, the number of lines with > 3 faults also increase. This is because, for lower values of Φ, the faults are randomly distributed in the memory array, whereas for higher values of Φ, clustering of faults occur due to higher correlation. Depending

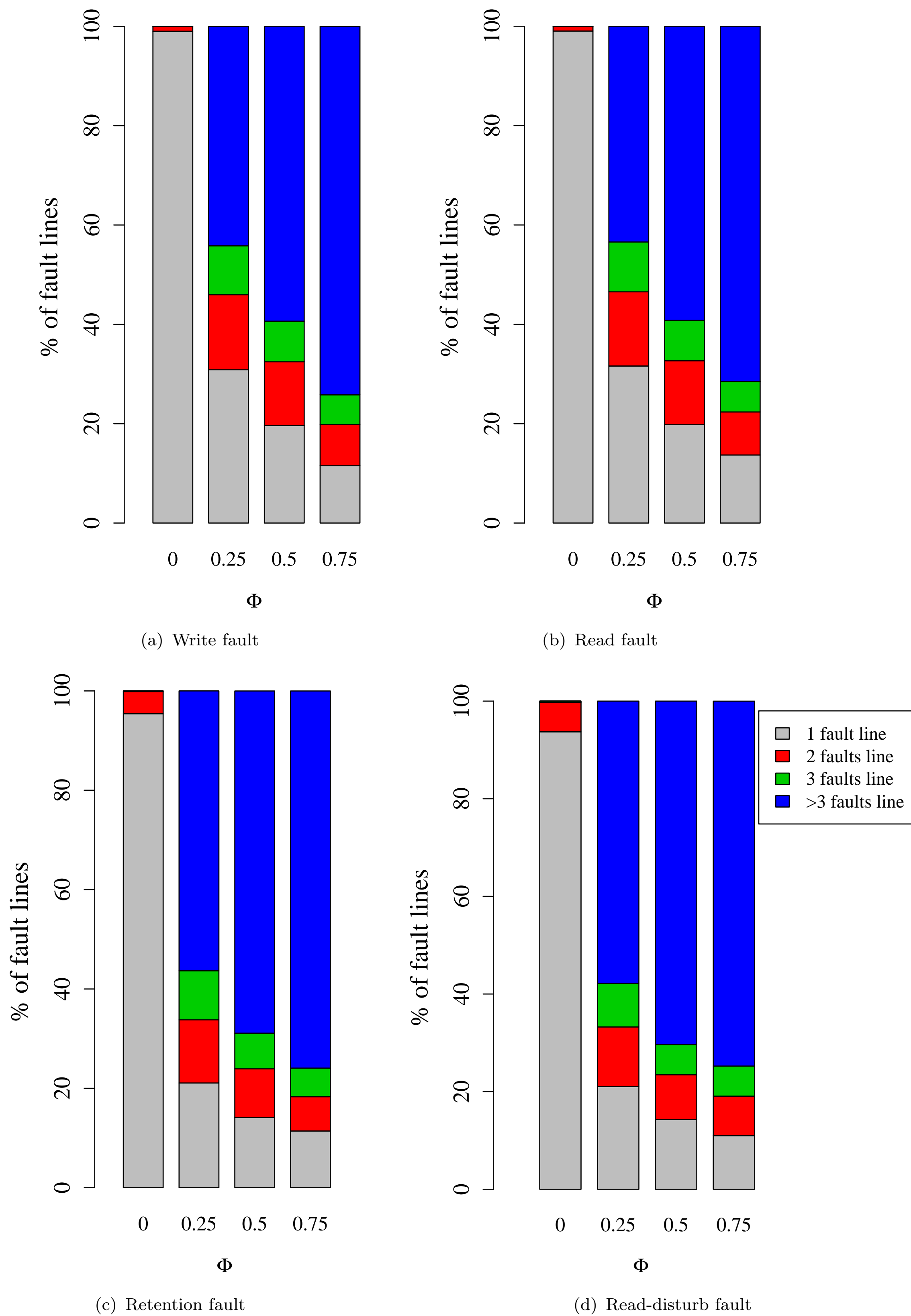


Figure 4.9: Line fault distribution for a 512×512 memory array at 25 °C or various values of correlation coefficient (Φ)

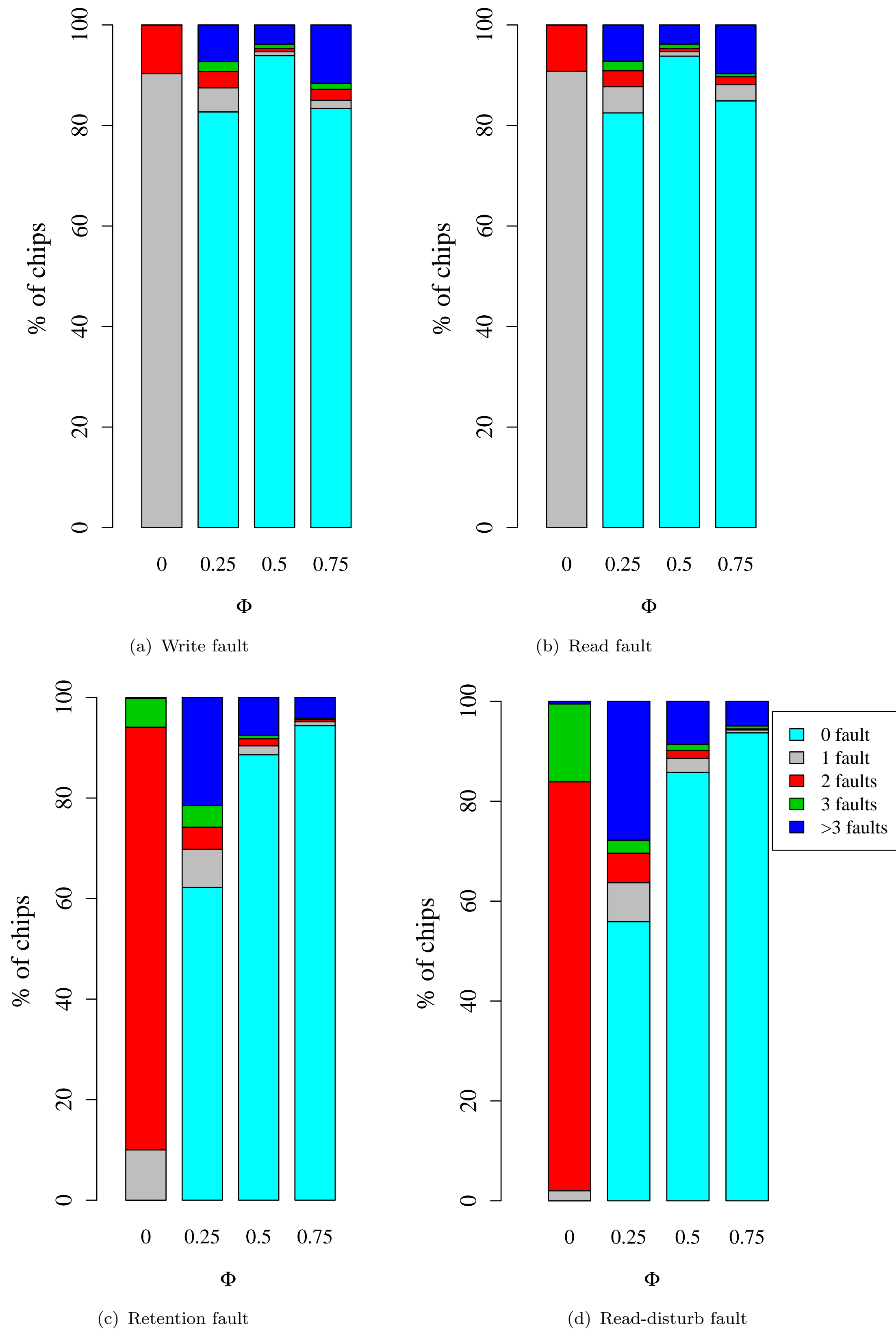
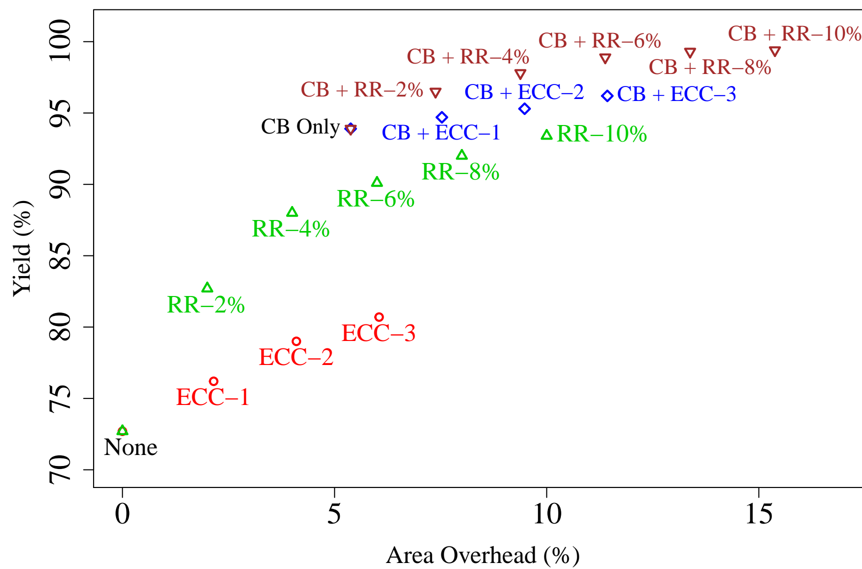
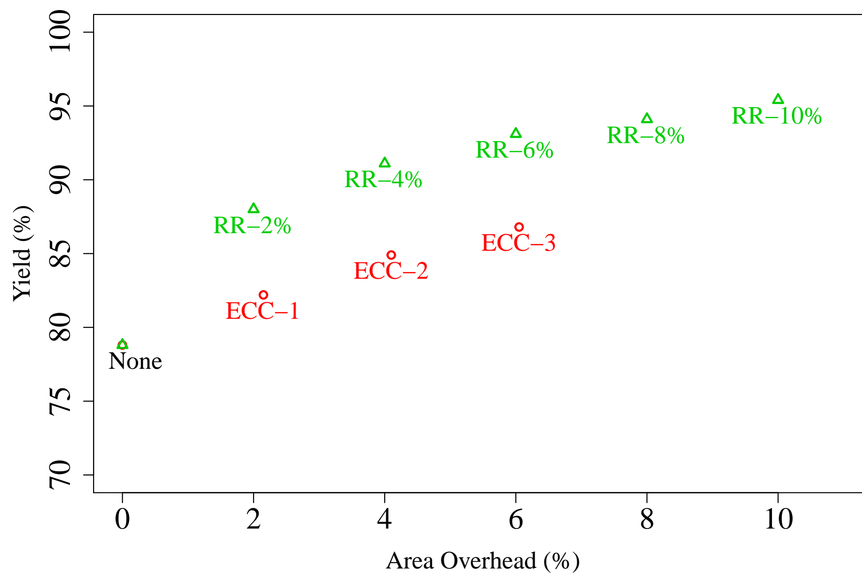


Figure 4.10: Percentage of chips with their fault types for a 512×512 memory array at 25 °C for various values of correlation coefficient (Φ)

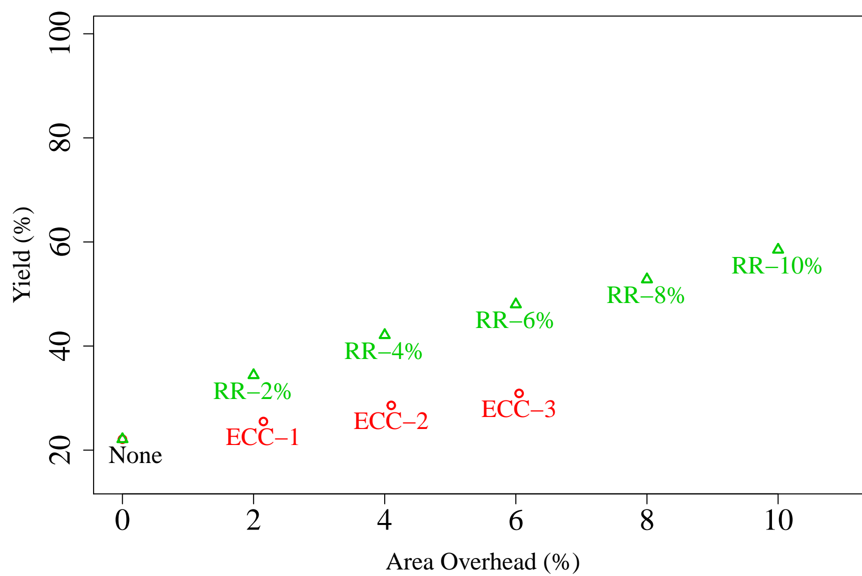


(a) Yield analysis for write fault at -25 °C

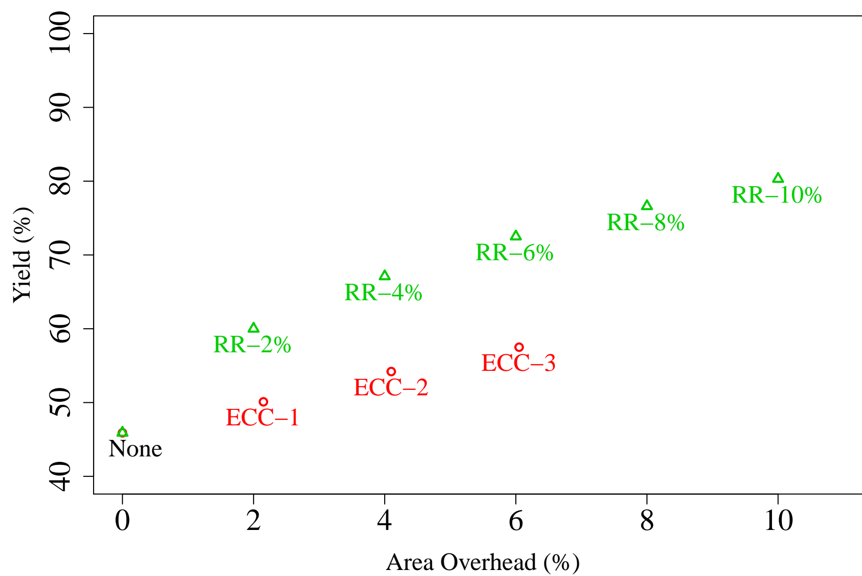


(b) Yield analysis for read fault at -25 °C

Figure 4.11: Yield improvement using different defect tolerance techniques versus their area overhead for write and read faults for $\Phi = 0.50$
(RR - Redundancy Repair; CB - Current Boost)



(a) Yield analysis for retention fault at 75 °C



(b) Yield analysis for read-disturb fault at 75 °C

Figure 4.12: Yield improvement using different defect tolerance techniques versus their area overhead for retention and read-disturb faults for $\Phi = 0.50$ (RR - Redundancy Repair)

on the clustering of faults and how many faults occur per line, different types of defect tolerant schemes (ECC versus row/column redundancy) would be more effective.

Fig. 4.10 shows the percentage of chips and the corresponding fault types for a 512×512 memory array. Here the percentage of chips with 0 faults indicates the yield. From the figure, it can be seen that when $\Phi = 0$, the yield is very low. In this case, most of the chips fail due to 1, 2 or 3 line faults. As Φ increases, the faults get clustered, which increases the probability of having a large number of faults on some chips, and fewer number of faults on some other chips. This means that as Φ increases, the probability of having chips with 0 faults also increases (see Fig. 4.10), thus increasing the yield. The amount of yield increase depends on the nature and actual distribution of faults. For instance, from Fig. 4.10, we see that the yield increases for all faults when Φ increases from a value of 0 to 0.5. However, for read and write faults, the yield decreases when Φ increases from a value of 0.5 to 0.75.

The yield improvement with different defect tolerance techniques with their respective area overhead costs is shown in Fig. 4.11 and Fig. 4.12. The storage area overhead for ECC is calculated from the number of ECC bits required to correct e errors and detect $e + 1$ errors as $10e + 1$ [77]. For the current boosting technique, the area and energy overhead for the additional circuitry are around 5.38% and 0.65% respectively, which is obtained from NVSim [5]. The results show that under the same area constraint, the current boosting technique is the most effective technique to mitigate write failures. However, there is a limit to the amount of current boosting possible, due to TDDB of the MTJ. Hence, we limit ourselves to around 10% current boost. Current boosting is not very effective to mitigate read faults, since an increase in current, although reduces the read decision faults, increases the probability of read-disturb (see Eq. 2.7). It can also be seen that for the same area overhead, the RR technique is much more effective for yield improvement as compared to the ECC technique. The best combination for yield improvement is based on current boosting and modest RR.

From Fig. 4.12, it can also be observed that the yield for reliability failures (retention and read-disturb) is comparatively lower, especially for retention faults, where the yield is around 22.1%. This means that 77.9% of the chips are likely to have reliability failures due to short retention time in the field. The yield can be improved by different defect tolerance techniques as shown in the figure, however even with RR of 10%, the yield improves to only around 58.5%. This observation is in line with those reported in other works such as [77], where retention failures are seen as a major reliability concern for STT-MRAM in advanced technology nodes.

The temperature dependence of yield for different fault types is shown in Fig. 4.13. The temperature effect is based on the thermal model integrated in the used MTJ model. This model uses the Neel-Brown model for the stochastic mode and the Sun model for the ballistic mode for the write operation. For the read operation, the conductance is driven by tunneling phenomena described by the models established by Simmon, Slonczewski and Brinkman. More details about this model can be found in [84, 85]. An increase in temperature generally increases the yield for permanent read and write faults and decreases the yield for reliability failures (retention and

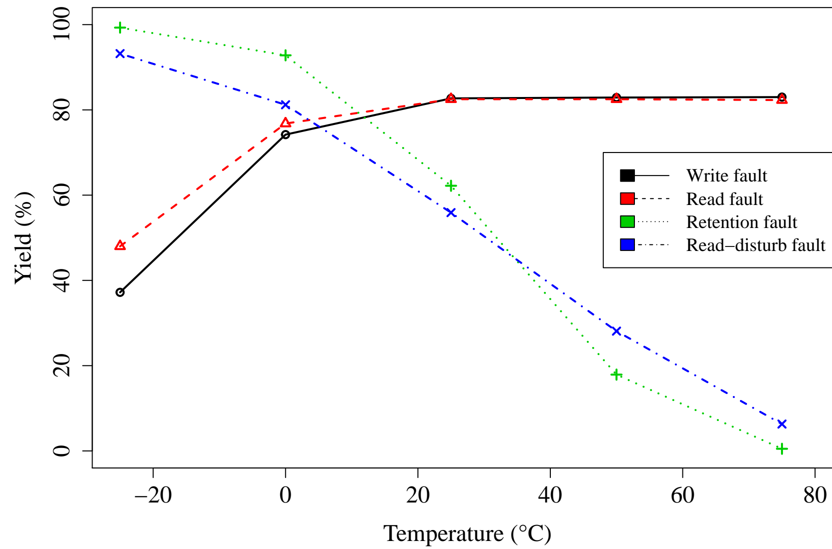


Figure 4.13: Temperature dependence of yield for various faults ($\Phi = 0.25$)

read-disturb). This is because the reliability faults are highly dependent on Δ (see Eq. 2.7 and Eq. 2.8). Hence, as temperature increases, Δ decreases resulting in increased reliability failures.

The combined yield analysis considering all faults is shown in Fig. 4.14. The yield is very low (almost zero) when Φ is zero. This is because, since the faults are randomly distributed, the probability of having at least one fault in a chip is very high. If $\mathrm{Y_c}$ is the yield due to combined faults and $\mathrm{Y_i}$ is the yield due to fault type i, then the following inequalities hold due to set intersection rules.

$$\mathrm{Y_c} \leq \min(\mathrm{Y_i}); \tag{4.3}$$

$$\mathrm{Y_c} \geq \prod_i \mathrm{Y_i} \tag{4.4}$$

The combined yield $\mathrm{Y_c}$ will be closer to $\min(\mathrm{Y_i})$ (Eq. 4.3) if the different types of faults $\mathrm{Y_i}$ are highly correlated, else $\mathrm{Y_c}$ will be closer to $\prod_i \mathrm{Y_i}$ (Eq. 4.4), if there is low correlation among $\mathrm{Y_i}$. From Fig. 4.15, it can be seen that $\mathrm{Y_c}$ is in between these boundaries, closer to $\min(\mathrm{Y_i})$, which suggests a good correlation among the different types of faults.

The effectiveness of various defect tolerance techniques in improving the combined yield $\mathrm{Y_c}$ (considering all faults) is shown in Fig. 4.16 for various values of Φ. When $\Phi = 0$ (faults are uncorrelated), ECC-2 and ECC-3 improves the yield considerably, whereas RR is not effective in yield improvement. This is because the faults are not clustered and are randomly distributed. In addition, the fact that ECC-1 is not effective and ECC-2 or above is required to improve the yield shows that most lines have 2 or more faults. However, as Φ increases (faults are more spatially correlated), the figure shows that RR is much more effective than ECC in improving the yield due to the clustering of faults. Hence, the right technique for yield improvement depends on the value of Φ and the operating temperature and our yield analysis framework can guide the designer in choosing the correct one.

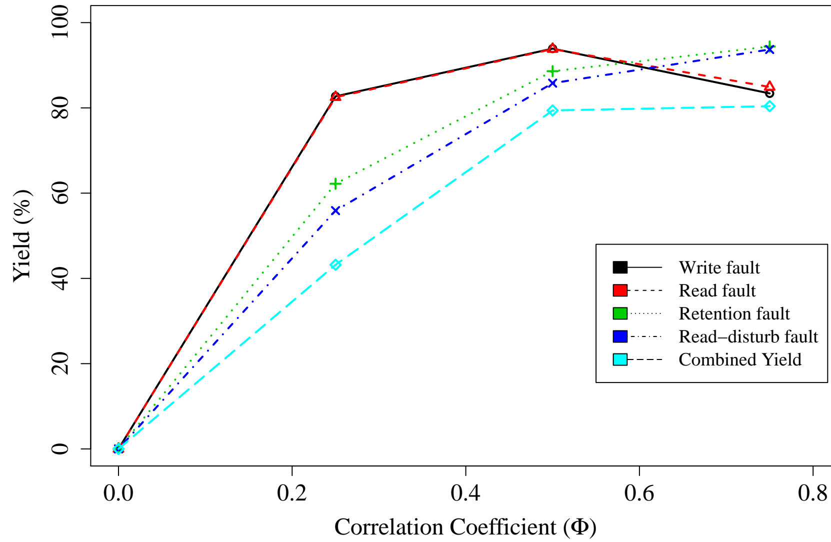


Figure 4.14: Yield analysis for various faults and combined yield considering all faults (Temperature = 25 °C)

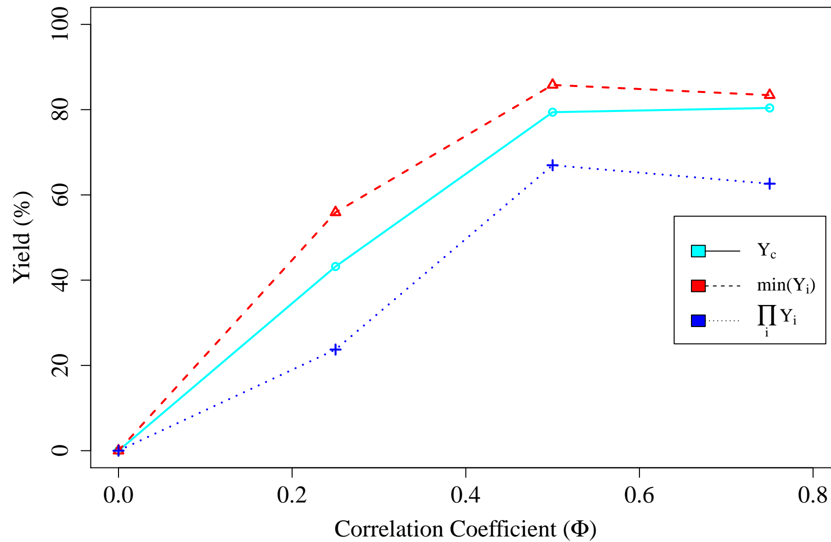


Figure 4.15: Combined yield and its boundaries (Temperature = 25 °C) [Y_c - Combined Yield; Y_i - Yield due to fault type i]

In Fig. 4.17, we show how the nominal design of the MTJ affects the yield. The figure presents the yield for different values of Δ. It can be seen that yield considering both read and write faults improves with decrease in Δ. For the retention and read-disturb faults, the yield is not affected by Δ scaling since we assume the same percentage variation for different Δ values. However, it should be noted that lowering the nominal Δ would negatively impact the read-disturb and retention probabilities (see Eq. 2.7 and Eq. 2.8). Hence the Δ should be fixed based on the specific application requirement.

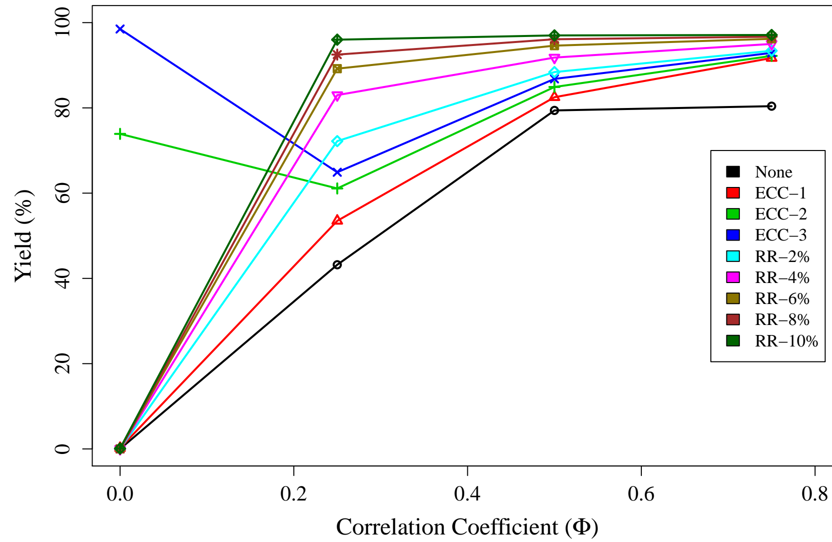


Figure 4.16: Comparison of defect tolerance techniques for yield improvement considering all faults for various correlation coefficients (Φ) (Temperature = 25 °C)

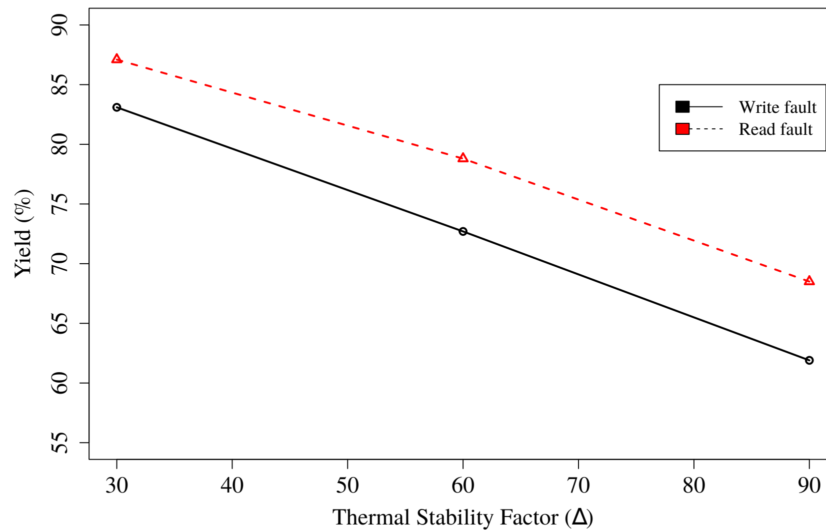


Figure 4.17: Yield analysis for different Δ values assuming same percentage variation (Φ = 0.5, Temperature = -25 °C). For setup, please refer to Section 4.3.1.

4.4 Summary

In this chapter, we proposed a framework for yield analysis of STT-MRAM memory arrays considering reliability and permanent faults due to parametric variations. We have considered the variations in the bit-cell and the peripheral components as well as the spatial correlation among the bit-cells in our analysis. The framework can analyze the yield for individual faults or a combination of different faults at various temperatures and correlation coefficients. It also allows the designer to perform a design-for-yield exploration and helps in choosing the right combination of defect tolerance techniques to improve the yield. The results show that, in addition to the

traditional yield improvement techniques (such as ECC and RR), unique techniques specific to STT-MRAM (such as current boosting) can be very effective in yield improvement with minimal overhead.

Chapter 5

Defect Injection, Fault Modeling and Test Algorithm Generation

5.1 Overview

The fabrication of STT-MRAM consists of a magnetic process to manufacture the MTJ in the BEOL and a CMOS process for the access transistor and other CMOS-based peripheral components. Hence, the yield of STT-MRAM technology is influenced by both these processes. Moreover, the MTJ related defects are different from those of the CMOS technology as these are developed using new processes with new materials. Hence, it is required to develop a rigorous testing methodology for STT-MRAM technology, especially when it is still in the yield ramp up phase.

Defects in STT-MRAM are fundamentally different from those of existing memory technologies. This is because, the working principles as well as development processes for these technologies are dissimilar. Additionally, due to manufacturing complexities and interdependencies of magnetic materials, MTJ devices are subject to various new defects. For instance, during the ion beam etching process, the sputtered atoms deposited at the MTJ sidewall may result in a short in the oxide layer [86]. On the other hand, an open defect can occur in MTJ because of interconnect imperfections [87]. In addition, the magnetic orientation of the MTJs can be fixed to a specific magnetization configuration, meaning that their magnetic orientation and thus their resistances cannot be changed [58, 88]. This may happen permanently because of manufacturing defects in the magnetic layers or due to loss of margin in the CMOS support circuitry, such as reduced switching current or duration [58]. Besides these, the impact of process variation can disturb the memory operations. Therefore, for STT-MRAM, a detailed defect injection and fault analysis needs to be done to develop efficient testing algorithms subsequently.

5.1.1 Related Work

Several works have studied about the MTJ related faults in the past [56, 76, 89–92], however, only a few have done the detailed fault modeling. For instance, the work in [92] has classified and analyzed the MRAM defects, and two MRAM specific faults are identified, i.e., the Multi-Victims fault and Kink fault. However, this work was done for the conventional MRAM technology where the magnetic switching happens due to the external magnetic field. This technology is different from the existing current perpendicular STT-MRAM in which the current has to flow through the MTJ stacks for the magnetic switching. Therefore, the MRAM specific faults including Multi-Victims faults and Kink faults are not applicable for the STT-MRAM technology. There are some more similar fault modeling works that have also been done using previous MRAM technologies [87, 93, 94].

The work presented in [75] has done fault modeling for STT-MRAM considering both parametric variations as well as spot defects such as opens and shorts. In this work, the formulation of the fault primitive occurrence is done at the cell-level, whereas the electrical faults are injected at the memory array-level. Nevertheless, their fault injection is only limited to the netlist-level and do not perform any analysis at layout-level. For a realistic fault analysis, it is important to consider the actual layout, especially for the defect injection. Moreover, this work does not perform fault analysis at different voltage and temperature corners

5.1.2 Contributions

We have done defect injection and fault modeling for STT-MRAM at the layout-level and performed fault analysis at various voltage and temperature corners. These analyses are then utilized to develop a testing framework for STT-MRAM [38, 39]. For this purpose, Synopsys AIFA (Advanced Inductive Failure Analysis) flow [95] is used which allows automated defect injection capability into memory layout (GDS) or/and in SPICE netlist, as well as it allows to extract fault models and generate test algorithms. Based on the obtained results Synopsys STAR Memory System [96] which was original meant to test embedded SRAMs, ROM, CAM, eFlash as well as external memories has been extended to test MRAMs also. All the obtained test mechanisms and test algorithms are implemented in STAR Memory System for efficient MRAM testing.

The impacts of test environment, namely the temperature and voltage, as well as process variations on the manifestation of defects are analyzed. We have considered both the spot defects in the layout and their manifestation as resistive opens and shorts in the netlist, as well as the impact of MTJ defects on the functionality of memory devices. This analysis shows the dynamic read fault behavior which requires multiple vectors for excitation and detection. This particularly happens in the case of inter-cell coupling faults. In addition, the write faults are very sensitive to the test voltage and temperature. Based on our defect analysis and fault modeling, we propose an efficient test algorithm to cover all STT-MRAM specific faults.

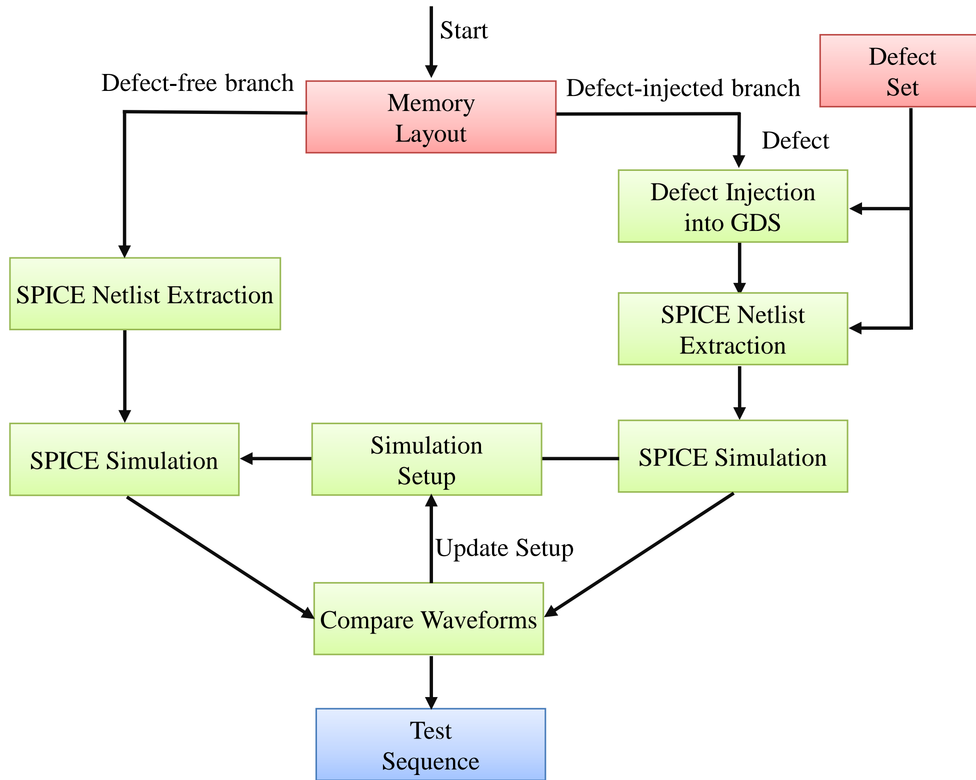


Figure 5.1: Advanced Inductive Failure Analysis (AIFA) flow

5.2 Fault Analysis Framework

5.2.1 Defect Injection Methodology

For making the defect injection and fault modeling on STT-MRAM less time consuming and more effective Synopsys AIFA automated flow [95] was used as shown in Fig. 5.1. With this flow finding the appropriate test sequences for detection of defects, as well as construction of the corresponding fault models becomes much more straightforward. The flow consists of the following steps:

1. As an input a set of target defects (Defect Set) is received for which the corresponding fault models should be developed and simulation environment setup parameters. Defect Set can be regularly enriched incorporating new types of physical defects. Simulation setup contains the initial set of test sequences that are considered for the physical defect, as well as other simulation settings, such as supply voltage, temperature, frequency, and considered resistance range in case of a resistive physical defect.

2. Each of the considered physical defects is injected either into the memory Graphic Data System (GDS) layout or the SPICE Netlist, depending on which one is preferable in the specific case.

3. Next two simultaneous SPICE simulations are performed, one without any defect injected (Defect-free branch) and second with the injected defect (Defect-injected branch), and for each simulation PASS/FAIL information is obtained, as well as signal waveforms.

4. If the result of simulation without the physical defect injection is FAIL, it means that there are inaccuracies in the simulation settings or used memory models which must be fixed. Simulation should be re-run until the result is PASS.

5. If the result of simulation with a physical defect injection is FAIL, then at least one test sequence has been identified that detects the physical defect. Otherwise, if the result is PASS, none of the provided test sequences was able to detect the physical defect so the given sequences need to be altered and simulation performed again. This step is done by the user (for example, a test engineer) by performing certain judgments based on the obtained results. Particularly, using the received waveforms and comparing with the similar waveforms obtained for the defect-free branch, one can conclude how to modify the sequence in order to be able to detect the injected physical defect. This process continues until a test sequence is discovered that allows the user to identify the physical defect or the user comes to the conclusion that the physical defect does not result into any fault.

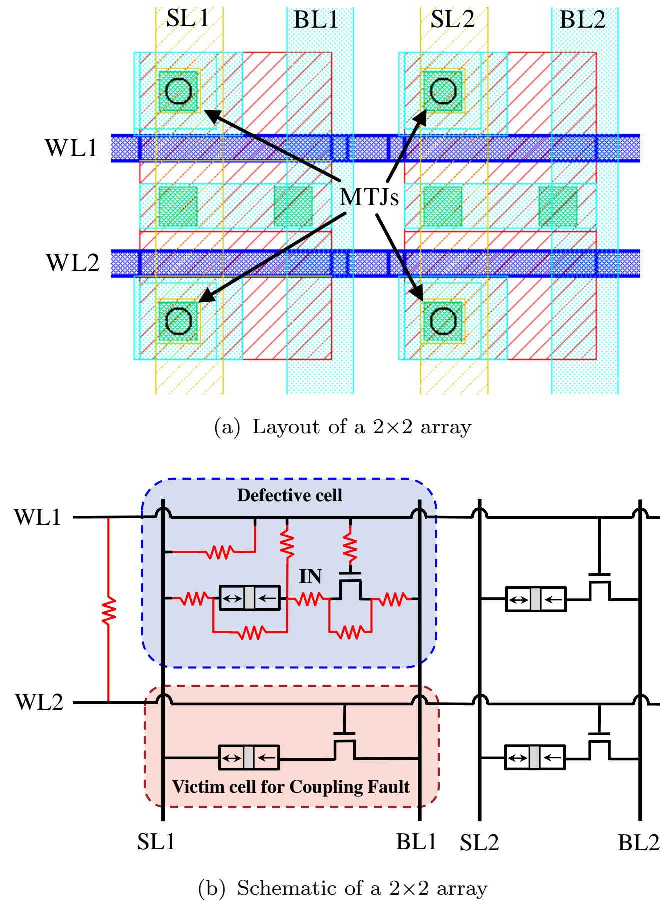
6. In the last step using the identified test sequences, the corresponding fault models are automatically derived.

5.2.2 Simulation Framework

Both intra-cell and inter-cell defects in STT-MRAM have been considered in our analysis. Some of probable open and short defects based on layout analysis and the corresponding fault models are summarized in Table 5.1. The analysis is layout-aware and we have only considered such defects which are probable based on STT-MRAM layout, as shown in Fig. 5.2. We considered all the open defects as probable since they can happen due to opens in metal contacts or vias. Among the shorts, we classify some of the defects as non-probable, since the corresponding nodes are far apart in the layout. Thus, we remove WL-SL and BL-SL shorts since they are in different layers. This is because the SL is connected to the MTJ, which is typically manufactured at

Defect Place	Location	Defect Type	Layout based analysis	Read Fault	Write Fault
Intra-cell defect	BL	Open	Open in via/contact	IRF0	TF0, TFI
	SL	Open	Open in via/contact	IRF0	TF0, TFI
	SL − IN	Short	MTJ short	IRF1	TF0, TFI
Inter-cell defect	WL1 − WL2	Short	-	IRF1	TF1

Table 5.1: Defects and Fault Models in STT-MRAM



(a) Layout of a 2×2 array

(b) Schematic of a 2×2 array

Figure 5.2: Layout based fault injection. The injected resistors are shown in red.

higher metal layers. Similarly, we remove IN-VDD and IN-GND shorts since the cells do not usually connect directly to these. We have also considered inter-cell coupling faults affecting different cells within the same column and across different columns. The most probable among them is the short between word-lines of adjacent cells as given in Table 5.1. In addition, we have considered MTJ specific defects. For all the probable cases, we use SPICE simulations to obtain the resulting read/write faults due to these defects.

The write faults are classified as *Transition Faults* (TF), TF1 and TF0. TF1 fault occurs when a cell storing a '0' value is unable to switch to '1' during the write process. Similarly, the inability of a cell storing '1' value to switch to '0' is classified as TF0 fault.

For the read operation, the faults are classified as Incorrect Read Faults (IRF), IRF1 and IRF0. IRF1 means that a cell storing a '1' is read as '0' by the read circuitry, whereas in an IRF0 fault, a cell storing '0' is read as '1'.

Once the read/write faults are obtained for each of the defects, we replace the opens/shorts with resistors and find the range of resistances for which the respective faults can be observed. We then repeat the above step for different operating conditions of voltage and temperature and obtain the corresponding resistance ranges.

In addition to the static faults mentioned above, we explore dynamic as well as inter-cell coupling faults in STT-MRAM. A dynamic fault requires multiple read/write operations to sensitize the fault. In this work, we check for dynamic faults by first writing to a cell and then reading multiple times from the same cell. In the case of dynamic faults, the first m reads are fault-free while the outcome of m+1 and subsequent reads are faulty. A coupling fault occurs when a defect in one cell (aggressor) results in faults in a neighboring cell (victim). In this work, the aggressor and victim cells are in the same column, or they can be in adjacent rows. We also check for the occurrence of both static and dynamic faults in case of inter-cell coupling faults.

5.2.3 Impact of Variability

To quantify the impact of variability, we perform the fault analysis for different voltage and temperature corners. In addition, we have considered the impact of process variation in the bit-cell as well as the periphery using statistical mismatch models for the MTJ and the CMOS components. In the case of static faults, the value of resistance required to sensitize the fault depends on the operating corner. For dynamic faults, the resistance range for which a dIRF-n fault happens depends on the operating corner. In addition, due to process variation, the value for n for each resistance also changes. This means that the maximum value of n for a resistance increases due to process variation. Since the test pattern should be fixed based on the maximum value of n, the process variation affects the test pattern as well.

5.3 Results

For our simulations, we have used the TSMC 65 nm SPICE models for the CMOS components and the PMA-MTJ model from [84, 85] with a radius of 20 nm. Based

Defect Type	Temperature	Voltage					
		Low		Nominal		High	
		TF0	TF1	TF0	TF1	TF0	TF1
BL open	Low	$10k\Omega$	$1k\Omega$	$10k\Omega$	$10k\Omega$	$10k\Omega$	$10k\Omega$
	Nominal	$10k\Omega$	$10k\Omega$	$100k\Omega$	$10k\Omega$	$100k\Omega$	$10k\Omega$
	High	$100k\Omega$	$10k\Omega$	$100k\Omega$	$100k\Omega$	$100k\Omega$	$100k\Omega$
SL open	Low	$10M\Omega$	$100k\Omega$	$10M\Omega$	$1M\Omega$	$10M\Omega$	$10M\Omega$
	Nominal	$10M\Omega$	$1M\Omega$	$10M\Omega$	$10M\Omega$	$10M\Omega$	$10M\Omega$
	High	$10M\Omega$	$10M\Omega$	$10M\Omega$	$10M\Omega$	$10M\Omega$	$10M\Omega$
SL − IN short	Low	100Ω	$10k\Omega$	100Ω	$1k\Omega$	10Ω	$1k\Omega$
	Nominal	100Ω	$1k\Omega$	100Ω	$1k\Omega$	100Ω	$1k\Omega$
	High	100Ω	$1k\Omega$	100Ω	100Ω	100Ω	100Ω
Open: Minimum resistance to sensitize the fault Short: Maximum resistance to sensitize the fault							

Table 5.2: Defect Injection and Write Fault Models at different operating conditions

Defect Type	Nominal Voltage, Nominal Temperature
BL open	$10k\Omega$
SL open	$1M\Omega$
SL $-$ IN short	$10k\Omega$
Open: Minimum resistance to sensitize the fault Short: Maximum resistance to sensitize the fault	

Table 5.3: Defect Injection and Read Fault Models

Defect Type	Nominal		Corner Analysis	
	TF0	TF1	TF0	TF1
BL open	$100k\Omega$	$10k\Omega$	$10k\Omega - 100k\Omega$	$1k\Omega - 100k\Omega$
SL open	$10M\Omega$	$10M\Omega$	$10M\Omega - 10M\Omega$	$100k\Omega - 10M\Omega$
SL $-$ IN short	100Ω	$1k\Omega$	$10\Omega - 100\Omega$	$100\Omega - 10k\Omega$
Open: Minimum resistance to sensitize the fault Short: Maximum resistance to sensitize the fault				

Table 5.4: Corner Analysis for Write Fault Models

on the probable faults, the resistance range for the defects which sensitize the faults is summarized in Table 5.2 and Table 5.3 for write and read faults respectively. We used a nominal voltage and temperature of 1.3V and 25°C respectively. For corner case analysis, we used one high voltage (1.4V) and one low voltage (1.2V). Similarly, we used a low temperature of -25°C and a high temperature of 75°C.

From Table 5.2 and Table 5.3, it can be seen that the resistance value of the defect to sensitize the fault depends on the defect type and location. Similarly, the resistance value also depends on the voltage and temperature corner for write faults. This is illustrated in Table 5.4, which shows the range of resistance variation for TF0 and TF1 faults based on corner analysis. For open defects, the worst operating condition is the corner in which the resistance value is the least. Similarly, for short defects, the worst operating corner has the largest resistance. From Table 5.2, we see that the worst operating corner for write faults is low temperature and low voltage (1.2V, -25°C). However, for read faults, our analysis shows that the voltage and temperature corners do not have a significant impact on the resistance. Hence, we show only the resistance values for the nominal case for read faults in Table 5.3. This is because the switching properties of the MTJ are highly dependent on the switching current and the temperature, and hence, the operating conditions have a major impact on writing into the MTJ device. However, the read operation is done through resistance sensing, using a pre-charge based sensing circuitry, which is less impacted by operating voltage and temperature.

Table 5.5 shows one case of observed dynamic fault in our analysis. Here, depending on the resistance, we observe that the first few reads pass, but then read faults are observed. This requires multiple (two, three or four) read operations to detect the fault, which means that the fault is a dynamic fault. Fig. 5.3 shows the corresponding waveform of such fault, where the third read is faulty. We categorize this type of

Defect Type	Resistance Range	Fault
SL − IN short	$2k\Omega - 3k\Omega$	dIRF1-2
	$4k\Omega - 14k\Omega$	dIRF1-3
	$15k\Omega - 29k\Omega$	dIRF1-4

Table 5.5: Dynamic Faults

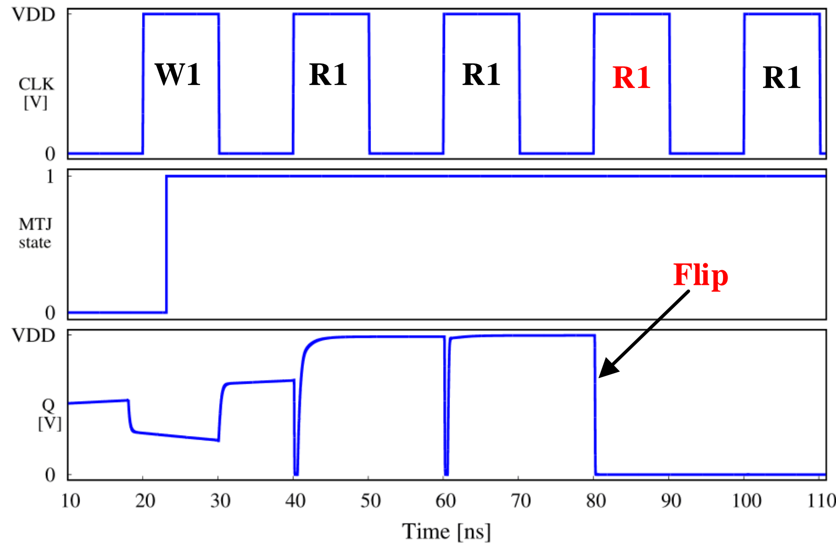


Figure 5.3: Waveform showing dIRF1-3 dynamic fault

faults as dynamic Incorrect Read Fault (dIRF). A dIRF-n is a dynamic IRF which requires at least n consecutive reads to excite, and hence detect, the faults. It means that with m < n reads, the read result is still fault-free. Please note that unlike the Read Destructive Faults and the dynamic version commonly observed in SRAM memories [97], the current required to flip (switch) MTJ is far larger (almost 3-4 times) than the read current, and hence, the read faults in STT-MRAM are not destructive.

Table 5.6 shows the corner analysis of an observed dynamic fault in our analysis. The table shows that the resistance range and the number of operations required to sensitize the fault (n) depends on the operating corner. For instance, a resistance of 12kΩ results in a dIRF1-3 fault in the nominal voltage and nominal temperature (NVNT) corner, whereas in the low voltage and high temperature corner (LVHT), it would cause a dIRF1-4 fault. For any value of n, the least resistance range is at the LVHT corner, hence this corner can be considered as the worst corner for dIRF1 fault. Please note that Table 5.6 does not have the resistance range in the LVLT corner since transition fault TF1 happens in this range, making the cell not writable.

In addition to the voltage and temperature corner, the dynamic read faults are also impacted by the process variation, as shown in Fig. 5.4. From the figure, we see that the value of n in dIRF-n faults can go to a maximum of 8 at the LVHT (worst case) corner in the presence of process variations. Within the same resistance range, the maximum value of n without process variation was 4 as given in Table 5.6. This

n	Resistance Range								
	LV LT	LV NT	LV HT	NV LT	NV NT	NV HT	HV LT	HV NT	HV HT
2	-	-	$1k\Omega-2k\Omega$	-	$2k\Omega-3k\Omega$	-	$2k\Omega-3k\Omega$	$1k\Omega-4k\Omega$	$1k\Omega-4k\Omega$
3	-	$3k\Omega-11k\Omega$	$3k\Omega-11k\Omega$	$3k\Omega-13k\Omega$	$4k\Omega-14k\Omega$	$1k\Omega-9k\Omega$	$4k\Omega-16k\Omega$	$5k\Omega-17k\Omega$	$5k\Omega-18k\Omega$
4	-	$12k\Omega-25k\Omega$	$12k\Omega-26k\Omega$	$14k\Omega-28k\Omega$	$15k\Omega-29k\Omega$	$10k\Omega-25k\Omega$	$17k\Omega-32k\Omega$	$18k\Omega-33k\Omega$	$19k\Omega-36k\Omega$
LV – Low Voltage; NV – Nominal Voltage; HV – High Voltage LT – Low Temperature; NT – Nominal Temperature; HT – High Temperature									

Table 5.6: Corner Analysis for dIRF1 dynamic fault (SL – IN short)

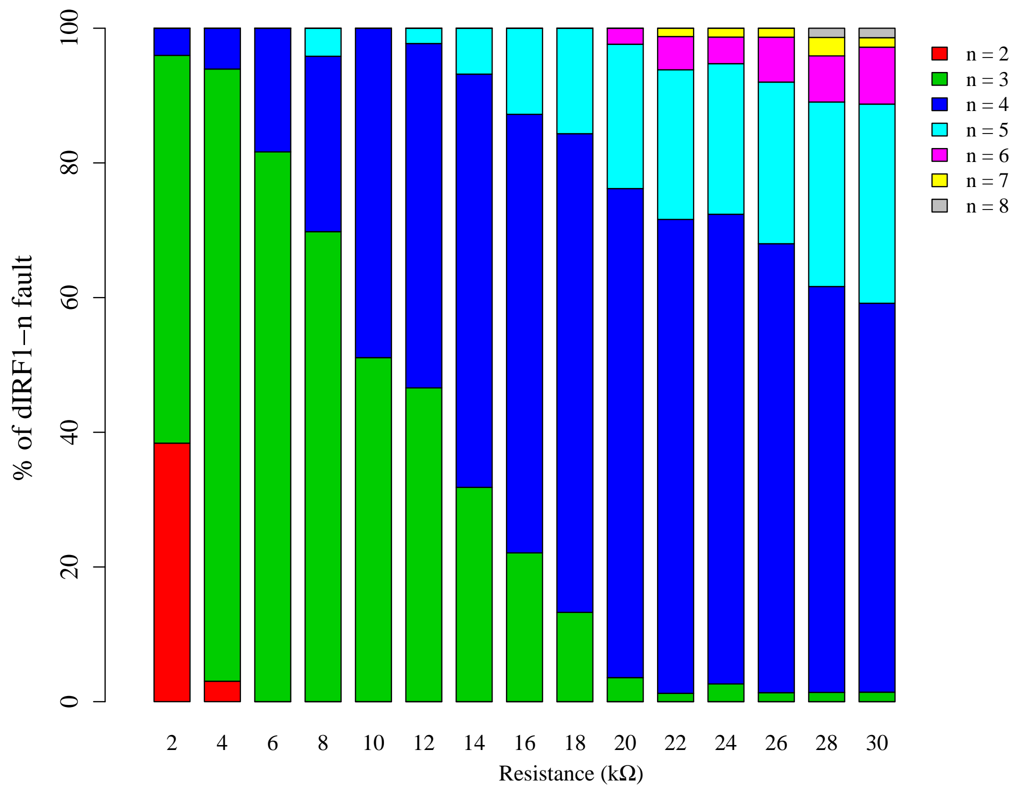


Figure 5.4: Dynamic fault with process variation at LVHT corner

n	Resistance Range								
	LV LT	LV NT	LV HT	NV LT	NV NT	NV HT	HV LT	HV NT	HV HT
2	-	-	$< 3k\Omega$	-	$< 3k\Omega$	$< 6k\Omega$	$< 3k\Omega$	$< 5k\Omega$	$< 8k\Omega$
3	$< 9k\Omega$	$< 7k\Omega$	$4k\Omega-16k\Omega$	$< 18k\Omega$	$4k\Omega-20k\Omega$	$7k\Omega-23k\Omega$	$4k\Omega-22k\Omega$	$6k\Omega-14k\Omega$	$9k\Omega-28k\Omega$
4	$10k\Omega-28k\Omega$	$8k\Omega-28k\Omega$	$17k\Omega-30k\Omega$	$19k\Omega-38k\Omega$	$21k\Omega-36k\Omega$	$24k\Omega-46k\Omega$	$23k\Omega-44k\Omega$	$15k\Omega-39k\Omega$	$29k\Omega-43k\Omega$
LV – Low Voltage; NV – Nominal Voltage; HV – High Voltage LT – Low Temperature; NT – Nominal Temperature; HT – High Temperature									

Table 5.7: Corner Analysis for dCFir1 (BL – IN short)

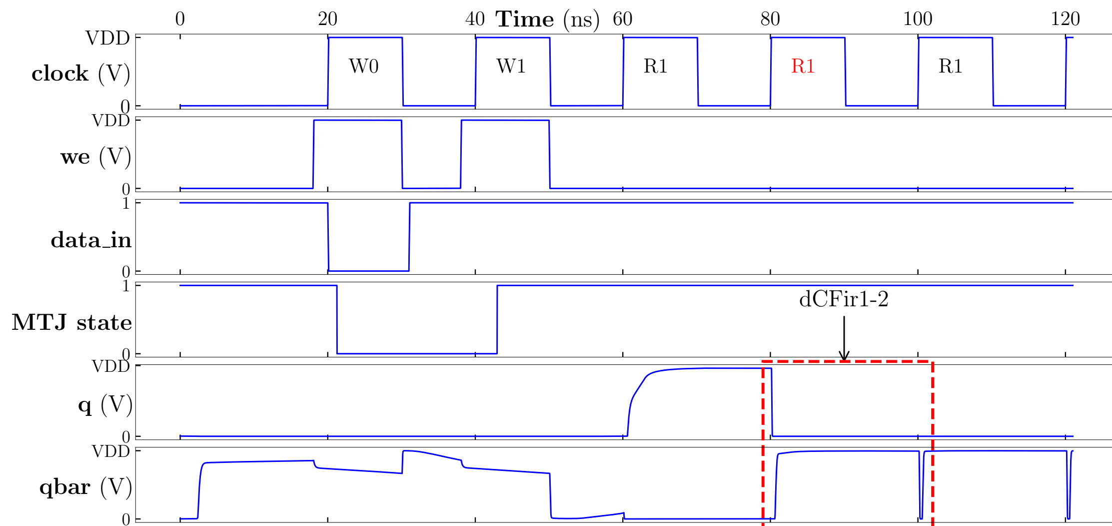


Figure 5.5: Waveform showing the signals of a victim cell affected by dCFir1-2

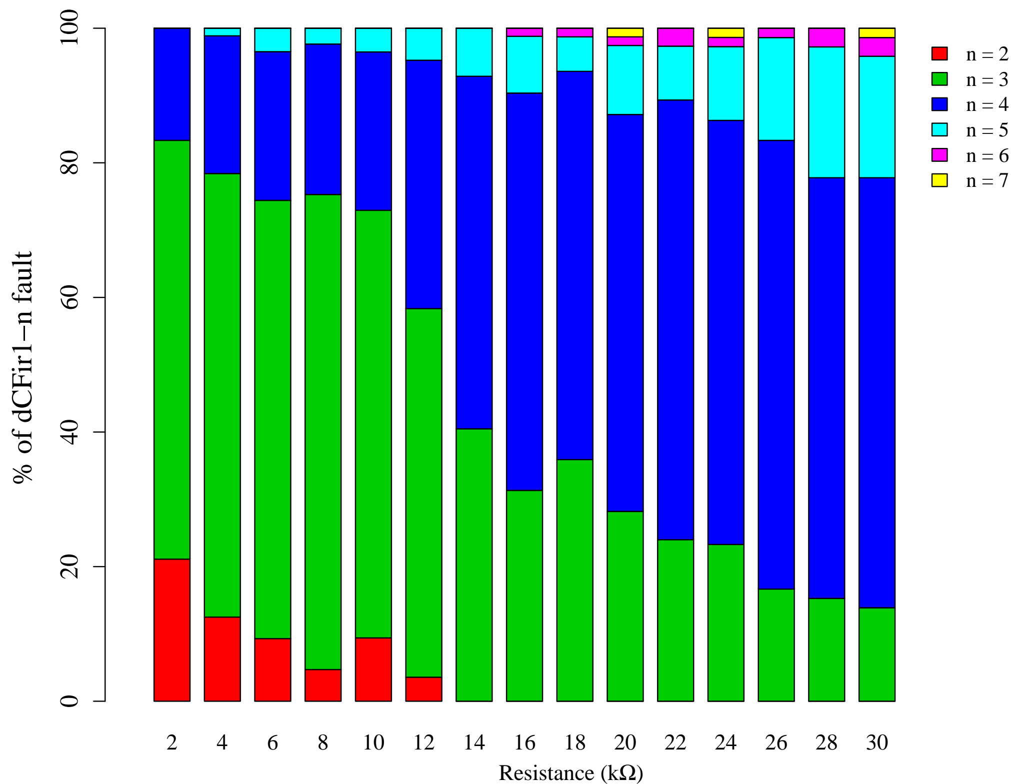


Figure 5.6: Coupling fault with process variation at LVHT corner

illustrates the significance of considering process variations into the analysis to design proper test patterns to detect dynamic faults.

In Table 5.7, we show one case of a coupling fault, namely dCFir (dynamic Incorrect Read Coupling Fault). Here, the coupling fault occurs due to a short between BL and IN. The waveform of a victim cell affected by this fault is given in Fig. 5.5, where the second read from the cell fails when the cell is storing a '1', indicating a dCFir1-2 dynamic coupling fault. Similar to the case with the dIRF1 fault, the resistance

range for the dCFir1 fault also depends on the operating corner as presented in the Table 5.7. The worst corner in this case is also the LVHT corner. In addition, the value of n can increase from 4 in the nominal case (Table 5.7) to 7 due to the impact of process variation as illustrated in Fig. 5.6.

We have also explored MTJ related defects, mainly the oxide thickness variations and the RA (resistance-area) product. These variations mainly affect the 'P' and 'AP' resistances of the MTJ, resulting in IRFs. A low of oxide thickness or RA results in lower values of the 'P' and 'AP' resistances. This results in IRF1, since the 'AP' resistance decreases below the sensing threshold. Similarly, a high value of oxide thickness or RA increases the 'P' and 'AP' resistances, resulting in IRF0. The other MTJ defects result in write faults.

5.4 Test Pattern Generation

Based on the results presented in the previous chapter the next step is construction of the efficient test algorithm for STT-MRAM that will provide a full coverage of the observed faults. Since STT-MRAM is CMOS compatible therefore, the set of test operations is the same with RAM memories and as a result conventional March test algorithms can be used for STT-MRAM testing. There are already a number of tools for building test algorithms that accept as an input a set of faults and return the corresponding test algorithm as a result. In most cases these tools require improvements whenever new types of faults are discovered. The approach suggested by AIFA and used in this paper is slightly different and is not directly connected to the set of faults.

Test Sequence 1 (22N):
⇑(W0);
⇑(R0,W1,R1, R1, R1, R1, R1, R1, R1, R1);
⇓(R1,W0,R0, R0, R0, R0, R0, R0, R0, R0);
⇓(R0).

The idea here is to construct the final test algorithm based on the set of identified test sequences during the fault modeling process (see Fig. 5.7). For example, the below Test Sequence 1 detects some of STT-MRAM-specific faults discussed above.

In a similar way, Test Sequence 2, Test Sequence 3, etc, are constructed for different classes of STT-MRAM faults and finally a unified test algorithm for testing all the considered faults is generated. This unified test algorithm is part of Synopsys STAR Memory System (SMS) product (see Fig. 5.8) which has been extended to test MRAMs [96]. It provides possibility to test different types of memories within the same unified test and repair infrastructure. As it can be seen from Fig. 5.8, one SMS Processor can test SRAMs (SMS Processor contains built-in self-test engine), another one MRAM, while the other SMS Processors can test Caches using Test Bus interface, and all these SMS Processors are connected to the common SMS Server.

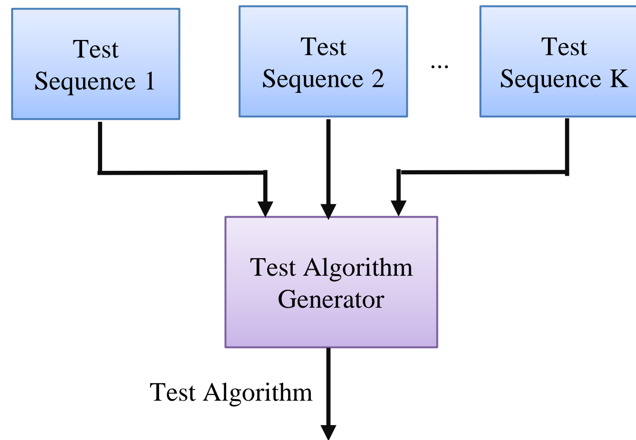


Figure 5.7: Test algorithm generation flow

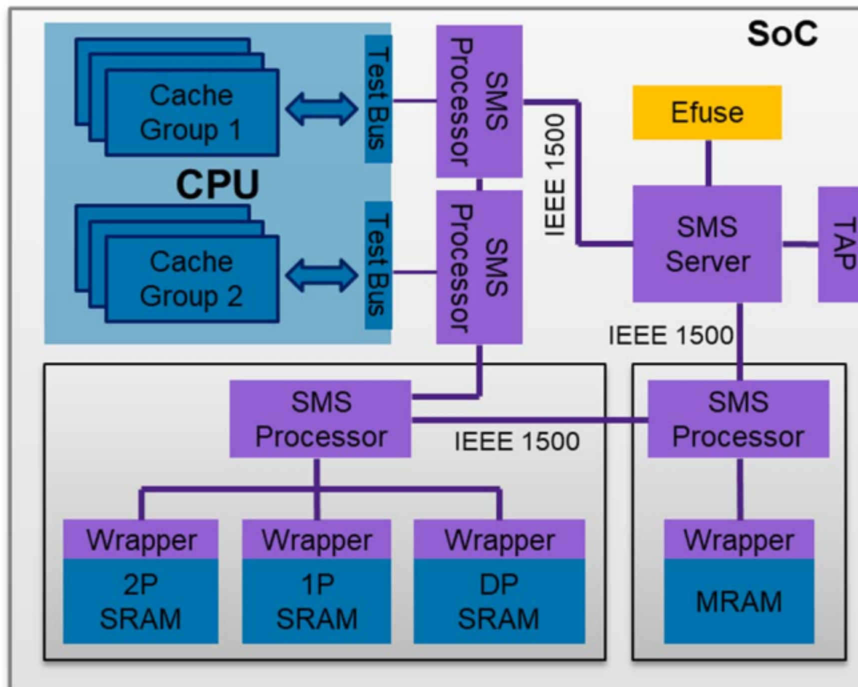


Figure 5.8: STAR Memory System

This infrastructure allows to organize flexible test scheduling and apply parallel test on different types of memories within the same test session.

5.5 Summary

Due to unique magnetic fabrication processes, STT-MRAM is subject to new failure mechanisms, defects and faults. In this chapter, a defect characterization and fault modeling methodology is presented for STT-MRAM, based on extensive defect injection campaign, by considering the netlist and layout, as well as test environments.

Both spot defects, manifesting as resistive opens and shorts, as well as MTJ specific defects have been evaluated. Based on the results of fault analysis, efficient test algorithms have been developed to cover the unique faults of STT-MRAM.

Chapter 6

Conclusions and Outlook

Conventional CMOS memories are facing severe challenges due to technology scaling, primarily due to the increased power consumption. In particular, the static or leakage power dominates the total power in advanced technology nodes. Various emerging non-volatile memories such as FeRAM, PCRAM, ReRAM and STT-MRAM are being considered as a replacement to CMOS memories to overcome their scaling limitations. These non-volatile memories have zero leakage power and can provide normally-off/instant-on capabilities. Among these, the STT-MRAM technology is the most promising since it has access speeds similar to those of SRAM, integration density similar to DRAM and is non-volatile like Flash memory. These unique features make STT-MRAM an interesting candidate for a universal memory, since it can potentially fit into every level of the memory hierarchy. Several recent industrial demonstrations from major memory chip manufacturers such as Samsung, Intel, TSMC and Everspin have established the feasibility of using STT-MRAM for both on-chip as well as standalone memory applications.

As the STT-MRAM technology becomes popular, it is important to analyze the impact of manufacturing variations and defects on its reliability. In particular, as the technology scales down, the manufacturing variations are increasing, and the associated reliability challenges could offset some of the benefits of adopting this technology. In addition, the impact of manufacturing variations exacerbates the stochastic switching behavior of the MTJ, which is the storage element in the STT-MRAM technology. Hence, it is essential to have tools and frameworks to quantify the effect of variations and defects on STT-MRAM memories, which can provide realistic estimates of the latency, energy, failure rates and yield of STT-based memories.

6.1 Conclusions

In this dissertation, various reliability aspects of STT-MRAM are investigated including the impact of process variation, voltage, temperature, extreme parametric variations and manufacturing defects. To quantify the impact of these variations, several tools and frameworks are developed. With the help of these tools, existing

reliability and yield improvement techniques as well as new mitigation techniques specific to STT-MRAM are analyzed and their effectiveness determined. These analyses can help the designers to evaluate the various trade-offs to be considered in the design of STT-MRAM and select the right combination of mitigation techniques to be implemented to keep the failure rates to a minimum, thus improving the reliability and yield.

The first part of this dissertation deals with variation-ware STT-MRAM analysis and design space exploration (Chapter 3). Here, we quantify the impact of variations in the bit-cell, peripheral circuitry as well as the stochastic switching of the MTJ on the access latency and energy of the entire memory system. The results show that nominal analysis heavily underestimates the overall latency and energy, highlighting the importance of variation-aware analysis.

In the second part of this dissertation, we analyze the impact of extreme parametric variations on both the permanent faults and the reliability failures (Chapter 4). Here, we consider variations in the bit-cells and peripherals as well as the correlation among neighboring cells to get the fault distribution map of the memory array to estimate the yield. The results show that unique techniques specific to this technology, such as current boosting, can be used in conjunction with traditional techniques, such as ECC and redundancy, to maximize the yield.

The final part of this dissertation focuses on defect injection and fault modeling of STT-MRAM (Chapter 5). The fault modeling framework considers both spot defects (resistive opens and shorts) as well as MTJ specific defects. Based on the fault analysis results, efficient test algorithms to cover the unique faults of this technology are developed.

6.2 Outlook

This dissertation focuses on the reliability challenges and yield improvement of STT-MRAM memory. As this memory gains widespread adoption, several new challenges need to be overcome. In particular, the security aspects related to this technology need to be thoroughly investigated. This is explored in one of our works [98]. In this work, we show that the unique fault mechanisms of this technology can be exploited by an adversary to deploy a hardware Trojan, which are stealthy modifications to the circuit intended to cause malicious effects. Furthermore, new computing paradigms, such as *Compute-in-Memory or CiM*, based on STT-MRAM, are becoming popular. The reliability aspects of these new paradigms also need to be examined, as shown in our work [99]. In addition, new spintronic-based memories such as SOT-MRAM are under research and analyzing their reliability aspects would become important in the near future. To this end, the tools and analysis frameworks developed in this dissertation can serve as the basis for further research on security and reliability aspects of emerging spintronic memories and new computing paradigms based on this novel memory technology.

Bibliography

[1] "40 Years of Microprocessor Trend Data." [Online]. Available: https://www.karlrupp.net/2015/06/40-years-of-microprocessor-trend-data/.

[2] K. Itoh, M. Horiguchi, and H. Tanaka, *Ultra-low voltage nano-scale memories*. Springer Science & Business Media, 2007.

[3] "Hynix." [Online]. Available: https://www.skhynix.com/eng/about/history2000.jsp.

[4] "As Nodes Advance, So Must Power Analysis." [Online]. Available: http://semiengineering.com/as-nodes-advance-so-must-power-analysis/.

[5] X. Dong, C. Xu, Y. Xie, and N. P. Jouppi, "NVSim: A Circuit-Level Performance, Energy, and Area Model for Emerging Nonvolatile Memory," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 31, no. 7, 2012.

[6] G. E. Moore *et al.*, "Cramming more components onto integrated circuits," 1965.

[7] G. E. Moore, "Cramming more components onto integrated circuits," *Proceedings of the IEEE*, vol. 86, no. 1, pp. 82–85, 1998.

[8] R. H. Dennard, F. H. Gaensslen, V. L. Rideout, E. Bassous, and A. R. LeBlanc, "Design of ion-implanted MOSFET's with very small physical dimensions," *IEEE Journal of Solid-State Circuits*, vol. 9, no. 5, pp. 256–268, 1974.

[9] N. S. Kim, T. Austin, D. Baauw, T. Mudge, K. Flautner, J. S. Hu, M. J. Irwin, M. Kandemir, and V. Narayanan, "Leakage current: Moore's law meets static power," *computer*, vol. 36, no. 12, pp. 68–75, 2003.

[10] "International Technology Roadmap for Semiconductors (ITRS)." [Online]. Available: https://www.semiconductors.org/wp-content/uploads/2018/06/0_2015-ITRS-2.0-Executive-Report-1.pdf.

[11] S. E. Thompson and S. Parthasarathy, "Moore's law: the future of si microelectronics," *Materials today*, vol. 9, no. 6, pp. 20–25, 2006.

[12] M. Hosomi, H. Yamagishi, T. Yamamoto, K. Bessho, Y. Higo, K. Yamane, H. Yamada, M. Shoji, H. Hachino, C. Fukumoto, *et al.*, "A novel nonvolatile memory with spin torque transfer magnetization switching: Spin-RAM," in *IEEE International Electron Devices Meeting (IEDM)*, pp. 459–462, 2005.

[13] T. Kawahara, "Scalable spin-transfer torque ram technology for normally-off computing," *IEEE Design & Test of Computers*, no. 1, pp. 52–63, 2010.

[14] H. Li and Y. Chen, "An overview of non-volatile memory technology and the implication for tools and architectures," in *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp. 731–736, 2009.

[15] M.-T. Chang, P. Rosenfeld, S.-L. Lu, and B. Jacob, "Technology comparison for large last-level caches (L^3Cs): Low-leakage SRAM, low write-energy STT-RAM, and refresh-optimized eDRAM," in *19th International Symposium on High Performance Computer Architecture (HPCA)*, pp. 143–154, 2013.

[16] A. D. Kent and D. C. Worledge, "A new spin on magnetic memories," *Nature nanotechnology*, vol. 10, no. 3, p. 187, 2015.

[17] H.-S. P. Wong and S. Salahuddin, "Memory leads the way to better computing," *Nature nanotechnology*, vol. 10, no. 3, p. 191, 2015.

[18] A. Driskill-Smith and Y. Huai, "STTRAM–A new spin on universal memory," *Future Fab Intl. Report*, 2008.

[19] S. A. Wolf, J. Lu, M. R. Stan, E. Chen, and D. M. Treger, "The Promise of Nanomagnetics and Spintronics for Future Logic and Universal Memory," *Proceedings of the IEEE*, vol. 98, no. 12, pp. 2155–2168, 2010.

[20] O. Golonzka, J.-G. Alzate, U. Arslan, M. Bohr, P. Bai, J. Brockman, B. Buford, C. Connor, N. Das, B. Doyle, *et al.*, "MRAM as embedded non-volatile memory solution for 22FFL FinFET technology," in *2018 IEEE International Electron Devices Meeting (IEDM)*, pp. 18–1, 2018.

[21] H. Sato, H. Honjo, T. Watanabe, M. Niwa, H. Koike, S. Miura, T. Saito, H. Inoue, T. Nasuno, T. Tanigawa, *et al.*, "14ns write speed 128Mb density Embedded STT-MRAM with endurance $>10^{10}$ and 10yrs retention@ 85 C using novel low damage MTJ integration process," in *2018 IEEE International Electron Devices Meeting (IEDM)*, pp. 27–2, 2018.

[22] K. Lee, R. Chao, K. Yamane, V. Naik, H. Yang, J. Kwon, N. Chung, S. Jang, B. Behin-Aein, J. Lim, *et al.*, "22-nm FD-SOI Embedded MRAM Technology for Low-Power Automotive-Grade-l MCU Applications," in *2018 IEEE International Electron Devices Meeting (IEDM)*, pp. 27–1, 2018.

[23] Y. Song, J. Lee, S. Han, H. Shin, K. Lee, K. Suh, D. Jeong, G. Koh, S. Oh, J. Park, *et al.*, "Demonstration of highly manufacturable STT-MRAM embedded in 28nm logic," in *2018 IEEE International Electron Devices Meeting (IEDM)*, pp. 18–2, 2018.

[24] S. Aggarwal, H. Almasi, M. DeHerrera, B. Hughes, S. Ikegawa, J. Janesky, H. Lee, H. Lu, F. Mancoff, K. Nagel, *et al.*, "Demonstration of a reliable 1 gb standalone spin-transfer torque mram for industrial applications," in *2019 IEEE International Electron Devices Meeting (IEDM)*, pp. 2–1, 2019.

[25] K. Lee, J. Bak, Y. Kim, C. Kim, A. Antonyan, D. Chang, S. Hwang, G. Lee, N. Ji, W. Kim, *et al.*, "1Gbit High Density Embedded STT-MRAM in 28nm FDSOI Technology," in *2019 IEEE International Electron Devices Meeting (IEDM)*, pp. 2–2, 2019.

[26] V. Naik, K. Lee, K. Yamane, R. Chao, J. Kwon, N. Thiyagarajah, N. Chung, S. Jang, B. Behin-Aein, J. Lim, *et al.*, "Manufacturable 22nm FD-SOI Embedded MRAM Technology for Industrial-grade MCU and IOT Applications," in *2019 IEEE International Electron Devices Meeting (IEDM)*, pp. 2–3, 2019.

[27] J. Alzate, U. Arslan, P. Bai, J. Brockman, Y. Chen, N. Das, K. Fischer, T. Ghani, P. Heil, P. Hentges, *et al.*, "2 MB Array-Level Demonstration of STT-MRAM Process and Performance Towards L4 Cache Applications," in *2019 IEEE International Electron Devices Meeting (IEDM)*, pp. 2–4, 2019.

[28] J.-H. Park, J. Lee, J. Jeong, U. Pi, W. Kim, S. Lee, E. Noh, K. Kim, W. Lim, S. Kwon, *et al.*, "A novel integration of STT-MRAM for on-chip hybrid memory by utilizing non-volatility modulation," in *2019 IEEE International Electron Devices Meeting (IEDM)*, pp. 2–5, 2019.

[29] G. Hu, J. Nowak, M. Gottwald, S. Brown, B. Doris, C. DEmic, P. Hashemi, D. Houssameddine, Q. He, D. Kim, *et al.*, "Spin-transfer torque MRAM with reliable 2 ns writing for last level cache applications," in *2019 IEEE International Electron Devices Meeting (IEDM)*, pp. 2–6, 2019.

[30] W. Gallagher, E. Chien, T.-W. Chiang, J.-C. Huang, M.-C. Shih, C. Wang, C.-H. Weng, S. Chen, C. Bair, G. Lee, *et al.*, "22nm STT-MRAM for Reflow and Automotive Uses with High Yield, Reliability, and Magnetic Immunity and with Performance and Shielding Options," in *2019 IEEE International Electron Devices Meeting (IEDM)*, pp. 2–7, 2019.

[31] Y.-C. Shih, C.-F. Lee, Y.-A. Chang, P.-H. Lee, H.-J. Lin, Y.-L. Chen, K.-F. Lin, T.-C. Yeh, H.-C. Yu, H. H. Chuang, *et al.*, "Logic Process Compatible 40-nm 16-Mb, Embedded Perpendicular-MRAM With Hybrid-Resistance Reference, Sub-μA Sensing Resolution, and 17.5-nS Read Access Time," *IEEE Journal of Solid-State Circuits*, vol. 54, no. 4, pp. 1029–1038, 2019.

[32] A. Antonyan, S. Pyo, H. Jung, and T. Song, "Embedded MRAM Macro for eFlash Replacement," in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–4, 2018.

[33] K. Garello, C. O. Avci, I. M. Miron, M. Baumgartner, A. Ghosh, S. Auffret, O. Boulle, G. Gaudin, and P. Gambardella, "Ultrafast magnetization switching by spin-orbit torques," *Applied Physics Letters*, vol. 105, no. 21, p. 212402, 2014.

[34] S. M. Nair, R. Bishnoi, M. S. Golanbari, F. Oboril, and M. B. Tahoori, "VAET-STT: A variation aware estimator tool for STT-MRAM based memories," in *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp. 1456–1461, 2017.

[35] S. M. Nair, R. Bishnoi, M. S. Golanbari, F. Oboril, F. Hameed, and M. B. Tahoori, "VAET-STT: Variation aware STT-MRAM analysis and design space exploration tool," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 37, no. 7, pp. 1396–1407, 2017.

[36] S. M. Nair, R. Bishnoi, and M. B. Tahoori, "Parametric failure modeling and yield analysis for STT-MRAM," in *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp. 265–268, 2018.

[37] S. M. Nair, R. Bishnoi, and M. B. Tahoori, "A Comprehensive Framework for Parametric Failure Modeling and Yield Analysis of STT-MRAM," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 27, no. 7, pp. 1697–1710, 2019.

[38] S. M. Nair, R. Bishnoi, M. B. Tahoori, G. Tshagharyan, H. Grigoryan, G. Harutyunyan, and Y. Zorian, "Defect injection, fault modeling and test algorithm generation methodology for STT-MRAM," in *International Test Conference (ITC)*, pp. 1–10, 2018.

[39] S. M. Nair, R. Bishnoi, M. Tahoori, H. Grigoryan, and G. Tshagharyan, "Variation-aware Fault Modeling and Test Generation for STT-MRAM," in *International Symposium on On-Line Testing and Robust System Design (IOLTS)*, pp. 80–83, 2019.

[40] S. Salehi, D. Fan, and R. F. Demara, "Survey of STT-MRAM cell design strategies: Taxonomy and sense amplifier tradeoffs for resiliency," *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, vol. 13, no. 3, pp. 1–16, 2017.

[41] D. Lee, S. K. Gupta, and K. Roy, "High-performance low-energy STT MRAM based on balanced write scheme," in *2012 ACM/IEEE International Symposium on Low Power Electronics and Design (ISLPED)*, pp. 9–14, 2012.

[42] S. Borkar, T. Karnik, and V. De, "Design and reliability challenges in nanometer technologies," in *Design Automation Conference (DAC)*, pp. 75–75, 2004.

[43] P. Friedberg, Y. Cao, J. Cain, R. Wang, J. Rabaey, and C. Spanos, "Modeling within-die spatial correlation effects for process-design co-optimization," in *International Symposium on Quality Electronic Design (ISQED)*, pp. 516–521, 2005.

[44] M. Orshansky, L. Milor, and C. Hu, "Characterization of spatial intrafield gate CD variability, its impact on circuit performance, and spatial mask-level correction," *IEEE Transactions on Semiconductor Manufacturing*, vol. 17, no. 1, pp. 2–11, 2004.

[45] B. E. Stine, D. S. Boning, and J. E. Chung, "Analysis and decomposition of spatial variation in integrated circuit processes and devices," *IEEE Transactions on Semiconductor Manufacturing*, vol. 10, no. 1, pp. 24–41, 1997.

[46] W. Kang, L. Zhang, J.-O. Klein, Y. Zhang, D. Ravelosona, and W. Zhao, "Reconfigurable codesign of STT-MRAM under process variations in deeply scaled technology," *IEEE Transactions on Electron Devices*, vol. 62, no. 6, pp. 1769–1777, 2015.

[47] J. Li, C. Augustine, S. Salahuddin, and K. Roy, "Modeling of failure probability and statistical design of spin-torque transfer magnetic random access memory (STT MRAM) array for yield enhancement," in *Proceedings of the 45th annual Design Automation Conference (DAC)*, pp. 278–283, 2008.

[48] W. Zhao, X. Zhao, B. Zhang, K. Cao, L. Wang, W. Kang, Q. Shi, M. Wang, Y. Zhang, Y. Wang, *et al.*, "Failure analysis in magnetic tunnel junction nanopillar with interfacial perpendicular magnetic anisotropy," *Materials*, vol. 9, no. 1, p. 41, 2016.

[49] K. Munira, W. H. Butler, and A. W. Ghosh, "A quasi-analytical model for energy-delay-reliability tradeoff studies during write operations in a perpendicular STT-RAM cell," *IEEE transactions on electron devices*, vol. 59, no. 8, pp. 2221–2226, 2012.

[50] D. Apalkov, A. Khvalkovskiy, S. Watts, V. Nikitin, X. Tang, D. Lottis, K. Moon, X. Luo, E. Chen, A. Ong, *et al.*, "Spin-transfer torque magnetic random access memory (STT-MRAM)," *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, vol. 9, no. 2, p. 13, 2013.

[51] C. W. Smullen, V. Mohan, A. Nigam, S. Gurumurthi, and M. R. Stan, "Relaxing non-volatility for fast and energy-efficient STT-RAM caches," in *17th International Symposium on High Performance Computer Architecture (HPCA)*, pp. 50–61, 2011.

[52] Y. Ye, F. Liu, M. Chen, S. Nassif, and Y. Cao, "Statistical modeling and simulation of threshold variation under random dopant fluctuations and line-edge roughness," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 19, no. 6, pp. 987–996, 2011.

[53] S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, and V. De, "Parameter variations and impact on circuits and microarchitecture," in *Proceedings of the 40th annual Design Automation Conference (DAC)*, pp. 338–342, 2003.

[54] M. J. Pelgrom, A. C. Duinmaijer, and A. P. Welbers, "Matching properties of MOS transistors," *IEEE Journal of solid-state circuits*, vol. 24, no. 5, pp. 1433–1439, 1989.

[55] K. J. Kuhn, M. D. Giles, D. Becher, P. Kolar, A. Kornfeld, R. Kotlyar, S. T. Ma, A. Maheshwari, and S. Mudanai, "Process technology variation," *IEEE Transactions on Electron Devices*, vol. 58, no. 8, pp. 2197–2208, 2011.

[56] R. Bishnoi, F. Oboril, and M. B. Tahoori, "Design of defect and fault-tolerant nonvolatile spintronic flip-flops," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 25, no. 4, pp. 1421–1432, 2016.

[57] W. Kang, L. Zhang, W. Zhao, J.-O. Klein, Y. Zhang, D. Ravelosona, and C. Chappert, "Yield and reliability improvement techniques for emerging nonvolatile STT-MRAM," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 5, no. 1, pp. 28–39, 2015.

[58] S. Motaman, S. Ghosh, and N. Rathi, "Impact of process-variations in STTRAM and adaptive boosting for robustness," in *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp. 1431–1436, 2015.

[59] W. Kang, W. Zhao, Z. Wang, Y. Zhang, J.-O. Klein, Y. Zhang, C. Chappert, and D. Ravelosona, "A low-cost built-in error correction circuit design for STT-MRAM reliability improvement," *Microelectronics Reliability*, vol. 53, no. 9, pp. 1224–1229, 2013.

[60] N. Sayed, M. Ebrahimi, R. Bishnoi, and M. B. Tahoori, "Opportunistic write for fast and reliable STT-MRAM," in *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp. 554–559, 2017.

[61] N. Sayed, F. Oboril, R. Bishnoi, and M. B. Tahoori, "Leveraging Systematic Unidirectional Error-Detecting Codes for fast STT-MRAM cache," in *VLSI Test Symposium (VTS)*, pp. 1–6, 2017.

[62] C.-L. Su, Y.-T. Yeh, and C.-W. Wu, "An integrated ECC and redundancy repair scheme for memory reliability enhancement," in *IEEE International Symposium on Defect and Fault Tolerance in VLSI Systems*, pp. 81–89, 2005.

[63] Z. Pajouhi, X. Fong, and K. Roy, "Device/circuit/architecture co-design of reliable STT-MRAM," in *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp. 1437–1442, 2015.

[64] X. Bi, Z. Sun, H. Li, and W. Wu, "Probabilistic design methodology to improve run-time stability and performance of STT-RAM caches," in *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pp. 88–94, 2012.

[65] E. I. Vatajelu, R. Rodriguez-Montañes, M. Indaco, P. Prinetto, and J. Figueras, "STT-MRAM cell reliability evaluation under process, voltage and temperature (PVT) variations," in *10th International Conference on Design & Technology of Integrated Systems in Nanoscale Era (DTIS)*, pp. 1–6, 2015.

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

[67] A. Ahari, M. Ebrahimi, F. Oboril, and M. Tahoori, "Improving reliability, performance, and energy efficiency of STT-MRAM with dynamic write latency," in *33rd IEEE International Conference on Computer Design (ICCD)*, pp. 109–116, 2015.

[68] S. J. Wilton and N. P. Jouppi, "CACTI: An enhanced cache access and cycle time model," *IEEE Journal of solid-state circuits*, vol. 31, no. 5, pp. 677–688, 1996.

[69] E. Eken, L. Song, I. Bayram, C. Xu, W. Wen, Y. Xie, and Y. Chen, "NVSim-VX^{s}: an improved NVSim for variation aware STT-RAM simulation," in *53nd ACM/EDAC/IEEE Design Automation Conference (DAC)*, pp. 1–6, 2016.

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

[71] E. Castillo, A. S. Hadi, N. Balakrishnan, and J.-M. Sarabia, *Extreme value and related models with applications in engineering and science.* Wiley Hoboken, NJ, 2005.

[72] I. M. Sobol, "Uniformly distributed sequences with an additional uniform property," *USSR Computational Mathematics and Mathematical Physics*, vol. 16, no. 5, pp. 236–242, 1976.

[73] A. Singhee and R. A. Rutenbar, "Why quasi-monte carlo is better than monte carlo or latin hypercube sampling for statistical circuit analysis," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 29, no. 11, pp. 1763–1776, 2010.

[74] W. Guo, G. Prenat, V. Javerliac, M. El Baraji, N. De Mestier, C. Baraduc, and B. Dieny, "SPICE modelling of magnetic tunnel junctions written by spin-transfer torque," *Journal of Physics D: Applied Physics*, vol. 43, no. 21, p. 215001, 2010.

[75] A. Chintaluri, H. Naeimi, S. Natarajan, and A. Raychowdhury, "Analysis of Defects and Variations in Embedded Spin Transfer Torque (STT) MRAM Arrays," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 6, no. 3, pp. 319–329, 2016.

[76] R. Bishnoi, M. Ebrahimi, F. Oboril, and M. B. Tahoori, "Read disturb fault detection in STT-MRAM," in *International Test Conference (ITC)*, pp. 1–7, 2014.

[77] H. Naeimi, C. Augustine, A. Raychowdhury, S.-L. Lu, and J. Tschanz, "STTRAM scaling and retention failure.," *Intel Technology Journal*, vol. 17, no. 1, pp. 54–75, 2013.

[78] N. Sayed, S. M. Nair, R. Bishnoi, and M. B. Tahoori, "Process variation and temperature aware adaptive scrubbing for retention failures in STT-MRAM," in *Asia and South Pacific Design Automation Conference (ASP-DAC)*, pp. 203–208, 2018.

[79] S. Hamdioui, Z. Al-Ars, and A. J. Van de Goor, "Testing static and dynamic faults in random access memories," in *VLSI Test Symposium*, pp. 395–400, 2002.

[80] Y. Zhang, B. Yan, X. Wang, and Y. Chen, "Persistent and Nonpersistent Error Optimization for STT-RAM Cell Design," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 36, no. 7, pp. 1181–1192, 2017.

[81] R. Bishnoi, M. Ebrahimi, F. Oboril, and M. B. Tahoori, "Improving write performance for STT-MRAM," *IEEE Transactions on Magnetics*, vol. 52, no. 8, pp. 1–11, 2016.

[82] N. Sayed, R. Bishnoi, F. Oboril, and M. B. Tahoori, "A cross-layer adaptive approach for performance and power optimization in STT-MRAM," in *Design, Automation & Test in Europe (DATE)*, pp. 791–796, 2018.

[83] Q.-K. Trinh, S. Ruocco, and M. Alioto, "Dynamic reference voltage sensing scheme for read margin improvement in STT-MRAMs," *IEEE Transactions on Circuits and Systems I*, vol. 65, no. 4, pp. 1269–1278, 2018.

[84] A. Mejdoubi, G. Prenat, and B. Dieny, "A compact model of precessional spin-transfer switching for MTJ with a perpendicular polarizer," in *28th International Conference on Microelectronics (MIEL)*, pp. 225–228, 2012.

[85] F. Bernard-Granger, B. Dieny, R. Fascio, and K. Jabeur, "SPITT: A magnetic tunnel junction SPICE compact model for STT-MRAM," in *Proceedings of the MOS-AK Workshop of the Design, Automation & Test in Europe (DATE)*, 2015.

[86] K. Sugiura, S. Takahashi, M. Amano, T. Kajiyama, M. Iwayama, Y. Asao, N. Shimomura, T. Kishi, S. Ikegawa, H. Yoda, *et al.*, "Ion beam etching technology for high-density spin transfer torque magnetic random access memory," *Japanese Journal of Applied Physics*, vol. 48, no. 8S1, p. 08HD02, 2009.

[87] J. Azevedo, A. Virazel, A. Bosio, L. Dilillo, P. Girard, A. Todri, G. Prenat, J. Alvarez-Hérault, and K. Mackay, "Impact of resistive-open defects on the heat current of TAS-MRAM architectures," in *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp. 532–537, 2012.

[88] M. Kuepferling, S. Zullino, A. Sola, B. Van de Wiele, G. Durin, M. Pasquale, K. Rott, G. Reiss, and G. Bertotti, "Vortex dynamics in Co-Fe-B magnetic tunnel junctions in presence of defects," *Journal of Applied Physics*, vol. 117, no. 17, p. 17E107, 2015.

[89] R. Bishnoi, F. Oboril, M. Ebrahimi, and M. B. Tahoori, "Avoiding unnecessary write operations in STT-MRAM for low power implementation," in *International Symposium on Quality Electronic Design (ISQED)*, pp. 548–553, 2014.

[90] I. Yoon, A. Chintaluri, and A. Raychowdhury, "EMACS: Efficient MBIST architecture for test and characterization of STT-MRAM arrays," in *IEEE International Test Conference (ITC)*, pp. 1–10, 2016.

[91] C.-L. Su, C.-W. Tsai, C.-W. Wu, C.-C. Hung, Y.-S. Chen, D.-Y. Wang, Y.-J. Lee, and M.-J. Kao, "Write disturbance modeling and testing for MRAM," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 16, no. 3, pp. 277–288, 2008.

[92] R. Carboni, S. Ambrogio, W. Chen, M. Siddik, J. Harms, A. Lyle, W. Kula, G. Sandhu, and D. Ielmini, "Modeling of breakdown-limited endurance in spin-transfer torque magnetic memory under pulsed cycling regime," *IEEE Transactions on Electron Devices*, vol. 65, no. 6, pp. 2470–2478, 2018.

[93] C.-L. Su, R.-F. Huang, C.-W. Wu, C.-C. Hung, M.-J. Kao, Y.-J. Chang, and W.-C. Wu, "MRAM defect analysis and fault modeling," in *International Conference on Test*, pp. 124–133, 2004.

[94] C.-L. Su, C.-W. Tsai, C.-W. Wu, J.-J. Chen, W.-C. Wu, C.-C. Hung, and M.-J. Kao, "Diagnosis for MRAM write disturbance fault," in *IEEE International Test Conference*, pp. 1–9, 2007.

[95] G. Harutyunyan, G. Tshagharyan, V. Vardanian, and Y. Zorian, "Fault modeling and test algorithm creation strategy for FinFET-based memories," in *32nd VLSI Test Symposium (VTS)*, pp. 1–6, 2014.

[96] K. Darbinyan, G. Harutyunyan, S. Shoukourian, V. Vardanian, and Y. Zorian, "A robust solution for embedded memory test and repair," in *Asian Test Symposium*, pp. 461–462, 2011.

[97] L. Dilillo, P. Girard, S. Pravossoudovitch, A. Virazel, S. Borri, and M. Hage-Hassan, "Dynamic Read Destructive Faults in Embedded-SRAMs: Analysis and March Test Solution," in *European Test Symposium (ETS)*, pp. 140–145, 2004.

[98] S. M. Nair, R. Bishnoi, A. Vijayan, and M. B. Tahoori, "Dynamic Faults based Hardware Trojan Design in STT-MRAM," in *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2020.

[99] S. M. Nair, C. Muench, and M. B. Tahoori, "Defect Characterization and Test Generation for Spintronic-based Compute-In-Memory," in *European Test Symposium (ETS)*, 2020.

Glossary

H_k Effective field anisotropy.

I_c Critical Current.

I_w Write Current.

K_B Boltzmann constant.

M_s Saturation magnetization.

T Temperature in Kelvin.

V Volume of the free layer.

V_{th} Threshold Voltage.

Δ Thermal Stability Factor.

r Radius of the MTJ.

AIFA Advanced Inductive Failure Analysis.

BEOL Back End Of Line.

CACTI An analytical model for access and cycle times of caches.

CAM Content Addressable Memory.

CDF Cumulative Distribution Function.

CMOS Complementary Metal-Oxide Semiconductor.

DRAM Dynamic Random Access Memory.

ECC Error Correcting Code.

FEOL Front End Of Line.

FeRAM Ferroelectric Random Access Memory.

FM Fault Masking.

GDS Graphic Data System.

GEV Generalized Extreme Value.

LER Line-Edge Roughness.

MRAM Magnetic Random Access Memory.

MTJ Magnetic Tunnel Junction.

NVM Non Volatile Memory.

NVSim An emerging non-volatile memory simulator used for performance, energy and area estimation.

NVSim-VX An improved NVSim for variation aware STT-MRAM simulation.

PCRAM Phase Change Random Access Memory.

PMA Perpendicular Magnetic Anisotropy.

RAM Random Access Memory.

RDF Random Dopant Fluctuation.

RER Read Error Rate.

ReRAM Resistive Random Access Memory.

ROM Read Only Memory.

RR Redundancy Repair.

SOT-MRAM Spin Orbit Torque Magnetic Random Access Memory.

SPICE Simulation Program with Integrated Circuit Emphasis, a general-purpose, open-source analog electronic circuit simulator.

SRAM Static Random Access Memory.

STI Shallow-Trench Isolation.

STT-MRAM Spin Transfer Torque Magnetic Random Access Memory.

TMR Tunneling Magneto-Resistance.

TSMC Taiwan Semiconductor Manufacturing Company, one of the major semiconductor foundries.

VAET-STT Variation Aware Estimator Tool for STT-MRAM.

VARIUS A tool for modeling process variation and resulting timing errors for microarchitects.

WER Write Error Rate.