Karlsruhe Institute of Technology

Reliability Modeling and Mitigation in Advanced Memory Technologies and Paradigms

Zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften

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

Dissertation
von
Mahta Mayahinia

Tag der mündlichen Prüfung: 12. July 2024

1. Referent: Prof. Dr. Mehdi B. Tahoori
Karlsruhe Institute für Technology (KIT)
2. Referentin: Prof. Dr. Lorena Anghel
Universität Grenoble-Alpes

Acknowledgement

Reflecting on this journey, I feel incredibly fortunate to have been surrounded by extraordinary individuals whose invaluable support made this dissertation possible.

First and foremost, I want to express my deepest gratitude to my advisor, Prof. Dr. Mehdi Tahoori. I could not have asked for a better mentor throughout my PhD journey. His insightful comments, invaluable discussions, and unwavering support were especially significant during challenging times.

I am deeply grateful for the opportunity to collaborate with remarkable teams from around the world. This experience has been a truly fortunate aspect of my PhD journey. I extend my heartfelt thanks to Prof. Dr. Francky Catthoor from IMEC. His guidance as my second advisor since the early days of my PhD has been invaluable. Additionally, I am thankful to my colleagues within the IMEC collaboration—Tommaso Marinelli, Hsiao-Hsuan (Samantha) Liu, and Zhenlin Pei—with whom I had the pleasure of collaborating on multiple research projects.

My gratitude also extends to my amazing colleagues in the Chair of Dependable Nano Computing (CDNC). Special thanks to our secretary, Ms. Iris Schröder-Piepka, for her exceptional ability to keep everything running smoothly. I am also thankful to Dr.-Ing. Dennis Gnad, Dr.-Ing. Jonas Krautter, and Dr.-Ing. Christopher Münch for all our discussions and their kindness in answering my questions, even as they pursued new professions outside KIT. I would like to acknowledge Soyed Tuhin Ahmed and Sergej Meschkov, my colleagues who started this PhD journey with me during the challenging times of the COVID-19 pandemic. I am grateful to Zhe Zhang and Surendra Hemaram, with whom I co-authored several research papers. My office mates, Dina Moussa and Haneen Seyam, made our shared time enjoyable and memorable. Vincent Meyers provided invaluable help in refining the German abstract of this dissertation. Mahboobe Sadeghipourrudsari's technical artistry greatly aided in the preparation of the slides for my doctoral examination. Lastly, I want to thank Sina Bakhtavari Mamaghani, Brojogopal Sapui, Seyedehmaryam Ghasemi, Priyanjana Pal, Shanmukha Mangadahalli Siddaramu, Paula Carolina Lozano Duarte, Ali Nezhadi Khelejani, Gürol Saglam, Maha Shatta, Tara Gheshlaghi, and Aradhana Dube. Their support has been greatly appreciated.

I would also like to extend my heartfelt gratitude to my parents, Mozhgan and Babak, and my brother Sina. Despite the distance, their unwavering support has always been a tremendous source of comfort. Finally, I want to thank my husband, Ali. His support and love were beyond words; without him, I could not have reached this point. I feel blessed to have such a family by my side.

In conclusion, I wish to express my profound admiration for the brilliant minds who believe in and tirelessly strive for gender equality. Their unwavering efforts are a genuine source of inspiration to me and countless others worldwide. This work is humbly and respectfully dedicated to them.

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 – einschließlich 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, 6. May 2024
Mahta Mayahinia

Abstract

Utilization of advanced memory technologies and paradigms improves the performance and energy efficiency of the memory cores. The emergence of new technologies such as Non-volatile Resistive Memories (NVM) and advanced-node (sub-10 nm) transistor technologies, besides the paradigm of Computation in Memory (CiM) are among the successful attempts in enhancing the memory core.

Nevertheless, the true performance and energy efficiency of the advanced memory technologies and paradigms cannot be unleashed unless by ensuring their functional reliability. Further, these memory technologies and paradigms have certain reliability challenges that require detailed analysis and solutions.

In the scope of this thesis, we focus on two reliability issues: *decision failure* and *Electromigration (EM)*. Due to the resistive nature of the NVM devices, they can realize the analog CiM further enhancing the performance and energy efficiency. Albeit, less robust analog computation besides the impact of the process variation on the less mature fabrication process of NVM devices leads to short-term reliability concerns known as decision failure. As a result of decision failure, the output of the Computation in Non-volatile Resistive Memories (NVM-CiM) can be sensed incorrectly.

EM is another reliability concern that we investigate in this thesis. EM is an aging phenomenon affecting the metal interconnects and results in the increase of their parasitic resistance. The impact of EM on the power lines has been investigated thoroughly in the existing literature, however, its effect on the memory signal lines requires greater attention. Specifically, since all types of memory technologies are affected by the EM. In the advanced technology nodes, the higher operational temperature and current densities further exacerbate the EM, degrading the latency and hence, the energy of the memory cores.

Our final target in this thesis is to provide effective reliability-improving solutions for decision failure and EM. Our key approach to achieving this goal is *cross-level* co-optimization. There is a strong correlation between the so-called system abstraction levels i.e., technology, circuit, architecture, and application. By pursuing such a cross-level approach, the reliability can be improved without inducing a vast overhead on the other system merits such as energy consumption and performance. In this direction, our contributions center around analyzing, designing, and testing for different memory technologies including NVM and Static RAM (SRAM).

Zusammenfassung

Der Einsatz fortschrittlicher Speichertechnologien und -paradigmen verbessert die Leistung und Energieeffizienz von Speicherkernen. Das Aufkommen neuer Technologien wie nichtflüchtiger Widerstandsspeicher (NVM) und fortschrittlicher Knotentransistortechnologien (unter-10 nm) sowie das Paradigma der Computation in Memory (CiM) gehören zu den erfolgreichen Ansätzen, den Speicherkern zu verbessern.

Die tatsächliche Leistung und Energieeffizienz der fortschrittlichen Speichertechnologien und -paradigmen kann jedoch nur dann entfaltet werden, wenn ihre Funktionssicherheit gewährleistet ist. Darüber hinaus haben diese Speichertechnologien und -paradigmen bestimmte Zuverlässigkeitsherausforderungen, die detaillierte Analysen und Lösungen erfordern.

Im Rahmen dieser Arbeit konzentrieren wir uns auf zwei Zuverlässigkeitsprobleme: Entscheidungsversagen und Elektromigration (EM). Aufgrund der Widerstandsbeschaffenheit der NVM-Geräte können sie das analoge CiM realisieren und so die Leistung und Energieeffizienz weiter verbessern. Allerdings führt die weniger robuste analoge Berechnung neben den Auswirkungen der Prozessvariation auf den weniger ausgereiften Herstellungsprozess von NVM-Geräten zu kurzfristigen Zuverlässigkeitsproblemen, die als Entscheidungsfehler bezeichnet werden. Aufgrund eines Entscheidungsfehlers kann die Ausgabe der Berechnung in nichtflüchtigen Widerstandsspeichern (NVM-CiM) falsch erfasst werden.

EM ist ein weiteres Problem der Zuverlässigkeit, das wir in dieser Arbeit untersuchen. EM ist ein Alterungsphänomen das sich auf die Metallverbindungen auswirkt und zu einem Anstieg ihres parasitären Widerstands führt. Die Auswirkungen von EM auf die Stromversorgungsleitungen wurde in der vorhandenen Literatur gründlich untersucht, jedoch erfordert die Auswirkung auf die Signalleitungen der Speicher größere Aufmerksamkeit. Insbesondere, da alle Arten von Speichertechnologien von EM betroffen sind. In den fortgeschrittenen Technologieknoten verschlimmern die höhere Betriebstemperatur und Stromdichte die EM noch weiter und verschlechtern die Latenzzeit und damit die Energie der Speicherkerne.

Unser letztes Ziel in dieser Arbeit ist die Bereitstellung von Lösungen zur wirksamen Verbesserung der Zuverlässigkeit gegen Entscheidungsversagen und EM. Unser zentraler Ansatz zum Erreichen dieses Ziels ist die ebenenübergreifende Co-Optimierung. Es besteht ein starker Zusammenhang zwischen den sogenannten Systemabstraktionsebenen, also Technologie, Schaltung, Architektur und Anwendung. Durch die Verfolgung eines solchen ebenenübergreifenden Ansatzes kann die Zuverlässigkeit verbessert werden, ohne dass ein großer Mehraufwand für andere Systemvorteile wie Energieverbrauch und Leistung entsteht. In dieser Richtung konzentrieren sich unsere Beiträge auf die Analyse, das Design und das Testen verschiedener Speichertechnologien, einschließlich NVM und Static RAM (SRAM).

Contents

Abstract . . . iii

Zusammenfassung . . . v

List of own publications included in this thesis . . . xi

List of other publications not included in this thesis . . . xiii

I. Preliminaries 1

1. Introduction . . . 3
- 1.1. Contributions and Research Directions . . . 3
  - 1.1.1. Analysis and mitigation of the decision failure in NVM with standard and computational functionality . . . 4
  - 1.1.2. Analysis and mitigation of EM in the memory signal lines . . . 4
- 1.2. Dissertation Outline . . . 5

2. Background . . . 7
- 2.1. Static RAM (SRAM) . . . 7
- 2.2. Non-volatile Resistive Memories (NVM) . . . 7
  - 2.2.1. Spin-Transfer Torque Magnetic RAM (STT-MRAM) . . . 8
  - 2.2.2. Redox-based RAM (ReRAM) . . . 8
  - 2.2.3. Phase Change Memory (PCM) . . . 8
  - 2.2.4. Ferroelectric Field-Effect Transistor (FeFET) . . . 9
- 2.3. Computation in Memory (CiM) . . . 9
  - 2.3.1. Computation in Non-volatile Resistive Memories (NVM-CiM) . . . 9
- 2.4. Different flavors of NVM-CiM . . . 10
  - 2.4.1. Boolean operations . . . 10
  - 2.4.2. Multiply-Accumulate (MAC) . . . 11
  - 2.4.3. (Binary) pattern similarity measurement . . . 11
- 2.5. Decision failure . . . 12
- 2.6. Electromigration (EM) . . . 13
  - 2.6.1. EM concept . . . 13
  - 2.6.2. EM modeling . . . 14

II. Contributions 17

3. Decision Failure Modeling and Mitigation through Reference Adjustment . . . 19
- 3.1. preliminaries . . . 20
  - 3.1.1. Magnetic Tunnel Junction (MTJ) variability and their impacts on CiM operations 20
  - 3.1.2. Related work . . . 20

3.2. Analyzing read decision failure probability 22
3.2.1. Read Decision Failure (RDF) in conventional STT-read 22
3.2.2. RDF in STT-CiM 23
3.2.3. Modeling the cell variability 23
3.3. Optimized reference design to minimize RDF 25
3.3.1. Optimizing reference resistance based on (R) cell modeling 25
3.3.2. Optimizing reference resistance based on (Tr+R) cell modeling 27
3.3.3. Improving the read reliability in STT-CiM 27
3.4. Results and discussion 28
3.4.1. Experimental Setup 29
3.4.2. RDF results for (R) cell modeling 29
3.4.3. Analysis and reference optimization based on (Tr+R) model 31
3.5. Conclusion 33

4. Decision Failure Modeling and Mitigation through Voltage Scaling 35
4.1. Preliminaries 35
4.1.1. Implememnting the comparator required by scouting logic 35
4.1.2. Temperature and voltage dependency of the resistive states in memristive devices 36
4.1.3. Related work 36
4.2. Voltage tuning for reliable memristive operations 38
4.2.1. Main Idea 38
4.2.2. RDF probability with voltage tuning 39
4.2.3. Impact of the temperature and (VDD, V_{WL}) on the entire memory array 39
4.2.4. Time-dependent resistance drift in ReRAM technology 39
4.3. Results and discussion 41
4.3.1. Experimental setup 41
4.3.2. RDF probability, power and performance of the memristive-based operations . 42
4.4. Conclusion 43

5. Reliability Analysis and Mitigation for Analog Computation-in-memory: from Technology to Application 45
5.0.1. Related works 45
5.1. Proposed reliability analysis framework 46
5.1.1. Obtaining the hardware error 47
5.1.2. Obtaining the software error 48
5.2. Results and Discussion 49
5.2.1. Technology-level analysis 49
5.2.2. Circuit-level analysis 50
5.2.3. Architecture- and Application-level analysis 50
5.2.4. Mitigation of the overall decision failure 53
5.3. Conclusion 54

6. Algorithm to Technology Co-Optimization for CiM-based Hyperdimensional Computing 55
6.1. Preliminary concept 56
6.1.1. Related work 56
6.2. Our proposed methodology 57
6.2.1. Design solutions at the algorithmic level 57
6.2.2. Design solutions at the hardware level 57
6.2.3. Algorithm to technology co-optimization 58

6.3. Results and discussion . . . 59
6.3.1. Hardware-level analysis . . . 59
6.3.2. Algorithm-level analysis . . . 60
6.4. Conclusion . . . 62

7. Time-dependent Electromigration Modeling for Workload-aware Design Space Exploration in STT-MRAM . . . 65
7.1. Related work . . . 65
7.2. Proposed methodology . . . 66
7.2.1. EM-induced Mean Time To Failure (MTTF) modeling under a time-variant current density . . . 66
7.2.2. Applying the EM modeling under a time-variant current density to a workload 66
7.3. Results and discussion . . . 68
7.3.1. Experimental setup . . . 68
7.3.2. Model fitting . . . 69
7.3.3. Workload-aware EM analysis . . . 69
7.3.4. Design space exploration . . . 70
7.4. Conclusion . . . 72

8. Analyzing the Electromigration Challenges of Computation in Resistive Memories . 73
8.1. Preliminaries . . . 73
8.2. Selecting the appropriate EM modeling . . . 73
8.2.1. Related work . . . 74
8.3. Proposed EM-analysis method for CiM . . . 74
8.3.1. Modifying the EM modeling for CiM . . . 74
8.3.2. Interconnect dimensions . . . 74
8.4. Results and discussion . . . 75
8.4.1. Simulation tool and setup . . . 75
8.4.2. EM analysis for CiM . . . 77
8.5. Mitigation of EM degradation for CiM . . . 79
8.6. Conclusion and future work . . . 79

9. An Efficient Test Strategy for Detection of Electromigration Impact in Advanced FinFET Memories . . . 81
9.1. Background and related work . . . 82
9.1.1. Advanced transistor technology . . . 82
9.1.2. Related work . . . 82
9.2. Proposed methodology . . . 82
9.2.1. Advanced Inductive Failure Analysis (AIFA) . . . 82
9.2.2. Degraded SRAM writability as a means to EM test . . . 83
9.3. Results and discussion . . . 85
9.4. Simulation setup . . . 85
9.4.1. Test sequence generation for the EM . . . 87
9.4.2. Applying the proposed EM test methodology on the SRAM array . . . 87
9.5. Conclusion . . . 88

10. Electromigration-aware Design Technology Co-Optimization for SRAM in Advanced Technology Nodes . . . 89
10.1. Related work . . . 89

10.2. EM-aware Design Technology Co-Optimization (DTCO) methodology 90
10.2.1. SRAM sub-array size 91
10.2.2. EM reliability analysis 91
10.3. Results and Discussion 93
10.3.1. Required periphery circuit for the SRAM sub-array 96
10.3.2. Impact of the access dependency on the current passing through the Word-Line (WL) and Bit-Line (BL) 96
10.3.3. Impact of the access dependency on the EM reliability 97
10.3.4. EM-aware DTCO analysis 97
10.4. Conclusion 97

11. Conclusion and Perspectives 99
11.1. Conclusion 99
11.2. Future Perspective 99

III. Appendix 101

Bibliography 103

List of Figures 113

List of Tables 117

Acronyms 119

List of own publications included in this thesis

[95] M. Mayahinia, C. Münch, and M. B. Tahoori, “Analyzing and mitigating sensing failures in spintronic-based computing in memory”, in *IEEE International Test Conference (ITC)*, 2021, pp. 268–277.

[109] M. Mayahinia, A. Jafari, and M. B. Tahoori, “Voltage tuning for reliable computation in emerging resistive memories”, in *IEEE 40th VLSI Test Symposium (VTS)*, 2022, pp. 1–7.

[110] M. Mayahinia, M. Tahoori, G. Harutyunyan, G. Tshagharyan, and K. Amirkhanyan, “An efficient test strategy for detection of electromigration impact in advanced finfet memories”, in *IEEE International Test Conference (ITC)*, 2022, pp. 650–655.

[111] M. Mayahinia, M. Tahoori, M. P. Komalan, *et al.*, “Time-dependent electromigration modeling for workload-aware design-space exploration in stt-mram”, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 41, no. 12, pp. 5327–5332, 2022.

[112] M. Mayahinia, M. Tahoori, M. Perumkunnil, K. Croes, and F. Catthoor, “Analyzing the electromigration challenges of computation in resistive memories”, in *2022 IEEE International Test Conference (ITC)*, IEEE, 2022, pp. 534–538.

[128] M. Mayahinia, H.-H. Liu, S. Mishra, Z. Tokei, F. Catthoor, and M. Tahoori, “Electromigration-aware design technology co-optimization for sram in advanced technology nodes”, in *2023 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, IEEE, 2023, pp. 1–6.

[139] M. Mayahinia, H. G. Hezayyin, and M. Tahoori, “Reliability analysis and mitigation for analog computation-in-memory: From technology to application”, in *IEEE 42nd VLSI Test Symposium (VTS)*, 2024.

[140] M. Mayahinia, S. Thomann, P. Genssler, C. Münch, H. Amrouch, and M. Tahoori, “Algorithm to technology co-optimization for cim-based hyperdimensional computing”, in *2024 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2024.

List of other publications not included in this thesis

[96] S. M. Nair, M. Mayahinia, M. B. Tahoori, *et al.*, "Workload-aware electromigration analysis in emerging spintronic memory arrays", *IEEE Transactions on Device and Materials Reliability*, vol. 21, no. 2, pp. 258–266, 2021.

[97] S. T. Ahmed, M. Mayahinia, M. Hefenbrock, C. Münch, and M. B. Tahoori, "Process and runtime variation robustness for spintronic-based neuromorphic fabric", in *2022 IEEE European Test Symposium (ETS)*, IEEE, 2022, pp. 1–2.

[99] L. Brackmann, A. Jafari, C. Bengel, *et al.*, "A failure analysis framework of reram in-memory logic operations", in *2022 IEEE International Test Conference in Asia (ITC-Asia)*, IEEE, 2022, pp. 67–72.

[103] S. Hemaram, M. Mayahinia, and M. B. Tahoori, "Adaptive block error correction for memristive crossbars", in *2022 IEEE 28th International Symposium on On-Line Testing and Robust System Design (IOLTS)*, IEEE, 2022, pp. 1–6.

[104] A. Jafari, M. Mayahinia, S. T. Ahmed, C. Münch, and M. B. Tahoori, "Mvstt: A multi-value computation-in-memory based on spin-transfer torque memories", in *2022 25th Euromicro Conference on Digital System Design (DSD)*, 2022, pp. 332–339.

[106] J. Krautter, M. Mayahinia, D. R. Gnad, and M. B. Tahoori, "Data leakage through self-terminated write schemes in memristive caches", in *2022 27th Asia and South Pacific Design Automation Conference (ASP-DAC)*, IEEE, 2022, pp. 666–671.

[113] Z. Pei, M. Mayahinia, H.-H. Liu, *et al.*, "Graphene-based interconnect exploration for large sram caches for ultrascaled technology nodes", *IEEE Transactions on Electron Devices*, vol. 70, no. 1, pp. 230–238, 2022.

[117] D. Wouters, L. Brackmann, A. Jafari, *et al.*, "Reliability of computing-in-memory concepts based on memristive arrays", in *2022 International Electron Devices Meeting (IEDM)*, IEEE, 2022, pp. 5–3.

[119] S. T. Ahmed, M. Mayahinia, M. Hefenbrock, C. Münch, and M. B. Tahoori, "Design-time reference current generation for robust spintronic-based neuromorphic architecture", *ACM Journal on Emerging Technologies in Computing Systems*, vol. 20, no. 1, pp. 1–20, 2023.

[123] S. Hemaram, S. T. Ahmed, M. Mayahinia, C. Münch, and M. B. Tahoori, "A low overhead checksum technique for error correction in memristive crossbar for deep learning applications", in *2023 IEEE 41st VLSI Test Symposium (VTS)*, IEEE, 2023, pp. 1–7.

[124] J. Henkel, L. Siddhu, L. Bauer, *et al.*, "Special session-non-volatile memories: Challenges and opportunities for embedded system architectures with focus on machine learning applications", in *Proceedings of the International Conference on Compilers, Architecture, and Synthesis for Embedded Systems*, 2023, pp. 11–20.

[129] M. Mayahinia, M. Tahoori, G. Tshagharyan, G. Harutyunyan, and Y. Zorian, "On-chip electromigration sensor for silicon lifecycle management of nanoscale vlsi", in *2023 IEEE European Test Symposium (ETS)*, IEEE, 2023, pp. 1–4.

[131] Z. Pei, M. Mayahinia, H.-H. Liu, *et al.*, "Emerging interconnect exploration for sram application using nonconventional h-tree and center-pin access", in *2023 24th International Symposium on Quality Electronic Design (ISQED)*, IEEE, 2023, pp. 1–1.

[132] Z. Pei, M. Mayahinia, H.-H. Liu, *et al.*, "Technology/memory co-design and co-optimization using e-tree interconnect", in *Proceedings of the Great Lakes Symposium on VLSI 2023*, 2023, pp. 159–162.

[133] V. Rietz, C. Münch, M. Mayahinia, and M. Tahoori, "Timing-accurate simulation framework for nvm-based compute-in-memory architecture exploration", *it-Information Technology*, vol. 65, no. 1-2, pp. 13–29, 2023.

[134] B. Sapui, J. Krautter, M. Mayahinia, *et al.*, "Power side-channel attacks and countermeasures on computation-in-memory architectures and technologies", in *2023 IEEE European Test Symposium (ETS)*, IEEE, 2023, pp. 1–6.

[136] H. Farzaneh, J. P. C. de Lima, A. Nezhadi Khelejani, *et al.*, "Sherlock: Scheduling efficient and reliable bulk bitwise operations in nvms", in *ACM 61st Design Automation Conference (DAC)*, 2024.

[138] P. R. Genssler, M. Mayahinia, S. Thomann, M. Tahoori, and H. Amrouch, "DropHD: Technology/algorithm co-design for reliable energy-efficient nvm-based hyperdimensional computing under voltage scaling", in *IEEE DATE*, 2024.

[142] Z. Zhang, M. Mayahinia, C. Weis, *et al.*, "Addressing the combined effect of transistor and interconnect aging in sram towards silicon lifecycle management", in *IEEE 42nd VLSI Test Symposium (VTS)*, 2024.

Part I.

Preliminaries

1. Introduction

A conventional computer architecture consists of two separate cores: *processing* and *memory*. However, mainly to mitigate the *memory wall problem*, a paradigm shift has occurred that aims to enhance the memory to also perform certain computation tasks, i.e., CiM. Therefore, memory, either with standard or computational functionality, has a pivotal impact on the overall performance, energy efficiency, and last but not least, reliability of the computer system.

Moore's law-driven reduction of the physical sizes of the Front-End-of-the-Line (FEoL) and Back-End-of-the-Line (BEoL) devices to achieve higher integration density, as well as the emergence of the Non-volatile Resistive Memories (NVM) have considerable impacts on the memory core. Shrinking the size of FEoL devices improves the performance and energy efficiency, however, shrinking the size of BEoL interconnect increases its parasitic resistance which results in a dominant effect of the BEoL interconnect on the performance, energy efficiency, and reliability of the computer system.

Besides, utilizing the NVM not only eliminates static power consumption but also enables analog CiM due to their resistive nature. Despite these advantages, an immature fabrication process of the NVM makes them prone to process variation which can result in an unreliable functionality. Therefore, advanced (sub-10 nm) memory technologies and paradigms come along with certain reliability challenges that need to be properly addressed. Otherwise, such reliability challenges can even diminish their performance and energy efficiency benefits.

A holistic view of the memory cores reveals strong interactions between different system abstraction levels from the high-level applications all the way down to the device technology. Therefore, a reliability improvement mechanism cannot be achieved unless by co-optimization of the so-called abstraction levels. The lack of such multi-level co-optimization is the main gap that this thesis aims to address. With such a holistic cross-level approach it can be assured that improving the reliability of the advanced memory technologies and paradigms can be achieved without an unacceptable vast overhead on their other design merits, such as energy efficiency and performance.

1.1. Contributions and Research Directions

In the broad range of reliability investigation, this thesis aims to focus on short-term decision failure and long-term Electromigration (EM) issues and provide effective solutions to mitigate them. Decision failure originates from the effect of the process variation on the FEoL and BEoL devices. Due to the decision failure, the content of the memory or the output of the CiM unit can be sensed incorrectly.

EM affects the BEoL interconnect and originates from the momentum transfer of the current-carrying electrons to the metal atoms. Such momentum transfer can force the metal atoms to migrate from their initial location and result in void nucleation. A sufficiently large void finally increases the resistance of the interconnect, which results in in-field failures.

In the scope of this thesis, firstly, we aim to accurately model and analyze the reliability of the memory cores working either as standard or computational memory. Secondly, we provide efficient reliability improvement methodologies in different abstraction levels and quantify their trade-offs on the overall computer system.

In this thesis, for standard memory operations, we take into account both the charge-based SRAM and the NVM. However, for the CiM use case, we only focus on the more energy-efficient analog type in NVM.

1.1.1. Analysis and mitigation of the decision failure in NVM with standard and computational functionality

In the case of NVM-CiM, we consider the realization of Boolean operation, Multiply-Accumulate (MAC), and similarity computation (using the Content Addressable Memory (CAM) structure). In the standard memory operation, only one but in the case of CiM, multiple memory rows need to be activated simultaneously. To perform the standard memory read as well as Boolean operation, we consider the concept of *scouting logic* [68]. In this concept, a voltage is applied to the crossbar organization of NVM, and the output current (or voltage) is compared with a known reference. The result of this comparison determines the output of the standard read or CiM. Due to the effect of the process variation on both the FEoL and BEoL devices, the output current (or voltage) does not have a fixed value and follows a statistical distribution. Hence, depending on the reference value, the output currents at the tails of distributions can sensed incorrectly. This reliability scenario is known as *decision failure*.

One of the impactful parameters on decision failure is the value of the reference. Hence, we propose a methodology to find the value of the reference to minimize the probability of decision failure. Moreover, the resistive distributions of the NVM devices (in this work, STT-MRAM) show temperature dependency, so, we also augment the optimized reference with temperature-aware self-adjustment capability. The details of this work have been published in [95] and included in Chapter 3 of this thesis.

Due to the voltage dependency of the resistive distributions of the NVM, we leverage the voltage tuning to mitigate the decision failure during the read and CiM functionalities. However, voltage tuning affects the power and latency of the scouting-based CiM circuitry. Therefore, we also investigate the Pareto optimization of the power and latency of the entire standard and CiM unit. STT-MRAM and ReRAM are the target technologies in this work. The publication version of this work is [109], and in the outline of this thesis, its details are presented in Chapter 4.

In both of these works (Chapters 3, 4) we mostly target the technology to circuit correlation. However, the algorithm that is computed in NVM-CiM units considerably affects the overall decision failure. To perform a holistic end-to-end decision failure analysis, we execute Boolean- and MAC-oriented workloads on the *gem5* infrastructure enabled with the capability of NVM-CiM. The two NVM technologies in this work are STT-MRAM and ReRAM. Moreover, we also take into account the effect of the BEoL interconnect and its process variation on the overall decision failure probability during Boolean and MAC operations. This work has also been published in [139] and discussed in Chapter 5.

For another end-to-end investigation, we focus on reliable improvement of the performance and energy efficiency of Hyperdimenssional Computing (HDC) inference through NVM-CiM using CAM structure. The main inference task of HDC is similarity computation and our methodology in this work is cross-level co-optimization of the various system abstraction levels namely technology, circuit, architecture, and algorithm. The more elaborated version of this work is available in publication [140] and Chapter 6.

1.1.2. Analysis and mitigation of EM in the memory signal lines

In the EM analysis, we show that in the advanced BEoL interconnect that carries higher current density, even the memory signal lines (WL and BL) are susceptible to EM [96]. An important insight of our analysis shows that the EM profile is highly affected by the memory access pattern i.e., the workload. Therefore, we extend our workload-aware EM analysis to effectively explore the design space. This work has been detailed in publication [111] as well as Chapter 7 of this thesis.

We also perform an investigation of EM in MAC-based NVM-CiM fabrics and show the exacerbation of its EM risk. By considering various NVM technologies (STT-MRAM, ReRAM, and PCM), we show a strong dependency of the EM profile on not only the number but also the position of the activated rows. Furthermore, we also leverage this position dependency to mitigate the risk of EM with a low-cost yet effective solution. Publication [112] and Chapter 8 of this thesis contain the details of this work.

By having a detailed analysis and modeling provided in [96] and Chapter 8, we aim to mitigate EM through various means. In the publication [110] and Chapter 9, we introduce a low-cost yet effective testing solution to detect the EM issue in the SRAM technology. Early detection of the EM before it results in catastrophic in-field failures is the main goal of this work and we achieve it by carefully tuning the testing condition.

Finally, improving the resiliency of the SRAM design toward EM is another topic that we investigate. At first, we show the exacerbation of EM in advanced VLSI nodes compared to older technology nodes. Further, we utilize the widely acknowledged DTCO methodology to achieve an optimized trade-off between the performance, energy efficiency, and EM-resilience of the SRAM module designed in advanced technology nodes. This work is further elaborated in publication [128] and Chapter 10.

1.2. Dissertation Outline

This thesis is organized into the chapters as follows:

- Chapter 2 reviews the essential background information required for this thesis.
- Chapter 3 discusses the effectiveness of the reference adjustment in mitigation of the decision failure in the concept of Scouting NVM-CiM.
- Chapter 4 proposes voltage tuning as a means for decreasing the decision failure and further discusses its impact on the overall efficiency of the NVM-CiM fabrics.
- Chapter 5 targets the decision failure in NVM-CiM and consider a technology-to-algorithm approach in this regard.
- Chapter 6 further shows the requirement of cross-level co-optimization by considering the HDC as the main application.
- Chapter 7 focuses on the EM modeling trade-offs in the case of memory signal lines.
- Chapter 8 raises an attention toward the exacerbated EM issue in the case of NVM-CiM-oriented MAC operation.
- Chapter 9 centers around an effective testing solutions for the effect of EM in the case of SRAM's BL.
- Chapter 10 is the work showing the necessity of EM-aware DTCO mechanism for SRAM.
- Chapter 11 concludes this thesis and points out the potential future works.

2. Background

2.1. Static RAM (SRAM)

Figure 2.1 shows the SRAM array with the required periphery circuits. As shown by Figure 2.1 (b), core storage in the SRAM technology is a cross-coupled inverters. Due to the positive feedback formed by the cross-coupled inverters, the binary date can be retained as long as the voltage source is applied. Moreover, accessing/manipulating (i.e., read and write) the content of the SRAM is possible by activating the access transistors through the WL. Further, the BL and Bit-Line-bar (BLB) are connected to the data (Q) and its complementary (Q-bar).

As shown by Figure 2.1 (b), the realistic model for the WL, and BL (BLB) interconnects consists of the resistive and capacitive (RC) parasitic. To precisely model the parasitic effect of the interconnect, an RC element can be considered for segments of the wire, corresponding to each SRAM cell. In the advanced sub-10 nm CMOS technology, the threshold voltage of the transistor elements is decreasing, resulting in higher performance and energy efficiency. However, due to the smaller interconnect cross-section of the fabricated interconnects, their resistance parasitic is relatively high. Such high resistance parasitic degrades the hold, read, and write noise margin of the SRAM. Amongst them write noise margin is degraded more severely. Therefore, the SRAM write operation is challenging in the advanced *resistance-dominated* interconnect, since the voltage of the write drivers is considerably lower due to the high resistance parasitic [107]. Moreover, as shown in Figure 2.1 (a) the address decoder, read and write circuitries, pre-recharger, as well as timing control module are the required periphery circuits for the proper functionality of the SRAM-based memory module.

2.2. Non-volatile Resistive Memories (NVM)

NVMs are devices with programmable resistance states that can provide at least two distinct resistive states namely LRS and HRS, respectively. There are multiple NVM technologies and among them, Spin-Transfer Torque Magnetic RAM (STT-MRAM), Redox-based RAM (ReRAM), Phase Change

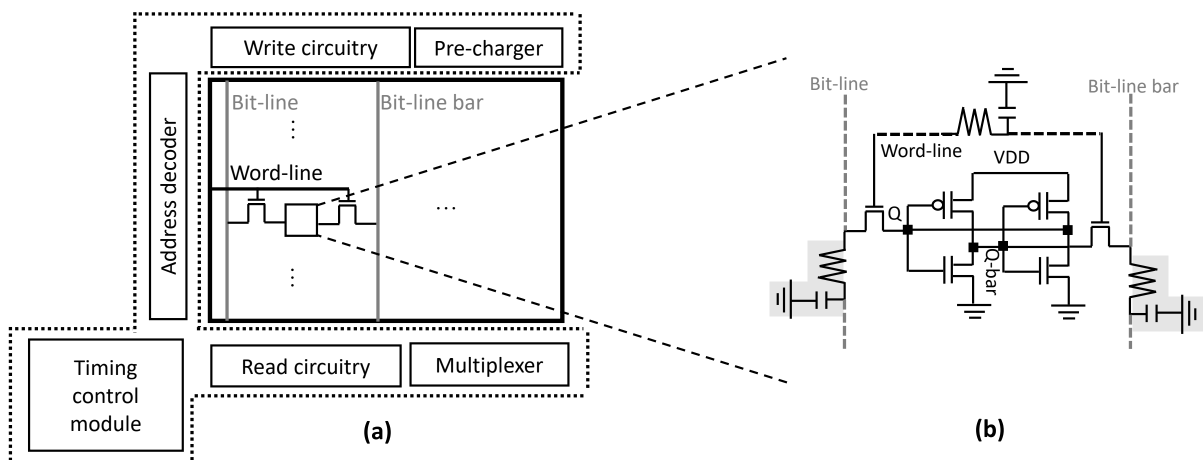


Figure 2.1.: (a) SRAM array with the required periphery circuit, (b) a 6-transistor SRAM cell.

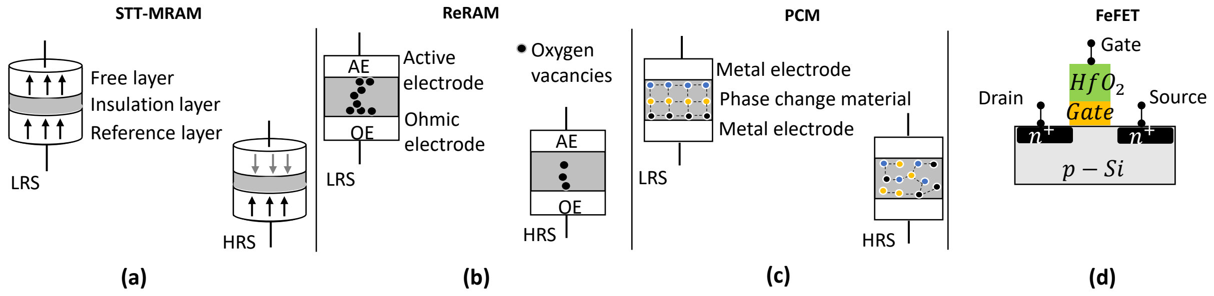


Figure 2.2.: NVM technologies, Low Resistive State (LRS) on top and High Resistive State (HRS) on the bottom, (a) STT-MRAM, (b) ReRAM, (c) PCM, and d) Ferroelectric Field-Effect Transistor (FeFET)

Memory (PCM), and Ferroelectric Field-Effect Transistor (FeFET) are those that utilized in the scope of this thesis. We discuss briefly the properties of the four technologies and their basic physics are also shown in Figure 2.2).

2.2.1. STT-MRAM

The main storage element of the STT-MRAM is the MTJ which consists of three layers (Figure 2.2 (a)). Two ferromagnetic layers sandwich an insulation layer. The magnetic orientation of the *free* layer can be rotated, while the magnetic orientation of the *pinned* layer is fixed. In the MTJ, the relative magnetic orientation of the free and pinned layers determines the resistive state. In the case of parallel ('P') magnetic orientation, the device is in the LRS, and otherwise, in the case of antiparallel ('AP') magnetic orientation, the device is in the HRS. To perform the write operation, a current needs to pass through the MTJ device. Write current in the direction of the free-to-fixed layer aligns the magnetic orientation of the free and pinned layers parallel, i.e., programs the device into the LRS. Whereas the opposite current direction programs the device into the HRS [38], [81]. The works presented in [74], [93] are two examples of industrially fabricated STT-MRAM chips by TSMC and Samsung, respectively.

2.2.2. ReRAM

As shown in Figure 2.2 (b), the ReRAM cell is also a stack of three layers. Two metallic (Ohmic and Active) electrodes sandwich an oxide-based insulation layer. Within the oxide-based insulation layer, the oxygen vacancies (V_O^{2+}) can form a conducting filament connecting the metal electrodes. If the conducting filament exists in the insulation layer, the device is in LRS, and otherwise, it is in HRS. To perform the write operation, a voltage needs to be applied to theReRAM device. Since the work functions of the Ohmic and Active electrodes are not equal, a certain voltage polarity shapes the filament (SET) and programs the device in LRS. Further, reversing the voltage polarity raptures the filament (RESET) and programs the device in HRS [71], [87]. ReRAM technology has also been successfully demonstrated by TSMC [32].

2.2.3. PCM

Similarly, as shown in Figure 2.2 (c), PCM also consists of three layers. Two metallic electrodes and, in between, a phase change material. The phase change material can be either in *crystalline* or *amorphous* phases which results in the LRS and HRS, respectively. To perform the write operation, an electric pulse needs to be applied to the PCM device. This electric pulse can provide the required Joule heating for changing the phase of the material. By applying a long-width, low-amplitude pulse, the phase change material can be heated below the melting temperature, which is appropriate for the amorphous-to-crystalline phase change; i.e., HRS to LRS switching. For the reverse switching, the phase

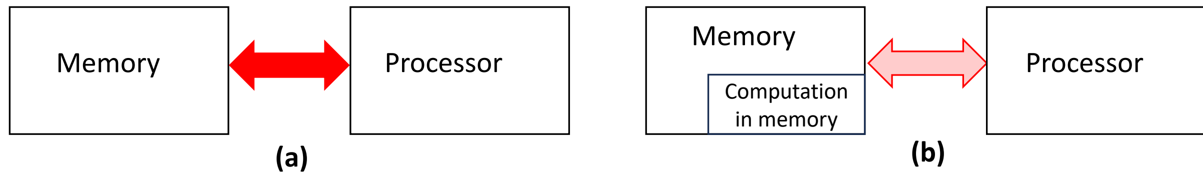


Figure 2.3.: (a) Conventional memory architecture pressing the memory wall problem, (b) mitigation of the memory wall problem by enabling the CiM

change material needs to be heated above the melting temperature, and then immediately cooled down. To fulfill this action, a short-width, high-amplitude electric pulse is required [42], [79]. IBM and Macronix have already shown successful PCM fabrication in [101].

2.2.4. Ferroelectric Field-Effect Transistor (FeFET)

As shown in Figure 2.2 (d), FeFET can be built by introducing a thick layer (10 nm) of hafnium oxide to the gate stack of conventional Field-Effect Transistor (FET) developing itself into a promising on-chip memory over the recent years [20]. As no additional materials are needed, the fabrication process is fully CMOS compatible which is a great advantage [20], [50]. The ferroelectric dipoles in the hafnium layer can be polarized through a strong voltage bias at the gate node, which aligns their direction. This polarization modulates the threshold voltage of the FeFET, and in fact, realizes two logic states giving FeFET its NVM capabilities. The successful industrial adoptions of FeFET have been shown by Intel [70] and GlobalFoundries [130].

A series connection of an NVM cell with a transistor provides an addressable memory unit. Moreover, despite the different physical phenomena that govern the NVM technologies, to perform the read operation, their resistance needs to be sensed. Therefore, by applying a voltage to these devices and measuring the current, their resistive state can be determined. Basically, the NVM technologies are mainly different in the write procedure.

2.3. Computation in Memory (CiM)

In the conventional computer architecture paradigm, the machine cycle encompasses multiple phases including fetching data and instructions from memory, instruction decoding, execution of instructions by the processing core, and the subsequent storage of results back in memory. The back-and-forth transfer of the data between the memory and processing cores gives rise to significant performance and energy overhead, commonly referred to as *memory wall* (Figure 2.3 (a)).

In response to the ever-growing demands of emerging big-data applications for a larger amount of data, higher performance, and energy efficiency, a paradigm shift has been taking place. This shift is directing computer architecture toward CiM [59], in which the memory cores are enhanced to perform certain computational tasks, mitigating the memory wall problem(Figure 2.3 (b)).

2.3.1. Computation in Non-volatile Resistive Memories (NVM-CiM)

CiM can be realized in various memory technologies including the traditional charge-based technologies, such as SRAM [55] and Dynamic RAM (DRAM) [39]. However, these memory technologies possess inherent volatility and static energy consumption. NVMs, however, are an alternative to charge-based memories. In terms of energy efficiency, NVMs do not require static energy to maintain stored data. Additionally, the resistive nature of NVMs aligns well with the concept of analog CiM, further improving

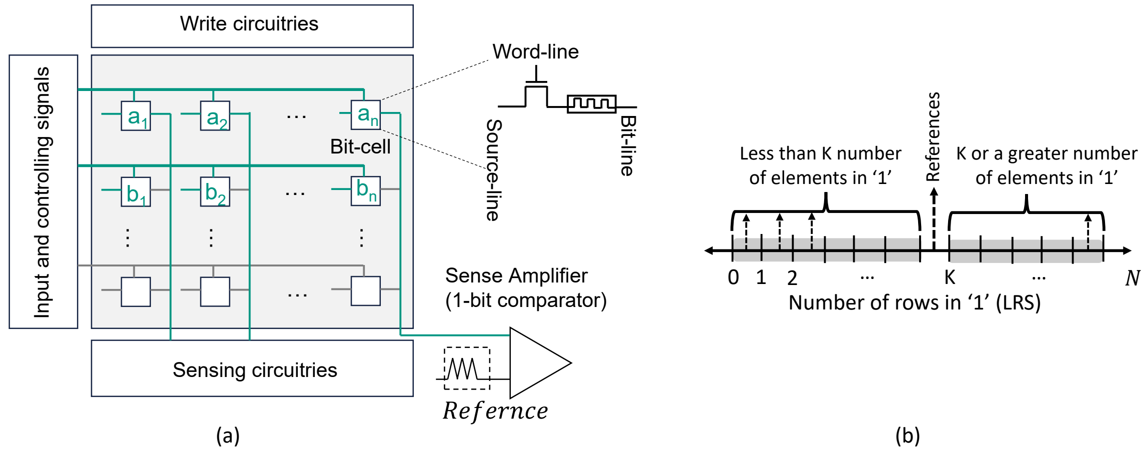


Figure 2.4.: a) Performing the Boolean operation using the concept of NVM-CiM on two binary vectors $\{a_1, a_2, ..., a_n\}$, $\{b_1, b_2, ..., b_n\}$, b) Realizing the threshold Boolean operation by adjusting the reference

the efficiency of the CiM. Due to these benefits, NVM-CiM has received great interest from the frontier VLSI industry [91], [92], [105], [121], [125], [135], [141].

2.4. Different flavors of NVM-CiM

NVM-CiM can be utilized to perform various functionalities including Boolean operations, Multiply-Accumulate (MAC), and (Binary) pattern similarity measurement. Despite the distinct differences in the implementation of these functionalities, they have the following three steps in common:

- *Step 1:* Activating multiple memory rows and/or columns at the same time.
- *Step 2:* Applying the input and controlling voltages to the memory array.
- *Step 3:* Sense the output current and transfer it into a digital corresponding, since the CiM module is a part of the digital computer.

In the next subsections, we will discuss the specifications of each NVM-CiM flavor.

2.4.1. Boolean operations

Boolean operation using NVM-CiM can be performed with the concept of the *scouting logic* [68]. In the concept of scouting logic, the analog output signal of the NVM-CiM module is compared with a reference signal, and the binary output will be determined based on the result of this comparison. Basically, this reference signal can be adjusted to realize different Boolean functions in the form of the threshold. This means the output of a threshold-K operation is '1' if and only if out of its N operands, at least K operands are '1'.

Figure 2.4 (a), shows the memory array utilized for the execution of the Boolean NVM-CiM. As the output of a Boolean operation is binary, a Sense Amplifier (SA) i.e., one-bit comparator can be used to fulfill the final digitization. As shown in Figure 2.4 (b), for a given fixed N, adjusting the reference can realize the different values of K in a threshold Boolean function. Please note that the extreme cases, i.e., $K = 1$ and $K = N$ are logical *OR* and *AND* operations, respectively.

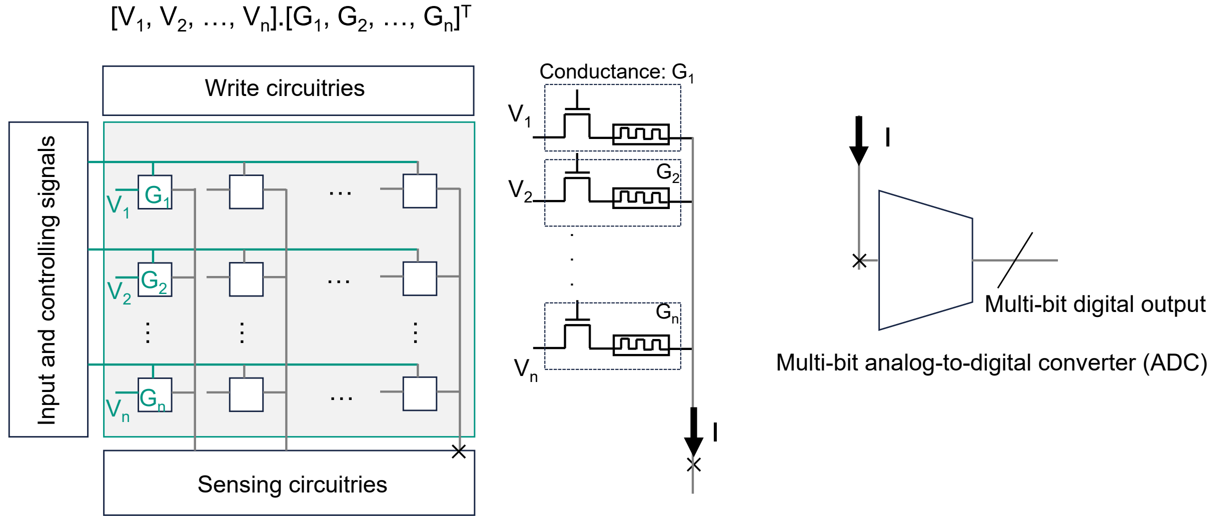


Figure 2.5.: Performing the MAC operation in the form of $[V_1, V_2, ..., V_n] \cdot [G_1, G_2, ..., G_n]^T$ using the concept of NVM-CiM

2.4.2. Multiply-Accumulate (MAC)

MAC is a widely used operation in artificial intelligence and neural network applications. This operation can be accelerated by performing in the crossbar organization of the NVM devices in an analog way. In the analog MAC operations, there are two sets of operands, the first set is applied as the voltage and the second set of inputs is stored as the conductance ($G = R^{-1}$) of the NVM devices. Theoretically, as shown in Figure 2.5, the current that passes through each NVM bit-cell is governed by Ohm's law ($I_x = V_x \cdot G_x$). Following the Krischof current law (KCL), the current passing through the common interconnect is the result of the MAC operation; $I \propto [V_1, V_2, ..., V_n] \cdot [G_1, G_2, ..., G_n]^T$. Figure 2.5 shows the concept of the analog MAC operation using the crossbar organization of the NVM devices.

To maintain the communication between this analog accelerator and the rest of the digital system, a Digital-to-Analog Converter (DAC) at the inputs and an Analog-to-Digital Converter (ADC) at the output are required. Please note that, unlike the Boolean operations, the output of the MAC operation is not binary. Hence, a multi-bit ADC is required for the final digitization.

2.4.3. (Binary) pattern similarity measurement

(Binary) pattern similarity measurement in NVM-CiM can be performed using the CAM structure, in particular, NVM-CAM. Contrary to RAM which is indexed by an address, CAM is queried by using data. Due to their high parallelism, CAM hardwares are used in high-performance applications.

For each CAM operation, there are two sets of data operands, *prototypes* (P) which are fixed, and *query* (Q) which is compared with *all* the prototypes in parallel. Since prototypes are not changing dynamically, they need to be written only once in NVM, with almost no static power consumption. Thus, NVM-CAM can be a good fit for this use case. Figure 2.6 shows the NVM-CAM which has been proposed in [36]. In this NVM-CAM structure, the original and complementary bits of the prototype patterns (P) are stored in the NVM, while the original and negated bits of the query pattern (Q) are applied as binary voltage levels to transistor gates as shown in Figure 2.6.

A binary '1' ('0') in the prototype vector is encoded as HRS (LRS), and '1' ('0') in the query vector as a high (low) voltage level. In case of a match, the already precharged match-line is discharged through the path consisting of HRS cells. In case of a mismatch, the discharge path through the LRS cells. Hence, at the time of sampling, the voltage of the match-line will be higher in the case of the match. Due to the limited HRS-LRS ratio in different NVM technologies, the length of the match-line cannot be increased

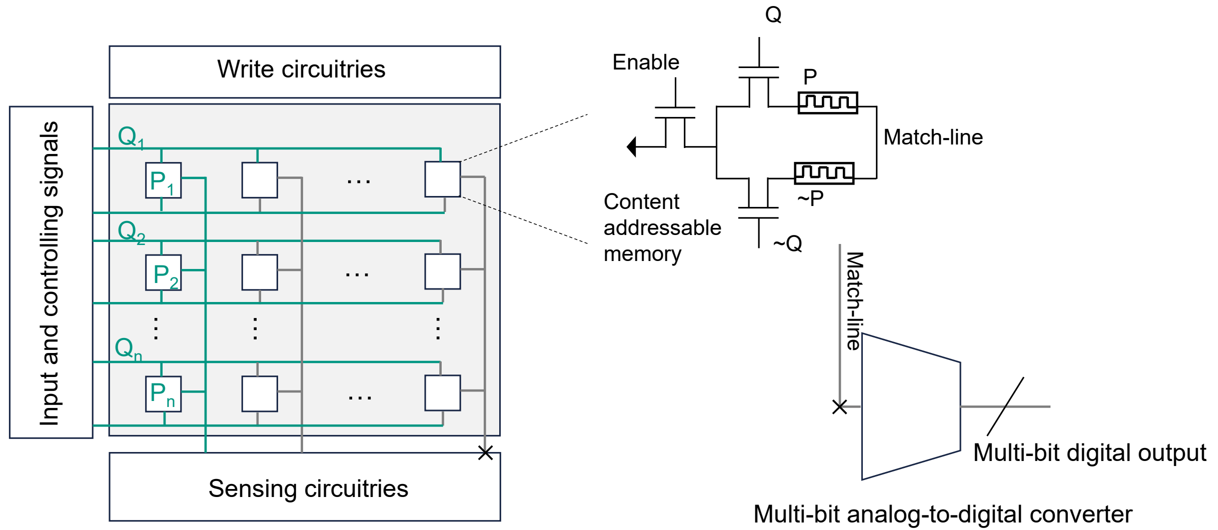


Figure 2.6.: Performing the (Binary) pattern similarity measurement using the concept of NVM-CiM, using the NVM-based CAM (NVM-CAM) structure comparing two binary vectors $\{Q_1, Q_2, ..., Q_n\}$, $\{P_1, P_2, ..., P_n\}$

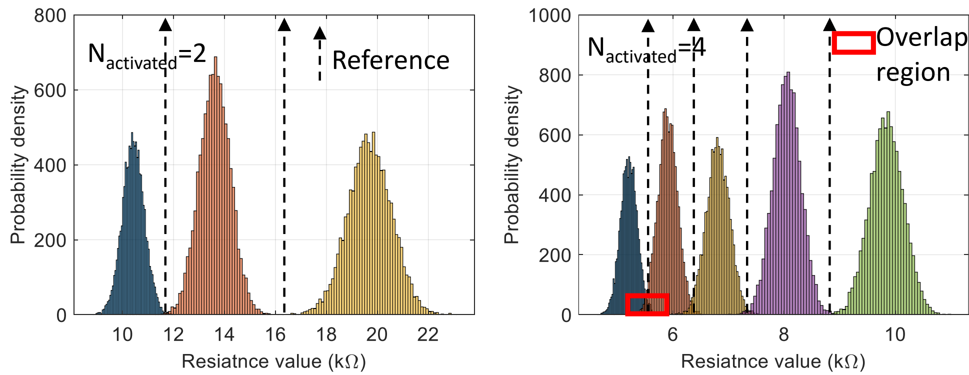


Figure 2.7.: Overlap regions originating from the impact of process variation resulting in an incorrect final output of NVM-CiM and its exacerbation in the case of increasing the number of operands

arbitrarily. Similar to the MAC operation, in the case of (Binary) pattern similarity measurement, the output i.e., the Hamming distance is a multi-bit value. Hence, a multi-bit ADC is required at the output.

2.5. Decision failure

The reliability of the NVM-CiM can be challenged mainly due to the following *technology*- and *circuit*-level factors. At the *technology level*, NVM devices are prone to process variation. Therefore, the values of the LRS and HRS for different NVM technology are not fixed and follow statistical distributions. Additionally, these distributions are not fixed and are temperature-dependent [38], [42], [71], [79], [87]. More specifically, a time-dependent resistance shift has also been confirmed for the ReRAM technology [30]. It means that the LRS and HRS distributions are time-variant. Unbalanced SET and RESET in ReRAM redistribute the oxygen vacancies. If excessive oxygen vacancies are accumulated at the switching interface, the HRS tends to move to the LRS. Conversely, depletion of the oxygen vacancies results in the gradual increase of the LRS [27]. Figure 2.7 shows the possible overlap between the supposedly distinct resistance levels, resulting in an incorrect final output. Besides, increasing the number of NVM devices participating in the computation exacerbates the issue.

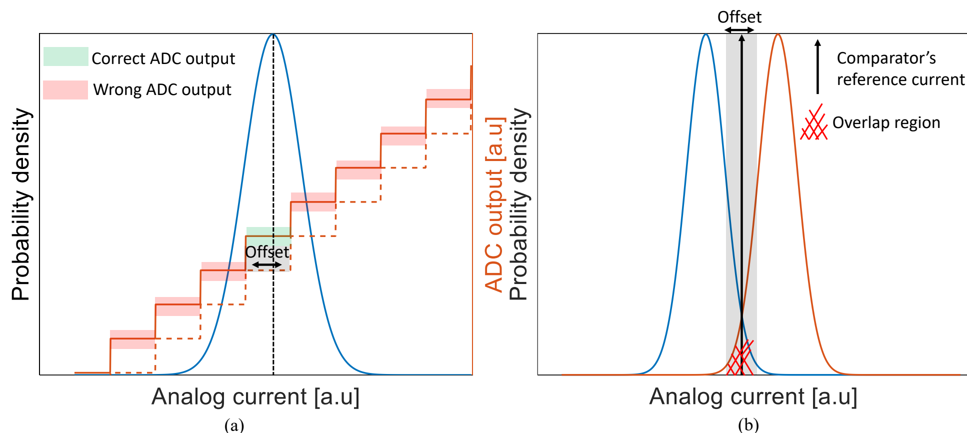


Figure 2.8.: Offset issue in a) ADC and b) one-bit comparator. contributing in the increase of decision failure

Moreover, at the *circuit level*, conversion of the analog signal correspondence is the prerequisite for all the discussed NVM-CiM flavors. Therefore, the ADC and comparator modules are crucial for the NVM-CiM. However, they are not perfect and show offset issues. Figure 2.8 (a) and (b) show the offset issue for ADC and comparator, respectively. The offset is a boundary in which the ADC or comparator cannot be reliably functional. Hence, reduction of the offset is desirable for reliability enhancement and can be achieved at the cost of higher energy and/or area.

Additionally, the interconnect network is not ideal and usually shows resistive and capacitive (RC) parasitic. Especially in the more advanced technology nodes, the reduction of the cross-sectional area increases the resistive parasitic [126]. In the NVM-CiM, this increase of the resistive parasitic interferes with the effective resistance of the NVM devices and impairs the sensing of the analog signal reaching the crossbar output. The above technology- and circuit-level factors can finally cause erroneous sensing of NVM states participating in the computation and resulting in an incorrect output of the NVM-CiM module which is known as *decision failure*.

2.6. Electromigration (EM)

Analyzing the EM as a long-term reliability issue affecting the BEoL interconnect is the other concept that we investigate in the scope of this thesis. The required background in the concept of EM will be presented in the following

2.6.1. EM concept

EM is one of the reliability concerns for the BEoL interconnects in the VLSI technology. Higher current density, as well as higher temperatures, exacerbates the EM risk. Hence, EM concern becomes more valid in the more advanced technology nodes, in which both the current density and the operating temperature are higher due to the smaller dimensions of the interconnect and higher device density and circuit complexity [6], [16].

The reason behind the EM phenomena is that in the metal interconnect, conducting electrons transfer the momentum to the metal atoms and force them to migrate. In the EM phenomena, a *void* is being formed at the specific location of the line that undergoes the highest stress. Due to the existence of the void, the resistance of the interconnect is increased and as a result, failure happens in the chip [65], [84].

VLSI interconnects can be categorized into *signal lines*, *power lines*, and *clock lines*. Signal lines are usually intra-cell and inter-cell links with the bidirectional current. Power lines, however, carry mostly unidirectional current and deliver the power supply and ground to the entire chip [85]. In the

older technology nodes, the signal lines have been assumed to carry (much) less current, besides, the bi-directionality of the current passing through the signal lines has a recovery effect for EM. Hence, most of the existing research work tried to capture the impact of the EM on the unidirectional power lines [35], [37]. Moreover, the presence of strongly periodic waveforms, such as the alternating positive and negative phases, minimizes the interference of EM signals on the clock lines. However, studying the impact of the EM in the signal lines is essential and the focus of this thesis. Since, in more advanced technology nodes, some of the tight pitch signal lines in especially dense memory arrays, are also susceptible to the EM [31], [33], [96].

EM reliability of the interconnects can be improved through device-level approaches such as strengthening the Copper (Cu)-cap interface or adding Manganese (Mn) or Aluminum (Al) doping in Cu [19], [22]. However, by taking into account the unique EM characteristics at the application level, this reliability concern can be effectively managed at the architecture level as well as the device level.

2.6.2. EM modeling

2.6.2.1. Empirical modeling

The MTTF estimation for a linear interconnect can be determined through *Black's equation* [1]:

$$t_{50\%} = \frac{A}{j^n} \cdot exp(\frac{E_a}{k \cdot T}) \tag{2.1}$$

In Black's equation, j is the current density, E_a is the activation energy, k is Boltzmann's constant, T is the temperature in *Kelvin*, n is a constant ($1 < n < 2$) and A is a fitting parameter. Although Black's equation has a simple formulation, it is an empirical equation with only a limited number of parameters. So, it cannot capture the effect of other parameters such as *residue stress*.

2.6.2.2. Physical modeling

Unlike empirical modeling, *physical modeling* of the EM, includes only physical parameters and can consider all the impactful parameters. EM physical modelings are based on the solution of Korhonen's initial boundary Partial Derivative Equation (PDE) [3]. Due to the atomic flow inside an interconnect, *compressive stress* and *tensile stress* are created at the anode and cathode end of the line, respectively. The stress gradient in the line is present due to the interaction of these two opponent stresses and can be modeled by Korhonnen's PDE in (Equation 2.2).

$$\frac{\partial \sigma}{\partial t} = \frac{\partial}{\partial x}[\kappa(\frac{\partial \sigma}{\partial x} - \frac{Ze\rho}{\Omega}j)] \tag{2.2}$$

In Korhonen's PDE, $\sigma(x, t)$ is the hydrostatic stress at the location x and time t, j is the current density, B is the effective bulk elasticity modulus, Ω is the atomic lattice volume, Z is the effective valance charge, ρ is the resistivity of the metal line, $\kappa = D_a B\Omega/kT$ and D_a is the atomic diffusivity. D_a is related to the activation energy and temperature through *Arrhenius* equation: $D_a = D_0 \cdot exp(-E_a/kT)$, where D_0 is the diffusion constant.

In the physical modeling of EM, the MTTF is split into two phases; *Nucleation* and *Growth* phases. In the Nucleation phases, the hydrostatic stress in the interconnect is governed by Korhonen's PDE. When the hydrostatic stress at any location of the line reaches a critical value (σ_{crit}), a void is nucleated. After the void nucleation at time t_{nucl}, the stress inside the line is relaxed [88].

Equation 2.3 shows a solution of Korhonen's PDE [41]. Here, l is the length of the metal line, and σ_{res} is the residual stress. σ_{res} is induced due to the cooling down of the process from the higher temperatures

to room temperature and as a result of a mismatch in coefficients of thermal expansion between different materials.

$$\sigma = \sigma_{res} + \frac{eZ\rho jl}{\Omega}\left(\frac{1}{2} - \frac{x}{l} - 4\sum_{n=0}^{\infty}\frac{cos(\frac{(2n+1)\pi x}{l})}{(2n+1)^2\pi^2}e^{-\kappa\frac{(2n+1)^2\pi^2}{l^2}t}\right) \tag{2.3}$$

In the growth phase, the void starts to grow. The resistance of the interconnect is not changing significantly until the *incubation time* (t_{incub}); the time at which the void reaches a critical size (Δl_{crit}) and covers the intersection of the interconnect. Incubation time can be obtained through Equation 2.4 [61]. After the incubation time, due to the current shunting by liner barriers, the resistance of the interconnect is increasing with a rate (Δr) [11], [13], [15]. Δr can be approximated as Equation 2.5 in which, ρ_{Ta} and ρ_{Cu} are the resistivity of the Tantalum (Ta)-based barrier metal (Ta/TaN) and Cu respectively. W is the line width, H is Cu thickness, and h_{Ta} is the barrier layer thickness. Finally, the moment at which the EM-induced failure is observed, is considered as the EM-induced MTTF. So, as shown by Equation 2.6, MTTF is the summation of nucleation time (t_{nucl}), incubation time (t_{incub}) and time required for the critical resistance increase (t_r). This physical-based EM modeling has been developed in collaboration with IMEC and published in [84].

$$t_{incub} = \frac{\Delta l_{crit} \cdot k \cdot T}{D_a eZ\rho j} \tag{2.4}$$

$$\Delta r = \left[\frac{\rho_{Ta}}{h_{Ta}(2H + W)} - \frac{\rho_{Cu}}{HW}\right] \cdot \frac{D_a}{kT} \cdot eZ\rho_{Cu} j \tag{2.5}$$

$$MTTF = t_{nucl} + t_{incub} + t_r \tag{2.6}$$

2.6.2.3. EM physical modeling under a time-variant current density

For an interconnect under the bidirectional current waveform, the direction of the hydrostatic stress is reversed which results in the EM recovery. To analyze the EM phenomena under the time-variant current density, the work in [4] has suggested using the *effective Direct Current (DC)* as formulated in Equation 2.7 in which, γ is the recovery factor, and $\bar{j}_+$ and $\bar{j}_-$ are the absolute current values in each direction. The effective direct current for the bidirectional current waveforms is based on time-averaging. Leveraging the average current model is also verified through real measurements [5]. However, the recovery factor (γ) is not a fixed value and needs to be modeled with regard to the current waveform [41].

$$j_{\text{DC effective}} = \bar{j}_+ - \gamma\bar{j}_- \tag{2.7}$$

To avoid the inaccuracy of assuming a fixed value for γ, the solution of Korhonen's PDE under a time-constant current density can be considered as Equation 2.3 [41]. Any arbitrary time-variant current density can be modeled as a *piecewise constant* function.The current density in (t_{i-1}, t_i) is considered to be a constant value of j_i. Considering a very short period for (t_{i-1}, t_i) increases the accuracy while requiring a complex computation. On the other side, relatively large periods result in an unacceptable low accuracy but not complex computation. Hence, selecting this period should be done such that a good enough accuracy can be obtained through a reasonably complex computation.

As Equation 2.8 shows, the impact of a time-variant current density on the hydrostatic stress can be captured by performing an integral on $j(t)$ over time, and the nucleation time (t_{nucl}) is the time that satisfies Equation 2.8 for $\sigma = \sigma_{crit}$.

$$\sigma = \sigma_{res} + \frac{4eZ\rho}{\Omega l}\kappa\sum_{n=0}^{\infty} cos\frac{(2n+1)\pi x}{l}e^{-\frac{(2n+1)^2\pi^2}{l^2}\kappa t} \times \int_0^t j(t)e^{\frac{(2n+1)^2\pi^2}{l^2}\kappa\tau}d\tau \tag{2.8}$$

2.6.2.4. EM modeling in a multi-segment interconnect

In a multi-segment interconnect, typically, the currents that pass through the different segments are not equal to each other. Hence, the discussed physical-based EM modeling (Equations 2.3-2.6) does not work for a segmented interconnect. To investigate the EM phenomenon in a multi-segment interconnect, authors in [43] have proposed *offline approach* for modeling the EM phenomenon in a multi-segment interconnect, by obtaining the hydrostatic stress in the steady-state ($t \rightarrow \infty$). The interconnect confinement implies that its 'mass' will not change due to the EM. Therefore, *mass conservative law* imposes a crucial *boundary condition* on the steady-state EM-induced hydrostatic stress modeling. According to [43], Equations 2.9-2.11 are valid in a multi-segment confined interconnect.

$$\sigma(x) = \frac{Ze\rho}{\Omega} jx + \sigma_{res} \tag{2.9}$$

$$\sigma_i^c - \sigma_j^a = \Delta\sigma_{ij} = \frac{-ez\rho(j_{ij} \times l_{ij})}{\Omega} \tag{2.10}$$

$$\sum_{i,j=1}^{k} \left(\sigma_i - \left[\sigma_t - \frac{ez\rho(j_{ij} \times l_{ij})}{2\Omega} \right] \right) l_{ij} = 0 \tag{2.11}$$

Equation 2.9 is obtained from Equation 2.3 if ($t \rightarrow \infty$). Equation 2.10 is the direct conclusion of Equation 2.9 for calculating the $\Delta\sigma$ between the cathode and anode end of the wire-segment ij, and finally, the satisfaction of Equation 2.11, in which the σ_t is the thermal stress, ensures that mass conservation of the confined interconnect. Though the steady-state EM modeling (Equations 2.9-2.11) can capture the EM-induced stress in a multi-segment interconnect, it cannot provide any information regarding the EM-induced MTTF.

Part II.

Contributions

3. Decision Failure Modeling and Mitigation through Reference Adjustment

A reliable read operation in STT-MRAM technology (STT-read) is challenging for several reasons. The resistive states of MTJs are affected by the *Process Variation* and they are characterized by distribution rather than a fixed value. Moreover, the difference between the resistive levels or Tunnel Magneto-Resistance Ratio (TMR) in MTJs is much smaller than other resistive memory technologies [21], [83]. On top of that, temperature variations also affect the resistance distribution of these resistive levels in an asymmetric manner. These challenges are even more pronounced in the concept of computation in STT-MRAM (STT-CiM) since multiple rows are activated at the same time to perform logic operations and the sensing margin in CiM is significantly reduced compared to standard memory sensing.

During the sensing phase, the equivalent resistance of the activated MTJ cells is evaluated by sensing the current sum passing through their common BL. Then, the sensed equivalent resistance should be compared with a reference resistance to detect the result of the read/CiM operation. The aforementioned challenges during the sensing (reading) operation can result in erroneous read, leading to unacceptable RDF.

In this chapter, we first analyze the probability of RDF during standard STT-read as well as STT-CiM operations using different models of bit-cell array and sensing circuitry. We then try to improve the reliability of the read process by revisiting the design of the reference circuitry. Our main contributions to this chapter are as follows:

- Performing a detailed statistical read failure analysis in STT-CiM by considering the impact of process and temperature variations on the bit-cell array as well as sensing circuitry. We consider two models, a simplified resistive model of MTJs and a more comprehensive model that also includes the transistor behavior.
- Proposing a self-adjustable temperature- and variation-aware design of the reference resistance by targeting the minimum failure rate, based on the two models of the CiM-array for RDF analysis.
- Analyzing the impact of the number of input operands of STT-CiM on the RDF probability, and the scalability characteristics of STT-CiM in terms of reliability.
- Investigating the impact of the redundant mapping of CiM operands to multiple bit-cells to minimize the read failure rate for scalable CiM operations, and comparing with device parameter adjustments through the fabrication process optimizations.

The rest of the chapter is organized as follows. Section 3.1 presents basic information about STT-MRAM-rean and STT-MRAM-CiM mechanisms followed by the previous efforts on improving the sensing operation in the STT-read. In Section 3.2, we present the details of our STT-CiM failure rate analysis and modelings. Section 3.3 explains the proposed methodology to design self-adjusted optimum reference resistance. Section 3.4 evaluates the efficiency of our proposed reference resistance, and finally, section 3.5 concludes the chapter.

N	1 (Read)	2	4	8
TMR^{*}_{CiM-N}	100%	32.45%	13.80%	6.42%

Table 3.1.: TMR^{*}_{CiM-N} (normalized to TMR of single-cell STT-Read operation) for different N at $25°C$

3.1. preliminaries

The crucial reliability challenge for CiM in resistive memories is a narrower sense margin between multiple resistive levels. For STT-CiM, the situation is worse, since the sense margin is narrow even for the standard memory read, and the resistive values of the MTJ are affected by both temperature and process variation.

For a standard STT-Read, the distance between the resistive level of the 'P' and 'AP' MTJ (R_P and R_{AP}, respectively) is represented by the TMR (see Equation (3.1)). For the case of CiM-N (a memory crossbar where N devices in a column are activated at the same time), the narrowest sense margin is the distance between the lowest resistive states (all the devices in the 'P' state *and* all but one devices in the 'P' state). For CiM-N, TMR^{*}_{CiM-N} is defined as the narrowest sense margin and can be calculated through Equation (3.2) by replacing R_{AP} and R_P with $\frac{R_P}{N-1}||R_{AP}$ and $\frac{R_P}{N}$, respectively, in Equation (3.1). Table 3.1 shows the TMR^{*}_{CiM-N} for different numbers of operands of the CiM operations (N). The narrowest sense margin decreases with the number of operands which severely impairs the reliability of the STT-CiM.

$$TMR = \frac{R_{AP} - R_P}{R_P} \tag{3.1}$$

$$TMR^{*}_{CiM-N} = \frac{\frac{N \times R_P}{(N-1) \times R_{AP} + R_P} \times R_{AP} - R_P}{R_P} \tag{3.2}$$

3.1.1. MTJ variability and their impacts on CiM operations

Process variation is a major concern in manufacturing MTJs [54]. Due to the process variation, the 'P' and 'AP' state resistivity are not fixed values and follow probability distributions where the 'AP' state resistance is more prone to the process variation than the 'P' state [62]. Additionally, the resistivity of the MTJ shows temperature-dependency which is also differing based on the state of the free layer ('P' or 'AP') [47]. Figure 3.1 shows the temperature-dependent distribution of the 'P' and 'AP' states. The 'P' state is nearly stable in the operational temperature range. The resistance level of the 'AP' state, however, is showing a Negative Temperature Coefficient (NTC) behavior; the mean (μ) and also the standard deviation (σ) of the 'AP' state is decreasing at higher temperatures [26]. The temperature dependency reduces the absolute sense margin, and distinguishing between these resistance states at higher temperatures becomes difficult. Consequently, the read failure rate in conventional STT-Read operation increases for higher temperatures [48]. STT-based CiM architectures are even more prone to this behavior as their sense margins are significantly smaller even at lower temperatures [82].

3.1.2. Related work

The development of reliable read-sensing mechanisms to correctly distinguish the states of a sensed MTJ has been the focus of many works in the recent past. The compound reference structure introduced in [8] is one of the most prominent designs. It consists of 4 MTJs organized in two parallel chains, each chain is a series connection of one 'P' and one 'AP' MTJ. The mean (μ) of the resistance of the compound structure is in the middle of the 'P' and 'AP' resistive levels. This structure allows the

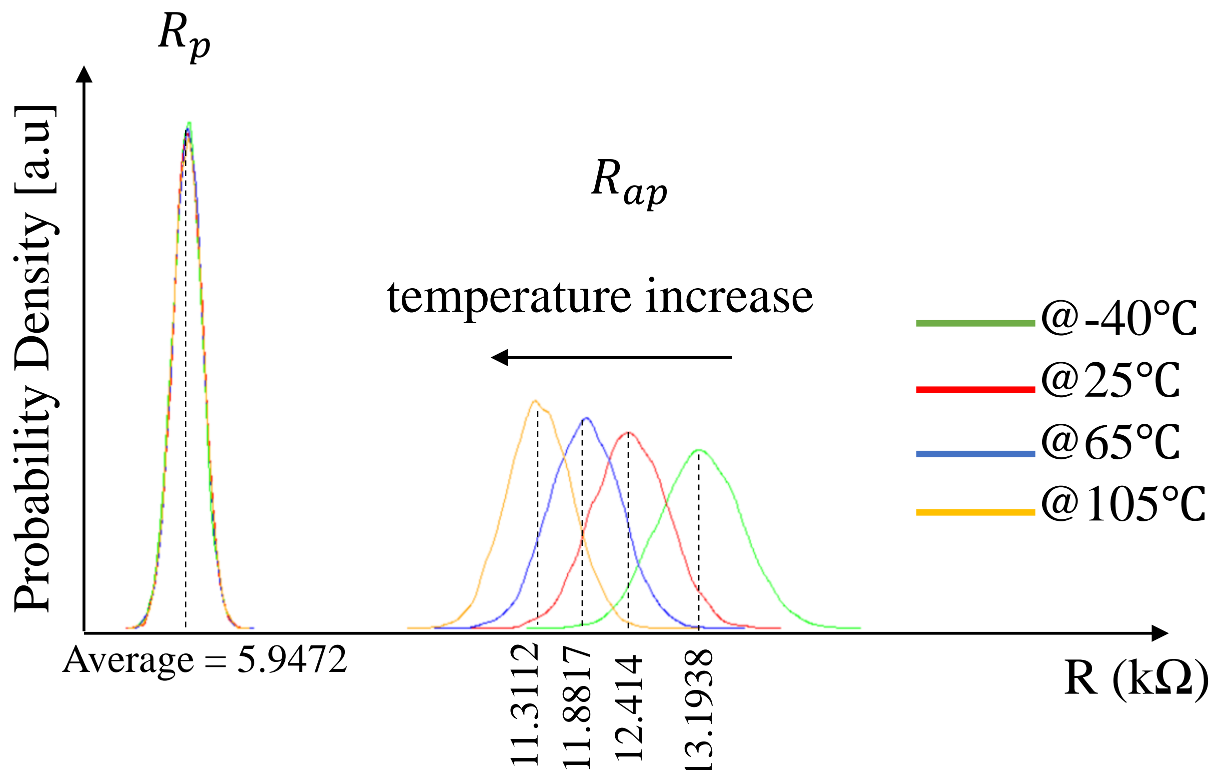


Figure 3.1.: Impact of the temperature on the resistance distributions of 'P' and 'AP' MTJ. For the setup, please refer to section 3.4

reference resistance to consider the process variation directly without additional area overhead to adjust for manufacturing variations. The 'P' and 'AP' states of MTJ are differently affected by the temperature. Therefore, the resulting reference resistor with resistance in the middle of the 'P' and 'AP' resistive states is not optimal with respect to the temperature. This leads to an unacceptably larger RDF probability for higher temperatures. A fixed reference resistance structure based on polysilicon (poly) is presented in [56]. Additional smaller resistors are put in series with the reference resistor, which can be dynamically bypassed. This so-called *trimming* is used to mitigate the process variation in a post-fabrication. An additional dummy bit-line is also used to include the parasitic behavior of the bit-line for cells located far away from the sense amplifier. A bias transistor is used in combination with an additional temperature adaptive bias signal generator to address a possible temperature shift during the operation.

The temperature dependency problem can be addressed with the help of an NTC resistor, as described in [73]. This NTC resistor is used in series with a 'P' state MTJ to adjust the reference point between the 'P' state and the 'AP' state. The NTC resistor is agnostic to the temperature-induced variation dependency of the MTJ and thus will only mitigate the nominal temperature-induced resistance shift. Also, other techniques such as under-driving the WL signal can be utilized to mitigate the impact of the higher temperature [67].

Voltage-based sensing is another approach to perform read operation in resistive memories [47], [57], [58]. An additional current generation with twice the read current is then used in combination with two MTJs, one in 'AP' and one in 'P', to generate the reference current. To adjust for bit-line loads, entire columns of these reference cells are used. By using a three-section reference scheme and splitting up those reference columns into groups that are activated depending on the needed bit-line, higher sense resolutions can be achieved [49].

All the solutions mentioned above focus on the average resistance shift of the reference and therefore offer only limited mitigation capabilities for the temperature-dependent change of MTJ variation. Additionally,

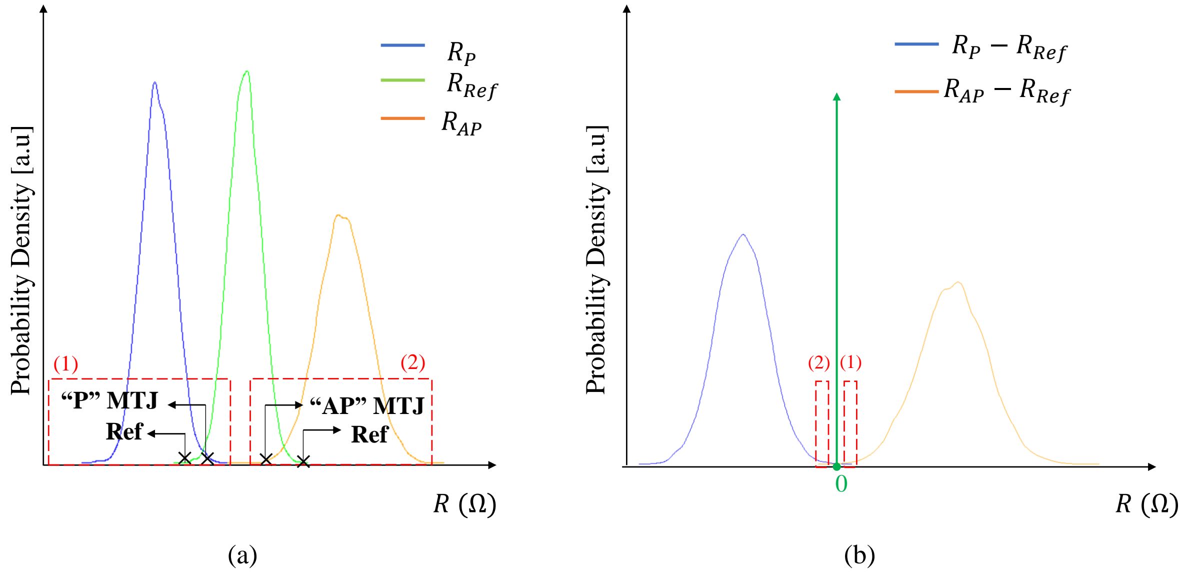


Figure 3.2.: (a) Resistance distributions of 'P' and 'AP' MTJ and also reference resistance, box (1) and (2) show the RDF case for 'P' MTJ and 'AP' MTJ, respectively (b) equivalent RDF cases by considering $R_P - R_{Ref}$ and $R_{AP} - R_{Ref}$ distributions

reliable sensing of STT-CiM is not addressed in any of the aforementioned works. However, the temperature-dependent MTJ variation is particularly challenging for CiM operations, as the sense margins for these operations are significantly smaller than the margins of conventional memory read operations.

3.2. Analyzing read decision failure probability

In this section, we provide a detailed RDF probability analysis for conventional STT-read and then generalize it for the case of STT-CiM under both process and temperature variations in the bit-cell array as well as the sensing circuitry. We start by analyzing RDF in conventional STT-read and then extend it to CiM operation. We also consider two models for analyzing the RDFs. First, we consider a simpler model in which only the resistive states of the MTJs and sensing resistances are considered. Then we extend the model by considering the effects of transistor variability in the bit-cells as well as the sensing circuitry.

3.2.1. RDF in conventional STT-read

Sensing (reading) an MTJ is accomplished by comparing its resistance value with a reference resistance. As Figure 3.2 (a) shows, similar to the sensedMTJs, the reference resistance itself does not have a fixed resistance, it rather has a statistical distribution. For a standard STT-read operation, an RDF happens in two cases:

1. The resistance of the 'P' MTJ becomes greater than the reference resistance (Figure 3.2 (a) box (1)); $R_P > R_{Ref} - M_{SA}$
2. The resistance of the 'AP' MTJ becomes less than the reference resistance (Figure 3.2 (a) box (2)); $R_{AP} < R_{Ref} + M_{SA}$

M_{SA} is the minimum required margin of the sense amplifier. For the simplicity of analysis, in the rest of the chapter, an ideal sense amplifier with $M_{SA} = 0$ has been taken into account. However, adjusting the equations to also consider the minimum required margin of the sense amplifier is possible.

The RDF probability ($\mathbf{P}_{RDF}$) can be calculated through Equation (3.3)[1] [18].

$$\begin{aligned} \mathbf{P}_{RDF} &= \mathbf{P}(\text{read as `AP'}|\text{stored as `P'}) \cdot \mathbf{P}(\text{stored as `P'}) \\ &+\mathbf{P}(\text{read as `P'}|\text{stored as `AP'}) \cdot \mathbf{P}(\text{stored as `AP'}) \end{aligned} \tag{3.3}$$

As it is also presented by Figure 3.2 (b), the probabilities $\mathbf{P}$(read as 'AP'|stored as 'P') (when R_P is greater than R_{Ref}) and $\mathbf{P}$(read as 'P'|stored as 'AP') (when R_{AP} is less than R_{Ref}) are equivalent to $\mathbf{P}(R_P - R_{Ref} > 0)$ and $\mathbf{P}(R_{AP} - R_{Ref} < 0)$, respectively.

3.2.2. RDF in STT-CiM

To generalize Equation (3.3) to also extend the RDF model to the CiM concept, we first need to define S, which is the stored state, and O which is the read output state corresponds to S. T his relation can be shown as $\mathbb{1}(O = S)$. The RDF happens if the read state is not the expected output of the stored state and the RDF probability can be calculated through Equation (3.4):

$$\mathbf{P}_{RDF} = \sum_{i=0}^{N} \mathbf{P}(O_i|S_i) \cdot \mathbf{P}(S_i); \quad if\ \mathbb{1}(O_i! = S_i) \tag{3.4}$$

In Equation (3.4), N is the total number of the states. In the logical STT-CiM operation with the involvement of N resistive memory cells as the operands, S_i is the stored state with i ($i = 0..N$) MTJs in the 'P' state and $N - i$ MTJs in the 'AP' state.

For the reference design, we target the RDF mitigation in the average and the probability of being programmed either in 'P' or 'AP' state is equal ($\mathbf{P}$(Stored as 'P') = $\mathbf{P}$(Stored as 'AP') = $\frac{1}{2}$).

3.2.3. Modeling the cell variability

The bit-cell of STT-MRAM can be modeled simply as only the resistance of the MTJ (*R modeling*) or by also taking the variability of the access transistor into account (*Tr+R modeling*). The overall resistive distributions and their temperature dependencies are different in these two models. The RDF probability analysis in section 3.2 can be applied for both types of cell modelings with proper adjustments of respective distributions. Later, we will show the implications of these modeling approaches in the estimation of RDF and optimization of the sensing scheme.

3.2.3.1. (R) Cell Modelling

The resistive distribution of only the 'P' and 'AP' MTJ cells at different temperature points (-40°C to 125°C) can be obtained through Monte Carlo analysis and then fitted to a non-negative normal distribution. Table 3.2 shows the normal distribution characteristics of the 'P' and 'AP' resistance levels.

3.2.3.2. (Tr+R) cell modelling

While the STT-MRAM bit-cell contains an access transistor in addition to the MTJ, it is essential to also consider the impact of the process and the temperature variation on the access transistor and subsequently on the overall resistance distribution of the full cell (Tr+R) model.

[1] It should be noted that in this equation, the probability of the *read disturbance* is not considered; read disturbance happens if the read current changes the state of the MTJ cell.

Table 3.2.: The normal distribution parameters of 'P' and 'AP' MTJs, for the setup, please refer to section 3.4

Temp(°C)	'P' MTJ(μ, σ) (kΩ)	'AP' MTJ(μ, σ) (kΩ)
-40	(5.9472, 0.2977)	(13.1938, 0.6234)
-20	(5.9554, 0.2981)	(12.9645, 0.6096)
0	(5.9621, 0.2984)	(12.7264, 0.5984)
25	(5.9678, 0.2987)	(12.414, 0.5817)
45	(5.9702, 0.2989)	(12.152, 0.5686)
65	(5.9703, 0.2989)	(11.8817, 0.5540)
85	(5.9680, 0.2987)	(11.6010, 0.5408)
105	(5.9630, 0.2985)	(11.3112, 0.5293)
125	(5.9552, 0.2981)	(11.0112, 0.5154)

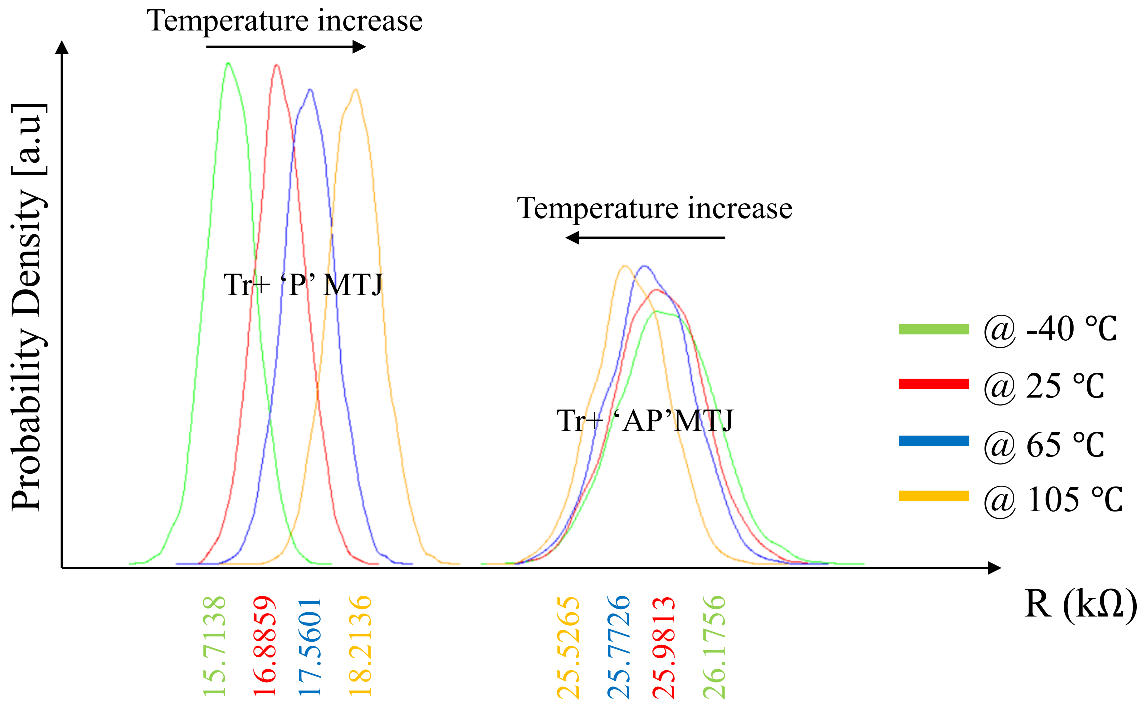


Figure 3.3.: Resistance distribution of the (Tr+R) cell model, for the setup, please refer to section 3.4

Figure 3.3 shows the (Tr+R) cell resistance distribution. Similar to the (R) modeling, the equivalent resistance in (Tr+R) modeling also follows the non-negative normal distribution. It can be seen that in the full cell, the (Tr+'P') MTJ shows a Positive Temperature Coefficient (PTC) behavior while the (Tr+'AP') shows NTC behavior with the lower temperature coefficient in comparison to the (Tr+'P') MTJ.

Modeling the equivalent resistance of the full cell is essential to optimize and design the reference circuitry based on (Tr+R) cell model. Figure 3.4 shows the equivalent resistance of the (Tr+'P') MTJ and (Tr+'AP') MTJ versus the resistance of the 'P' and 'AP' MTJ loads at four different temperatures. In general, the equivalent resistance of the (Tr+R) cell is a function of the load (R_{load}) and the temperature (T) ($R_{eq-Tr+R} = f(R_{Load}, T)$). We can approximate $R_{eq-Tr+R} = f(R_{Load})$ using linear regression at each temperature. By having this linear regression, finding the required load on the reference side will be possible.

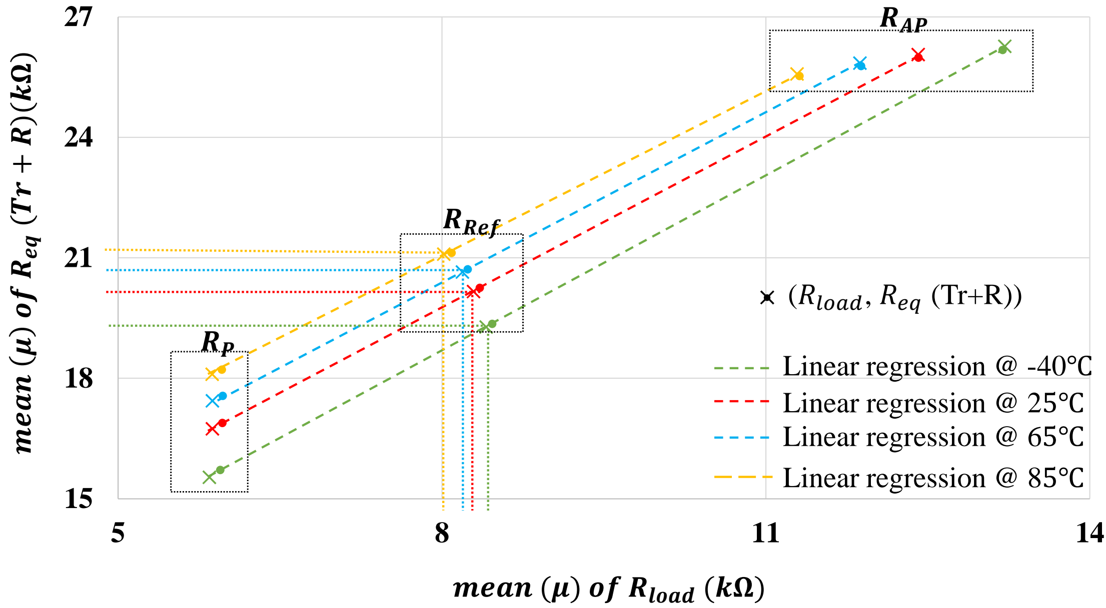


Figure 3.4.: Linear approximation of the equivalent resistance of the full cell (Tr+R) vs the 'P' and 'AP' MTJs as the load at 4 different temperatures, for the setup, please refer to section 3.4

3.3. Optimized reference design to minimize RDF

In this section, based on the statistical RDF analysis discussed in section 3.2 and with the consideration of the (R) and the (Tr+R) cell modelings, we try to minimize the RDF probability (Equation (3.4)) by designing proper reference circuitry to track the temperature behavior. We introduce our proposed reference structure to minimize the RDF probability across all temperatures and then extend our methodology to design a reference resistance for multi-operand CiM.

3.3.1. Optimizing reference resistance based on (R) cell modeling

The following steps need to be taken to design the optimal reference resistance structure for the (R) cell modelings.

1. In each temperature, obtain the $\mu_{ref_{Opt}}$ (optimum mean for the reference resistance) to minimize the Equation (3.4). The $\mu_{ref_{Opt}}$ is located between (not necessarily middle) of the means (μ) of two distinct resistive levels of the input data and is obtainable through numerical simulation. To conduct such numerical simulation, we calculate the RDF probability (Equation (3.4)) for a range of reference values between the means (μ) of two distinct resistive levels of the input data. The reference value that results in the lowest RDF probability is then selected as the mean of the optimum reference resistance ($\mu_{ref_{Opt}}$).

2. Find the linear regression between the values of $\mu_{ref_{Opt}}$ (in different temperatures) and $\mu_{T-variant}$ (mean of the temperature-variant block), such as the 'AP' MTJ with the mean of μ_{AP} or a block consisting of the parallel connection of the 'P' and 'AP' MTJ with the mean of $\mu_{P||AP}$.
$\mu_{ref_{Opt}} = \frac{n}{m} \times \mu_{T-variant} + C$

3. Realize the coefficient of the linear regression ($\frac{n}{m}$) through a sufficient number of temperature-variant blocks in series and parallel (an m parallel connection of chain blocks, each chain with n temperature-variant blocks).

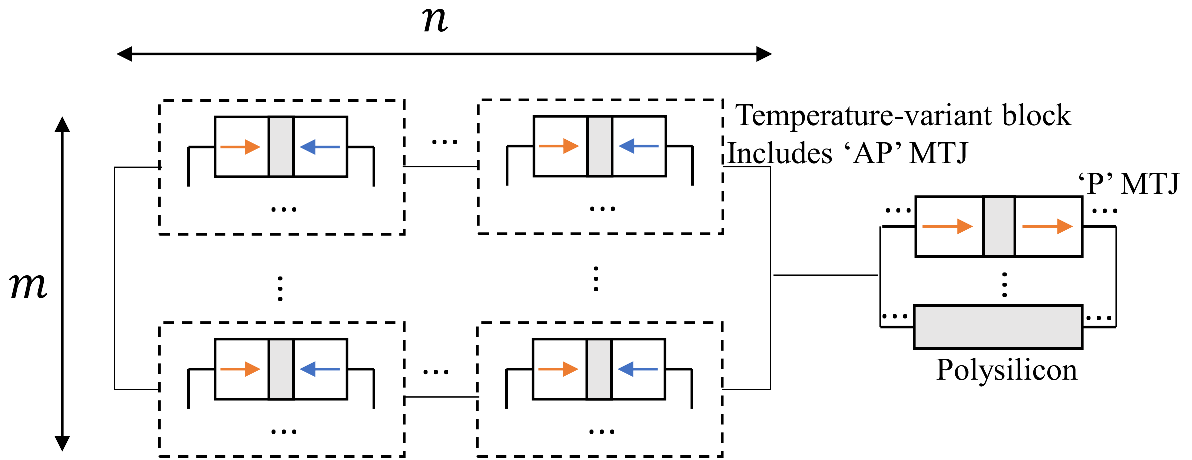


Figure 3.5.: The conceptual design of the proposed reference resistance

4. Realize the constant intercept of the linear regression (C) through the connection of the ‘P’ MTJs or parallel connection of the ‘P’ MTJ and polysilicon films.

5. Compare the finalized reference resistance (step 4) with the optimum mean (μ) of the reference resistance (step 1) and check if they are almost equal. In the case of inequality, step 4 should be revisited.

The rationale for such a structure is the following. Since the only temperature-dependent component is the ‘AP’ MTJ, the temperature-variant block should include the ‘AP’ MTJs. Both the ‘P’ MTJ and the polysilicon are temperature-invariant elements, the resistance of the polysilicon can be adjusted through sizing, and ‘P’ MTJ has the smallest standard deviation (σ) among the ‘P’ MTJ, ‘AP’ MTJ, and polysilicon. The parallel connection of two normally distributed resistances has a standard deviation (σ) less than the smallest element. Therefore connecting ‘P’ MTJ and polysilicon in parallel reduces the standard deviation (σ) and results in a lower RDF probability. Figure 3.5 shows the conceptual design of the optimum reference resistance.

Based on the $\mu_{ref_{Opt}}$ and also the μ_P and μ_{AP}, the optimum reference resistance for the conventional STT-read by considering (R) cell modeling can be realized by Equation (3.5). To implement the linear regression in Equation (3.5), the slope 0.2214 can be approximated by 0.2 and the temperature-variant term can be realized through 5 parallel-connected “AP” MTJ elements. To realize the constant temperature-invariant term, a parallel connection of 2 chains, each containing a series connection of 1 ‘P’ MTJs and 1 poly with the dimensions of $W = 400\ nm$, $L = 3.695\ um$ can be used. Such poly resistance provides the non-negative normal resistance distribution with ($\mu = 5.6064\ k\Omega$, $\sigma = 0.3684\ k\Omega$). Leveraging the parallel connection for the realization of the constant temperature-invariant term results in decreasing the overall standard deviation (σ) of the reference structure.

$$\mu_{ref_{Opt-STT-read}} = 0.2214 \times \mu_{AP} + 5.4107 \tag{3.5}$$

3.3.1.1. Extension for CiM referencing

The proposed 5-step methodology of designing the optimum reference structure can be generalized also for CiM operation. In the case of a 2-operand CiM operation, for instance, two reference resistances are required, one to distinguish between $[P||P]$ and $[P||AP]$ (R^{OR}_{Ref}) and the other one to distinguish between

$[P||AP]$ and $[AP||AP]$ (R_{Ref}^{AND})[2]. In each temperature, the worst-case RDF probability happens when distinguishing between the lowest resistance levels, i,e. *all activated MTJs in 'P'*, and *(all but one) MTJs in 'P' and only one MTJ in 'AP'*. In Steps 1 of our methodology, the mean of the optimum R_{Ref}^{OR} ($\mu_{ref_{Opt-OR}}$) is located between $[\frac{P}{N}]$ and $[\frac{P}{N-1}||AP]$. In step 2 the linear regression can be computed between the optimum $mu_{ref_{Opt-OR}}$ and a temperature-variant element such as two parallel MTJs one in 'P' state and the other one in 'AP' state ($P||AP$). Equation (3.6) shows the linear regression to realize R_{Ref}^{OR} for the CiM-2 operation and by considering (R) cell modeling. The temperature-variant part of the Equation (3.6) can be realized through three parallel blocks each consisting of a parallel 'P' and 'AP' MTJ. Also, for the constant temperature invariant part, a parallel connection of a 9 'P' MTJs chain, one chain consisting of 4 'P' MTJs and 8 chains consisting of 3 'P' MTJs can be used to realize the temperature-invariant part. The realization of the temperature-invariant part has been done only with the 'P' MTJ to decrease the standard deviation (σ). Also, the asymmetry in the realization of the temperature-invariant part is introduced in step 5 of the proposed methodology to decrease the difference between the $\mu_{Ref_{Opt}}$ and the $\mu_{Ref_{Realized}}$.

$$\mu_{ref_{Opt-OR}} = 0.3648 \times \mu_{P||AP} + 1.9378 \tag{3.6}$$

3.3.2. Optimizing reference resistance based on (Tr+R) cell modeling

The 5-step methodology for designing the reference resistance based on (R) modeling can also be used for the (Tr+R) cell modeling. However, to also capture the impact of the access transistor variation, the following modifications are required:

- In step 1, after obtaining the $\mu_{ref_{Opt}}$, the optimum mean of the load ($\mu_{load_{Opt}}$) which is connected in series with the access transistor should be computed. As introduced in section 3.2.3.2 and depicted in Figure 3.4, in each temperature, the pair of ($\mu_{load_{Opt}}$, $\mu_{ref_{Opt}}$) is located on the linear regression corresponding to that temperature.
- In steps 2, 3 and 4, $\mu_{load_{Opt}}$ should be realized.
- In step 5, multiple iterations may be required to ensure that the realized (Tr+R) reference model converges to the optimum reference resistance. Since the resistivity of the 'AP' MTJ is voltage-dependent.

Please note, that in the concept of scouting logic, the reference resistance is only a part of the sensing circuitry. To study different aspects of the sensing mechanism, such as power and performance, considering the entire sensing circuitry including the sense amplifier and bit-cells is required. Such studies are out of the scope of this work.

3.3.3. Improving the read reliability in STT-CiM

The activation of multiple cells in parallel is required to perform the computation in the memory. In the case of resistive memories, the means (μ) of the different resistive levels (corresponding to different data) become closer to each other, and as a result, the RDF probability increases significantly. To decrease the RDF probability, we investigate two approaches namely: 1. Changing the device parameters of the MTJ using process engineering and 2. Encoding (mapping) the data (operands) in multiple redundant cells.

[2] R_{Ref}^{OR}, R_{Ref}^{AND} are not necessarily two separate structures and to decrease the hardware overhead, a re-configurable integration can be designed to efficiently implement them.

3.3.3.1. Changing the device parameters of the MTJ

Adjusting two device-level parameters of the MTJ namely Resistance-Area product (RA) and TMR can also decrease the RDF probability. Changing the stack properties such as increasing the MgO layer thickness can increase the RA [24]. An increase in RA on the one hand, causes the larger mean (μ) in the 'P' and 'AP' MTJ resistive distributions which is beneficial for decreasing the RDF probability. On the other hand, the larger the RA, the larger the standard deviation (σ) of the 'P' and 'AP' MTJ resistive distributions. Therefore increasing the RA has two opposite impacts on the RDF probability. Our results, however, show that increasing the RA will ultimately increase the RDF probability. Increasing the RA is also undesirable from the write energy point of view. To perform a write operation on devices with higher resistivity, a higher write current or voltage should be applied to the device.

By increasing the TMR, the resistance distribution of the 'P' MTJ remains untouched, and both the mean (μ) and the standard deviation (σ) of the 'AP' MTJ resistance distribution increase. Increasing the TMR can significantly decrease the RDF probability of the STT-read. It is also effective for decreasing the RDF probability of CiM operations. Similar to the RA, the higher the TMR, the higher the write energy, and the higher the risk of the device break-down. Therefore, increasing the RA for the sake of the RDF probability improvement should be done by considering the aforementioned consequences and trade-offs.

3.3.3.2. Encoding data (operands) in the multiple redundant cells

To improve read reliability for the CiM operations, instead of storing one bit of data in one bit-cel (Tr+R), we can encode it in multiple parallel redundant cells. The parallel structure decreases the standard deviation (σ) and consequently, reduces the RDF probability. Figure 3.6 (a) shows the redundant structure of n_{Red}(Tr+R). With this method, the periphery circuitry needs to be adjusted to allow such redundant mapping, and then, the specific mapping can be achieved either by the CiM compiler or through the Memory Management Unit (MMU) at the microarchitecture level.

It must be noted that in general, software support is required for the CiM realization, and Instruction Set Architecture (ISA) should be augmented with the required instruction for offloading. However, the n_{Red}(Tr+R) encoding method, does not necessarily require any further changes in the ISA. For each memory access instruction, the MMU maps each single-row address to the particular number of rows that must be activated at the same time. The number of simultaneously activated rows depends on the type of instruction (standard memory access vs. CiM) and can be handled by the MMU.

The design of the reference side should also be changed according to the number of redundant cells. The number of activated calls on the data side and on the reference side should be equally balanced. Hence, encoding the data in redundant cells requires the reference to also be implemented with redundant cells. Figure 3.6 (b) shows the reference side, while the data side utilizes the redundancy to decrease the RDF probability. Based on our models, the structure of the optimum reference load is not affected by the n_{Red}(Tr+R) encoding. The increase in the parallelism degree on the reference side as well as the data side, is beneficial in decreasing the standard deviation (σ).

3.4. Results and discussion

In this section, we present the RDF analysis results for both the (R) and (Tr+R) models and present the optimized reference design based on these models. Additionally, we present the impact of device parameters and redundant mapping on read reliability of CiM operations.

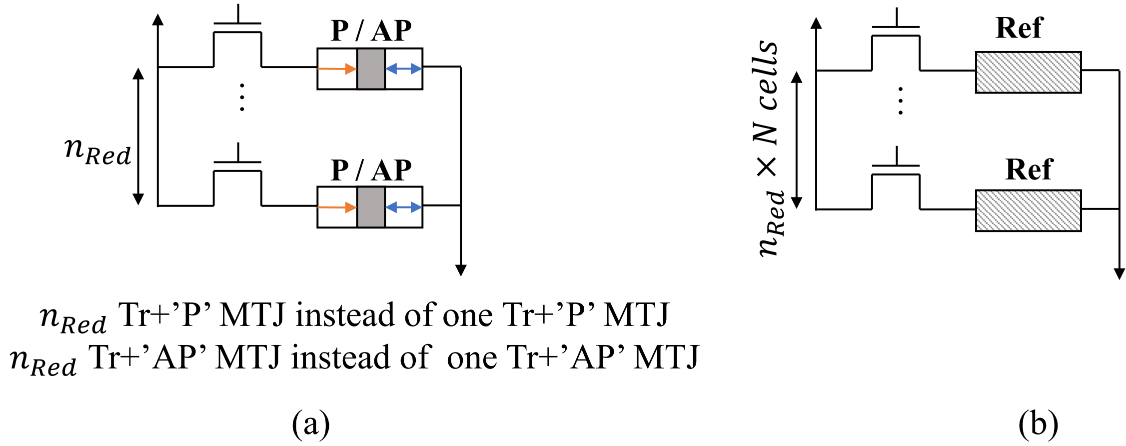


Figure 3.6.: (a) Encoding the data in the redundant cells and (b) a general structure of the reference cells for CiM operation with N number of operands and encoding the data in n_{Red} redundant cells

Table 3.3.: simulation setup tools and parameters

Simulation tool	Cadence Virtuoso
Technology node for CMOS	TSMC 40 nm
MTJ model [38]	radius = 20 nm Free/Oxide layer thickness = 1.3/1.48 nm RA = $7.5\Omega\mu m^2$ Nominal TMR = 150%
Monte Carlo analysis	1k in each temperature
Polysilicon model	P+Poly resistor without silicide; sheet resistance = $598.65\frac{\Omega}{square}$
Statistical analysis language	R

3.4.1. Experimental Setup

The experimental setup details are shown in Table 3.3. The distributions of the (Tr+R) cell modeling have been adjusted based on the STT-Read error probability of approximately $1E-5$ at $25°C$ and $125°C$, as reported in [57].

3.4.2. RDF results for (R) cell modeling

We compare the different structures of the reference resistance, namely the proposed structure, MTJ-based compound structure [8] and only polysilicon resistor [57] from the RDF probability and the area point of view. It must be noted that for the CiM purpose, the compound reference structure is indeed the generalization of the STT-MRAM compound structure. Figure 3.7 (a) shows the compound structure for the standard STT-MRAM, while (b) and (c) show the generalized structure for R^{OR}_{Ref} and R^{AND}_{Ref}.

Figure 3.8 shows the RDF probabilities of these reference structures. In Figure 3.8 (a) one MTJ (STT-Read) and in Figure 3.8 (b) two MTJs are activated simultaneously (CiM-2). To adjust the resistive value of the polysilicon reference resistance, our target is to minimize the maximum error at the highest temperature, so we adjust its dimensions for the resistive value equal to the value of the optimum reference resistance in the $125°\ C$ corner. Among the three different reference structures, the highest RDF probability (in the range of $5.68E-4$ to $6.54E-4$) happens for polysilicon resistor-only, since this

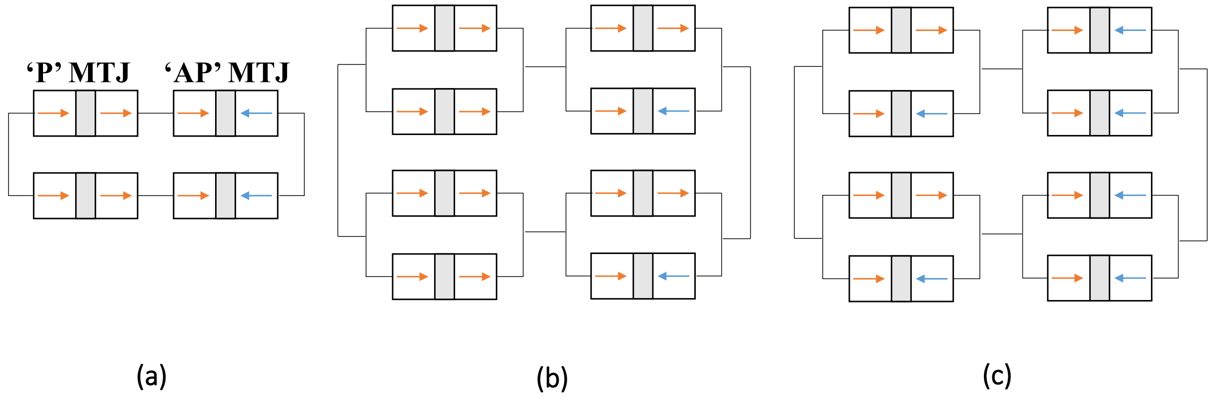


Figure 3.7.: Compound reference structure (a) [8] for the standard STT-read, (b) and (c) generalizing to the 2-operand CiM for R_{Ref}^{OR} and R_{Ref}^{AND}, respectively

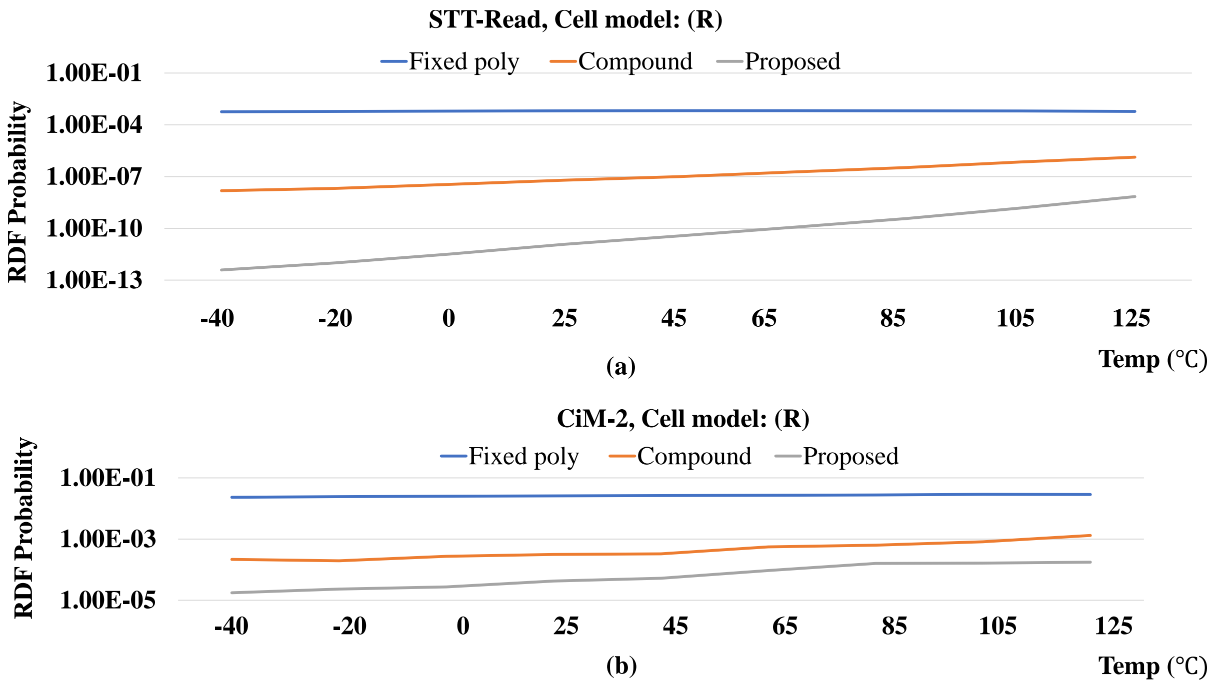


Figure 3.8.: RDF probability for the proposed structure, compound, and only poly as the reference resistance. (R) model has been considered for the cell. (a) STT-read (b) CiM-2

element is unable to track the temperature and also has a relatively high standard deviation (σ). Table 3.4 summarizes the area and the average RDF probability for the activation of one and two MTJs.

As we can see in Table 3.4, the proposed reference structure can improve the average RDF probability significantly. However, for the case of 2-operand CiM, the improvement of the RDF through the proposed design for the reference resistance is not as significant. For the case of CiM-4 and CiM-8, the mean (μ) of the extended compound structure is almost equal to the $\mu_{Ref_{Opt}}$. In other words, the extended compound structure can be used to implement the proposed reference. Therefore the reference side for CiM-4 and CiM-8 has the extended compound structure. Figure 3.9 shows the RDF probability with respect to the temperature for the different numbers of CiM operands.

Table 3.4.: Average RDF probability and area of the different reference resistance structures for STT-Read (1 MTJ activated) and CiM-2 (2 MTJs activated) based on (R) cell modeling

# of Activated MTJs	Structure of the reference resistance	Average RDF probability	area (# of MTJs, L of polysilicon um) W of polysilicon = 400 nm
1	Compound [8]	3.08×10^{-7}	(4,-)
1	Only poly [57]	6.23×10^{-4}	(-, 5.115)
1	**Proposed**	9.89×10^{-10}	(7, 3.695×2)
2	Compound [8]	4.15×10^{-4}	(8,-)
2	Only poly [57]	2.62×10^{-2}	(-, 2.16)
2	**Proposed**	8.37×10^{-5}	(34, -)

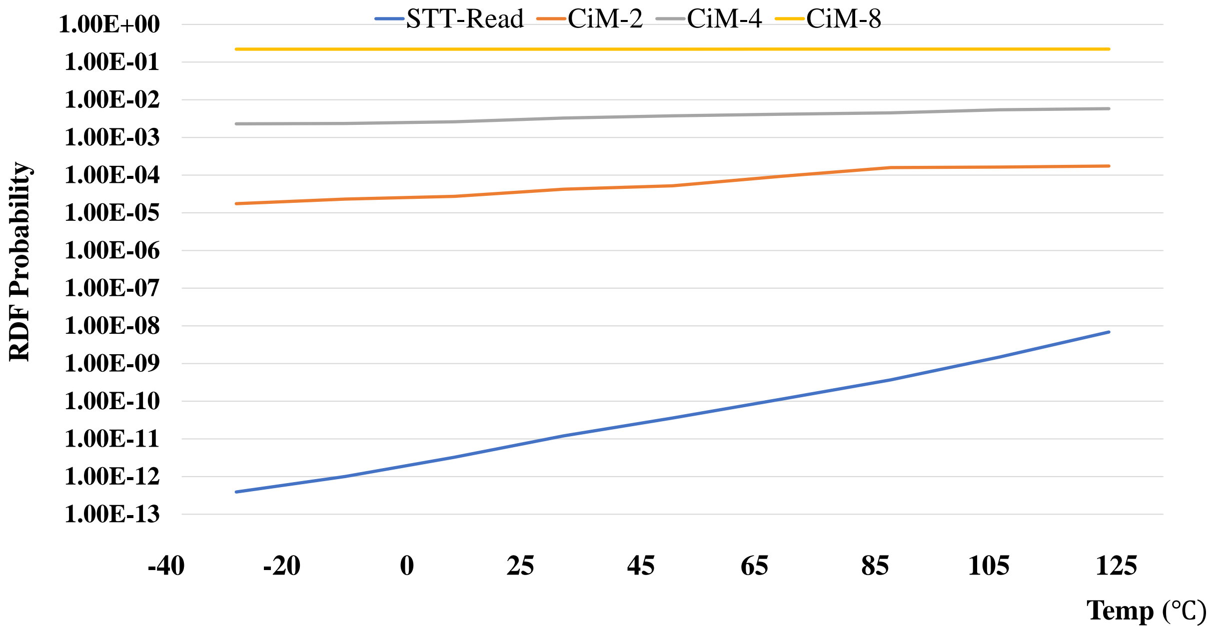


Figure 3.9.: RDF probability for different number of CiM operands based on (R) cell model, the reference resistance structure for STT-Read and CiM-2 are the proposed structure and for CiM-4 and CiM-8 is the compound structure

3.4.3. Analysis and reference optimization based on (Tr+R) model

In this work, we have presented and investigated two models of cell variability namely (R) and (Tr+R). Here we investigate how these models are used for the *reference optimization* and then *RDF estimation* perform. Considering the variation of the access transistor in the full cell model for both the optimization and RDF estimation steps, results in the most optimized and realistic values. The reference structure is almost the same for both (R) and (Tr+R) cell modeling. In other words, considering transistor variations does not affect the reference structure design. However, considering the transistor variations in the estimation of the RDF probability is crucial, since (R) cell modeling for the RDF analysis significantly

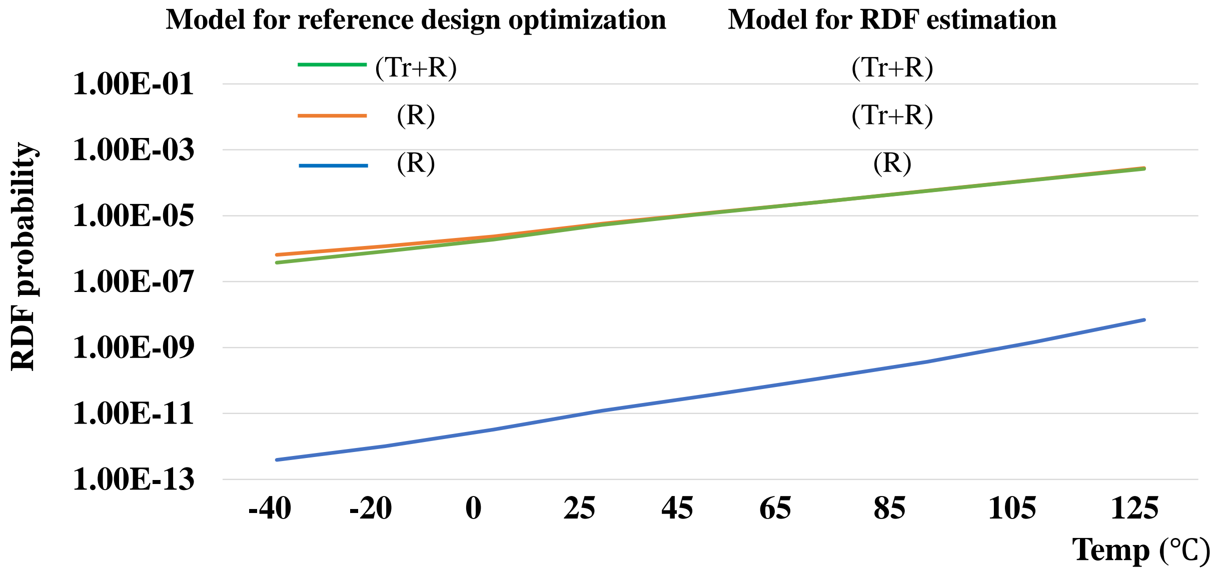


Figure 3.10.: RDF probability with different approaches of the reference optimization and RDF estimation using (R) and (Tr+R) cell modelings

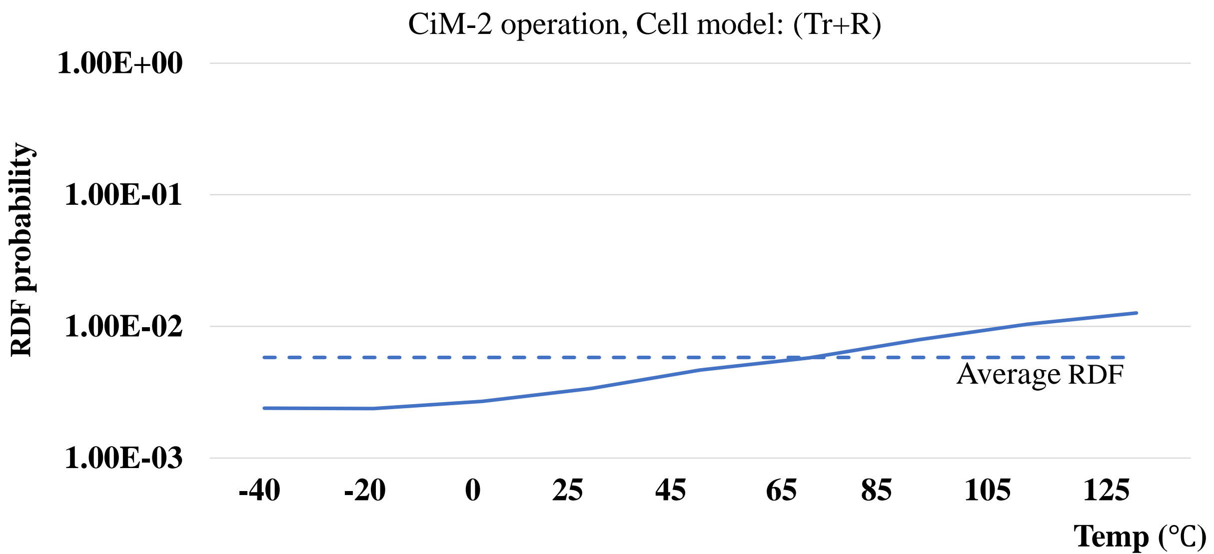


Figure 3.11.: RDF probability for the CiM-2 operation, the cell model is considered to be (Tr+R)

underestimates the RDF values, causing higher failures in the field. Figure 3.10 shows the RDF probability in consideration of different modelings in the reference optimization and RDF estimation phases.

3.4.3.1. Utilizing (Tr+R) cell modeling for STT-CiM

According to [57] the target for the STT-read is around $1E-5$. However, the results of Figure 3.11 show that the average RDF probability of the CiM-2 operations is $5.8E-3$. As already discussed in section 3.3.3.2, encoding (mapping) the data (operands) in multiple redundant cells can be a solution to improve the RDF, especially in the STT-CiM cases.

Figure 3.12 shows the minimum number of cells for redundant mapping based on different CiM operations with different number of operands, to maintain the target RDF probability. The higher

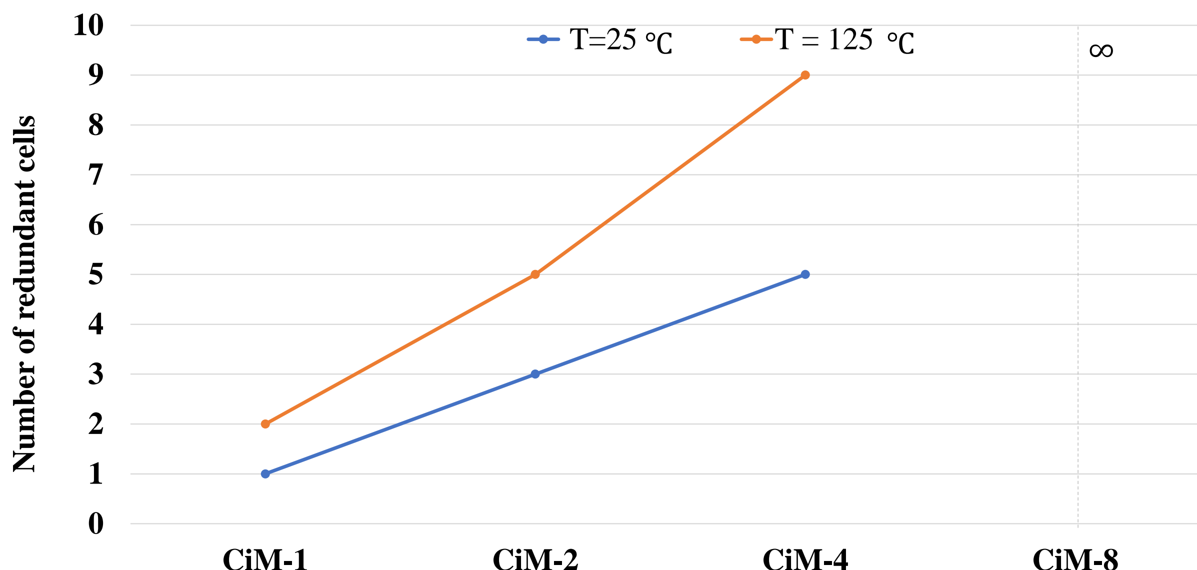


Figure 3.12.: The minimum number of the redundant (Tr+R) cells versus different numbers of the CiM operands to achieve target RDF probability of STT-read

the temperature and the number of CiM operands, the more redundant cells are required for the RDF probability to reach the target RDF probability.

For the CiM-8 however, even utilizing redundant cells fails to decrease the RDF probability. In CiM-4 operation, for instance, increasing the number of redundant cells from 1 to 2 results in 7.97x and 3.96x RDF probability improvement at $25°C$ and $125°C$, respectively. However, for the case of CiM-8 at $125°C$ even increasing the number of redundant cells to 9 results in no RDF probability improvement. Hence, we can state that the method of *encoding the data in the redundant cells* cannot improve the RDF probability beyond CiM-4 operations.

We have also evaluated adjusting the MTJ device parameters, namely RA and TMR based on tuning the material stack and fabrication process [24]. The TMR can also be increased up to 220% in the room temperature [73]. The baseline RA and TMR are 7.5 and 150%, respectively [38]. As Table 3.5 shows, increasing the RA increases the minimum RDF probability. The impact of increasing the TMR is shown in Table 3.6. Increasing the TMR is highly effective for RDF probability improvement of the STT-Read operation. However, as it can be seen in Table 3.6, similar to the redundant cell encoding, increasing the TMR fails to decrease the RDF probability of CiM-8 operation. Therefore, we can see that the high RDF probability, limits the scalability of the *scouting STT-CiM*.

3.5. Conclusion

STT-CiM is a potential solution to improve the performance of data-intensive applications. However, the smaller off/on ratio and asymmetrical process, and temperature variations of the resistance states of the MTJs can severely impact the reliability of CiM operations. In this chapter, we have highlighted the impact of the reference resistance on the reliability of *scouting STT-CiM*. We first have performed a detailed statistical read decision failure analysis. By designing a proper referencing structure for CiM sensing as well as redundancy in the bit-cell, we have shown that the failure probability of CiM operations can be reduced significantly. Our results and analysis show that using a simpler resistive model for bit-cells is sufficient for the optimization of reference circuitry. However, a more comprehensive model that includes the impact of access transistor variations is required for accurate failure rate analysis. We have also investigated the impact of device parameter tuning and redundant mapping on the failure rate of

Table 3.5.: The minimum RDF probability for different values of the RA. The TMR is 150% and the cell model is (Tr+R)

T = 25°C				
RA Value ($\Omega\mu m^2$)	STT-Read	CiM-2	CiM-4	CiM-8
7.5	$1.86E-8$	$7.75E-5$	$1.59E-3$	$2.2E-1$
15	$2.32E-8$	$1.43E-4$	$1.86E-3$	$2.19E-1$
30	$7.64E-8$	$2.91E-4$	$2.86E-3$	$2.19E-1$
T = 125°C				
7.5	$4.86E-8$	$1.01E-3$	$4.76E-3$	$2.2E-1$
15	$2.94E-6$	$9.69E-4$	$5.12E-3$	$2.2E-1$
30	$6.63E-6$	$1.57E-3$	$6.45E-3$	$2.22E-1$

Table 3.6.: The minimum RDF probability for different values of the TMR. The RA is 7.5 $\Omega\mu m^2$ and the cell model is (Tr+R)

T = 25°C				
TMR	STT-Read	CiM-2	CiM-4	CiM-8
150%	$1.86E-8$	$7.75E-5$	$1.59E-3$	$2.2E-1$
220%	$2.17E-12$	$2.14E-6$	$2.25E-4$	$2.19E-1$
T = 125°C				
150%	$4.86E-8$	$1.01E-3$	$4.76E-3$	$2.2E-1$
220%	$4.76E-9$	$3.99E-5$	$1.1E-3$	$2.19E-1$

CiM with more operands. .Our analysis shows that there is a limit to the scalability of CiM operations based on scouting logic when implemented in spintronic technologies.

4. Decision Failure Modeling and Mitigation through Voltage Scaling

Due to various variability effects, the distinct resistive levels of both STT-MRAM and ReRAM are not fixed values and follow statistical distributions. Both the STT-MRAM and ReRAM are affected by the *device-to-device variation*. Additionally, ReRAM technology is also affected by the *cycle-to-cycle variation*, which means, after each memory-write operation, the resulting resistive value of a ReRAM follows a statistical distribution. Read noise and time-dependent resistance drift have also been observed in the ReRAM technology, and they are other reasons behind the distributed resistive levels of the ReRAM devices [51], [87]. Moreover, the temperature and biasing voltages can alter the distribution of the resistive levels of these technologies in an asymmetric way. Besides, the relatively small distance between the distinct resistive levels in the STT-MRAM technology, also makes the read operation challenging in this technology.

In the concept of scouting logic, the already narrow read sense margin in STT-MRAM becomes even narrower, leading to the higher failure rate during the STT-CiM operation [77]. In the ReRAM technology, although the sensing margin is quite large at the beginning, due to the presence of the read noise, especially Random Telegraph Noise (RTN) [51], [87] and the time-dependent resistance drift [30], the sense margin becomes narrower over time, leading to the CiM failures.

In this chapter, we thoroughly study and quantitatively compare the CiM reliability in these two memristive technologies in the form of the RDF probability. We also investigate the impact of the temperature and voltage biasing on the RDF probability. Then, we propose a low-cost yet effective *voltage tuning approach* to improve the reliability of the CiM operations in both the STT-MRAM and ReRAM technologies. We also perform a temperature- and technology-aware analysis on the impact of the voltage tuning on the RDF probability, power consumption, performance, and the scalability of the CiM operations in both technologies.

The rest of this chapter is organized as follows. Section 4.1 covers the required preliminaries, followed by Section 4.2, which presents the core of our proposed voltage tuning approach and analysis of the RDF probability. Section 4.3 presents the results, and finally, section 4.4 concludes the chapter.

4.1. Preliminaries

4.1.1. Implememnting the comparator required by scouting logic

As already discussed in Section 2.4.1, the prerequisite of the scouting logic is a comparison with a reference signal. Moreover, the reference signal (reference-side) should be adjustable to select the type of logical operation. One way to implement the scouting-CiM is using the *pre-charge SA* (Figure 4.1 (b)) [25]. Due to the inequality of the resistive load at the data-side and the reference-side of the pre-charge SA, the already pre-charged nodes (*OUT* and the *OUT-BAR*) are discharging with different time constants (τ_{dis}) and the node (either OUT or OUT-BAR) which has a smaller τ_{dis} is discharged until the ground and the other one with the larger τ_{dis}, is charged until VDD. In other words, both the output and its complementary can be generated by the pre-charge SA.

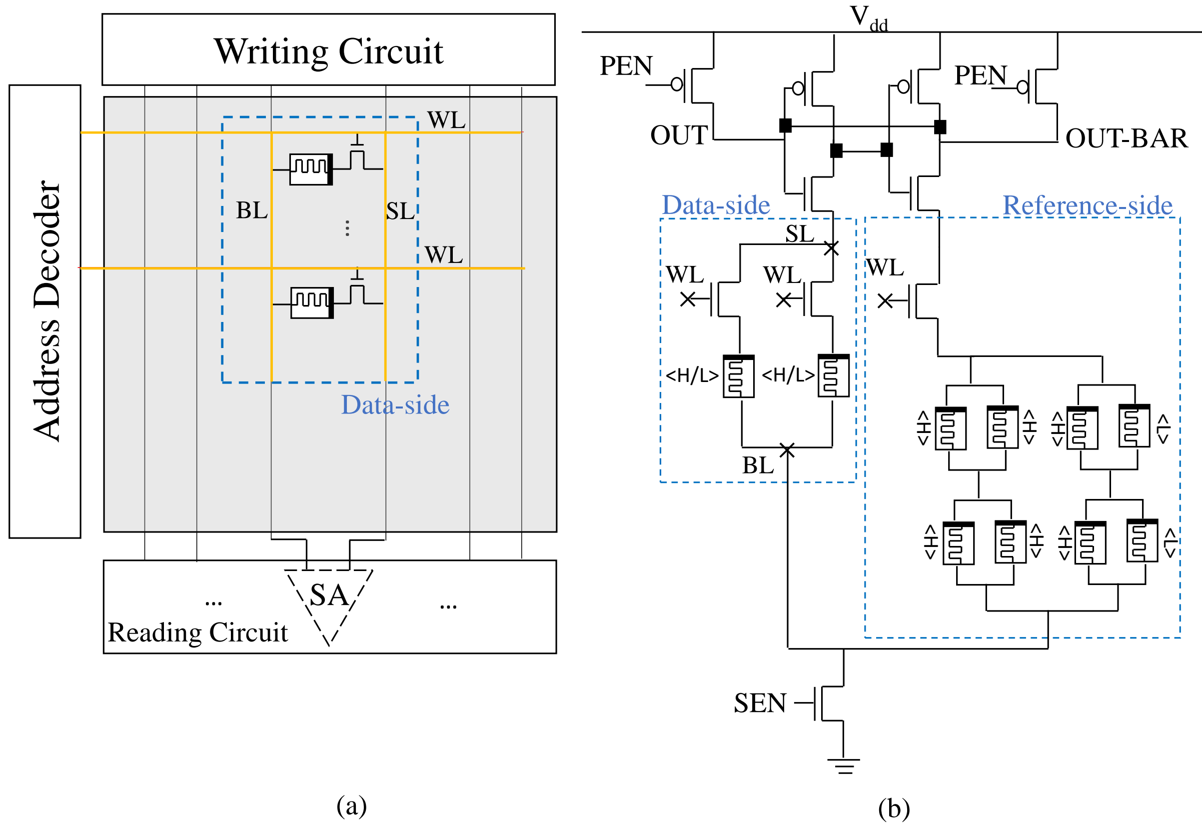


Figure 4.1.: a) Using the memristive array to perform CiM, b) Pre-charge SA to evaluate scouting-CiM AND (NAND) operation, <H> and <L>: HRS and LRS respectively

4.1.2. Temperature and voltage dependency of the resistive states in memristive devices

The statistical distribution of the HRS in both the STT-MRAM and ReRAM technologies is affected by the temperature and biasing voltage, while the distribution of the LRS in both of these technologies is roughly stable against the temperature and the biasing voltage. In a memristive *bit-cell* structure, the access transistor as well as the memristive device, is affected by the temperature- and voltage-dependent process variations. Figure 4.2 shows the effect of the temperature, while Figure 4.3 presents the impact of the V_{WL} and VDD (applied to the gate and drain of the access transistor respectively) on the equivalent resistance distribution of the STT-MRAM and ReRAM bit-cell. These results are obtained using the simulation setup outlined in Table 4.1.

4.1.3. Related work

Previous research work on improving the reliability of the memristive-read operation can be classified into three classes:

- *Trim-based approaches:* In these approaches, there are small resistors series with the reference resistor, by adjusting the values of the trim resistances, the resistivity of the reference resistance can be adjusted. For instance, authors in [56] have proposed a trimmable polysilicon-based reference for the STT-MRAM.
- *Temperature-tracking approaches:* Using the resistive elements with the temperature-dependent resistances, is beneficial to mitigate the impact of the temperature on the RDF probability. For

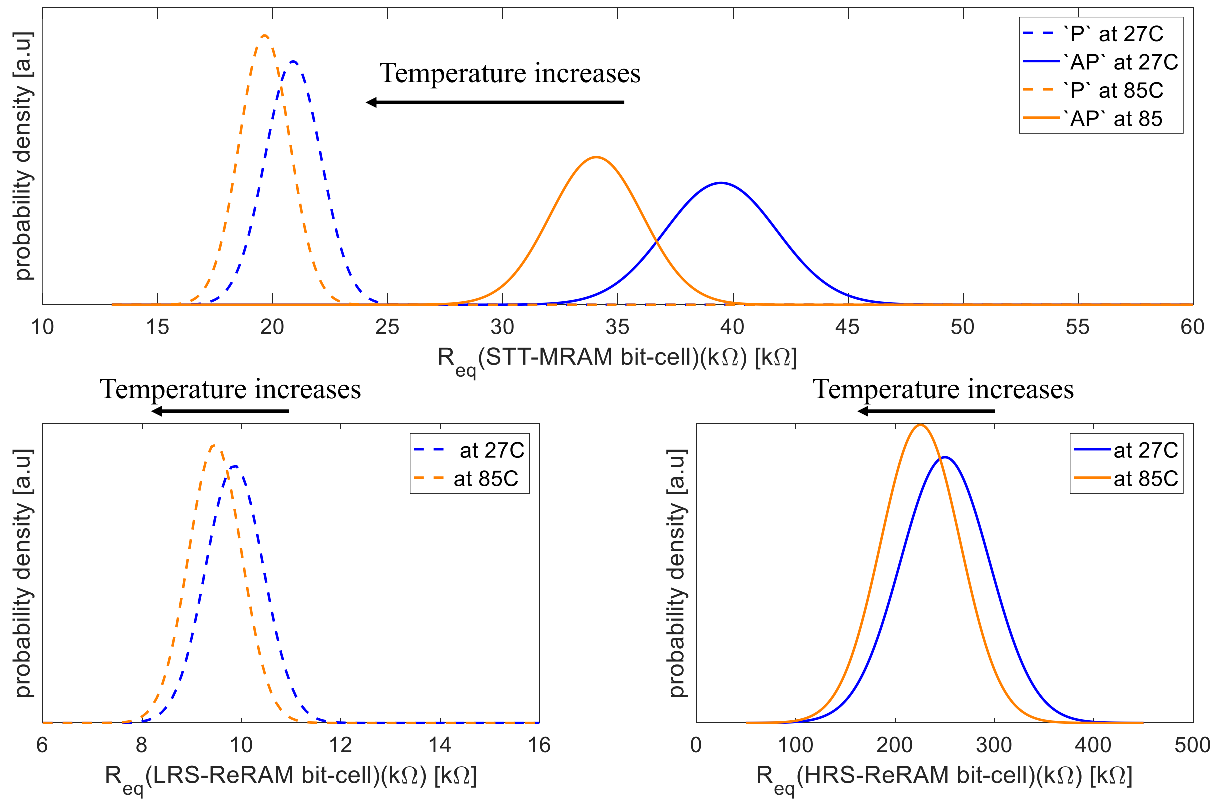


Figure 4.2.: Effect of the temperature on the resistance distributions of LRS and HRS for the STT-MRAM and ReRAM bit-cells at the nominal (VDD, V_{WL})

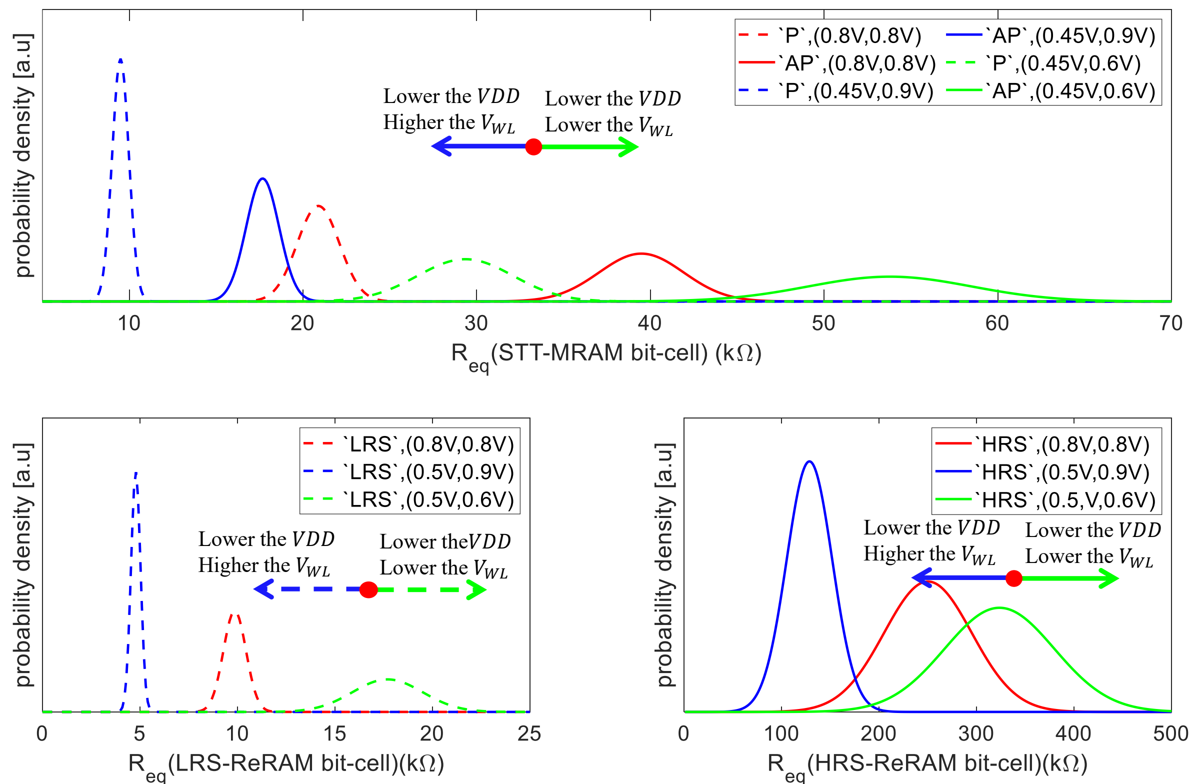


Figure 4.3.: Impact of the voltage bias on the resistance distributions of LRS and HRS (at 27 $°C$ for the STT-MRAM and ReRAM bit-cells, the distributions for the nominal (**VDD,** V_{WL}) are shown by *red*

instance, authors in [73] have proposed using an NTC reference based on an MTJ in series with an NTC resistor.

- *Load adjustment techniques:* The load of each bit-line can be adjusted by leveraging an extra reference column. To design such a reference column, authors in [49] have suggested using the combination of one 'P' and one 'AP' MTJs in the structure of the reference generator column, and hence, they have achieved the higher resolution by revisiting the reference generating.

The general method of improving the sensing reliability, as stated above, is not dedicated to STT-MRAM and can also work for other memristive technologies, including ReRAM. Although reliable computation in memristive devices requires different considerations, it has not been addressed in the previous work.

Authors in [94] have also reported the inference accuracy degradation due to the ReRAM retention, and have tried to compensate it by re-training the fully connected layers with the small portion of data. However, since the write operation is rather costly in the memristive devices, here, we focus on scouting logic which is based on the read operation.

In [82], a test algorithm to detect the STT-CiM faults was proposed. Besides the fault modeling, authors in [72] have performed yield analysis and explored fault-tolerant solutions for STT-CiM. However, due to the different nature of manufacturing defects from variability-induced RDF, such a test algorithm cannot properly work for the RDF.

4.2. Voltage tuning for reliable memristive operations

4.2.1. Main Idea

The voltage-dependency of resistance distribution of the memristive bit-cell can be an effective way to mitigate the RDF in the STT-MRAM and ReRAM technologies. However, as has been already discussed in section 4.1.2, improving the RDF probability by tuning the VDD and V_{WL} in the specific memristive technology needs to be done by considering the temperature and the number of the involved cells in the scouting logic. To implement such a voltage tuning mechanism, we propose a *programmable level shifter* [10]. The ambient temperature (which can be sensed by a temperature sensor) and the number of the activated cells (1 for memristive-read and 'N_{OP}' for the scouting logic) are used for determining the voltage levels of the VDD and V_{WL}. The VDD is a global signal and needs to be changed for the entire memory array. However, the level of the V_{WL} needs to be shifted locally in the address decoder module.

As the variation parameters of the memristive bit-cell in the STT-MRAM and ReRAM technologies are affected by the temperature as well as the voltage bias, we propose voltage tuning of the design parameters: *VDD* and V_{WL} (see Figure 4.1 (b)). In other words, we try to compensate for the effect of the temperature (as an uncontrollable environmental parameter), and multiple inputs of scouting-CiM operation on increasing the RDF probability by tuning the *VDD* and V_{WL}. According to our variation analysis, the lower the *VDD* and the higher the V_{WL} result in narrower distributions for binary resistive states in both the STT-MRAM and ReRAM bit-cells. Hence, with such corner biasing, the RDF probability can be effectively decreased for multiple inputs of scouting-CiM and at different temperatures. Please note that the operating corner of the (VDD, V_{WL}) is only applied at the sensing time, and writing circuitry is not affected by the proposed voltage tuning approach.

Since such voltage tuning affects the latency of memristive-CiM operation as well as its power consumption, it is a design trade-off to reduce RDF probability with negligible impact on the performance and power. Therefore, the voltage tuning needs to be done such that it meets the RDF requirements while minimizing the impacts on performance and power consumption.

4.2.2. RDF probability with voltage tuning

The typical method for obtaining the RDF probability is to conduct the Monte Carlo analysis on the entire pre-charge SA. However, as the target failure rate for the logical-CiM operation is considered to be less than 1e-4, a minimum number of the Monte Carlo analysis needs to be at least 10k, and for some combinations of (VDD, V_{WL}) which result in narrower variations, orders of magnitude more Monte Carlo analysis is required to capture a non-zero number of RDF probabilities. Such Monte Carlo analysis for the multiple pairs of (VDD, V_{WL}) is extremely time-consuming and effortful.

To decrease the number of Monte Carlo analyses, we propose a 4-step statistical approach as follows:

1. Perform the Monte Carlo analysis for each state of the STT-MRAM and ReRAM bit-cells (LRS or HRS), at different temperatures (high and low) and for each pair of (VDD, V_{WL}).
2. Fit the Monte Carlo samples of the first step to a known statistical distribution which is a normal distribution. Based on *Cullen and Frey graph*, the assumption of a normal distribution is a good enough fit for both the STT-MRAM and ReRAM bit-cells at different temperature and voltage combinations.
3. Obtain the statistical distribution of parallel connection of N_{OP} memristive bit-cells. N_{OP}-memristive-CiM operation has $N_{OP}+1$ states; S_0*: N_{OP} bit-cells in the LRS and 0 bit-cells in the HRS.* $S_{N_{OP}}$*: 0 bit-cells in the LRS and N_{OP} bit-cells in the HRS.*
4. By targeting the reference resistance to be located between the two adjacent resistive states of S_a and S_b, we sweep the reference resistance R_{Ref} from the mean (μ) of S_a to mean (μ) of S_b and for each R_{Ref}, calculate the RDF probability through Equation 4.1,

$$
\begin{aligned}
P(RDF)_{R_{Ref}} = & \sum_{i=0}^{a} \Big(P(S_i) \times (1 - CDF(S_i|_{x=R_{Ref}})) \Big) \\
& + \sum_{j=b}^{N_{OP}} \Big(P(S_j) \times (CDF(S_j|_{x=R_{Ref}})) \Big)
\end{aligned}
\tag{4.1}
$$

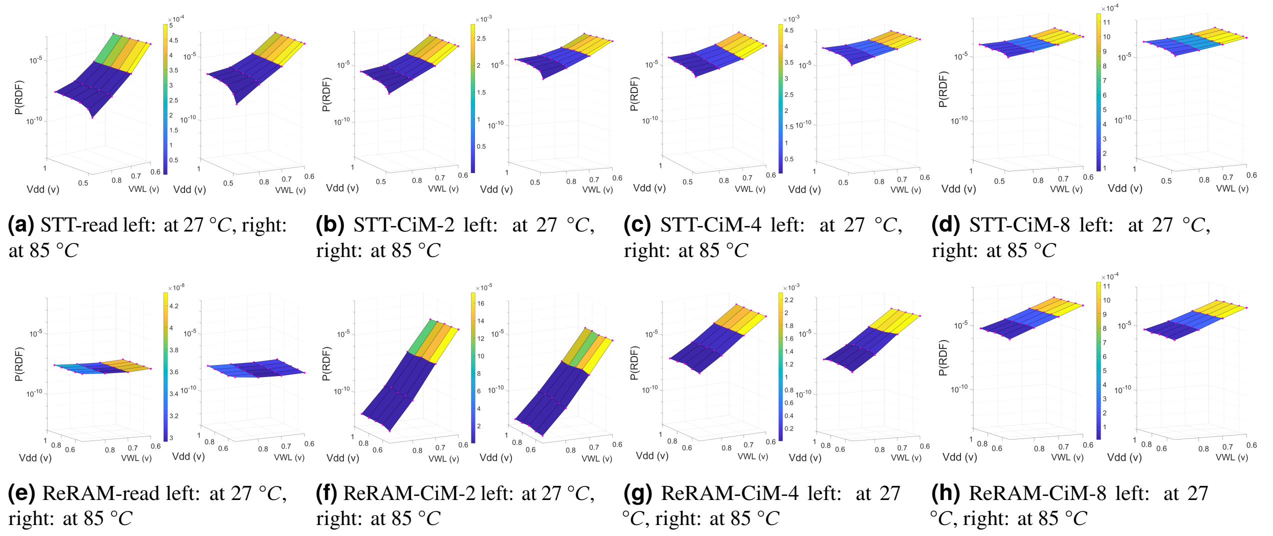
According to Equation 4.1, for a N_{OP}-CiM in which the reference resistance is located between S_a and S_b resistive levels, RDF happens if the resistance of either $S_0..S_a$ becomes greater than the reference resistance (R_{Ref}) or the resistance of either $S_b..S_{N_{OP}}$ becomes less than R_{Ref}. The minimum RDF probability through Equation 4.1 has been considered as the RDF probability corresponds to each voltage combination and temperature.

4.2.3. Impact of the temperature and (VDD, V_{WL}) on the entire memory array

Temperature and VDD which are globally distributed through the entire memory chip, have a great impact on the power consumption and performance of the entire memristive chip. The value of V_{WL}, however, only affects the power consumption and performance of the pre-charge SA. The circuit-level impact of the temperature and (VDD, V_{WL}) on the pre-charge SA can be analyzed through the electrical-level simulation (with SPICE). To investigate the performance and power consumption of the entire STT-MRAM and ReRAM array, we use NVSim [28].

4.2.4. Time-dependent resistance drift in ReRAM technology

Both the LRS and HRS distributions are broadening over time (standard deviations (σ) increases over time). The mean (μ) of LRS and HRS are also shifting to the higher and lower values, respectively. Such resistance drift will be saturated, in other words, the rate of the resistance shift is decreasing over time [30].



(a) STT-read left: at 27 °C, right: at 85 °C
(b) STT-CiM-2 left: at 27 °C, right: at 85 °C
(c) STT-CiM-4 left: at 27 °C, right: at 85 °C
(d) STT-CiM-8 left: at 27 °C, right: at 85 °C
(e) ReRAM-read left: at 27 °C, right: at 85 °C
(f) ReRAM-CiM-2 left: at 27 °C, right: at 85 °C
(g) ReRAM-CiM-4 left: at 27 °C, right: at 85 °C
(h) ReRAM-CiM-8 left: at 27 °C, right: at 85 °C

Figure 4.4.: Impact of the voltage tuning and the temperature on the RDF probability in STT-MRAM and ReRAM technologies, STT-CiM operation: logical AND (NAND), ReRAM-CiM operation: logical OR (NOR), the measured RDF probability are shown with '*'

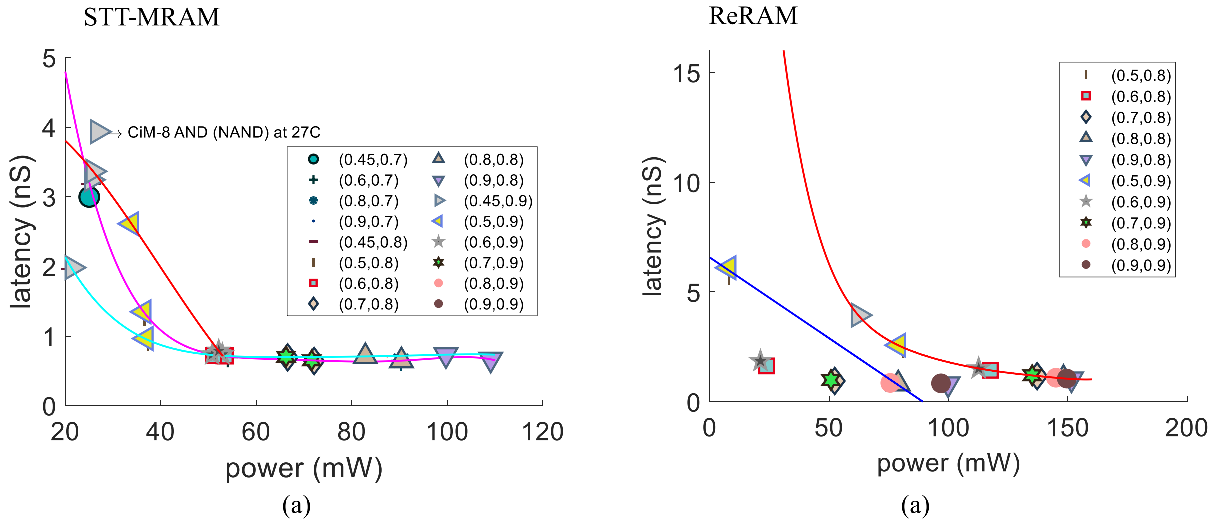


Figure 4.5.: Pareto front (a) CiM-2,4,8 operations for an array of 256kB STT-CiM, (b) CiM-8 operations for an array of 256kB ReRAM-CiM, for various (VDD (V), V_{WL} (V)) combinations

To model such resistance shift, we have assumed exponentially increasing time-dependent resistance for the mean (μ) and standard deviation (σ) of the LRS as well as the (σ) of the HRS, and exponentially decreasing time-dependent resistance for the (μ) of HRS: ($R(t) = R_{Sat} - (R_{Sat} - R_0) \cdot e^{\frac{-t}{\tau_r}}$). The R_{Sat} is the final value of the modeled parameter while R_0 is its initial value, and τ_r is the time constant of the model. According to [30], the change due to the resistance drift phenomenon depends on device-level characteristics of the specific ReRAM technology, such as the configuration of the conducting filament. We have considered τ_r as a period, in which the modeled resistive value has changed for 5%.

Table 4.1.: Simulation setup tools and parameters

Simulation tool	Cadence Virtuoso
Technology node for CMOS	GF 22FDX
Standard VDD for CMOS	0.8 V
MTJ model [38]	-Radius = 20 nm - Barrier Material = MgO - Free layer thickness = 1.3 nm - Oxide layer thickness = 1.48 nm - RA = $7.5\Omega\mu m^2$ - Nominal TMR = 150%
VCM-based device model: *JART VCM v1b Read variability* [71], [87]	-Radius of the filament = 45 nm - Length of the disc region = 0.6 nm - Initial oxygen vacancies concentration in the disc [$10^{26}/m^3$] for LRS = 3, for HRS = 0.009 - (Maximum and Minimum) oxygen vacancy concentration in the disc [$10^{26}/m^3$] = (20, 0.008)
Monte Carlo analysis	1k at each temperature and (VDD, V_{WL})
Array-level estimation of the power and delay	NVSim [28]

4.3. Results and discussion

4.3.1. Experimental setup

We have used the tools and model parameters as indicated by Table 4.1. To study the process variation-induced distribution of the STT-MRAM bit-cell, we have used the variation model in [38], and for the ReRAM bit-cell, we have considered 10% Gaussian distribution on the four parameters as follows: the radius of the filament, length of the disc region, maximum and minimum oxygen vacancy concentration in the disc. According to the ReRAM model, which we have used (outlined in Table 4.1), the aforementioned parameters are those that are affected by the process variation.

To analyse the impact of the V_{WL} and VDD on different features of the memristive module, for both the memristive technologies, V_{WL} is swept from 0.6 V to 0.9 V. For the STT-MRAM and ReRAM technologies, the VDD is swept from 0.45 V to 0.9 V and 0.5 V to 0.9 V, respectively. According to the CMOS library (outlined in Table 4.1), the standard VDD is 0.8 V, and it is not allowed beyond 0.9 V. On the other hand, decreasing the VDD outside the aforementioned range prevents the circuit from proper operation.

Due to the lower distance of the binary resistive levels in the STT-MRAM, compared to the ReRAM, we have adjusted the bit-cell resistive distribution at (VDD, V_{WL} = 0.8 V, 0.8 V) according to the target RDF probability for STT-read operation: P(RDF) < 1e-5 at 25°C and 125 °C as reported in [58]. For the array-level simulation with NVSim, we have considered a 256kB STT-MRAM and ReRAM array, with a 64-byte data width and sub-array organization of 256 rows × 128 columns.

4.3.2. RDF probability, power and performance of the memristive-based operations

Figure 4.4 shows the RDF probability of STT-MRAM and ReRAM for standard memory-read and scouting CiM operations as a function of (VDD, V_{WL}) at low and high temperatures. Due to the extremely narrow sense margin of the scouting STT-CiM operation, only logical AND (NAND) operation can meet the RDF probability threshold of 1e-4. However, for the scouting ReRAM-CiM operation, both the logical AND (NAND) and OR (NOR) operations are feasible from the reliability point of view. Since the RDF probability of the ReRAM-AND (NAND) is orders of magnitude less than the ReRAM-OR (NOR), the worse-case RDF probability of logical ReRAM-OR (NOR) has been considered in Figure 4.4. Unlike the STT-MRAM technology, the ReRAM technology shows promising scalability for implementation of the scouting CiM in which the output is generated in the periphery circuitry (stateless logic aka CiM-Periphery [86]). On the other hand, due to the lower write energy of the STT-MRAM compared to ReRAM (through electrical-level simulation on a single cell, 0.43 pj compared to 1.1 nj), STT-MRAM can be considered as a better candidate for implementation of the CiM operation, in which the output result is generating within the crossbar array (stateful logic aka CiM-Array [86]). In other words, the technology of the memristive device needs to be aligned with the requirements of the CiM architecture.

As already discussed in section 4.1.2, increasing the temperature has two opposing impacts on the RDF; first, decreasing the distance between the means (μ) of the distinct resistive levels, and second, narrowing the distributions by decreasing the standard deviations (σ) of the distinct resistive levels. According to Figure 4.4, for the STT-MRAM technology, the first (destructive) impact of the temperature is dominant, hence, the higher the temperature, the higher the RDF probability. While, for the ReRAM technology, the second (constructive) impact of the temperature is dominant and we can see the decrease of the RDF probability while the temperature is increasing.

The impact of the number of the CiM-operands (N_{OP}) on the RDF probability is similar to the effect of the high temperature and it is technology-dependent as well. The higher the number of operands, in STT-MRAM technology is destructive from the RDF point of view. For the ReRAM technology, however, the RDF probability is not monotonically increasing with N_{OP} and we observe a minimum RDF probability value is for the ReRAM-CiM-2 operation.

Variation-optimized voltage tuning can significantly improve the RDF probability. However, the impact of such voltage tuning on the performance and the power consumption of the entire memory (CiM) module needs to be investigated. In other words, there is a trade-off between power consumption, performance, and the RDF reliability. Hence, the concept of *Pareto optimum* can be applied to the power consumption and the latency of the entire memory module while meeting the RDF reliability criterion. Figure 4.5 shows the Pareto points for the STT-CiM operations and ReRAM-CiM-8 at the high and low temperatures. The result of Figure 4.5 is based on the memristive array simulation with NVSim [28]. According to Figure 4.5, by considering the RDF probability as the design constraint, higher performance can be achieved at the cost of higher power consumption.

We have also analyzed the impact of time-dependent resistance drift in ReRAM. Due to this phenomenon, the RDF probability degrades over time. Table 4.2 shows the rate of RDF probability degradation for the nominal and variation-optimized bias voltage combination for ReRAM-CiM-8 operation. We can see that voltage tuning can not only improve the RDF probability at each point of time but also it can significantly decelerate the RDF probability degradation rate.

Table. 4.3 shows an overview of the impact of voltage tuning on the standard memory operation as well as the CiM-8 operation in STT-MRAM and ReRAM technologies at 27 $°C$. For the memory-read operation in both STT-MRAM and ReRAM technologies and the ReRAM-CiM operations, moving to the variation-optimized voltage combination improves the RDF probability and saves power while it has a negative impact on the performance. However, for the STT-CiM-8 operation, the trade-off between the RDF probability, power, and performance needs more investigation. The direct CiM-8 operation by activating 8 STT-MRAM bit-cells at a time is not possible, since it cannot meet the RDF probability

Table 4.2.: Rate of the RDF probability degradation for ReRAM-CiM-8 operation

RDF degradation rate in a period of 100 $\times \tau_r$		
Voltage biasing	**T = 27**$°C$	**T = 85**$°C$
Nominal (VDD, V_{WL} = 0.8 V, 0.8 V)	$\frac{3.7e-6}{100\times\tau_r}$	$\frac{1.9e-6}{100\times\tau_r}$
Variation-optimized (VDD, V_{WL} = 0.5 V, 0.9 V)	$\frac{0.21e-6}{100\times\tau_r}$	$\frac{0.36e-6}{100\times\tau_r}$
Improvement	17.62×	5.28×

Table 4.3.: Array-level power and latency for both the STT-MRAM and ReRAM technologies in nominal and the variation-optimized voltage combination, considering the memory-read, and CiM-8 at 27 $°C$ for the array-size of 256 kB

Implementation	Power	Latency
STT-read (With *nominal* voltage biasing)	93.10 mW	624 ps
STT-read (With *variation-optimized* voltage biasing)	47.96 mW	1.72 ns
STT-CiM-8, broken to CiM-2 (With *nominal* voltage biasing)	139.40 mW	21.89 ns
STT-CiM-8, direct	27.60 mW	3.92 ns
ReRAM-read (With *nominal* voltage biasing)	96.51 mW	602 ps
ReRAM-read (With *variation-optimized voltage biasing*	81.68 mW	1.01 ns
ReRAM-CiM-8 (With *nominal* voltage biasing)	78.87 mW	810.01 ps
ReRAM-CiM-8 (With *variation-optimized* voltage biasing)	7.9 mW	6.09 ns

constraint. To implement the STT-CiM-8 operation at the nominal voltage combination, it should be broken into a series of STT-CiM-2 operations, however, this method requires multiple STT-write operations which results in high power and performance penalty. Performing the STT-CiM-8 operations at the variation-optimized voltage combination, however, can improve power and performance for 5.05× and 5.58×, respectively, while still meeting RDF constraints.

4.4. Conclusion

Reliable memristive-CiM implementation is challenging due to variability effects during the manufacturing and the run-time operation of both the memristive and CMOS elements. In this chapter, we have performed a thorough reliability analysis of scouting logic and based on that, proposed a voltage-tuning approach with the help of the level shifter. We have also analysed the impact of this approach on the power and performance of the entire memristive-CiM module. We have shown the superiority of the ReRAM over the STT-MRAM for scouting logic implementation while, the higher ReRAM-write energy compared to STT-MRAM, limits the ReRAM technology for stateful CiM implementations.

5. Reliability Analysis and Mitigation for Analog Computation-in-memory: from Technology to Application

The performance promise of NVM-CiM can be jeopardized by inherent reliability challenges. As discussed in Section 2.5, there are technology and circuit-level factors that exacerbate the reliability of NVM-CiM. In the analog NVM-CiM, there is a strong interaction between different levels of abstraction, from the low-level technology all the way to the high-level application. Similarly, analyzing and mitigating its reliability issues also needs an end-to-end investigation. The holistic technology to application reliability investigation of the analog NVM-CiM is the shortcoming of the existing literature [75], [80], [95], [98], [100], [109].

In this chapter, we continue focusing on the reliability of the analog NVM-CiM in terms of *decision failure* which makes the output of the NVM-CiM kernels incorrect. Therefore, we propose a systematic flow from the technology level all the way to the application level, targeting error generation and propagation across multiple design stacks and abstraction levels. At the technology level, we model the effects of process variation on the NVM devices. At the circuit level, we additionally consider the effects of circuit-level parameters including interconnect parasitic, sensing offset, crossbar size, and the number of activated operands on the decision failure. Finally, we statistically model the consequent error. To monitor the propagation of the error to the higher levels of abstraction, namely architecture, and application, we consider the full-system architecture simulation by extending the well-used gem5 framework with analog NVM-CiM. Moreover, we perform fault injection on this CiM-enabled gem5 infrastructure by executing real-world applications. This way, we can quantitatively evaluate the masking capability of the high-level applications. Our comprehensive reliability analysis enables us to explore effective hardware- and application-level knobs to decrease decision failure during the analog NVM-CiM.

The rest of this chapter is organized as follows, in Section 5.0.1, we review the existing related work. In Section 5.1, we discuss our proposed methodology, followed by Section 5.2 which includes our results. Finally, Section 5.3 concludes the chapter.

5.0.1. Related works

Improving the reliability of the NVM-CiM has been the focus of multiple research works in recent years. However, the main shortcoming of the related work is the lack of end-to-end analysis. Authors in [95] targeted the hardware-level decision failure of scouting Boolean NVM-CiM and proposed fine-tuning of

Table 5.1.: Summary of the related work

Work	Analysis			
	Technology-level	**Circuit-level**	**Architecture-level**	**Application-level**
[95], [109]	+	+	-	-
[75], [80], [98], [100]	+	+	-	+
This work	+	+	+	+

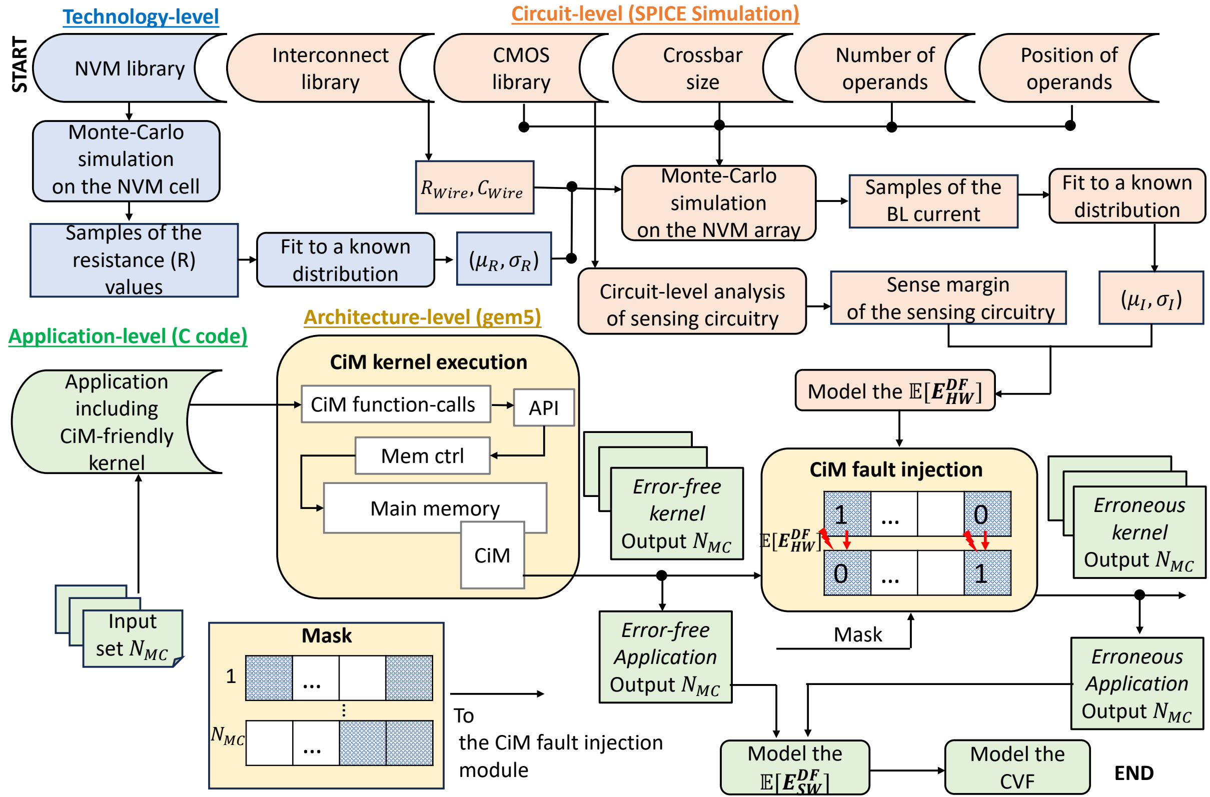


Figure 5.1.: The end-to-end flow for reliability analysis during NVM-CiM, μ and σ show the mean and standard deviation of a statistical distribution, N_{MC} shows the number of Monte Carlo simulations, $\mathbb{E}[E_{HW}^{DF}]$ and $\mathbb{E}[E_{SW}^{DF}]$ show the hardware- and software-level expected errors, respectively

the reference signal to mitigate the decision failure. The idea of voltage tuning was the main contribution of the work presented in [109] to mitigate the decision failure during scouting Boolean NVM-CiM. Both [95] and [109] limited their investigation to the technology and circuit levels and did not link the circuit to the architecture and application levels.

Further, works presented in [75], [80], [98], [100] considered the MAC-based NVM-CiM and the reliability target for these works is IR drop. Although these works comprehensively analyzed the technology, circuit, and application levels, they did not consider the architectural level analysis. Moreover, these works used a (one-bit) comparator as their sensing circuitry, and they did not consider the effect of the (multi-bit) ADC. However, selecting the ADC can improve the overall inference accuracy compared to the comparator [104], [118]. In this chapter, we perform a comprehensive end-to-end analysis of the reliability of the NVM-CiM targeting decision failure. Table 5.1 summarizes the related work.

5.1. Proposed reliability analysis framework

In this section, we propose an end-to-end framework for the reliability analysis of NVM-CiM targeting decision failure. In this regard, we first model the $\mathbb{E}[E_{HW}^{DF}]$ by taking into account multiple technology- and circuit-level parameters. At the architecture level, we extend the well-used gem5 infrastructure with the analog NVM-CiM to enable the full-system end-to-end analysis. Additionally, the extended gem5 framework is augmented with the Monte Carlo fault injection. The error value ($\mathbb{E}[E_{HW}^{DF}]$) is the prerequisite of the fault injection that is already modeled at the circuit level. Finally, to evaluate the propagation of the $\mathbb{E}[E_{HW}^{DF}]$ to the software (SW) level, i.e., obtaining the $\mathbb{E}[E_{SW}^{DF}]$, we run high-level CiM-friendly applications on the CiM-enabled gem5 infrastructure. The $\mathbb{E}[E_{SW}^{DF}]$ is the expected relative

error at the software output. Figure 5.1 shows our end-to-end reliability analysis flow and its details will be discussed in the following subsections.

5.1.1. Obtaining the hardware error

To obtain the $\mathbb{E}[E_{HW}^{DF}]$, we perform the electrical-level Monte Carlo simulation (using SPICE) and fit the samples of the BL current to a known distribution. The distribution of the BL current depends on the multiple technology- and circuit-level parameters. At the technology level, the effect of the process variation on the NVM devices is dominant, resulting in their resistance distribution. At the circuit level, the number of activated rows, the offset of the ADC/comparator, and the line parasitic resistance are the pivotal parameters. Besides, due to the line parasitic resistance, other parameters such as the position of the activated operands in the crossbar relative to the sensing circuitry, as well as the crossbar size also affect the $\mathbb{E}[E_{HW}^{DF}]$. In the following, we elaborate on each parameter.

5.1.1.1. NVM technology

Generation of the CiM output in the periphery circuitry only includes sensing of the current [86]. Therefore, we abstract each NVM technology by its resistance distribution in LRS and HRS. By performing an electrical-level simulation (using SPICE) on each NVM technology and fitting the resistance samples to a known distribution, the $(\mu_{LRS}, \sigma_{LRS})$, $(\mu_{HRS}, \sigma_{HRS})$ can be obtained.

5.1.1.2. Number of the activated rows

Increasing the number of operands during both MAC and Boolean operations increases the $\mathbb{E}[E_{HW}^{DF}]$ since the to-be-distinguished current levels are getting closer when more rows are activated.

5.1.1.3. Offset of the ADC/comparator

As discussed in Section 2.5, the reduction of the offset decreases the $\mathbb{E}[E_{HW}^{DF}]$, however, at the cost of more energy and/or area.

The principle of the NVM-CiM is sensing the cumulative resistance from the operands involved in the computation. The increase in the line parasitic resistance interferes with this resistance sensing by impairing the sensing margin. However, the effect of line parasitic resistance on the $\mathbb{E}[E_{HW}^{DF}]$ is not straightforward, and the final impact of the accumulated line parasitic resistance on the $\mathbb{E}[E_{HW}^{DF}]$ needs comprehensive electrical-level simulations (using SPICE). In the following subsection, we discuss the effect of parameters related to line parasitic resistance affecting the $\mathbb{E}[E_{HW}^{DF}]$.

5.1.1.4. Interconnect technology

The smaller the cross-sectional area of the interconnect, the larger its parasitic resistance. So, in more advanced technology nodes, the line parasitic resistivity increases [126].

5.1.1.5. Crossbar size

The larger the size of the crossbar, the higher the entire line resistance.

5.1.1.6. Position of the activated operands

If the activated rows are closer to the sensing circuitry, the effective parasitic line resistance is smaller consequently decreasing the $\mathbb{E}[E_{HW}^{DF}]$.

In the case of using ADC during MAC operation, the boundaries, i.e., the quantization levels are fixed. For the scouting Boolean operation, however, the reference is selected between two input states and varies depending on the different discussed hardware configurations. Therefore, the trend of the error values is not similar in these two CiM kernels.

5.1.2. Obtaining the software error

To investigate the propagation of $\mathbb{E}[E_{HW}^{DF}]$ at both the architecture and application levels, our *initial step* is to enhance and extend the architectural simulator framework. In this chapter, we develop our framework based on CiM-enabled gem5 as detailed in [133]. Within this framework, the CiM module is integrated into the main memory and signaled by the memory controller. Also, the main memory is reconfigured to perform CiM by issuing a write into a specific memory address.

We extend the NVM-CiM module to support both MAC and Boolean operations, allowing programmers to flexibly specify the number of operands. Moreover, we modify the Application Programming Interface (API) (written in inline assembly and C code) to generate the necessary configuration to support the execution of the extended CiM instructions. With these adjustments, programmers can use the corresponding *CiM function calls* in their kernels to allow it to be executed on the CiM module. Additionally, we augment the NVM-CiM module with realistic hardware details such as memory size and timing information.

In the *second step*, we need to add the fault injection capability to our NVM-CiM gem5 framework. As discussed in section 2.5, the main source of decision failure is the accumulated impact of process variations in NVM devices and circuit-level imperfections. Therefore, simulating the effect of decision failure by implementing bit-flips on the input operands is an imprecise model. To address this, we perform the fault injection on the NVM-CiM outputs.

As shown in Figure 5.1, performing the fault injection to account for the decision failure requires two basic pieces of information: the error value and the precise output segments that are susceptible to errors. The error values are the $\mathbb{E}[E_{HW}^{DF}]$ which can be abstracted from the hardware level and the *random mask* technique helps to identify the output parts that undergo the fault injection.

5.1.2.1. Mask generation during MAC and scouting Boolean operation

In the case of MAC operation, it is more likely that the right side of the results, i.e., the least significant bits are affected by the $\mathbb{E}[E_{HW}^{DF}]$. In the case of the scouting Boolean operation, however, all the generated bits are independent of each other. Hence, they are equally likely to be affected by the $\mathbb{E}[E_{HW}^{DF}]$.

5.1.2.2. Measuring the masking capability of the entire application

A CiM kernel is part of an application. So, to measure the masking capability of the entire application to the decision failure, we can execute a sufficient number of architectural-level Monte Carlo (N_{MC}) fault injections while executing CiM-friendly applications. The $\mathbb{E}[E_{SW}^{DF}]$ is the expected *relative* inequality between error-free and erroneous outputs of the application that can be formulated as Equation 5.1.

$$\mathbb{E}[E_{SW}^{DF}] = \frac{1}{N_{MC}} \times \sum_{i=1}^{N_{MC}} \frac{|\text{error-free}_i - \text{erroneous}_i|}{\text{error-free}_i} \tag{5.1}$$

To measure the masking capability of the CiM-friendly application, we define CiM Vulnerability Factor (CVF) as Equation 5.2, in an analogous way that the Architectural Vulnerability Factor (AVF) was defined [9]. CVF is simply the probability that the error in the output of the CiM module causes a

Table 5.2.: End-to-end simulation setup and tools

Technology-level parameters	
MTJ model [38]	-Radius = 20 nm - Barrier Material = MgO - RA = 7.5Ωμm^2 - Nominal TMR = 150% - HRS, LRS = 11.56 kΩ, 5.967 kΩ
ReRAM model: JART VCM Read variability [71], [87]	-Radius of the filament = 45 nm - Length of the disc region = 0.6 nm - HRS, LRS = 105.51 kΩ , 1.975 kΩ
Circuit-level parameters	
Simulation tools	Cadence virtuoso
CMOS technology	Global foundry 22FDX
Interconnect technologies [112], [128]	Older: 22nm, more advanced: 5nm
Architectural-level parameters	
Instruction set architecture (ISA)	x86
CPU model	- In order - With buffered four-stage pipeline - Frequency: 1GHz
Memory size	256 MB
Application-level parameters	
Benchmark consists of MAC kernel	2mm from PolyBench
Benchmark consists of Boolean kernel	Database query

noticeable error (according to Equation 5.1) in the output of the program. A smaller CVF indicates a stronger masking capability for an application.

$$CVF = \frac{\mathbb{E}[E_{SW}^{DF}]}{\mathbb{E}[E_{HW}^{DF}]} \tag{5.2}$$

5.1.2.3. Hardware-level parameters affecting the CVF

Out of the five hardware-level parameters that affect the $\mathbb{E}[E_{HW}^{DF}]$ (1, 2, 4-6 in Section 5.1.1), only the number of the operands can be observed at the architecture level, i.e., the CVF can solely provide insight into the number of operands. The NVM and interconnect technology, along with the size of the crossbar are predetermined during the circuit-level design phase and modeled in $\mathbb{E}[E_{HW}^{DF}]$.

5.2. Results and Discussion

5.2.1. Technology-level analysis

For our end-to-end simulations, we use the tools and parameters outlined in Table 5.2. For the NVM technology, we use the measurement-verified Verilog-A model of two representative technologies of STT-MRAM and ReRAM. STT-MRAM provides relatively small $\Delta\mu_R (= \mu_{\mathrm{HRS}} - \mu_{\mathrm{LRS}})$ and $\sigma_{\mathrm{LRS,\,HRS}}$

while ReRAM provides large $\Delta\mu_R$ and $\sigma_{\text{LRS, HRS}}$. The $\Delta\mu_R$ and $\sigma_{\text{LRS, HRS}}$ for the other technologies such as PCM lies between the STT-MRAM and ReRAM [42].

5.2.2. Circuit-level analysis

We perform the circuit-level simulation using SPICE and parameters outlined in Table 5.2 to quantitatively account for the impact of the following circuit-level parameters on $\mathbb{E}[E_{HW}^{DF}]$.

5.2.2.1. RC parasitic of interconnect

For the interconnect technology, we use older 22 nm and more advanced 5 nm nodes. The smaller the node, the higher the parasitic resistance.

5.2.2.2. MAC operation

For the ADC during the MAC operation, we select a 9-bit ADC with a voltage source of 1 Volt from [114]. There is a trade-off between the offset, $\mathbb{E}[E_{HW}^{DF}]$, and the energy consumption. The larger the offset, the larger the $\mathbb{E}[E_{HW}^{DF}]$. Reducing the offset is achievable at the cost of more energy consumption. The offset-energy trade-off of the selected ADC from [114] is already optimized.

5.2.2.3. Scouting Boolean operation

For the comparison during the scouting Boolean operation, we perform the electrical-level simulation using SPICE on a comparator consisting of two back-to-back inverters. Figure 5.2 shows the trade-off between the offset, $\mathbb{E}[E_{HW}^{DF}]$, and the energy consumption during the NVM-CiM for scouting Boolean operation. As shown in Figure 5.2, with large $\Delta\mu_R$ in the case of ReRAM, the target expected error values can be met even with a much larger offset compared to the STT-MRAM. Hence, the energy consumption of the comparator can also be relaxed in the case of ReRAM. According to Figure 5.2, selecting the offset as $\sim 3\%$ can provide an optimum offset-energy trade-off.

For both the STT-MRAM and ReRAM technologies, we can use the same comparator with different respective references. Hence, the latency and energy of the comparator are almost independent of the NVM technology.

5.2.2.4. Obtaining the $\mathbb{E}[E_{HW}^{DF}]$

Figure 5.3 show the *worst-case* $\mathbb{E}[E_{HW}^{DF}]$ during the MAC operation and under the impact of five discussed parameters (1, 2, 4-6 in Section 5.1.1). As the ADC with fixed quantization levels is utilized for the MAC operation, due to the smaller σ_R values in the STT-MRAM, this technology results in a lower $\mathbb{E}[E_{HW}^{DF}]$ compared to the ReRAM. Besides, the higher number of operands exacerbates the $\mathbb{E}[E_{HW}^{DF}]$.

Moreover, Figure 5.4 shows the *worst-case* $\mathbb{E}[E_{HW}^{DF}]$ for the scouting Boolean operation (offset: 3%). For the Boolean kernel, the reference is selected between two distinct distributions. Hence, any parameter that increases the sensing margin decreases the $\mathbb{E}[E_{HW}^{DF}]$. Therefore, selecting an NVM technology with higher $\Delta\mu_R$, and decreasing the number of operands can decrease the $\mathbb{E}[E_{HW}^{DF}]$. Besides, using the interconnect technology with lower parasitic resistance, decreasing the crossbar size, and positioning the activated rows closest to the sensing circuitry are the other means to decrease the $\mathbb{E}[E_{HW}^{DF}]$.

5.2.3. Architecture- and Application-level analysis

To obtain the $\mathbb{E}[E_{SW}^{DF}]$, we conduct architectural-level fault injection and execute CiM-friendly applications within the CiM-enabled gem5 framework. We utilize the *2mm* benchmark from the PolyBench suite,

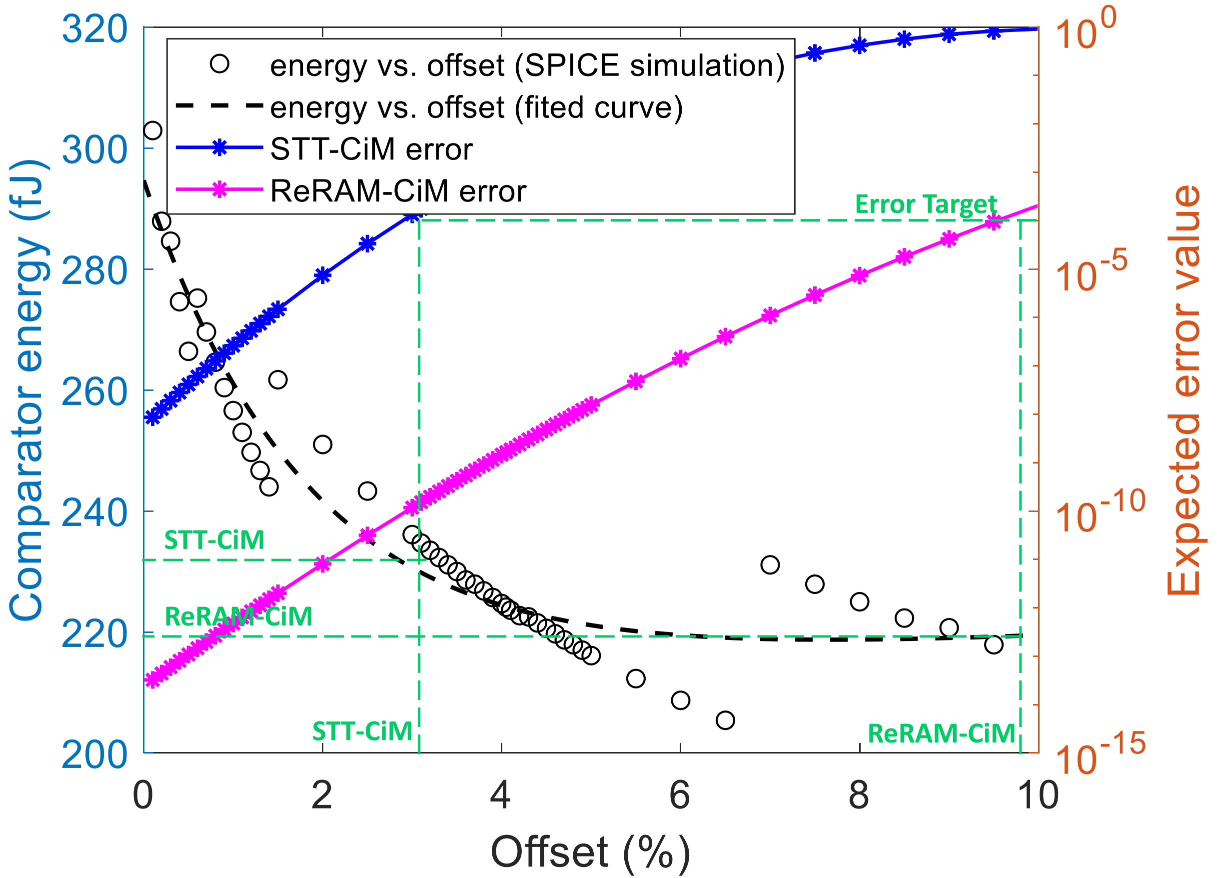


Figure 5.2.: The trade-off between energy, offset, and expected error values, for scouting Boolean NVM-CiM with two operands, crossbar size: 128 × 128, interconnect node: 5 nm, to achieve a comparable $\mathbb{E}[E_{HW}^{DF}]$ with an equal number of operands, the position of operands for STT-MRAM-CiM and ReRAM-CiM are closest to and farthest from the sensing circuitry, respectively

which performs matrix-matrix multiplication and its CiM-friendly kernel being the MAC operation. Additionally, we perform a generic *database query* application. In this *database query* application, we define a set of conditions and count the number of records that meet either all of the conditions (true logical AND) or at least one of the conditions (true logical OR). The CiM-friendly kernel in this application is the Boolean operation.

Table 5.3 shows the $\mathbb{E}[E_{HW}^{DF}]$, CVF, $\mathbb{E}[E_{SW}^{DF}]$, and normalized execution time during the execution of the *2mm* benchmark. By increasing the size of the square matrix, the *2mm* benchmark can more significantly benefit from the NVM-CiM. The larger the matrix size, the decreasing trend for the $\mathbb{E}[E_{SW}^{DF}]$, and the faster execution compared to the non-CiM (conventional architecture with no NVM-CiM module) execution.

Additionally, Table 5.4 shows the $\mathbb{E}[E_{HW}^{DF}]$, CVF, $\mathbb{E}[E_{SW}^{DF}]$, and normalized execution time for the *database query* application considering the scouting Boolean OR between the conditions (having the worst-case $\mathbb{E}[E_{SW}^{DF}]$). Compared to the *2mm* benchmark, the execution time of the *database query* is accelerated less significantly. Although the CVF during this application is small, it cannot compensate for the high $\mathbb{E}[E_{HW}^{DF}]$. During the NVM-CiM execution of the *database query*, we do not increase the number of conditions above eight. Based on Figure 5.4, increasing the number of operands beyond eight is not feasible due to the extremely high $\mathbb{E}[E_{HW}^{DF}]$.

In Table 5.3 and 5.4, we consider both the STT-MRAM and the ReRAM technologies, however, only with interconnect node of 22 nm and crossbar size of 128×128. As discussed in Section 5.1.2.3, the CVF is only impacted by the number of the operands, and the effect of other hardware parameters is modeled in

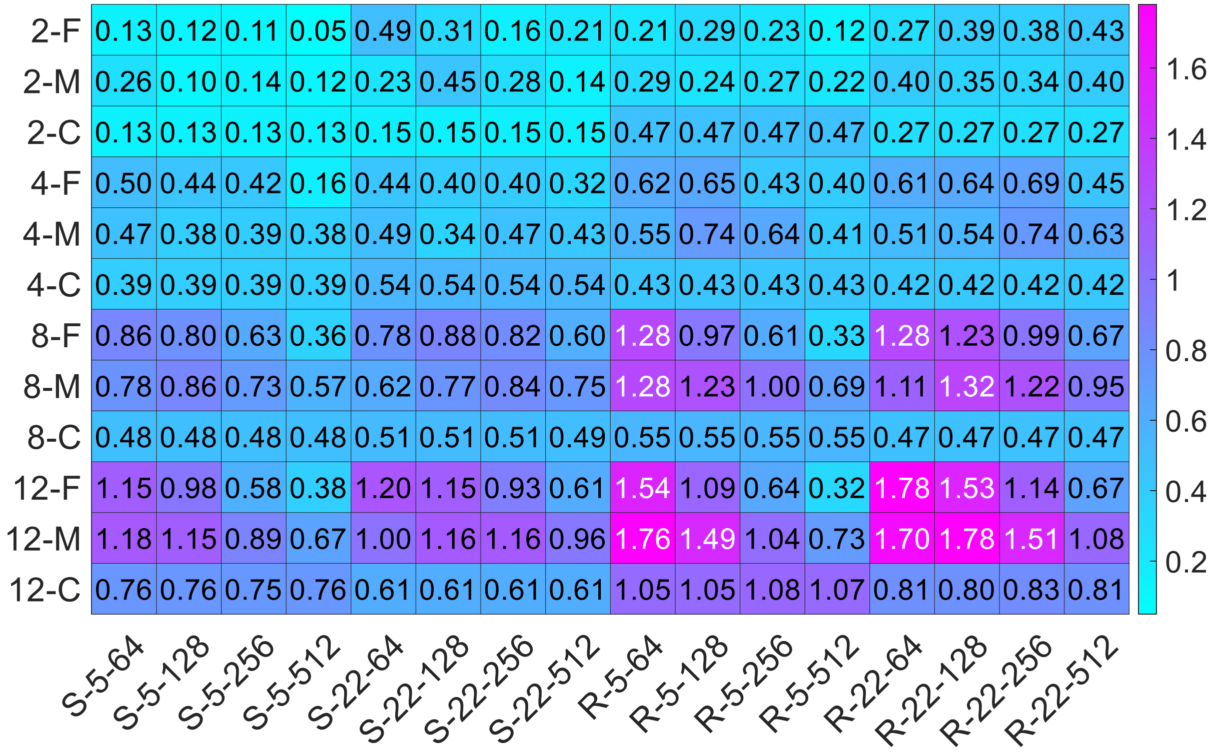


Figure 5.3.: Heat map of $\mathbb{E}[E_{HW}^{DF}]$ under different hardware parameters during the MAC operation, X labels: S/R, STT-MRAM, ReRAM. 5/22, interconnect node. 64/128/256/512, crossbar size. Y labels: 2/4/8/12, number of operands. F/M/C, Farthest, Middle, and Closest positions

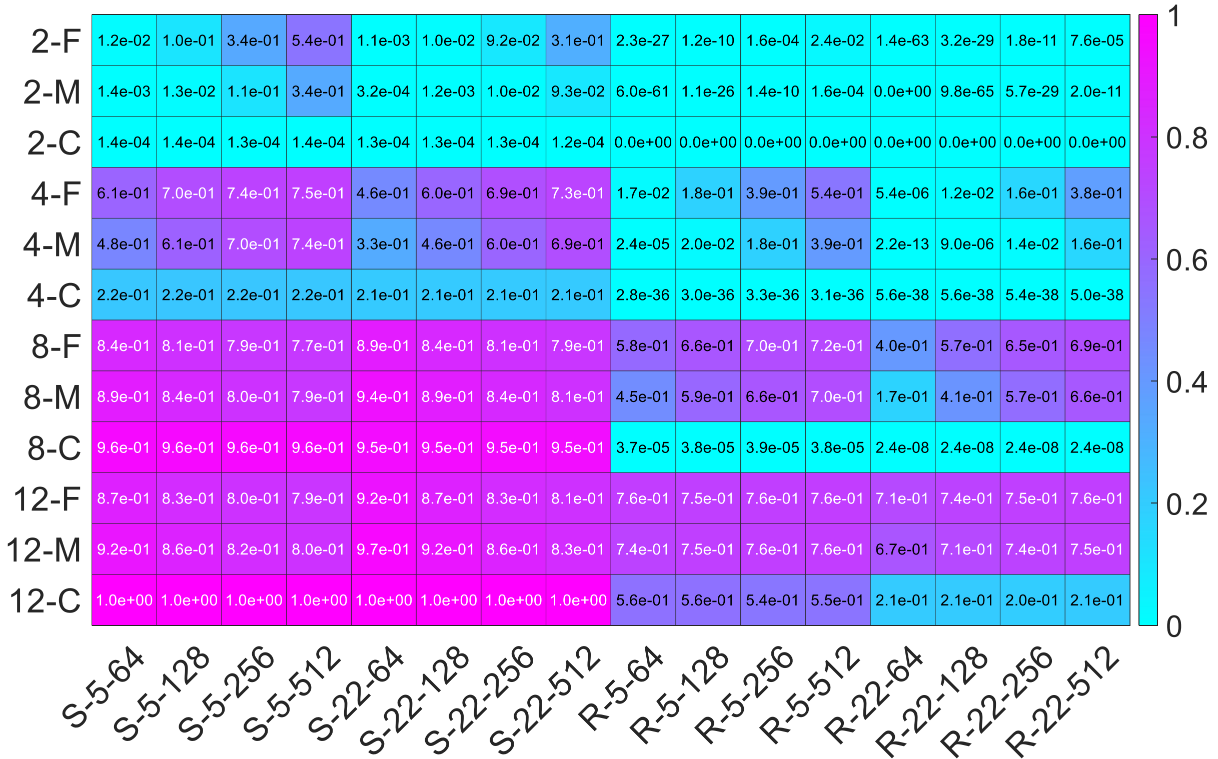


Figure 5.4.: Heat map of $\mathbb{E}[E_{HW}^{DF}]$ under different hardware parameters during NVM-CiM for scouting Boolean operation, the offset of the comparator is 3%, for X and Y labels, refer to the caption of Figure 5.3

Table 5.3.: The application-level analysis of error-masking capability and execution time for *2mm* benchmark consisting the MAC kernel, using 22 nm interconnect and crossbar size of 128×128

Number of the operands (Size of the matrices)	Average* $\mathbb{E}[E_{HW}^{DF}]$		CVF	Average* $\mathbb{E}[E_{SW}^{DF}]$		Execution time normalized to non-CiM
	STT-MRAM	*ReRAM*		*STT-MRAM*	*ReRAM*	
2	0.305	0.337	0.977	0.298	0.329	0.98
4	0.428	0.536	0.736	0.315	0.394	0.86
8	0.718	1.01	0.326	0.234	0.329	0.65
12	0.972	1.37	0.170	0.165	0.233	0.54

* Average on the different positions, farthest, middle, and closest with respect to the sensing circuitry

Table 5.4.: The application-level analysis of error-masking capability and execution time for *database query* consisting the Boolean kernel, using 22 nm interconnect, and crossbar size of 128×128, number of records in the database: 8192

Number of the operands (Number of conditions)	Average* $\mathbb{E}[E_{HW}^{DF}]$		CVF	Average* $\mathbb{E}[E_{SW}^{DF}]$		Execution time normalized to non-CiM
	STT-MRAM	*ReRAM*		*STT-MRAM*	*ReRAM*	
2	3.78e-3	1.08e-29	0.150	5.67e-4	1.62e-30	0.82
4	4.24e-1	4.15e-3	0.189	8.01e-2	7.84e-4	0.89
8	8.92e-1	3.27e-1	0.169	1.51e-1	5.53e-2	0.89

* Average on the different positions, farthest, middle, and closest with respect to the sensing circuitry

$\mathbb{E}[E_{HW}^{DF}]$ (1, 2, 4-6 in Section 5.1.1). The $\mathbb{E}[E_{SW}^{DF}]$ for any hardware configuration presented in Figure 5.3 and 5.4 can be obtained by multiplying its $\mathbb{E}[E_{HW}^{DF}]$ into the respective CVF value from Table 5.3 or 5.4.

5.2.4. Mitigation of the overall decision failure

Based on our findings, our focus lies on enhancing overall decision failure. Particularly for scouting Boolean operations, increasing the number of operands increases the $\mathbb{E}[E_{SW}^{DF}]$ (see Table 5.4). For the MAC operation, though the increase in the number of operands exacerbates the $\mathbb{E}[E_{HW}^{DF}]$, as shown in

Table 5.5.: Decreasing $\mathbb{E}[E_{SW}^{DF}]$ by hierarchical execution of *database query* application and its latency overhead, using STT-MRAM, 22 nm interconnect, and with the crossbar size of 128×128

Hardware configuration	Mapping configuration	Average* $\mathbb{E}[E_{HW}^{DF}]$	CVF	Reduction of the $\mathbb{E}[E_{SW}^{DF}]$ (Compared to Direct CiM-8)	Execution time normalized to non-CiM
Direct CiM-8	NVM-CiM with 8 inputs	8.92e-1	0.169	0%	0.89
Hierarchal CiM-4 and CiM-2	NVM-CiM with 4 inputs	4.24e-1	0.189	46.8%	0.92
Hierarchal 3-levels of CiM-2	NVM-CiM with 2 inputs	3.78e-3	0.233	99.4%	0.95

* Average on the different positions, farthest, middle, and closest with respect to the sensing circuitry

Table 5.3 for the *2mm* benchmark, the decreasing trend of CVF is considerably more significant, i.e., the masking capability of this application is stronger than the increase of the $\mathbb{E}[E_{HW}^{DF}]$.

Especially for STT-MRAM-CiM where $\mathbb{E}[E_{SW}^{DF}]$ can be significant for more than two operands, implementing a hierarchical approach for scouting Boolean operations is promising. For an 8-input Boolean STT-MRAM-CiM operation, three mapping configurations are explored: *direct CiM-8*, *hierarchical one level of CiM-4 and one level of CiM-2*, and *hierarchical three levels of CiM-2*.

Table 5.5 shows the CVF, the $\mathbb{E}[E_{HW}^{DF}]$, the reduction of the $\mathbb{E}[E_{SW}^{DF}]$, and the normalized execution time for these three different mapping configurations. According to Table 5.5, using the hierarchical mapping configurations slightly increases the CVF. However, as the $\mathbb{E}[E_{HW}^{DF}]$ in the case of STT-MRAM-CiM-2 and STT-MRAM-CiM-4 is considerably less than the STT-MRAM-CiM-8, these alternative mapping configurations can eventually achieve the error reduction in $\mathbb{E}[E_{SW}^{DF}]$. Based on our full-system simulation using the CiM-enabled gem5 framework, as outlined in Table 5.5, hierarchical execution of Boolean operations and making necessary adjustments in the application using the corresponding CiM kernels comes with the cost of higher execution time. This can be explained by the increase in the number of CiM instructions in the case of using alternative hierarchical configurations.

5.3. Conclusion

In this chapter, we have performed a comprehensive reliability analysis on the emerging paradigm of CiM by utilizing NVM. At the technology level, we have investigated the effects of NVM technology. At the circuit level, we have statistically modeled the expected decision failure error by taking into account different sensing circuitries required by the MAC and scouting Boolean operations. Additionally, we have considered other circuit-level variables such as resistive parasitic of the interconnect, the size of the crossbar, position, and the number of operands. We have also extended our analysis to the level of the application and investigated the propagation of the hardware-level errors to the application level, and quantitatively measured the masking capability of the applications.

6. Algorithm to Technology Co-Optimization for CiM-based Hyperdimensional Computing

HDC is an emerging brain-inspired algorithm that has gained considerable interest in the past years [76]. Not only for its capability of one-shot learning but also because of its robustness against noise in its computations. HDC is based on the concept of Hypervector (HV), i.e., vectors with a very large dimension [17]. Using these large HVs as the basic processing elements provides sufficient redundancy and thus considerable robustness against noise. In addition, HDC's simple mathematical operations can be easily implemented in hardware. Hence, combining HDC and NVM-CiM enables data-intensive yet efficient machine learning.

A core operation of HDC is the computation of the similarity between two HVs. The different classes are represented by *prototype HVs*. During inference, the similarity of an unlabeled *query HV* to all the prototype HVs is computed and the most similar class is selected as the label. In hardware, this inference operation is performed by the Associative Memory (AM). In the literature, CiM-based AM have been implemented [52], [78], [115], [138]. At the circuit level, CAM-based cells were employed to realize the AM. By employing NVM technologies in this CAM structure, prototype HVs are only written once, which can save energy. However, the length of the HVs challenges the NVM-CAM implementations. The noise in the computations overcomes even the robustness of HDC preventing a reliable operation [36].

The reliability of NVM-CiM is affected by two main factors. First, analog computing is inherently susceptible to noise. Second, the NVM technologies are still immature and hence more prone to manufacturing variability. Although hardware design techniques can partially improve the reliability of the NVM-CiM, there is a limit that the HDC inference algorithm can tolerate. Vastly incorrect similarity computations will impact the inference accuracy by selecting the incorrect class as the most similar to the query HV. Previous work did not precisely model the manufacturing variability of NVM at the technology level and, at the same time, consider its effect on the inference accuracy at the algorithm level. Further, the necessity of such an algorithm to technology co-optimization to achieve energy-efficient yet reliable HDC computations was not fully recognized in previous work. A holistic cross-level design and evaluation are essential to find the most efficient solutions.

In this chapter, we introduce the framework for this algorithm to technology co-optimization of energy consumption and inference accuracy for an HDC accelerator based on various NVM-CiM technologies. Moreover, by optimizing analog and digital computing and developing a hierarchical similarity computation, we cover the entire technology stack to improve the energy efficiency of the hardware-based HDC inference with a negligible reduction in the inference accuracy.

Our novel contributions within this chapter are as follows:

(**i**). Taking into account four different NVM technologies, including FeFET, PCM, STT-MRAM, and ReRAM to show the interaction of the device with the hardware and algorithm in the co-optimization of the entire HDC design space. (**ii**). Optimizing the analog and digital computing to implement similarity computation on large HVs directly in the hardware through a scalable hierarchical AM realized by NVM-CAM cells. (**iii**). Incorporating the proposed hardware design and its implication at the algorithmic level and investigating its impact on the HDC inference accuracy.

The rest of this chapter is organized as follows; Section 6.1 reviews the essential background and previous related work. Section 6.2 explains our core methodology, followed by Section 6.3, which presents and discusses our results. Finally, Section 6.4 concludes the chapter.

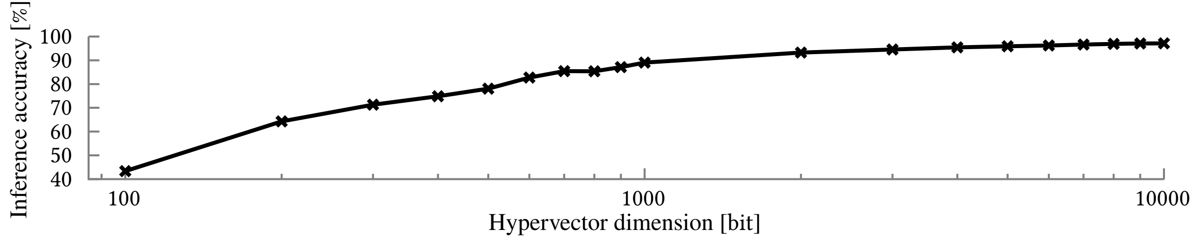


Figure 6.1.: Golden baseline inference accuracy of language recognition

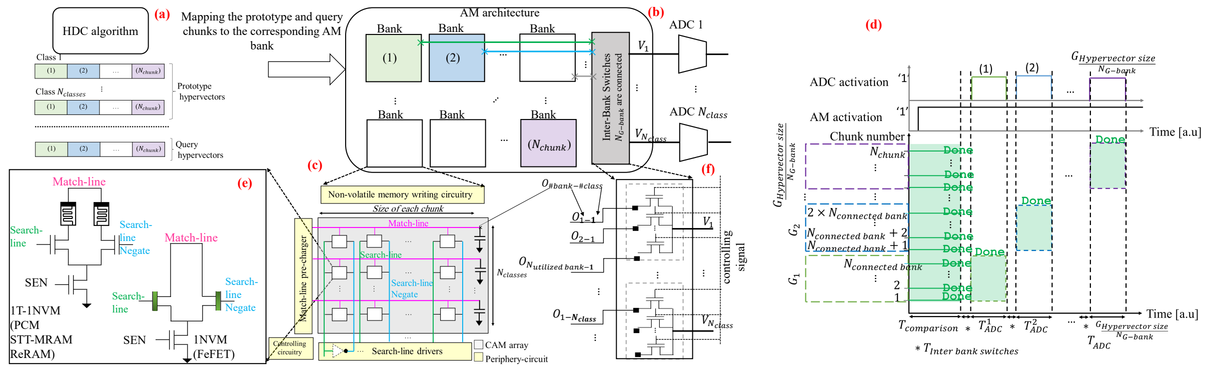


Figure 6.2.: Overview of the proposed method. (a) Split the HVs into chunks at the algorithmic level. (b) Multi-bank AM architecture to maintain the parallel search capability. (c) The array-level realization of the CAM structure. (d) The timing diagram shows the hierarchical calculation of the similarity measure. (e). The differential structure of the resistive CAM for 1T-1NVM (PCM, STT-MRAM, and ReRAM) and for 1T structure (FeFET). (f) Inter-Bank switches for scalability of the CAM-based accelerator for HDC

6.1. Preliminary concept

6.1.0.1. Hyperdimenssional Computing (HDC)

HDC is a rapidly growing alternative to classical machine learning algorithms [17]. The applicability of HDC has successfully been shown for language recognition [45], image and pattern classification [53], [90], circuit reliability [102], and more [44], [122], [137]. By using HVs with dimensions in the thousands, HDC and its operations are designed to create patterns and later recognize them. Moreover, the large dimension creates redundancy in the HVs, giving HDC its strong resilience to noise and errors. For the individual elements of the HVs, several different data types can be used, such as real or integer numbers, and for the hardware-based implementation, binary is the most compatible. To map real-world objects to the hyperspace, a small set of operations is used and their concrete implementations depend on the used data type. The utilized encoding scheme is application-dependent and the literature offers several, together with the analysis of the data type [76].

6.1.1. Related work

Together with a PCM-based n-gram encoder, a PCM-based AM has been proposed and fabricated in [78]. The CiM-oriented MAC version of the XNOR has been performed in an analog manner to compute the similarity of the prototype and query HVs. Although this work has also utilized analog computing, the amount of digital computing and ADC activations are still considerable, resulting in high energy consumption [63].

In [52], besides a fully digital AM hardware, leveraging the resistive CAM structure has been taken into account as well. Moreover, a fully analog version of the AM hardware based on the resistive CAM

has also been introduced. However, incorporating the technology-aware manufacturing variability in their simulation-based hardware model is not precisely performed, leading to less realistic results.

FeFET technology is also used in literature to construct AM by using CAM-based arrays [66], [115]. While [66] ignores the impact of the final analog to digital step, [115] proposes a “synaptic comparator” utilizing intermediate states of the FeFET device. The problem with both [66], [115] is that the small CAM size ranging from 8 to 15 bits, imposes high hardware overhead.

HDC for knowledge graphs accelerated by FeFET-based CiM has been discussed in [120]. Although [120] has benefited from chunk-wise analog computing, the compromise between multiple abstraction levels, namely, technology, circuit, architecture, and the algorithm has not been considered holistically.

Employing NVM-CAM-based design for accelerating the HDC inference implies strong interaction between the algorithm and the hardware, mainly due to the mixing of the digital and analog computations. Therefore, co-optimizing these interactions is crucial to reach an energy-efficient and highly accurate HDC inference. For such design space exploration, a co-optimization methodology and a full-stack framework are required. Such a framework needs to be flexible and realistic, i.e., sufficiently incorporated with the circuit-level and manufacturing variability details. The lack of the aforementioned methodology and framework is the shortcomings of the current literature.

6.2. Our proposed methodology

HDC can achieve high inference accuracy if the HV size is large enough as shown in Figure 6.1. Reducing the size of the HV (< 100) results in an extremely low HDC inference accuracy of $\sim 40\%$. However, the scalability of the NVM-CAM is severely limited by the lower HRS-LRS ratio of NVM compared to SRAM, as well as the impact of manufacturing variability. In this section, for the sake of energy efficiency, we aim to employ the NVM-CAM structure (Figure 6.2 (e)) for similarity computation. On top of that, we propose a full-stack solution to overcome the challenge of short NVM-CAM comparison length. Furthermore, we propose a methodology to co-optimize the algorithm and hardware design spaces to achieve a reliable and energy-efficient HDC inference accelerator.

6.2.1. Design solutions at the algorithmic level

Generating the HVs (i.e., encoding) is performed at the algorithmic level. The data type and the length of these HVs are crucial features and need to be chosen based on specific hardware characteristics. For the data type, binary is more compatible with the hardware. Also, the size of the HVs needs to be large to maintain high inference accuracy.

To co-optimize the inference algorithm based on the underlying hardware design and NVM technology, the large HVs need to be broken into *chunks* at the algorithm level (Figure 6.2 (a)). As we will discuss in Section 6.2.2, efficient mapping of the HVs chunks to the NVM-CAM accelerator requires some extra information, which needs to be provided by the algorithm. This information determines if the chunk is from a prototype or a query and what its position is in the original HV. The role of the HDC inference accelerator hardware is to process the chunks such that the final output of the similarity computation will not be affected by chunking.

6.2.2. Design solutions at the hardware level

The key feature of the AM is the capability of parallel search execution. To maintain the parallel search capability, the first requirement is to design a *multi-bank* AM architecture. Banks are memory sub-systems that can work independently, at the same time. As it is shown by Figure 6.2 (b), in a chunk-wise manner

HDC inference, the number of chunks (N_{chunk}) is the size of the HV divided over the maximum length of the match-line. Each chunk of the prototypes and query HVs are mapped to the corresponding AM bank.

The array level design of each bank is shown in 6.2 (c), the number of match-lines in each AM bank is equal to the number of the classes. So, in one AM bank, the corresponding query chunk is compared to the corresponding chunk of *all* the prototypes. As shown in Figure 6.2 (d), at each activation shot of the proposed AM architecture, at first, the similarity computation of the entire query (all the chunks in their corresponding banks) and prototype HVs can be performed. Nevertheless, assigning the query to the most similar class, equivalent to the highest inverse Hamming distance in the digital domain, requires the accumulation of similarity measures over all the chunks. To realize the cumulative similarity measure in the AM, we introduce the *inter-Bank switches* (Figure 6.2 (f)). Inter-Bank switches connect the group of match-lines (in total N_{G-bank}) corresponding to each class across the AM banks.

6.2.2.1. Hierarchical similarity measure computation

According to Figure 6.2 (d), after parallel similarity computation in the multi-bank AM, the first group of banks (G_1) is connected via the inter-Bank switches, and the output voltage of the inter-Bank switches is the analog representation of the similarity measure of G_1, which is converted to digital bits with the help of an ADC. Please note that the higher the output voltage of the inter-Bank switch, the higher the output of the ADC, and the higher the similarity measure. However, the output of the ADC is not the exact inverse Hamming distance.

In the next time step, the configuration of the inter-Bank switches for the next group, and activation of the ADCs will be performed. This procedure is repeated for all the groups onward

$$T_{\text{ADC}}^{1} : G_1, ..., T_{\text{ADC}}^{\frac{HV\ size}{N_{G-bank}}} : G_{\frac{HV\ size}{N_{G-bank}}}$$

Configure switches for the next group, activate ADC, and repeat for all groups.

For each prototype, the total similarity measure for the entire length of the HV can be obtained by summing up the digital representation of the similarity measure for each group, i.e., the output of the ADC. Eventually, the most similar class can be determined based on a winner-take-all mechanism, which simply compares the digital values corresponding to each class. As shown in Figure 6.2 (f), the inter-Bank switches are implemented by connecting each match-line output capacitance, to the source of an n-type transistor that acts as a switch. The drain of all the switch transistors corresponding to the same class are connected to a common node that is marked as V_x *(x: 1..*N_{class}*)* in Figure 6.2 (f), which is the input of the corresponding ADC.

6.2.3. Algorithm to technology co-optimization

In each abstraction level of the design space, from the high-level inference algorithm all the way down to the technology level, there are parameters that need to be co-optimized to reach an energy-efficient yet reliable hardware-based HDC accelerator with acceptable inference accuracy.

In the **algorithm level**, the size of the HV is a crucial parameter; the larger the HV, the higher the inherent HDC robustness. However, for larger HVs, the analog hardware design supporting that, imposes a considerable overhead in terms of the overall inference latency and energy, which is mostly due to ADC activations.

In the **hardware architecture level**, the number of connected banks through the inter-Bank switches (N_{G-bank}) is crucial. Increasing N_{G-bank} moves the output voltage levels of the inter-Bank switch (i.e., analog representation of the similarity measure) closer together. As these voltages are inputs to the ADC, it is important that the voltage corresponding to the most similar class can be differentiated from the other classes at least by one ADC quantization level.

Table 6.1.: Simulation setup tools and parameters

Simulation tool	Cadence Virtuoso
Technology node for CMOS	Global Foundries 22FDX
Standard VDD for CMOS	0.8 V
Temperature	27°C
Interconnect Parameters	- Barrier: Ta-based, thickness: 3 nm - Horizontal/vertical dielectric = 2.55/3.9 - RC π−model=11.95 Ω, 16.63 aF (per CAM cell)
FeFET model [127]	- Material = $Hf_{0.5}Zr_{0.5}O_2$ - Ferro layer thickness = 10 nm - HRS, LRS = 1 TΩ, 19.50 kΩ
PCM model [79]	- Fabrication node = 90 nm - PCM material thickness = 100 nm - HRS, LRS = 9 MΩ, 20 kΩ
MTJ model [38]	- RA = 7.5 Ω μm^2 - Nominal TMR = 150 % - HRS, LRS = 11.56 kΩ, 5.967 kΩ
ReRAM model: JART [71], [87]	- Filament radius = 45 nm - Disc region length = 0.6 nm - HRS, LRS = 105.51 kΩ, 1.975 kΩ

The existing trade-off: On the other side, in the presence of manufacturing variability, the output voltage of the CAM match-line corresponding to the most similar class can be incorrectly sensed less than other classes and produces wrong similarity computation. While increasing the N_{G-bank} impairs the sensing reliabiliy due to a decrease in the sense margin, larger N_{G-bank} can decrease the standard deviation of the analog voltage distribution and makes the distribution narrower, which can improve the distinguishability. Despite this effect, however, for larger N_{G-bank}, the impact of shrinking the sense margin is more dominant and negatively affects the inference accuracy. In general, large N_{G-bank} is desirable since it decreases the number of ADC activations and hence, improves the latency and energy efficiency, however, at the cost of impaired sensing reliability. Therefore, the robustness of the HDC algorithm against its computation noise determines the maximum value for N_{G-bank}.

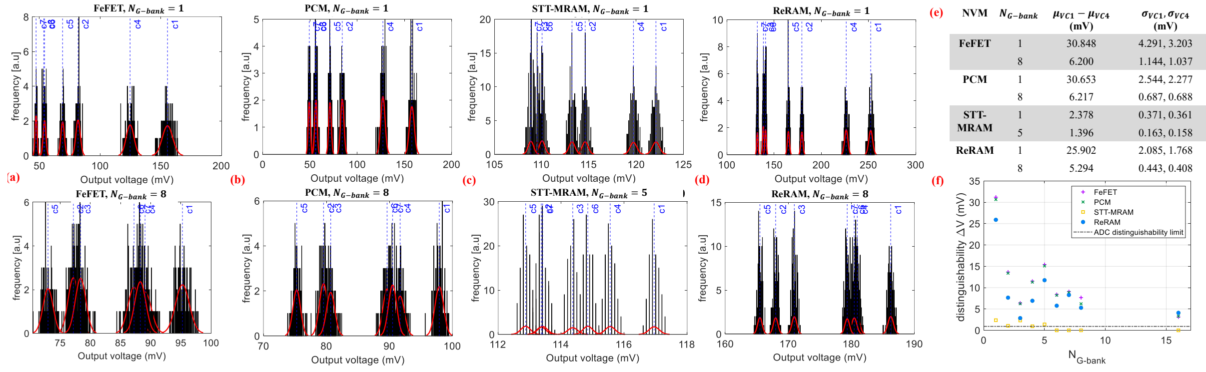
At last in the **technology** level, various NVM technologies can be utilized. The HRS-LRS ratio as well as the effect of the manufacturing variability are important technology-level aspects. The Maximum N_{G-bank}, which is an architectural parameter, is determined based on the employed NVM technology. Large HRS-LRS ratio (e.g., FeFET, PCM, and ReRAM technologies) is beneficial to increase the maximum N_{G-bank}. While a small HRS-LRS ratio (e.g., STT-MRAM technology) results in a smaller maximum N_{G-bank}.

The inference accuracy and energy consumption of the NVM-based HDC inference accelerator are determined by the aforementioned algorithm, hardware, and technology level parameters. The co-optimization methodology adjusts these parameters such that energy saving can be achieved at a small inference accuracy loss.

6.3. Results and discussion

6.3.1. Hardware-level analysis

We perform a detailed SPICE-based electrical-level simulation of the CAM array (6.2 (c), 6.2 (e)), as well as the inter-Bank switches (Figure 6.2 (f)). We consider the effect of the manufacturing variability



NVM	N_{G-bank}	$\mu_{VC1} - \mu_{VC4}$ (mV)	$\sigma_{VC1}, \sigma_{VC4}$ (mV)
FeFET	1	30.848	4.291, 3.203
	8	6.200	1.144, 1.037
PCM	1	30.653	2.544, 2.277
	8	6.217	0.687, 0.688
STT-MRAM	1	2.378	0.371, 0.361
	5	1.396	0.163, 0.158
ReRAM	1	25.902	2.085, 1.768
	8	5.294	0.443, 0.408

Figure 6.3.: (a-e) The expected ΔV with respect to the N_{G-bank} for different NVM technologies, (f) the distinguishability versus N_{G-bank}, (C_x, are the prototype classes, in the evaluated data set: C_1, C_4 are the most and the second most similar classes, respectively

for all of the explored NVM technologies, FeFET, PCM, STT-MRAM, and ReRAM. The HRS and LRS of these technologies are following a normal distribution (N ~ (mean (μ), standard deviation (σ)).

The resistive and capacitive (RC) parasitic of the metallic interconnect need to be considered for a realistic hardware model. We tune the electrical-level parameters of our proposed hardware design based on the trained data from HDC applications. Our FEoL and BEoL parameters are outlined in Table 6.1.

For the 10-bit ADC, we employ a voltage source of 950 mV [108]. Hence, the ADC quantization level is 0.928 mV and this value is the minimum voltage difference (ΔV) required for the most similar class to be distinguishable from the others.

Figure 6.3 (a-e) show the analog voltage distribution at the input of the ADC for two extreme cases of N_{G-bank} = 1, N_{G-bank} = 8 for FeFET, PCM, and ReRAM, and N_{G-bank} = 5 for STT-MRAM since it has the lowest HRS-LRS ratio. Due to the manufacturing variability and generally smaller HRS-LRS ratio compared to SRAM, in all the explored NVM technologies, the length of the CAM in each bank is limited to 64.

As discussed in Section 6.2.2, a larger N_{G-bank} decreases the sense margin, which brings the distinct voltage levels closer together. At the same time, it also decreases the standard deviation (σ) making the distribution narrower (see Figure 6.3 (a-e)). The latter effect, although not dominant for larger N_{G-bank}, mitigates the error of the similarity computation to some extent. Figure 6.3 (f) shows the decreasing trend of the sense margin while increasing the N_{G-bank}. Due to the interconnect parasitics, the decreasing trend of sense margin depends on the to-be-compared prototype and query HVs, and hence, is not always monotonic. For STT-MRAM and $N_{G-bank} > 5$, the sense margin is below the ADC quantization level. For FeFET, PCM, and ReRAM, N_{G-bank} can be as high as 16. However, leveraging N_{G-bank} = 16 at the algorithmic level results in a poor inference accuracy and thus we limit N_{G-bank} to eight.

6.3.2. Algorithm-level analysis

The first step toward algorithmic-level explorations is abstracting the hardware. In this regard, for different *block sizes* ($64 \times N_{G-bank}$), we model the nominal output voltage (mean (μ)) and its standard deviation (σ) as a function of inverse Hamming distance. Please note that the nominal output voltage and its standard deviation can be obtained through *single-run* and *Monte Carlo* simulations, respectively. By knowing the normal distribution of the output voltage, we can calculate the probability of the occurrence of falling into each ADC quantization level. Based on the block size, the output of the ADC in the digital domain corresponds to an inverse Hamming distance with a one-to-one mapping scheme.

To investigate the influence of the manufacturing variability on the different technologies, we evaluate HDC with several applications, including language recognition and gesture recognition with Electromyo-

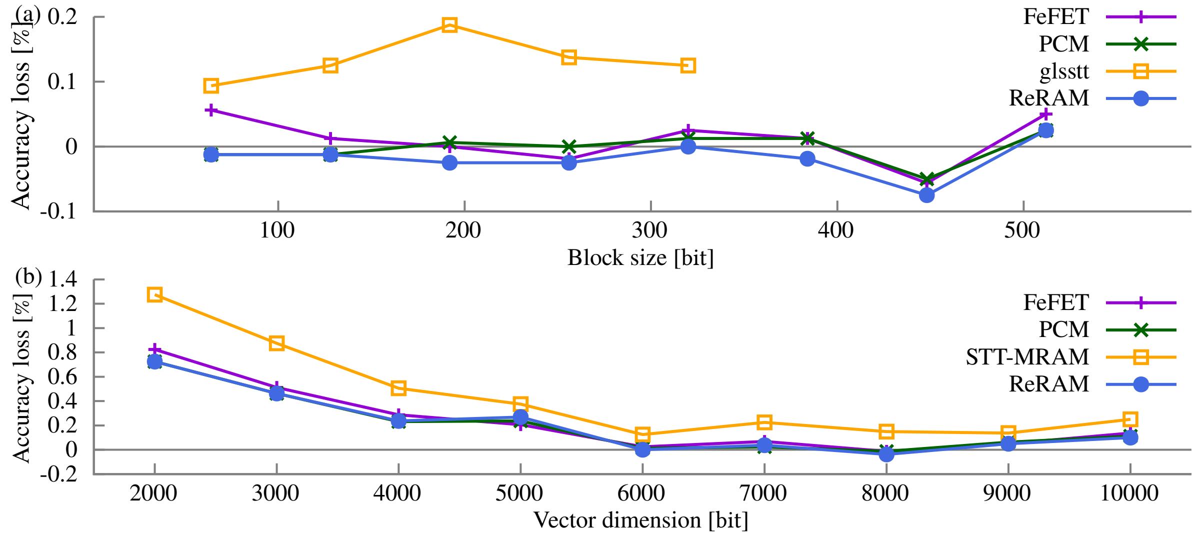


Figure 6.4.: Median inference accuracy loss of the tested technologies. (a) The relation of accuracy loss over block size at a hypervector (HV) dimension of 6000 bits (b) The relation of accuracy loss over HV dimension at a block size of 320 bits

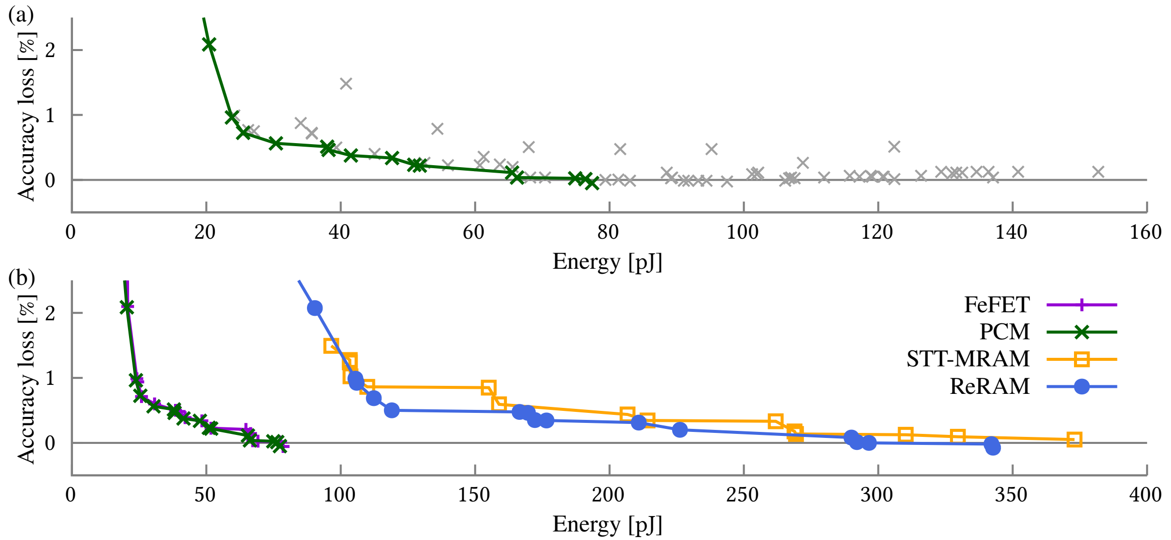


Figure 6.5.: Pareto optimal analysis of the different HV and block sizes with respect to accuracy loss and energy consumption. (a) The Pareto front of PCM at the connected dots and all non-optimal configurations in the background as a scatter plot. (b) Comparison of the different technologies

graphy (EMG). First, the inference accuracy is calculated in software as a golden baseline, then the abstracted hardware models are injected into the inference algorithm. We calculate the loss by subtracting hardware-based accuracy from the variation-free software-based baseline.

In Figure 6.4 (a), it can be seen the relation of accuracy loss versus the block size at the HV size of 6000 bits for the language dataset. As we can see in Figure 6.4 (a), the inference accuracy loss is not monotonically increasing with the block size. The competing effects of increasing the block size (or increasing the N_{G-bank}) on decreasing the sense margin and at the same time, decreasing the standard deviation (σ) of the voltage distributions is the reason behind the observations of the local extrema in Figure 6.4 (a). Besides, hardware implementation and manufacturing variability are not necessarily destructive for inference accuracy. In some cases, the models would classify incorrectly. Yet, the

randomness of the non-ideal hardware corrects these errors and actually improves the inference accuracy. Hence, for some of the (technology, block size) pairs in Figure 6.4 (a), negative accuracy loss can also be observed.

Due to the low HRS-LRS ratio in STT-MRAM, only a block size up until 320 bits is feasible. Moreover, STT-MRAM technology exhibits a 0.1% to 0.2% higher accuracy loss than all the other technologies which are gravitating around 0% loss. Relatively worse inference accuracy for the case of STT-MRAM is also explainable by its relatively smaller sense margin as shown in Figure 6.3 (c, e, f).

Figure 6.4 (b) shows an almost decreasing trend of inference accuracy loss with larger HVs. Moreover, as is depicted in Figure 6.4 (b), by selecting higher HV dimensions, the negative impact of the NVM technology is vanishing. Typically, larger HVs benefit from more redundancy, leading to stronger robustness against noise. Therefore, the effect of the technology on the inference accuracy loss is reduced compared to smaller HVs.

By combining the inference accuracy numbers with the energy of the proposed NVM-CAM-based accelerator for the HDC inference, we evaluate the NVM technologies as Pareto fronts. Figure 6.5 (a) shows the optimal configurations in green and non-optimal ones in gray. The point outside the plotted area is the 1024 bits block with a high accuracy loss of 5.53%. In Figure 6.5 (b), the different technologies are depicted together and their Pareto fronts follow a similar pattern. PCM and FeFET technologies are more energy efficient than ReRAM and STT-MRAM.

As discussed in Section 6.2.3, the size of the HVs at the algorithm level and the block size (or N_{G-bank}) at the hardware architecture level are crucial parameters for our proposed HDC accelerator. The most energy-efficient configurations are small HVs (from the algorithm level) with large block sizes (at the hardware architecture level), which minimize ADC activation cycles at the cost of higher accuracy loss due to more analog computing. The EMG dataset confirms these conclusions.

According to our Pareto analysis, increasing the HV size improves the inference accuracy, at the cost of higher energy consumption. However, increasing the HV size beyond a certain value stagnates the inference accuracy, i.e., an increase in the energy does not have a considerable effect on the inference accuracy. Horizontal jumps on the Pareto fronts in Figure 6.5 (b) are showing the increase of the HV size.

As earlier discussed in Section 6.3.1, although the block size of 1024-bit ($N_{G-bank} = 16$) is possible, however, the accuracy loss is too high which cannot be compensated with the energy saving. The Pareto analysis also shows that the block size of 1024-bit, is not a good choice in the context of HDC.

Table 6.2 shows the summary of results in our proposed CAM-based HDC accelerator and its comparison with the related work. For all the NVM technologies, by co-optimizing the digital and analog computing, the block size of 64-bit can increase to 320-bit, which means ADC reduction for 5x. Compared to the non-optimized hardware design (with the block size of 64-bit, equal to the length of NVM-CAM), the reduction of the energy consumption for the explored NVM technologies are as follows: FeFET: 3.92x, PCM: 3.95x, STT-MRAM: 2.65x, and ReRAM: 2.54x. While the inference accuracy loss, average on various NVM technologies, is only 0.125%.

6.4. Conclusion

In this chapter, we have explored accelerating the HDC inference by leveraging the NVM-CiM paradigm. We have introduced a hardware and algorithm co-optimization methodology and further used it to design a multi-bank AM based on NVM-CAM. The short length of the NVM-CAM is a challenge imposed by NVM technology while utilizing it for the similarity computation of the HVs. To overcome this challenge, we have proposed a chunk-wise HDC inference at the algorithm level, modified multi-bank AM, and the CiM architecture, and employed a hybrid analog and digital computing scheme for similarity computation at the circuit level. We have demonstrated that the co-optimization of the energy and inference accuracy results in considerably less energy consumption (3.27x), with an inference accuracy loss of only 0.125%.

Table 6.2.: Summary of the results and comparison with the related work, the size of the hypervector is 10k

Work	Technology CMOS/NVM	Accuracy of the hardware model	Energy per class [pJ]	Accuracy loss
[52]*	45 nm/ReRAM	+	2.2	0.5%
[78]	65 nm/PCM	+++	1180.0	0.4%
This work **	22nm/FeFET	++	76.7	0.1%
	22 nm/PCM	++	76.4	0.1%
	22 nm/STT-MRAM	++	135.9	0.2%
	22 nm/ReRAM	++	144.3	0.1%

* The fully analog implementation based on the resistive CAM
** The block size is 320 bit

7. Time-dependent Electromigration Modeling for Workload-aware Design Space Exploration in STT-MRAM

The interconnect network in a VLSI system can be categorized into two classes: Power lines and Signal lines. Unlike the power lines that carry mostly unidirectional large amounts of currents, the signal lines have been assumed to carry (much) less current, besides, the bidirectionality of the current passing through the signal lines has a recovery effect for EM. However, EM immortality of signal lines cannot necessarily be held in the more advanced technology nodes.

For instance, Despite the promising characteristics of the STT-MRAM, high read and write currents are challenging for this technology, and hence, make the BL) of the STT-MRAM susceptible to the EM.

In this chapter, we perform the workload-aware EM study in the STT-MRAM-based second/last level (L2) cache. For this goal, we first, try to find and verify an appropriate EM modeling at the physical level, and then, introduce the concepts of *workload segment* and *representative segment current*. By the aforementioned modeling and concepts, we are able to perform EM studies on the traces of realistic workload from the SPEC CPU2017 benchmark suite. These traces are generated by reusing an extended version of the *gem5* simulator that we have already used in [96].

To perform the workload-aware EM studies, the EM modeling needs to capture the time-variant current density. The current passing through the memory signal lines is generally time-dependent and different for each application. Besides, an accurate workload-aware EM modeling can be leveraged to perform EM-based design space exploration to avoid over-design and under-design.

The contributions of this chapter are as follows: **(i).** Proposing and verification of an EM physical model under the time-variant current density. **(ii).** Proposing two new methods to abstract the workload-aware EM-physical modeling and quantitatively compare them from the *accuracy* and the *timing overheads* points of view by applying them to the traces of ten realistic workloads. **(iii).** Leveraging the workload-aware EM model to perform design space exploration to minimize the design overhead while meeting the EM reliability requirements.

The rest of this chapter is organized as follows, in Section 7.1 we discuss the existing related work and proceed with the proposed method in Section 7.2. In Section 7.3 we present and discuss the results and finally, Section 7.4 concludes the chapter.

7.1. Related work

The EM-induced failure in the power lines has gained a lot of attention since they carry unidirectional and relatively high current density [35], [85], [89]. For instance, the work in [35] has used a tree-based structure for the interconnect and physical-based modeling for the EM. Further, [89] has proposed an optimization algorithm, that guarantees the reliability of the power networks more efficiently. Also, [85] has presented a detailed analysis on how to calculate the *steady-state stress* and discussed the design insight that can be provided by the steady-state stress before performing a detailed *dynamic* analysis.

Few previous works also have taken into account the EM effects on the memory BL [31], [33], [96]. In [31], [33] the EM susceptibility of the SRAM BL under a uniform workload has been studied, and the empirical Black's equation (Equation 2.1) has been used for the EM modeling. In a uniform workload,

the number of Write ‘1’ and Write ‘0’ are equal, hence, the EM-induced failure is caused only by the read operation. The first workload-aware EM studies using the physical-based EM modeling has been performed by [96]. Instead of a uniform workload, [96] has considered the realistic workload traces and highlighted the significant impact of the workload on the EM-induced failure. However, the modeling of the current that passes through the BL is not realistic enough and results in unrealistic MTTFs as well.

7.2. Proposed methodology

7.2.1. EM-induced MTTF modeling under a time-variant current density

As discussed in Section 2.6.2.3 the impact of the time-variant current density in the growth phase of EM progression has not been discussed in [41]. To capture the impact of $j(t)$ in the growth phase, similar to the nucleation phase, we also try to perform the integration on the piecewise constant $j(t)$ over time. For a time-variant current density, t_{incub} and t_r are the times that satisfy Equation 7.1 and Equation 7.2, respectively. In Equation 7.2, ρ_{Ta} and ρ_{Cu} are the resistivity of the Ta-based barrier metal (Ta/TaN) and Cu, respectively. Please note that in Equation 7.2, a 20% increase of the line resistance has been considered as the failure criteria. W, H and h_{Ta} are the line width, Cu thickness, and the barrier layer thickness, respectively. The EM-induced MTTF under a time-variant current density can also be calculated through $MTTF = t_{nucl} + t_{incub} + t_r$.

$$\Delta l_{crit} \cdot k \cdot T = D_a e Z \rho \times \int_0^{t_{incub}} j(t).dt \tag{7.1}$$

$$0.2 \cdot R_{initial} = [\frac{\rho_{Ta}}{h_{Ta}(2H + W)} - \frac{\rho_{Cu}}{HW}] \cdot \frac{D_a}{kT} \cdot eZ\rho_{Cu} \times \int_0^{t_r} j(t).dt \tag{7.2}$$

To verify the correctness of the EM-induced MTTF modeling under a time-variant current density (Equation 2.8, 7.1, and 7.2), as instructed in [65], we use variation-aware physical-based modeling and confirm that the fitted MTTF distribution is in good agreement with the measured MTTF at all the reported temperatures ($230°C$, $280°C$, and $330°C$) in [60].

7.2.2. Applying the EM modeling under a time-variant current density to a workload

Piecewise constant approximation of each BL’s current density is built according to the workload trace. As a typical workload trace may contain millions of operations, the number of pieces of BL’s current density can be extremely large. Performing the integration over time on this huge number of pieces is not practical as it is extremely time-consuming and effortful. Moreover, predicting the EM-induced MTTF (in the range of months or even years) from seconds of workload execution requires abstraction and approximation on the time-variant EM model. To overcome the aforementioned challenges, we define two concepts of *workload segment* and *representative segment current*.

7.2.2.1. Workload segmentation

To extract a workload trace with a reasonable size, a *segment* of the workload needs to be executed. The workload segment should be determined in a way that the entire workload footprint can be predicted through the extracted trace. In this work, the most representative chunk of each benchmark is extracted with the help of the *SimPoint* utility [7]. For this experiment, we have generated the STT-MRAM-based L2 traces, by execution of a chunk of workload containing 100 million instructions.

Workload execution represents only a few seconds, however, to observe the workload-dependent EM phenomena, months or even years of workload execution are required. To approximate the impact of

relatively long periods of workload execution, we can assume that the workload segment is been repeatedly running over time. Such approximation is possible since we make use of common statistical methods (through SimPoint) to determine the intervals with the highest relevance: this, together with a sufficiently long interval size, ensures that the workloads are realistic.

7.2.2.2. Representative segment current

Instead of taking into account a time-dependent current density waveform, we propose assigning a *constant representative current density* for each BL. As this current is calculated based on one segment execution of a workload, it is denoted by I_{seg}. By this assignment, instead of performing the effort-full integration over time-dependent current density ($j(t)$) in Equation 2.8, 7.1, and 7.2, we can proceed with the constant j, and at the end, the MTTF ($MTTF = t_{nucl} + t_{incub} + t_r$) remains the same. Here, we discuss three different methods to calculate the I_{seg} for each BL.

Method CAI: Current Averaging by Ignoring the idle cycles of the trace: This method has already been introduced in [96]. To calculate I_{seg} with this method, we can use Equation 7.3.

$$I_{seg} = \frac{I_{W1} \cdot n_{W1} + I_R \cdot n_R - I_{W0} \cdot n_{W0}}{n_{W1} + n_R + n_{W0}} \tag{7.3}$$

In Equation 7.3, $I_{W1/0}$ and I_R are *write '1'/'0'* and *read* currents, respectively. I_R and I_{W1} are usually considered to be in the same direction, while I_{W0} has the opposite direction. In a workload segment, the number of *write '1'*, *write '0'* and *read* are denoted by n_{W1}, n_{W0} and n_R, respectively.

Method CAI of I_{seg} calculation results in an over-conservative (too short) MTTFs, since it ignores the idle cycles. In the idle cycles, the current passing through the BL is assumed to be zero, and hence, the hydrostatic stress in the wire is relaxed. According to the STT-MRAM-based L2 traces which we generate for 10 workloads from SPEC CPU2017 benchmark suite, an average of 77% of cycles are the idle cycles. The other accuracy problem with Method CAI for I_{seg} calculation is that the read and write latency in the STT-MRAM are not equal and this asymmetry in write times is not considered in Equation 7.3.

Method CAC: Current Averaging by Considering the idle cycles of the trace: In this method, we try to address the two discussed shortcomings of Method CAI for calculating the I_{seg}. To include both the idle cycles and asymmetric read and write latencies, we modify Equation 7.3 as given in Equation 7.4.

$$I_{seg} = \frac{I_{W1} \cdot n_{W1} \cdot C_W + I_R \cdot n_R \cdot C_R - I_{W0} \cdot n_{W0} \cdot C_W}{n_{W1} \cdot C_W + n_R \cdot C_R + n_{W0} \cdot C_W + n_{idle}} \tag{7.4}$$

In Equation 7.4, C_R and C_W are the required cycles for read and write, respectively, and n_{idle} is the number of the idle cycles.

Method RCA: Calculating the Representative Current by Adding up the $\Delta\sigma_{seg}$: Figure 7.1, shows the hydrostatic stress of the bit-line versus execution cycles for an arbitrary sequence of operations. As the duration of each operation is relatively small, the hydrostatic stress for read and write operations are *linearly* changing and in the idle cycles, hydrostatic stress remains constant. The $\Delta\sigma_{total}$ under this sequence of operations can be calculated through the *piecewise-linear approximation*; i.e., the $\Delta\sigma_{total}$ is the summation of $\Delta\sigma$ due to the execution of the corresponding operation (read or write).

To extend this method to a workload segment, we propose a *'2+1'- step* procedure.

In *step 0*: we first calculate the $\Delta\sigma$ for different STT-MRAM operations (write 1/0 and read). To do this *pre-characterization*, we use Equation 2.8, however, with a time-constant current density. The values of the j and t correspond to the current density and latency of each of STT-MRAM operations and, $\Delta\sigma_{\text{operation}}$ is the change of the hydrostatic stress due to the execution of the specific operation.

In *step 1*: we use the piecewise-linear approximation to calculate the total stress due to the execution of a workload segment: $\Delta\sigma_{seg} = \Delta\sigma_{W1} \cdot n_{W1} + \Delta\sigma_R \cdot n_R - \Delta\sigma_{W0} \cdot n_{W0}$.

In *step 2*: I_{seg} needs to be found in a way that the interconnect under this I_{seg} experiences the $\Delta\sigma_{seg}$ which is calculated in *step 1*. To find the I_{seg}, for a set of currents, we can calculate the corresponding $\Delta\sigma$

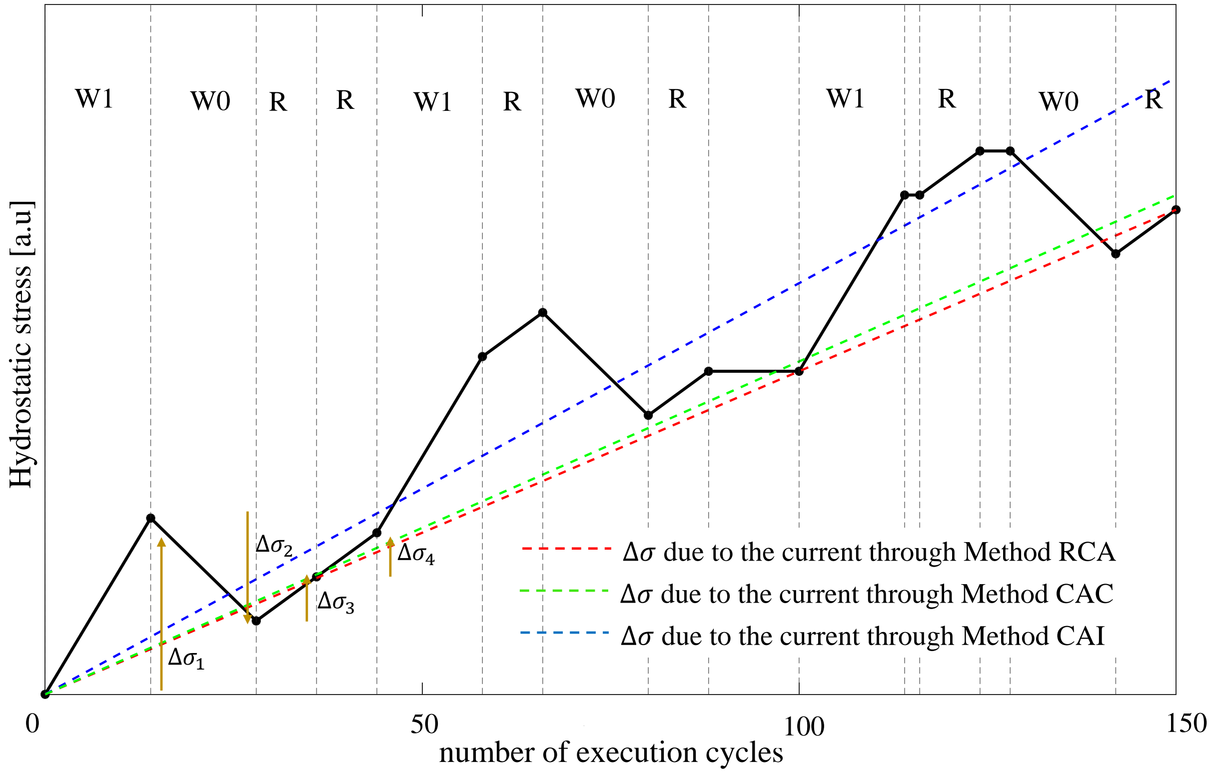


Figure 7.1.: Change of the hydrostatic stress due to the execution of an arbitrary sequence of operations and equivalent currents through methods RCA, CAC, and CAI corresponds to this sequence, cycles without a label are idle cycles

in the duration of the segment execution and a current which it's corresponding $\Delta\sigma$ becomes the closest to the $\Delta\sigma_{seg}$ is selected as the I_{seg}.

Unlike the Method CAI and CAC of calculating the I_{seg} which only depend on the trace and the write and read currents of the STT-MRAM, the Method RCA also considers the interconnect parameters such as length (l) and resistivity (ρ).

As we can see in Figure7.1, Method RCA of calculating the I_{seg} results in the most realistic MTTF. The $\Delta\sigma$ due to the I_{seg} through the Method CAC is also close to the realistic case. However, the Method CAI results in significantly higher $\Delta\sigma$ mainly due to the ignorance of the idle cycles. By adjusting the device-level and electrical-level parameters, the aforementioned methods of calculating the I_{seg}, can also be generalized to other flavors of spintronic memories.

7.3. Results and discussion

7.3.1. Experimental setup

We use the extended version of the *gem5* to extract the STT-MRAM-based L2 traces and Table 7.1 shows the parameters of the system simulated on the *gem5*. The applications which are executed on the *gem5* are all realistic workloads from the SPEC CPU2017 benchmark suite and cover a representative set of different applications from a wide range of domains.

Similar to [96], the STT-MRAM is organized in a 512×512 crossbar structure and the BL interconnect has a length of 200 nm for each cell, hence, the total length of the BL will be 512×200 nm. The width and height of the BL are also 32 and 64 nm, respectively. According to [64], the write '1', write '0', and read current for the considered STT-MRAM Process Design Kit (PDK) are roughly +120, -70, and 50 μA respectively. These current values are also aligned with other PDK like [38].

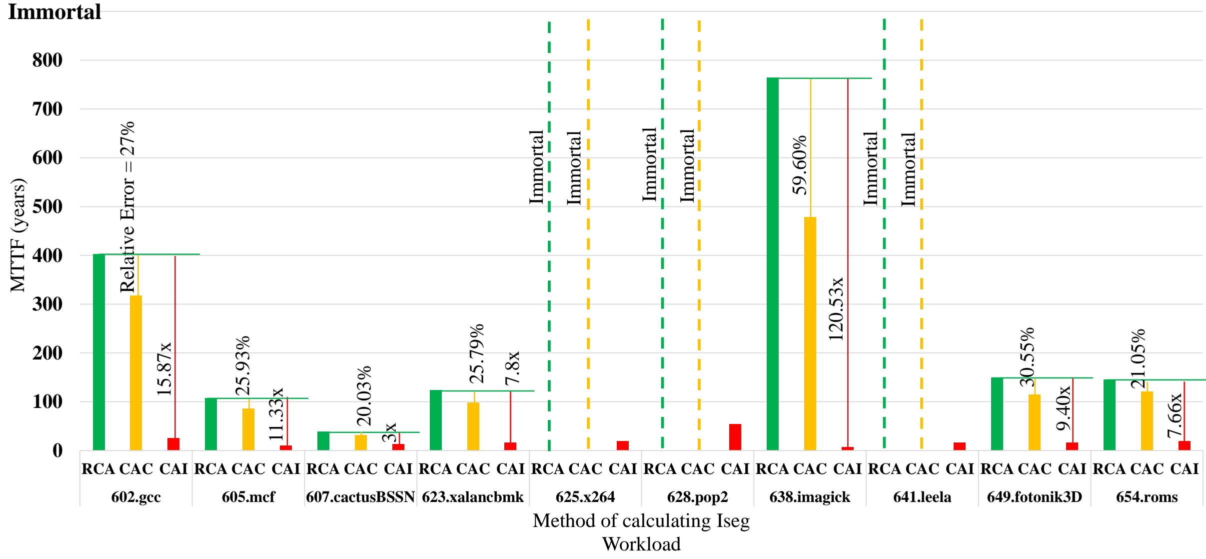


Figure 7.2.: The worst-case EM-induced MTTF for different workloads for ten different applications, using 3 different methods for calculating the I_{seg}: RCA, CAC and CAI

Table 7.1.: System-level parameters for *gem5* simulation

CPU characteristics			
Model	ARM HPI (high performance in-order)		
Clock frequency	1.4 GHz		
Cache hierarchy characteristics			
Cache line size		64 B	
Level/Tech.	Size	(Read/Write) cycles	(Assoc./Number of banks)
L1I/SRAM	16 kB	(2/ 2)	(2/ -)
L1D/SRAM	16 kB	(2/ 2)	(4/ -)
L2/STT-MRAM	256 kB	(8/ 14)	(16/ 4)

7.3.2. Model fitting

To evaluate the goodness of fitting the EM-induced MTTF through the model: *EM-induced MTTF under a time-variant current density* to a log-normal distribution, we use the Kolmogorov-Smirnov test (KS test)[1]. The KS test shows the goodness of the modeled MTTF fitted to a log-normal distribution, also, the obtained MTTF through the model is in good agreement with the real measurement reported at different temperatures in [60].

7.3.3. Workload-aware EM analysis

According to [96], at the temperature of $100°C$ and the aforementioned BL's dimensions, the values of σ_{res} and σ_{crit} are respectively 189.95 and 235.13 MPa. D_0 is assumed to be $7.7e-10\ m^2s^{-1}$ and B is

[1] According to [65], the mean (μ) and standard deviation (σ) of the log-normal distribution of MTTF are temperature-dependent and temperature-independent, respectively. So KS test which is more aware of the center of the distribution is preferred to the Anderson-Darling test. The maximum number of the devices under test in [60] is 9, so the Chi-Square test cannot work properly as it requires a large dataset.

13.4 Gpa. E_a is 0.85 eV, Z is assumed to be 1 and ρ_{Cu} is 43.7 $\Omega \cdot nm$. The barrier layer of the interconnect is assumed to be Ta-based with a thickness of 3 nm. Also, ρ_{Ta} in the operational condition is $3e4\ \Omega \cdot nm$.

The current that passes through the STT-MRAM BL is highly workload-dependent, for instance, the highest current density (shortest MTTF) and the lowest current density (longest MTTF) happens for the workload *607.cactusBSSN* and *625.x264*, respectively.

Figure 7.2 presents the worst-case MTTF for a variety of realistic workloads, while the I_{seg} is calculated through different methods. The error of the Method RCA is considered to be zero and the MTTF through Method RCA is the baseline to calculate the error for methods CAI and CAC. Unlike Method CAI which even results in an order of magnitude shorter MTTFs, the average error for Method CAC is only 22.45%.

By considering the method RCA, the shortest MTTF is 38.05 years and happens for BL number 251 and under *607.cactusBSSN* workload. This MTTF is much longer than the target specification and implies an over-conservative design of the BL. This relatively large margin between the predicted and the target EM-induced MTTF can be exploited to reduce the BL width, which results in a more area-efficient memory array design.

For a better comparison, we take into account the timing overhead of different methods of calculating the I_{seg}. Methods CAI and CAC require a simple calculation (Equation 7.3 and Equation 7.4) for each BL's current under each workload, while in Method RCA, for each BL, a range of currents should be tested and one which results in almost equal $\Delta\sigma_{seg}$ selected as the I_{seg}. The average (across different workloads) required time for methods CAI and CAC is 9 ms and 5 ms, respectively, while for Method RCA, the average required time is 248 s. Therefore, there is a trade-off between the accuracy and timing overhead of the discussed methods. Although the Method RCA is the most accurate, for large memory arrays, it is clearly too time-consuming. Considering the relatively small error rate of Method CAC, as well as its quite efficient timing overhead, this method seems to be good enough for calculating the I_{seg} and then, approximating the workload-dependent MTTF.

7.3.4. Design space exploration

The width of the interconnect, not only affects the current density (j) but also alters other involved EM parameters. According to [69], effective bulk modulus (B), copper resistivity (ρ_{Cu}), diffusion constant (D_0) residual and critical stress (σ_{res} and σ_{crit}) are actually width-dependent parameters. To study the impact of the wire width on the EM-induced MTTF, we need to first calculate the I_{seg}.

The I_{seg} through the Method RCA is also width-dependent while the I_{seg} through the Method CAC only depends on the STT-MRAM current values and the workload trace. It is worth mentioning that the calculation of the I_{seg} through the method RCA as well as its corresponding design space exploration flow is more accurate since it also considers the impact of dimension changes on the parameters of the EM models and hence, it also becomes more time-consuming. Proceeding with the Method RCA for the sake of design space exploration becomes almost infeasible, especially for larger memory arrays.

Figure 7.3 shows the impact of width on the MTTF for both flows corresponding to I_{seg} calculation through methods RCA and CAC. To plot this graph, the highest current density (happens for BL 251 and under *607.cactusBSSN* workload) has been considered. As we can see in Figure 7.3, the accuracy of the MTTF calculation through Methods CAC for obtaining I_{seg} is in good agreement with the Method RCA, and can result in a good enough design space exploration in a relatively small runtime, and Method RCA can be used at the end, for the most accurate EM analysis of the chosen design configuration. With such a hybrid methodology, the time required for the design space exploration can be significantly reduced, while the precision remains unaffected.

As Figure 7.3 shows, by increasing the width, MTTF is increasing, however, in a non-monotonic manner. To explain this non-monotonic behavior, according to [69], increasing the interconnect width results in decreasing of all the affected parameters (j, B, ρ_{Cu}, D_0, σ_{res} and σ_{crit}). Decreasing all the

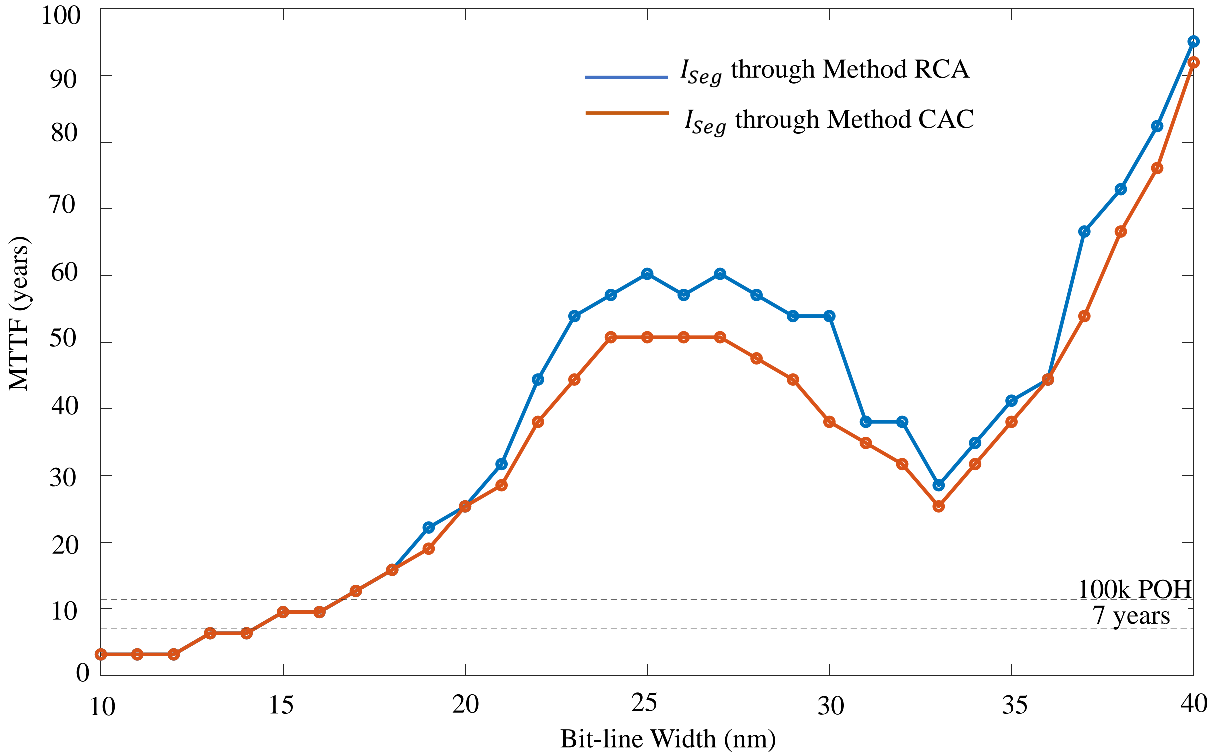


Figure 7.3.: The impact of the BL width on the MTTF for methods RCA and CAC of calculating the I_{seg}

Table 7.2.: MTTF for different STT-MRAM arry organization

Array organization	Nucleation time	Growth time	Total
64×64	–	–	Immortal
128×128	13.31 years	21.25 years	34.56 years
256×256	13.31 years	22.52 years	35.83 years
512×512	15.85 years	22.20 years	38.05 years

parameters, except σ_{crit}, yields the increase of the MTTF, while, a decrease in the σ_{crit} has the opposite impact and decreases the MTTF.

To meet different criteria of the EM reliability, MTTF equals 100k Power-on Hours (POH) and 7 years, the BL width of 17 nm and 15 nm (instead of 32 nm) respectively, would be sufficient. Hence, we can see that accurate and workload-aware EM modeling can result in more area-efficient design.

The length of the interconnect, on the other hand, has an impact on the EM-induced MTTF. The STT-MRAM array can also be organized into arrays of 64×64, 128×128, and 256×256. Based on our assumptions for the BL's dimensions and the temperature, and by considering the BL with the highest current density, except for the 64×64 organization, in which the BL's length is below the *Blech length* of 15.3 μm, all the other organizations are susceptible to EM, and the workload-aware analysis can also be applied to these organizations. Table 7.2 shows the MTTF for different STT-MRAM array organizations. On top of the EM reliability, deciding on the size of the memory array needs to be made by considering the throughput, area, and power of the memory module.

7.4. Conclusion

In this work, we have performed the EM analysis on the BL of the STT-MRAM array and shown the workload dependency of the EM susceptibility in these BLs. To include all the effective parameters of the design, we have used physical-based EM-induced MTTF modeling for time-dependent current density. Moreover, we have addressed the complexity of applying the EM physical-based modeling for the time-dependent current density and resolved this complexity by two concepts of *workload segment* and *representative segment current*. According to our detailed investigation, with our proposed methods, the workload-aware MTTF estimation can be done relatively fast with an acceptable error. We have also highlighted the design space exploration through the proposed methods and how it can be used in the co-optimization of the design parameters and EM reliability.

8. Analyzing the Electromigration Challenges of Computation in Resistive Memories

Analog MAC is the main kernel for the implementation of neural networks in deep learning applications [46]. Therefore, the long-term reliability of the analog MAC concept needs to be taken into account as well. One of the most pronounced reliability issues, particularly in the tight-pitch interconnects is EM [16].

Typically, performing the MAC requires activation of the multiple memory rows at the same time, which increases the current density in the common memory BL between the activated rows. So, EM is even more pronounced in the CiM concept and analog MAC operations. Besides the higher current density, the row activation pattern (the specific cells participating as CiM operands) significantly affects the EM profile of the memory BL.

In this work, we first analyze the EM phenomenon in different CiM-oriented MAC technologies. We extend the existing physical-based EM modeling to match the BL characteristics in the NVM crossbar. We also investigate the impact of the BL and array dimensions on its EM profile and show that, compared to the standard memory operation, in the case of analog MAC, activation of the memory rows with a specific pattern can decrease the EM-induced MTTF by 3.58x on average. We also explore the row activation pattern as an effective method to mitigate EM degradation in the analog MAC paradigm. EM-aware activation patterns can, on average, improve the EM-induced MTTF by 4.93x.

Our contributions to this chapter are as follows: **(i).** The first work, to the best of our knowledge, analyzes the EM challenges in CiM-oriented MAC. **(ii).** Considering the most popular CiM fabrics based on different NVM technologies, including STT-MRAM, ReRAM, and PCM. **(iii).** Analyze the EM phenomenon in CiM fabric by extending the existing physical-based EM modeling to match with the BL characteristics of the CiM. **(iv).** Address the row activation pattern as the means to mitigate EM in the analog MAC operation.

The rest of this chapter is organized as follows, Section 8.1 covers the required preliminary information and reviews the existing related work, and Section 8.3 discusses the method for EM analyzing in CiM. Section 8.4 presents the results of the EM analysis, followed by Section 8.5 which addresses the mitigating solution for the EM reliability problem of the CiM fabric. Finally, Section 8.6 concludes the chapter.

8.1. Preliminaries

8.2. Selecting the appropriate EM modeling

The discussed EM modeling in Sections 2.6.2.2 and 2.6.2.3 are not fully appropriate for conducting EM analysis on memory BLs. As it is shown in Figure 8.1, a memory BL that is extracted from the physical layout, consists of multiple *segment*s. Each segment corresponds to a memory cell and can be modeled as a resistive and a capacitive element: RC. In a multi-segment interconnect, typically, the currents that pass through the different segments are not equal to each other. Since the discussed physical-based EM modeling in Sections 2.6.2.2 and 2.6.2.3captures the constant current density (j), it does not work for a segmented interconnect. The proper physical-based EM modeling for the memory BL is the steady-state

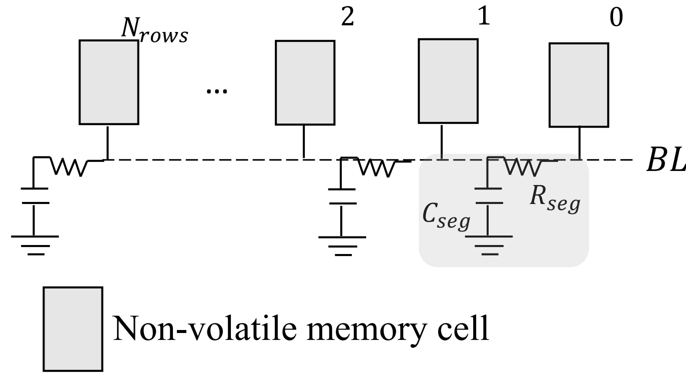


Figure 8.1.: A segmented BL of a NVM crossbar consists of RC segments corresponding to each memory bit-cell

EM modeling that is elaborated in Section 2.6.2.4. However, not providing any information regarding EM-induced MTTF is an issue with the steady-state EM modeling.

8.2.1. Related work

The current passing through the memory BL is bidirectional and has lower magnitudes. Therefore, it was initially assumed that the BLs are not vulnerable to the EM. However, especially in the advanced technology nodes where the BLs are designed in tight pitch sizes, their EM resilience is not a valid assumption. For instance, authors in [31], [33] have shown the EM vulnerability of the SRAM's BL, and authors in [96], [111] have shown the EM vulnerability of the STT-MRAM's BL under the realistic workloads. As motivated at the beginning of this chapter, though the EM analysis in the case of CiM is more challenging, it has not been studied in the previously existing work.

8.3. Proposed EM-analysis method for CiM

8.3.1. Modifying the EM modeling for CiM

As already discussed in Section 8.2, none of the existing EM modeling (details in Section 2.6.2) are suited for extracting the information regarding the EM-induced MTTF in a multi-segment memory BL. To resolve the shortcomings of the existing physical-based EM models, in this chapter, we propose a hybrid model based on the existing models. In this hybrid model, we first obtain the steady-state EM-induced stress on different segments of a multi-segment interconnect (Equations 2.9-2.11) [43]. In a multi-segment interconnect, we can assume that the void is nucleated at the location with the highest steady-state hydrostatic stress. In other words, the *maximum* steady-state hydrostatic stress determines the EM-induced MTTF. Now we find a *constant current density*, which can induce the same maximum steady-state hydrostatic stress to the multi-segment interconnect, and also use this *constant current density* as an input to the *timing-aware EM modeling* (Equation 2.6). The output of timing-aware EM modeling is the EM-induced MTTF. Figure 8.2 shows our proposed EM-modeling for a multi-segment memory BL during the CiM operations.

8.3.2. Interconnect dimensions

Interconnect length: The shorter (longer) the interconnect, the less (more) vulnerable to the EM, and the interconnects shorter than the threshold of *Blech length* are known to be EM-immortal [2]. On the other hand, however, the longer interconnects have larger resistive parasitic loads, which decreases the current values ($I = \frac{V}{R}$), and hence, it is beneficial for EM reliability. *Interconnect cross-section*: The

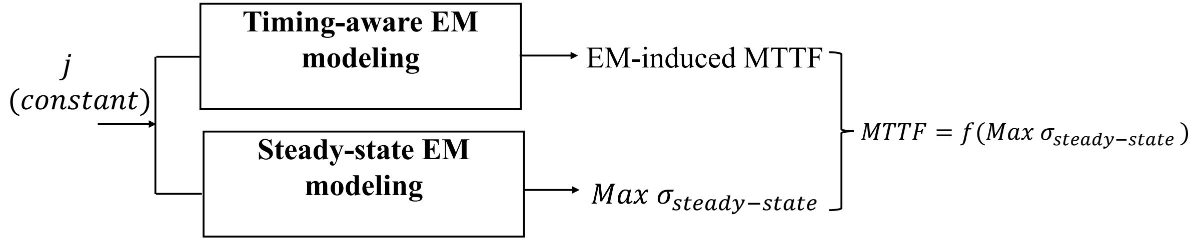


Figure 8.2.: The proposed EM modeling based on the existing models

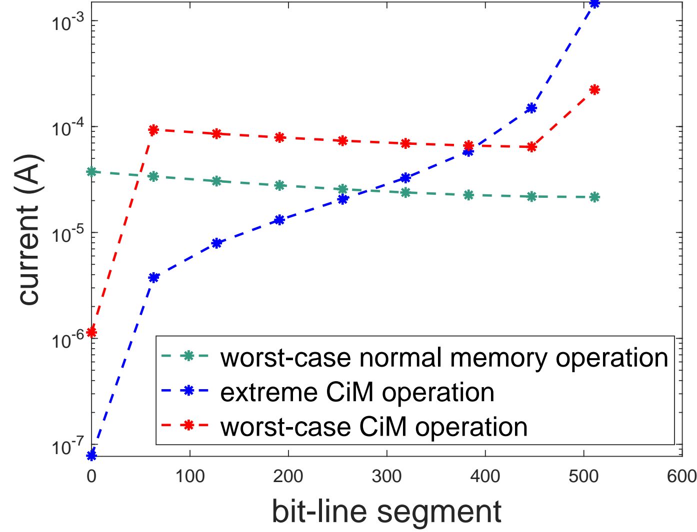


Figure 8.3.: Current distribution across the BL segments for different case studies in a ReRAM-based crossbar, $N_{rows} = 512$, for the STT-MRAM and PCM technologies, the current trends in the three case-studies are similar, the segment 0 and 511 are the farthest and closest segments to the sensing circuitry, respectively, number of activated rows in worst-case CiM operation is 64

cross-section of the interconnect has also a significant impact on its EM profile. The larger cross-section area (A) of the interconnect, on one hand, implies a smaller current density ($j = \frac{I}{A}$), and hence, longer EM-induced MTTF. However, on the other hand, the larger the interconnect cross-section the lower the resistive parasitic loads and the higher the current values. Moreover, some of the EM-related physical parameters are also width-dependent; decreasing the interconnect width, increases the j, the effective bulk elasticity modulus (B), copper resistivity (ρ_{Cu}), diffusivity constant (D_0), residual stress (σ_{res}), and σ_{crit} [69]. Reduction of the j, B, ρ_{Cu}, D_0, σ_{res}, increases the EM-induce MTTF, while decreasing the σ_{crit} is shortening the EM-induced MTTF.

8.4. Results and discussion

8.4.1. Simulation tool and setup

We use the simulation tool and models as indicated in Table 8.1. To approximate the BL's RC parasitic and the BL segment length corresponding to each cell, we use the NVSim wire models [28]. The parameters of the BL interconnect are indicated in Table 8.1. Please note that although the voltage source (VDD) affects the EM profile, deviating this parameter from the standard value can result in an unreliable transistor operation. Also, the operating temperature has a considerable impact on the EM profile, and the higher the temperature, the more exacerbated the EM risk. We choose 100 $°C$ as the chip temperature because high-performance memory operations are targeted.

Table 8.1.: Simulation setup tools and parameters

Simulation tool	Cadence Virtuoso
Technology node for CMOS	GF 22FDX
Standard VDD for CMOS	0.8 V
Interconnect parameters:	- Barrier: Ta-based, thickness: 3 nm - Ta resistivity (ρ_{Ta}) = 3e4 $\Omega.nm$ - Horizontal/vertical dielectric = 2.55/3.9 - Height of the BL = 41.8 nm - BL segment length = 118 nm
MTJ model [38]	-Radius = 20 nm - Barrier Material = MgO - RA = 7.5$\Omega\mu m^2$ - Nominal TMR = 150%
VCM-based device model: *JART VCM v1b Read variability* [71], [87]	-Radius of the filament = 45 nm - Initial oxygen vacancies concentration in the disc for LRS = 3, for HRS = 0.009 $\times 10^{26}/m^3$
PCM device model [79]	- Structure = Mushroom cell - Radius of bottom electrode = 20 nm - Doping of the phase change material = $Ge_2Sb_2Te_5$ (d-GST)
Temperature	100 $°C$

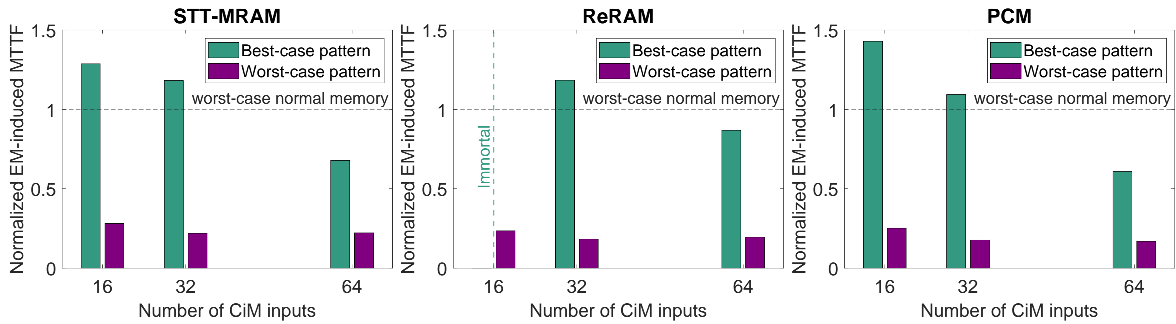


Figure 8.4.: Impact of the row activation pattern on the EM-induced MTTF, N_{rows} = 256, Width = 22 nm, normalized to MTTF of the *worst-case standard (normal) memory operation* in the respective technology, *best- (worst-) case activation pattern*: activated rows are the closest (farthest) addresses to the sensing circuitry

As already discussed in Section 8.3.2, besides the j, the B, ρ_{Cu}, D_0, σ_{res}, and σ_{crit} are width-dependent parameters, and their value in each investigated width, are extracted from [69], and the Z has been considered to be 1.

Table 8.2.: Impact of the BL length ($N_{rows}\times$ segment's length corresponding to each cell (118.8) nm) on the EM-induced MTTF, Width = 22 nm, and (R = 11.95 Ω, C=33.26 aF), for worst-case CiM operation, the number of activated rows is $\frac{N_{rows}}{8}$

N_{rows}	Normalized EM-induced-MTTF for **STT-MRAM**		
	worst-case standard memory	*extreme CiM*	*worst-case CiM operation*
128	Immortal	0.27	0.20
256	0.97	0.49	0.21
512	1	0.84	0.34

N_{rows}	Normalized EM-induced-MTTF for **ReRAM**		
	worst-case standard memory	*extreme CiM*	*worst-case CiM operation*
128	Immortal	0.30	0.14
256	0.87	0.59	0.16
512	1	1.23	0.31

N_{rows}	Normalized EM-induced-MTTF for **PCM**		
	worst-case standard memory	*extreme CiM*	*worst-case CiM operation*
128	Immortal	0.20	0.25
256	1.05	0.37	0.19
512	1	0.61	0.27

8.4.2. EM analysis for CiM

8.4.2.1. Impact of the row activation pattern on the EM profile of the CiM fabric

In a standard memory operation with a single row activated and CiM operation with multiple rows activated, the steady-state EM-induced hydrostatic stress depends on the address of the activated row(s) and its (their) data. In a standard NVM read operation, the current which passes through the BL segments *before* and *after* the activated row has the opposite direction. The existence of the current in the opposite directions alleviates the EM degradation. Therefore, the highest stress is induced (on the BL) when the farthest row from the sensing circuitry is being read (address 0). In this way, all the segments are actually located *after* the activated row, and the currents in all the segments are in the same direction. Obviously, programming the NVM cell in LRS results in higher current values than in HRS.

In a CiM operation, however, the *pattern* of the row activation has a significant impact on EM-induced hydrostatic stress. Similar to a standard memory operation, in a CiM operation, the highest (lowest) stress is induced, when the *block* of the activated cells are located the farthest (nearest) from (to) the sensing circuitry.

Figure 8.3 presents the current distribution with respect to the number of the BL segments in three special cases, *worst-case standard memory operation* which is reading out the address 0, programmed in LRS, *extreme CiM operation*, which is the activation of all the rows programmed in LRS, and *worst-case CiM operation*, which is the activation of the rows in a specific pattern.

Based on our results, in a NVM crossbar of 512 rows, a worst-case CiM operation happens (from the EM point of view) when the activated block of rows is located farthest from the sensing circuitry. Moreover, in a crossbar of 512 rows, the highest stress is induced if 64 rows are activated in a worst-case pattern. We roughly generalize this number of activated rows (N_{rows}) to other non-volatile memory crossbars with different N_{rows}; by activation of $\frac{N_{\text{rows}}}{8}$ in the worst-case pattern, the highest EM stress is induced on the BL. As it is presented in Figure 8.3, the *extreme CiM operation* case, the currents of the

Table 8.3.: Impact of the BL width on the EM-induced MTTF, N_{rows} = 512, for worst-case CiM operation: number of activated rows is 64, the RC parasitic values: (R = 13.86 Ω, C = 32.39 aF), (R = 11.95 Ω, C = 33.26 aF), (R = 10.26 Ω, C = 34.18 aF), (R = 9.44 Ω, C = 35.18 aF) for Width = 21..24 nm, respectively

Width (nm)	Normalized EM-induced-MTTF for **STT-MRAM**		
	worst-case standard memory	*extreme CiM*	*worst-case CiM operation*
20	0.67	0.67	0.28
21	0.83	0.79	0.31
22	1	0.84	0.34
23	1.23	1.02	0.38
24	1.33	1.05	0.38

Width (nm)	Normalized EM-induced-MTTF for **ReRAM**		
	worst-case standard memory	*extreme CiM*	*worst-case CiM operation*
20	0.73	0.99	0.27
21	0.85	1.12	0.29
22	1	1.23	0.31
23	1.16	1.35	0.32
24	1.20	1.34	0.31

Width (nm)	Normalized EM-induced-MTTF for **PCM**		
	worst-case standard memory	*extreme CiM*	*worst-case CiM operation*
20	0.64	0.48	0.22
21	0.80	0.55	0.24
22	1	0.60	0.27
23	1.24	0.70	0.30
24	1.34	0.71	0.30

closest segments to the sensing circuitry are quite large, however, it has relatively small currents in the farthest segments from the sensing circuitry. Therefore, *extreme CiM operation* does not necessarily lead to the highest EM-induced stress. For the *worst-case CiM operation*, the currents of the farthest segments from the sensing circuitry are larger than the *extreme CiM operation*, hence, it can result in a more exacerbated steady-state EM-induced hydrostatic stress.

In both the standard memory and CiM operations, the *write* operation needs to be taken into account. For the write operation, we assume that a constant write current is passing through all the segments. These constant write currents impose an additive EM-induced hydrostatic stress, which has been considered in the EM-induced MTTFs. For the STT-MRAM ([96]), ReRAM ([71]) and PCM ([79]) technologies, the constant write current are considered as 50, 23, and 50 μA respectively. Based on the proposed hybrid EM modeling in Section 8.2, we obtain the EM-induced MTTF for different design choices.

8.4.2.2. Impact of the BL length on the EM profile of the CiM fabric

Table 8.2 shows the impact of the BL length on the *normalized EM-induced MTTF* for the STT-MRAM, ReRAM, and PCM technologies. In this regard, we consider 128, 256, and 512 as the number of rows in the crossbar structure of the respective technologies. In the case of N_{rows} = 128, all three NVM technologies, show EM-immortality for the standard memory operations. However, in *worst-case CiM operation*, the EM-induced MTTF is reduced significantly in all the N_{rows}. Compared to the standard

memory operation (*worst-case*), across three NVM technologies, and N_{rows} = 256, 512, the *worst-case CiM operation* reduce the EM-induced MTTF by 4.24x. This confirms that the EM reliability in the CiM paradigm is significantly exacerbated.

8.4.2.3. Impact of the BL width on the EM profile of the CiM fabric

Table 8.3 shows the impact of the BL width on the *normalized EM-induced MTTF* for the STT-MRAM, ReRAM, and PCM technologies. In this regard, we sweep the BL width from 20..24 *nm*. Increasing the width generally improves the EM-induced MTTF. However, due to its impact on multiple EM-involved parameters (already discussed in 8.3.2), EM-induced MTTF is not always monotonically increasing with the increase of interconnect width. Compared to the standard memory operation (*worst-case*), across three NVM technologies, and Width = 20..24 *nm*, the *worst-case CiM operation* reduce the EM-induced MTTF for 3.31x.

8.5. Mitigation of EM degradation for CiM

According to Table 8.2, for all the STT-MRAM, ReRAM, and PCM technologies, N_{rows} = 256 results in a very low EM-induced MTTF. Figure 8.4, shows the impact of the activation pattern on the EM-induced MTTF in STT-MRAM, ReRAM, and PCM technologies for different numbers of the CiM inputs. Compared to the worst-case (activation of a block farthest from the sensing circuitry) pattern, the best-case (activation of a block closest to the sensing circuitry) pattern can improve the EM-induced MTTF by about 4.93x, averaged over all technologies. Realization of the best-case activation pattern can be done by modifying the *memory write allocator* module such that, for the CiM purpose, it allocates the memory rows based on the EM-aware best-case activation pattern. Please note that as the analog MAC operation is not sensitive to the position of the activated rows, the proposed best-case activation pattern does not change the intended functionality of the MAC operation. These results show that the choice of data layout and activation pattern has a significant impact on alleviating EM degradation.

8.6. Conclusion and future work

In this work, we have highlighted the EM reliability issue in the paradigm of CiM-oriented MAC. Studying the EM phenomenon during the CiM operation is challenging, mainly because of the higher current density, and pattern dependency of the row activation. We have modified the existing physical-based EM modeling for the segmented BL structure. We have identified *worst-case CiM operation*, which compared to the standard memory operation, decreases the EM-induced MTTF by an average of 3.58x in STT-MRAM, ReRAM, and PCM technologies across different interconnect design choices. Based on our results, with the help of EM-aware *activation pattern*, the EM-induced MTTF can be increased by an average of 4.93x. Our results have shown that besides the device- and electrical-level design choices, the workload has a significant impact on the EM reliability of the CiM fabrics, and workload-aware EM analysis and mitigation can be a potential future work.

9. An Efficient Test Strategy for Detection of Electromigration Impact in Advanced FinFET Memories

As already discussed in Section 2.6, in the advanced technology nodes, the memory signal lines, including SRAM's BL are susceptible to the aging impact induced by the EM. In the case that SRAM is utilized in a safety-critical application e.g., automotive, the requirements of zero Defective Parts Per Million (DPPM) as well as the functional safety standard of ISO26262 [23] mandate proper detection and mitigation of such interconnect aging failures. As a result, the testing infrastructure is not only responsible for detecting manufacturing defects occurring during the fabrication phase but is also expected to detect and report EM failures in the field. Ideally, an early detection is desirable before the part fails in the normal functional mode, to avoid safety violations. Particularly in automotive applications Advanced Driver Assistance Systems (ADAS), the impact of EM is pronounced, which can severely impact the overall Failure In Time (FIT) rate of the system. The extensive use of electronic components in the automotive industry mandates the failure rate of individual components to be extremely low. This is further exacerbated by the demands for high reliability due to strict regulations, such as the functional safety standard ISO26262. In addition, extreme operating environments, and in particular, the very wide operating temperature range of electronic components in automotive applications can severely aggravate the EM impact and increase the FIT rate. Due to the expected long operational lifetime of the automotive systems, they are more subject to interconnect aging effects, and hence, the need for detection and correction of EM-induced failures in the field is extremely important. The lack of proper detection and correction of EM-induced failures in the memory components of automotive systems can lead to catastrophic failures.

In this chapter, we aim to develop an efficient test strategy for early detection of the EM effect on the BL of the advanced FinFET-based SRAM memory, which is a long and tight-pitch interconnect, and most susceptible to EM impact. The effect of the EM is proportional to the length of the interconnect, the longer the interconnect, the more severe the EM impact, and short interconnects, more precisely, shorter than the threshold of Blech length are even EM-immortal [2].

To develop the aforementioned test strategy, we use the fact that the SRAM writability is degrading by increasing the resistive parasitic of the BL [107]. The EM-induced fault can manifest itself by an increase of the BL parasitic resistance. By leveraging the dependency between the consequent memory operations, as well as tuning the operational conditions, including the frequency, electrical-level voltage values, and temperature, we can intensify the sensitivity of the SRAM writability to the BL parasitic resistance. Basically, the earlier (the smaller amount of the resistance increase) the detection of the EM effect, the more efficient the EM test strategy. Although exploring the EM phenomenon in the memory modules has been already discussed in the literature, this chapter is the first work that investigates the testing solution for the EM. Our contributions to this chapter are as follows. **(i)**. Identify the SRAM writability degradation as an effective means for the EM defect detection. **(ii)**. Sensitize the SRAM writability by tuning the operating temperature and voltage values. **(iii)**. Utilizing the dependency between the consequent SRAM operation to further improve the efficiency of the EM test strategy.

The rest of this chapter is organized as follows; Section 9.1 reviews the essential background and related work, followed by Section 9.2, which discusses the core methodology for our EM test. Section 9.3 presents the results of our work, and finally, Section 9.5 concludes the chapter.

9.1. Background and related work

9.1.1. Advanced transistor technology

Due to growing leakage and short-channel problems of conventional planar MOSFET transistors, at some point, it became impossible to comply with Moore's law by further scaling down the transistor sizes. As a result, FinFET technology was introduced as a solution and deployed into production starting with sub-20 nm feature sizes. The distinguishing characteristic of a FinFET transistor is that the conducting channel consists of thin vertical silicon 'Fins' wrapped around by gate electrodes. The effective channel width is determined by the height and width of the fin and the number of fins. Over the years, FinFET transistors of different generations and types with different characteristics have been released, which helped to shrink the technology up to 3 nm. Nevertheless, FinFET has also reached its limit and Gate-All-Around (GAA) FET technology with its various variations (MBCFET, Nanowire/Nanosheet GAAFET, RibbonFET) is being actively considered as a possible successor. In all the discussed advanced CMOS technologies, the threshold voltage of the transistor elements is decreasing, resulting in higher performance and energy efficiency. However, due to the smaller interconnect cross-section of the fabricated interconnects, their parasitic resistance is relatively large. The SRAM write operation is challenging in the advanced resistance-dominated interconnect, since the voltage of the write drivers is considerably degraded due to the large resistance parasitic, i.e., the SRAM writability is degraded in the advanced technology nodes [107].

9.1.2. Related work

Analyzing the EM phenomenon in the memory module, particularly on the memory BL, has already been explored in the previous related work. The authors in [31], [33] have shown that EM can be a potential reliability challenge for the SRAM BLs, by also modeling the parasitic RC of the wires.

The discussed related works have accomplished the so-called *offline* EM-induced failure characterization. Therefore, such characterization cannot provide enough failure information based on unique memory working conditions such as size, workload activities, and temperature fluctuations. To the best of our knowledge, no previous work discussed testing for EM in memories, and this chapter is the first work addressing such an issue.

9.2. Proposed methodology

9.2.1. Advanced Inductive Failure Analysis (AIFA)

The first step toward the test solution development is the Advanced Inductive Failure Analysis (AIFA). For this step, the behavior of two circuit instances, namely *fault-free* and *defect-injected*, need to be clearly distinguished (see Figure 9.1). In this regard, for the fault-free instance, the SPICE netlist can be directly extracted from the circuit layout. For the defect-injected instance, however, the defect can be injected either into the extracted SPICE netlist or directly into the Graphic Design System (GDS) file before netlist extraction. By having the two SPICE netlists (corresponding to the fault-free and the defect-injected instances), performing the SPICE simulation with the identical setup, including frequency, voltage, temperature, and the range of parasitic effects, is the next step in AIFA flow. Eventually, after performing the SPICE simulations, and comparing the respective waveforms, the fault is manifested and can be further modeled [29], [34].

The type of the defect-injection needs to be decided based on the impact of the underlying phenomenon. Typically, in the case of advanced FinFET memories, defect injection can be performed at four different levels. Level 1 is the transistors (either FinFET or GAA-Fet). Fin open, or the short connection between

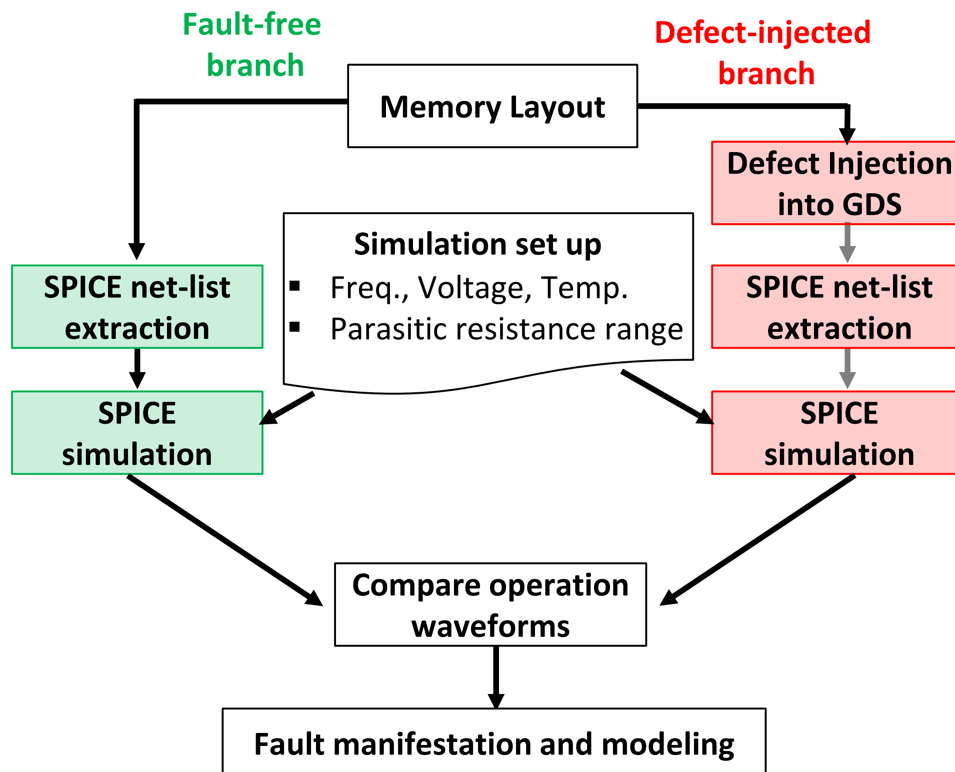


Figure 9.1.: Advanced Inductive Failure Analysis (AIFA)

the fin and gate is categorized in this level. Level 2 is the memory cell (for SRAM, it would be the structure of a 6-transistor cell as illustrated in Figure 2.1 (b)), any short/open connections between the internal nodes are a defect in this level. Level 3 is the memory array, weaker or stronger connections in the array-level interconnects are under this level. Finally, the memory periphery circuits can also be affected by defects, and injecting the defect into the periphery circuit is classified in level 4. As already discussed in Section 2.6, EM increases the resistance of the interconnect, hence, EM-induced defect injection is the level 3 defect injection.

9.2.2. Degraded SRAM writability as a means to EM test

Our focus in this work is to test the SRAM BL for the effect of the EM. Compared to the WL, SRAM BL is more frequently accessed and, hence, typically more vulnerable to the EM.

The larger the BL parasitic resistance, the more challenging the SRAM write operation, i.e., the more degraded the writability. In this work, we leverage the degraded SRAM writability to test the SRAM BL for the effect of the EM. Since the proposed EM test is done at the logical level, the EM can be sensitized and hence tested if it results in write faults. So, the main purpose is to magnify the impact of EM on the write operation to cause a fault. The following parameters affect the SRAM writability.

- *Distance between the cell and drivers*, the longer the distance between the WL and BL drivers and the SRAM cell, the more degraded the heritability. Therefore, the farthest cell from the drivers has the most degraded heritability. In fact, such SRAM cell is the first one to experience the writing failure due to the EM.

- *Direction of the write operation*, due to the RC parasitic of the WL, the SRAM switching write operation (0→1 or 1→0) do not have a symmetric writability. The type of write operation that

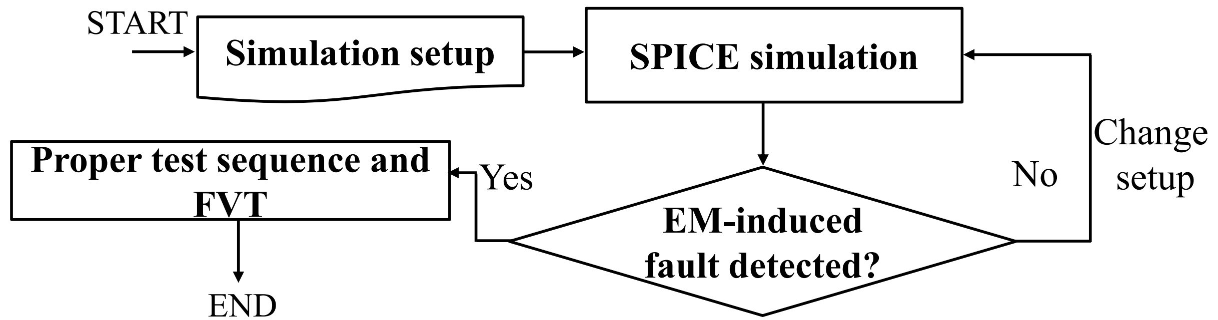


Figure 9.2.: The flowchart for our proposed EM test

results in a more degraded write margin can improve the sensitivity of the proposed EM test. In our simulation setup, write 1→0 is more critical than write 0→1.

- *Temperature*, the higher the temperature, the more degraded the SRAM writability.
- *Voltage source (VDD)* of the SRAM array, the higher the VDD of the SRAM array, the more difficult to overdrive the internal positive feedback of the SRAM cell, hence, more degraded the write margin.
- *Frequency*, the higher the SRAM frequency, which indicates the shorter latency for the write operation, results in a more degraded SRAM writability.

Based on the aforementioned discussion, for our EM test, we consider the 'write 1→0 in the SRAM cell farthest from the driver', as the target test instruction, which can be further sensitized by increasing the test temperature, and VDD of the SRAM array. Unsuccessful write with the above specifications indicates an increase of the BL resistance (for our case, due to the EM). In fact, our proposed EM test methodology is able to predict the effect of the EM in the resistance increase phase of the EM progression, i.e., after the fully-grown void covered the cross-section of the interconnect.

9.2.2.1. Dependency between the SRAM instructions

Besides the discussed impactful parameters on the SRAM writability, the dependency that exists between the consequent SRAM operations can also be another means for the further degradation of the SRAM writability. To ensure the reliability of the consequent SRAM operations, usually, after each operation, both the BL and BLB are pre-charged to a certain level of voltage ($V_{\text{pre-charge}}$). However, the final voltage of the BL and BLB cannot reach the exact value of ($V_{\text{pre-charge}}$), and there is a residue voltage on the BL and BLB. The final voltage of BL and BLB are operation-dependent, and such operation dependency creates a coupling between the SRAM operations. Therefore, having a proper longer sequence before the target write operation can magnify the EM defect and cause the target write operation to fail even with a smaller EM-induced resistance increase, i.e., earlier detection before functional failure. However, a longer sequence means a longer test time, so there is a trade-off between early detectability and EM test costs. To quantitively measure the dependency between the SRAM operations, we introduce difficulty to operation (D.t.O) [Volt] (Equation 9.1). The D.t.O is defined as the Euclidean distance between the two voltage points (V_{BL} @ time: t^-, V_{BLB} @ time: t^-) and (V_{BL} @ time: t, V_{BLB} @ time: t). The larger D.t.O indicates the more difficult the SRAM operation at time t.

$$D.t.O = \sqrt{(V_{BL}@\text{ time} : t^- - V_{BL}@\text{ time} : t)^2 + (V_{BLB}@\text{ time} : t^- - V_{BLB}@\text{ time} : t)^2} \tag{9.1}$$

Table 9.1.: Simulation setup parameters

SRAM array-level parameters	
SRAM array organization	64 Rows and 76 Columns
CMOS technology	Global Foundry 22 nm
BL parasitic per cell	R = 10 Ω, C = 55.68 aF
WL parasitic per cell	R = 10 Ω, C = 212.2 aF
VDD (periphery circuit)	750 mV
VDD (SRAM array)	(750 or 800) mV
Pre-charge level (BL and BL-bar)	VDD (750 mV)
Write voltage	High: VDD (750 mV) Low: -50 mV
Temperature	125°C
SRAM frequency	1250 MHz
Interconnect parameters	
ρ_{Cu} [40]	9.5e-8 Ω.nm
ρ_{Ta} [40]	3e-5 e-8 Ω.nm
Width (W) and height (H) of Cu interconnect	W = 41 nm H = 20.5 nm
Height of thin Ta-based barrier (h_Ta)	0.9 nm
D_a [69]	6.072e-21 ($m^2s^(-1)$)
Z	1

At the time t^- i.e., the end of the previous SRAM operation, the voltage of BL and BLB is close to (not exactly reaches) $V_{\text{pre-charge}}$. Further, the time t indicates the start of the new SRAM operation, and the voltage of the BL and BLB changes based on the type of the SRAM operation. For instance, for write 1→0 the BL and BLB need to be charged to the low (0 V, or even a negative value to assist the SRAM write operation [14]), and high (VDD, or even slightly higher) voltage levels respectively. Figure 9.2 shows the flowchart of our proposed EM testing strategy. We can start by considering a particular simulation setup including frequency (F), voltage levels (V), and temperature (T), as well as a test sequence, and perform the SPICE simulation on the extracted defect-injected netlist (please see Section 9.2.1). If the write operation (1→0) at the farthest cell from the drivers shows the failure, it means that the selected FVT parameters and test sequence are sufficient. Otherwise, the test setup needs to be updated. Updating the test setup can be performed by manipulating the test sequence or changing the FVT parameters. Please note that, although following the flowchart in Figure 9.2 guarantees the observation of the EM-induced defect, it cannot ensure whether the test sequence is optimized or not.

9.3. Results and discussion

9.4. Simulation setup

Table 9.1 shows our simulation setup and parameters. For the SRAM array, the VDD for the periphery circuit, the pre-charge level of BL and BLB, as well as the high-level voltage of the write driver are considered to be 750 mV. While the low-level voltage of the write driver has a negative value to assist the SRAM write operation. The VDD of the SRAM array is a variable and selected as either 750 mV or 800 mV. Also, the test temperature is relatively high, which results in a more sensitized SRAM write

Time	$t^{5(-)}$	$t^{4(-)}$	$t^{3(-)}$	$t^{2(-)}$	t^{-}	t
Operation	R1	Write 0→0	R0	Write 1→1	R1	Write 1→0
Address	--	--	--	--	--	Farthest from drivers

-- optional cell address

Op	D.t.O (mV)	Op	D.t.O (mV)	Op	D.t.O (mV)	Op	D.t.O (mV)	Op	D.t.O (mV)
R0	749.1	R0	0.905539	**R0**	**800.1005**	R0	0.905539	R0	799.1
R1	**750.1005**	R1	0.905539	R1	799.1	R1	0.905539	**R1**	**800.1005**
W 0→0	747.2	**W 0→0**	**2.8**	W 0→0	800.0049	W 0→0	2.8	W 0→0	797.2
W 0→1	749.1002	W 0→1	1.081665	W 0→1	799.4005	W 0→1	1.081665	W 0→1	799.1002
W 1→0	749.4005	W 1→0	1.081665	W 1→0	799.1002	W 1→0	1.081665	W 1→0	799.4005
W 1→1	750.0052	W 1→1	2.8	W 1→1	797.2	**W 1→1**	**2.8**	W 1→1	800.0049

Figure 9.3.: The EM test generation by using the concept of D.t.O

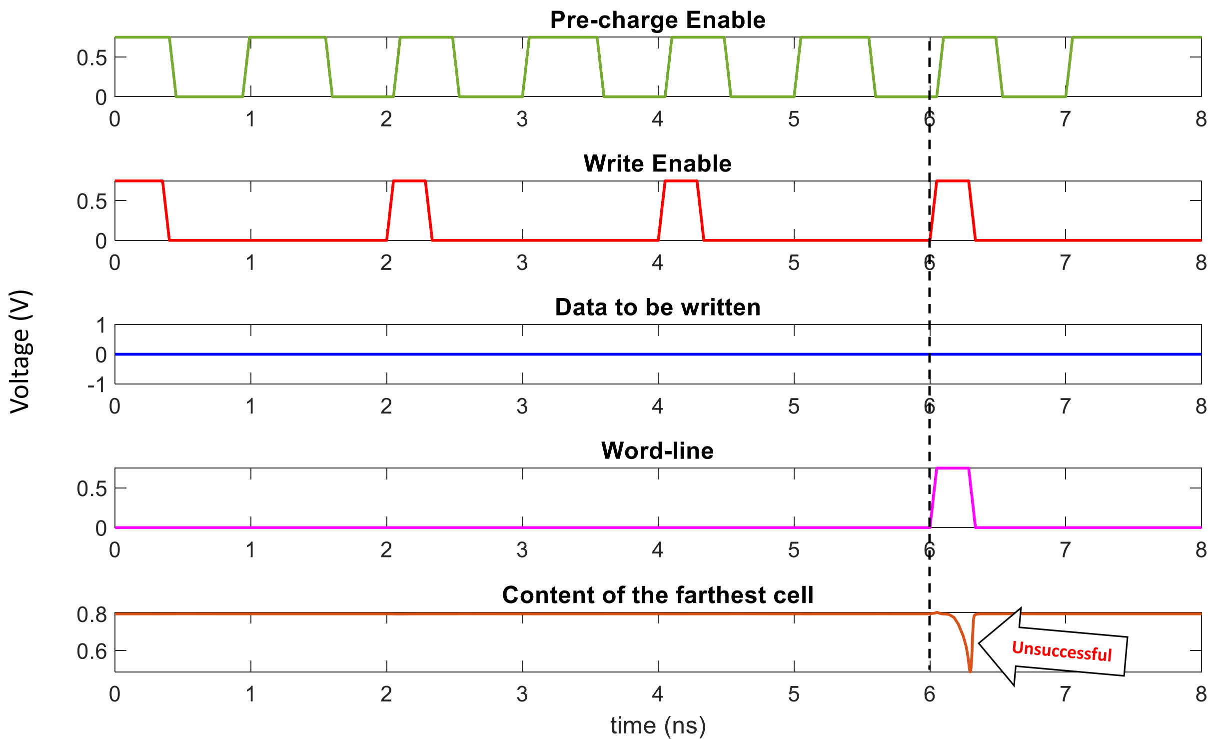


Figure 9.4.: The waveform as the output of the SPICE simulation on the defect injection netlist

operation. Moreover, our target SRAM sub-array has 64 Rows (0..63) and 76 Columns (0..75). Hence, the farthest SRAM cell from the drivers is located at Row number 63 and Column number 75.

Table 9.2.: Impact of the VDD of SRAM array and number of memory instructions on the early detection of the proposed EM test

VDD of SRAM array	Number of memory operations in the EM test vector	Required BL's resistance increase for detecting the EM effect	Required time for detecting the EM effect (Months)	
			Absolute	Relative
750 mV	1	175.7 Ω	$T_{Nucl+Incub}$+11.1	0
	2	147.6 Ω	$T_{Nucl+Incub}$+9.30	-1.8
	3	145.2 Ω	$T_{Nucl+Incub}$+9.15	-1.95
	4	145.2 Ω	$T_{Nucl+Incub}$+9.15	-1.95
	5	145.2 Ω	$T_{Nucl+Incub}$+9.15	-1.95
	6	145.2 Ω	$T_{Nucl+Incub}$+9.15	-1.95
800 mV	1	44.8 Ω	$T_{Nucl+Incub}$+2.41	-8.69
	2	27.8 Ω	$T_{Nucl+Incub}$+1.49	-9.61
	3	**26 Ω**	$\boldsymbol{T_{Nucl+Incub}}$**+1.40**	**-9.70**
	4	26 Ω	$T_{Nucl+Incub}$+1.40	-9.70
	5	26 Ω	$T_{Nucl+Incub}$+1.40	-9.70
	6	26 Ω	$T_{Nucl+Incub}$+1.40	-9.70

9.4.1. Test sequence generation for the EM

Figure 9.3 shows the generation of the EM test vector based on the D.t.O metric. As discussed in Section 9.2.2.1), the larger the D.t.O, the more difficult the write operation. By this approach, we continue to find a test sequence, terminated to write 1→0 in the SRAM cell farthest from the driver. Besides, Figure 9.4 shows the waveform of the EM-induced defect injected SRAM instance under the proposed EM test sequence.

Besides, Figure 9.3 shows a periodic behavior in the generated test structure. By moving backward in the time, the 'Read 1' operation has appeared at t=5 ns, and again at t=1 ns. So, to generate the EM test sequence, using the D.t.O concept for five times is enough, and further extending the test sequence can be done by only replicating the sequence period.

9.4.2. Applying the proposed EM test methodology on the SRAM array

Table 9.2 presents the results of our proposed EM test methodology. Please note, that the number of the SRAM operations in the proposed test sequence are counted by moving backward in the time and starting from the core instruction of Write 1→0 at Row 63 and Column 75 (farthest cell from the drivers). For instance, one SRAM operation means only write 1→0 operation on the farthest cell from the driver, while two SRAM operation indicates (Read 1 at Row (0) and Column 75 – Write 1→0 at Row 63 and Column 75), and so on.

The required BL's resistance increase for the manifestation of the EM defect can be obtained through the SPICE simulation. Further, to convert the amount of the resistance increase to the detection time, we use Equation 2.5 with the interconnect parameters outlined in Table 9.1. Moreover, obtaining the detection time also needs the current density (j) as the input. The assumed j in Equation 2.5, is the equivalent current by taking into account all the SRAM operations and following the method introduced in [112].

The interesting point in Table 9.2 is that for both the values of VDD of SRAM array (750 or 800 mV), the early detection of the proposed EM test stagnated after three instructions of the proposed EM test sequence. In this way, increasing the test instruction beyond three only increases the test cost and has no benefit regarding EM detection.

This observation highlights the possibility of the optimized EM test sequence, which results in lower-cost EM test solutions. Our optimized EM test methodology can detect the EM effect 9.7 months earlier than

the non-optimized EM test. To the best of our knowledge, this work is the first work that aims to avoid EM-induced field failure and proposes a low-cost yet effective EM testing strategy for the SRAM BL.

9.5. Conclusion

In the advanced technology nodes, SRAM BLs are susceptible to the EM, mainly due to the higher current density and temperature. Therefore, the in-field EM test, as mandated by the ISO26262 in automotive applications, as well as other safety-critical domains, is required before the failure in the functional mode. In this chapter, for the first time, we have proposed an EM test methodology for advanced SRAM memories in safety-critical applications. We have used the SRAM writability degradation as a means for the EM test and identified the farthest cell from the drivers as the cell with the most degraded writability. Moreover, we have further sensitized this degraded writability by increasing the temperature and VDD of the SRAM array. We have also leveraged the dependency of the SRAM operations to improve and optimize our proposed EM test for early EM detection.

10. Electromigration-aware Design Technology Co-Optimization for SRAM in Advanced Technology Nodes

Workload dependency is a challenge in the investigation of the EM profile for the memory signal lines. The impact of the EM on the memory WL, BL, and BLB heavily depends on the current passing through these lines. The type (read or write), data, and address of the operation change the EM reliability profile of the WL and BL since these factors alter the current passing through the memory lines. Besides the effect of the WL's and BL's current on their EM profile, they have a significant effect on the overall energy and performance. Therefore, there is a trade-off between these design merits and DTCO needs to be performed by carefully considering these trade-offs.

In sub-10 nm technology, it has become increasingly challenging to extend the conventional scaling laws. Classical Dennard's lambda scaling rules have broken already. What is left is the economical cost scaling law of Moore's law, and the node terminology for that roadmap is defined by industrial foundries once a new 'node' is put in production. As a workable alternative, advanced technology research has focused on so-called DTCO iterations, where critical parameters in the device and wire fabrication are co-optimized with important (circuit) design parameters. Without such a crucial co-optimization, technology scaling would have stopped already. Hence, in the research phase, the concept of a node does not make sense any longer, and it is not used even by scientists. Instead, the values of the above-mentioned critical parameters determine the state-of-the-art to compare different results.

In this work, we consider two SRAM technologies, because of the predominant wire focus of this chapter, we will mainly focus on Critical Dimension (CD) of a wire, and sometimes pitch (distance between 2 wires) [116]. Specifically, we compare two cases represented by the 22 nm and the 12 nm smallest BL CD, which are respectively referred to as 22 nm and sub-5 nm technology nodes by industry.

Our contributions in this chapter are as follows: **(i)**. Show the trend of the EM risk exacerbation with respect to technology scaling in SRAM WL and BL. **(ii)**. Perform a detailed analysis of the electrical-level characteristics of the SRAM design and show the access dependency of the EM profile. **(iii)**. Propose an EM-aware DTCO for the advanced SRAM design based on 12 nm CD.

The rest of this chapter is organized as follows. In Section 10.1, we review related work and continue with elaboration on our core methodology in Section 10.2. Section 10.3 contains the results and finally, the conclusion of this chapter will be presented in chapter 10.4.

10.1. Related work

The EM susceptibility of the memory BL has been investigated in several related works. In [31], [33], the susceptibility of the SRAM's BL has been studied. The EM modeling that has been used in this work is Black's equation that obtains the EM-induced MTTF [1]. Due to the empirical nature of Black's equation, it cannot capture the direct effect of different design parameters such as the wire length (l), and hence, such formulation cannot be used during EM-aware DTCO. Moreover, although the BL has been modeled as a segmented wire with a distributed RC network, its effect on the unequal segment current has not been taken into account.

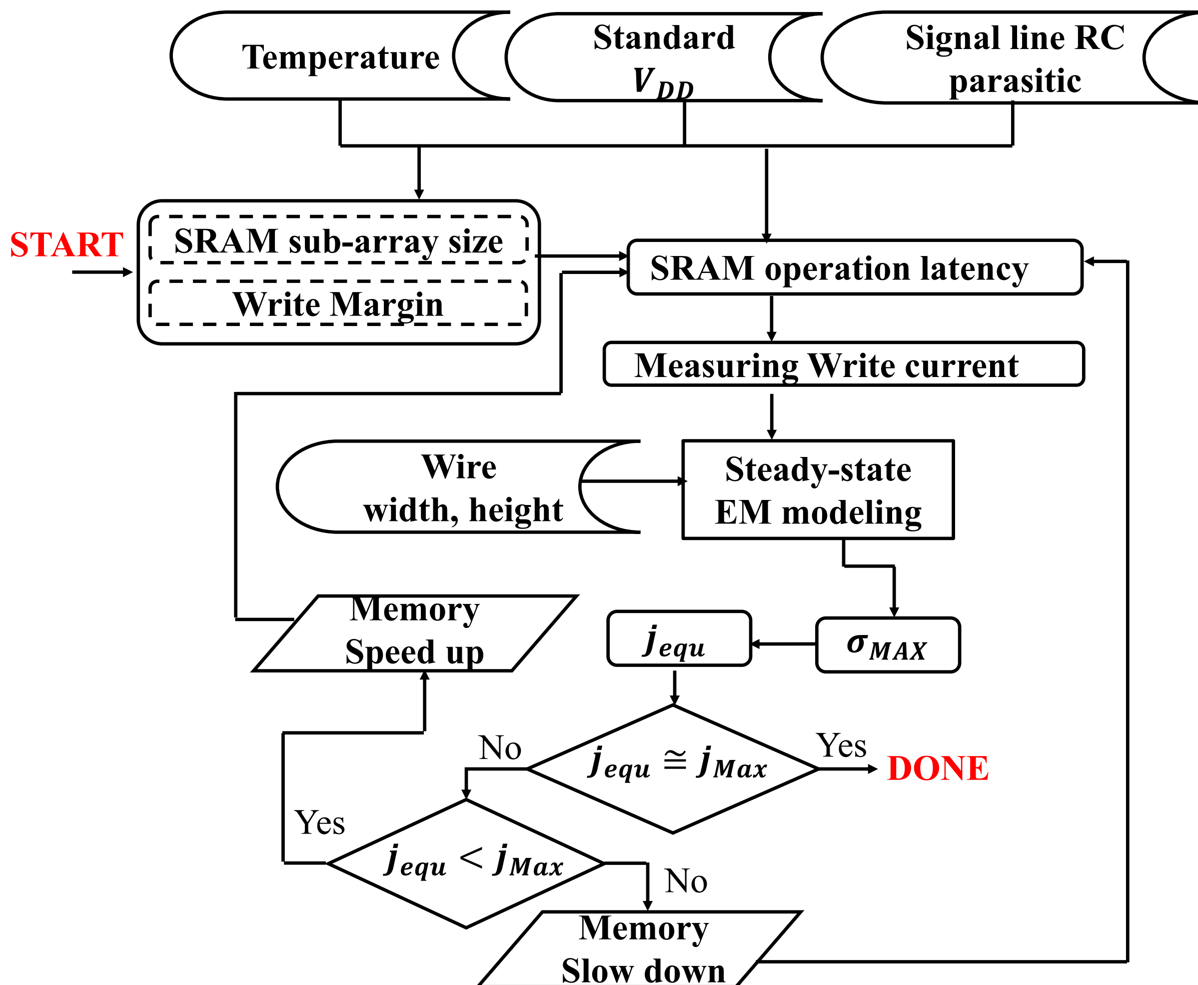


Figure 10.1.: The proposed EM-aware DTCO methodology for SRAM fabricated in 12 nm CD

On the other side, [96], [111] investigate the effect of the EM on the BL of the STT-MRAM. The authors have considered the physical-based EM modeling and analyzed the data-dependent impact of the EM on the BL of the STT-MRAM. However, due to ignorance of the wire RC parasitic, the address dependency and effect of the segmented wire have not been studied in [96], [111]. To the best of our knowledge, in this chapter, we are the first to comparatively investigate the EM reliability of the WL and BL of the SRAM under strong access-dependent conditions. Hence it is not possible for us to perform a quantitative comparison with related work. So, only qualitative statements are provided above. According to the *Blech length constraint*, the longer the line, the more susceptible to EM [2]. Although other long lines exist in the large SRAM macros (such as interconnect networks), the main focus of this work is to investigate the impact of EM on the WL and BL. Due to the severe space limitation inside the SRAM sub-array, WL and BL have the tightest pitch size, and hence, are more susceptible to EM.

10.2. EM-aware DTCO methodology

Figure 10.1 shows our proposed EM-aware DTCO methodology. The *temperature*, *standard voltage*, *wire parasitic* and *BEoL wire dimensions* are the input for our proposed EM-aware DTCO mechanism. Our aim is to find the largest SRAM size, working at the highest performance, however, by ensuring the satisfaction of the EM reliability criteria, which is the maximum current density (j_{max}).

10.2.1. SRAM sub-array size

Obtaining the *SRAM sub-array size* (number of rows and columns) is crucial for DTCO, typically, having large memory sub-arrays are beneficial for the overall performance and energy efficiency, however, large resistive parasitic results in SRAM writing failure. In resistance-dominated advanced tight-pitch interconnect, the increased BEoL resistance per unit length has degraded the SRAM write margin [107]. Applying the negative write voltage driver is one of the mechanisms that can improve the SRAM writability [14]. We also utilize the negative write driver and try to maximize the SRAM sub-array size by using this *write assist mechanism*.

To have a fair comparison between the SRAM designs based on 22 nm and 12 nm CD, we consider the same sub-array size in both technologies, which is determined based on the maximum possible size in SRAM, based on 12 nm CD. During the proposed EM-aware DTCO, we increase the SRAM sub-array size until it is limited by the SRAM writability.

10.2.2. EM reliability analysis

For the Cu-based BEoL interconnect corresponds to the 12 nm CD, which is also longer than $5\mu m$, the maximum allowed current density (j_{max}) (For 10 years EM lifetime) is drastically decreasing to $\sim 1\frac{MA}{cm^2}$ [69].

Increasing the current density far beyond this value increases the risk of EM failure. Nevertheless, investigating this criterion in the interconnects with *RC* parasitic can be misleading. To further elaborate, the current for the segmented (each segment corresponds to an RC element per SRAM cell) SRAM WL and BL is not equal for all the segments and the value of the current is heavily access-dependent. Hence, generalizing the current passing through one segment to the entire line may either overestimate or underestimate the EM risk.

10.2.2.1.EM-aware equivalent constant current density

To address the issue of non-constant current in the interconnect segments, we propose to use an *EM-aware equivalent constant current density (j_{equ})*. If a realistic segment-dependent current density induces $\sigma^{\text{MAX}}_{\text{steady-state}}$, the ($j_{equ}$) is a constant *segment-independent* current density that induces the same $\sigma^{\text{MAX}}_{\text{steady-state}}$ on the wire.

Monitoring the j_{max} needs to be done based on the access that results in the highest current passing through the line. The SRAM write operation typically results in a much higher current than the read operation. The change of the BL (or BLB) voltage during the read operation is relatively smaller than the write operation. In read operation, $\Delta V_{\text{either BL or BLB}} \geq SM_{\text{sense amplifier}}$ (SM: sense margin). However, the write operation requires the change of the BL (or BLB) voltage to 0 V. Moreover, utilizing the negative write driver to improve the writability of the SRAM, particularly in the 12 nm CD, results in an even higher voltage drop on the BL. So, due to the larger current during the SRAM write, we consider this operation for our proposed EM-aware DTCO mechanism.

10.2.2.2.Trade-off between the EM reliability and performance

To investigate the EM reliability, we take into account the maximum induced stress at the steady-state through Equation (2.9)-(2.11), which requires the current of each segment. The current passing through the WL and BL of the SRAM is access dependent; i.e., it varies depending on the access (read or write), data, and address. Taking into account the address dependency implies activating each memory address and measuring the currents, which is extremely effortful. Moreover, besides the access dependency of the write current, measuring the current of all the segments of the SRAM sub-array is not practical. To overcome this issue, we consider all four transitions in the write operation: $0 \rightarrow 0$, $0 \rightarrow 1$, $1 \rightarrow 0$, and 1

Table 10.1.: Parameters for co-simulation of the energy and EM reliability

Parameter	**22 nm CD**	**12 nm CD**
Temperature ($°C$)	100	
Array organization	256 Rows and 256 Columns	
Standard VDD (mV)	800	700
Wire length per cell (nm)	BL = 231 WL = 220 [28]	BL = 90* WL = 168*
Wire (Width/Height) (nm)	BL: (22/41.8) WL: (22/41.8) [28]	BL: (12/12)** WL: (13/26)**
Copper resistivity (ρ_{Cu}) ($\Omega \cdot nm$) [40]	6.8e-8	9.5e-8
Wire ($R(\Omega), C(aF)$) per cell	BL: (35.18, 89.92) WL: (24.09, 61.59) [28]	BL: (74.47, 37.7)* WL: (13.07, 107)*

* information extracted through the layout
** information extracted through the model fitting

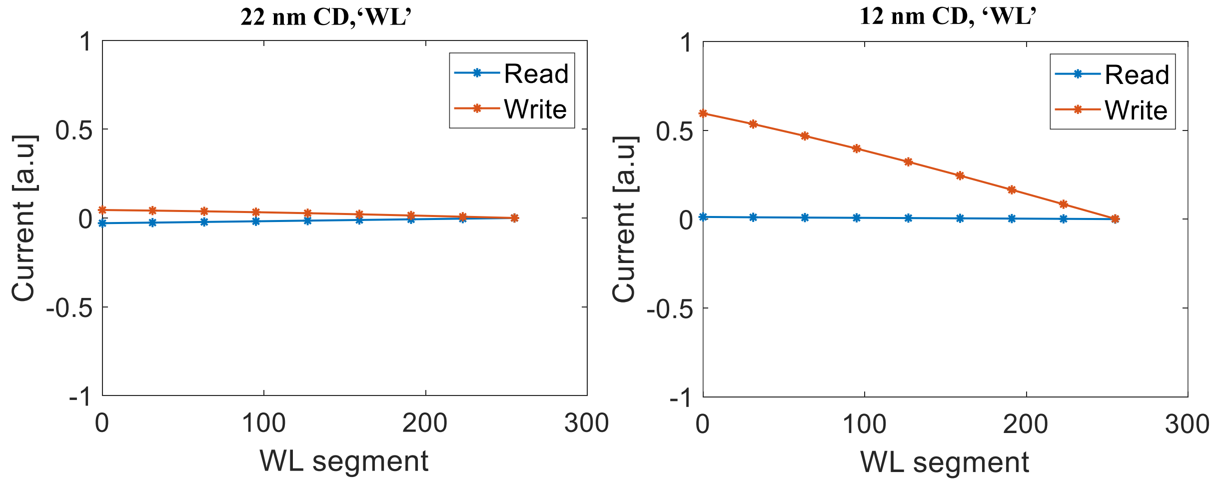


Figure 10.2.: The current versus the WL segment during the read and write operations in two 22 nm and 12 nm CD technologies

$\rightarrow$ 1, however, only on a limited number of selected rows. These selected rows are located at first (closest to the write driver and the pre-charger, and farthest from the sense amplifier), middle, and end (farthest from the write driver and pre-charger, and closest to the sense amplifier) of the sub-array. Moreover, for each activated row, we only sample from the current of a few segments, and *extrapolate* the rest, by fitting the current samples to a polynomial curve.

By having the current-segment information, we can obtain the steady-state σ_{max} and eventually, the j_{equ}. If the j_{equ} is less than the j_{max}, it implies that there is still room to decrease the SRAM latency. Otherwise, the latency of the SRAM needs to be compromised to meet the EM reliability criterion: for the interconnect corresponds to 12 nm CD, $j_{max} \leq\sim 1\frac{MA}{cm^2}$. To find the optimized point for the latency and EM-reliability, we tune the SRAM operating cycle in such a way that the current density of both the WL and BL is close to $\sim 1\frac{MA}{cm^2}$.

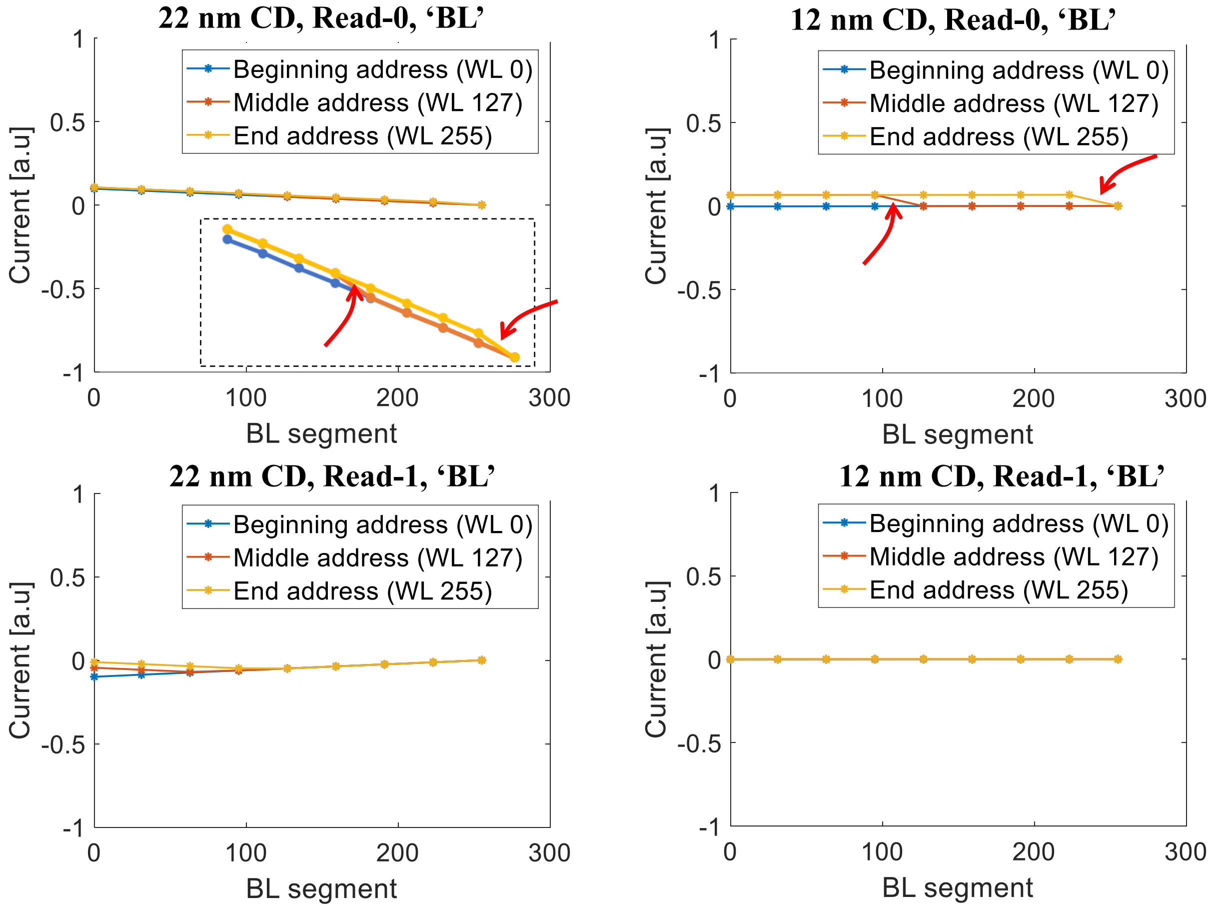


Figure 10.3.: The current versus the BL segment during the read operation in two SRAM designs based on 22 nm and 12 nm CD, red arrows show the sudden decrease of the current at the activated segment during read-0

Table 10.2.: Summary of the results for performing the EM-aware DTCO on the SRAM based on 22 *nm* and 12 *nm* CD

Parameter	**22 nm CD**	**12 nm CD**
EM reliability criteria (j_{max}) [69]	$\leq\sim 3.5\frac{Ma}{cm^2}$	$\leq\sim 1\frac{Ma}{cm^2}$
EM-aware DTCO analysis	Not required ($j \sim 0.589\frac{Ma}{cm^2}$)	Required
Achieved j after DTCO (WL)	–	$0.884\frac{Ma}{cm^2}$
Achieved j after DTCO (BL)	–	$1.04\frac{Ma}{cm^2}$

10.3. Results and Discussion

The technology parameters corresponding to 22nm and 12nm CDs which are used for our analysis are summarized in Table 10.1. The temperature has been selected by targeting the high-performance cache units. Moreover, as it has already been discussed in Section 10.2.1, the size of the SRAM sub-array has been determined based on the 12 nm CD, which has a degraded write margin of ~-570 mV, compared to the write margin of 0 V in 22 nm CD technology. In the 12 nm CD, the wire length per cell, as well as the RC parasitic have been extracted from the layout. While the wire width and height have been obtained by

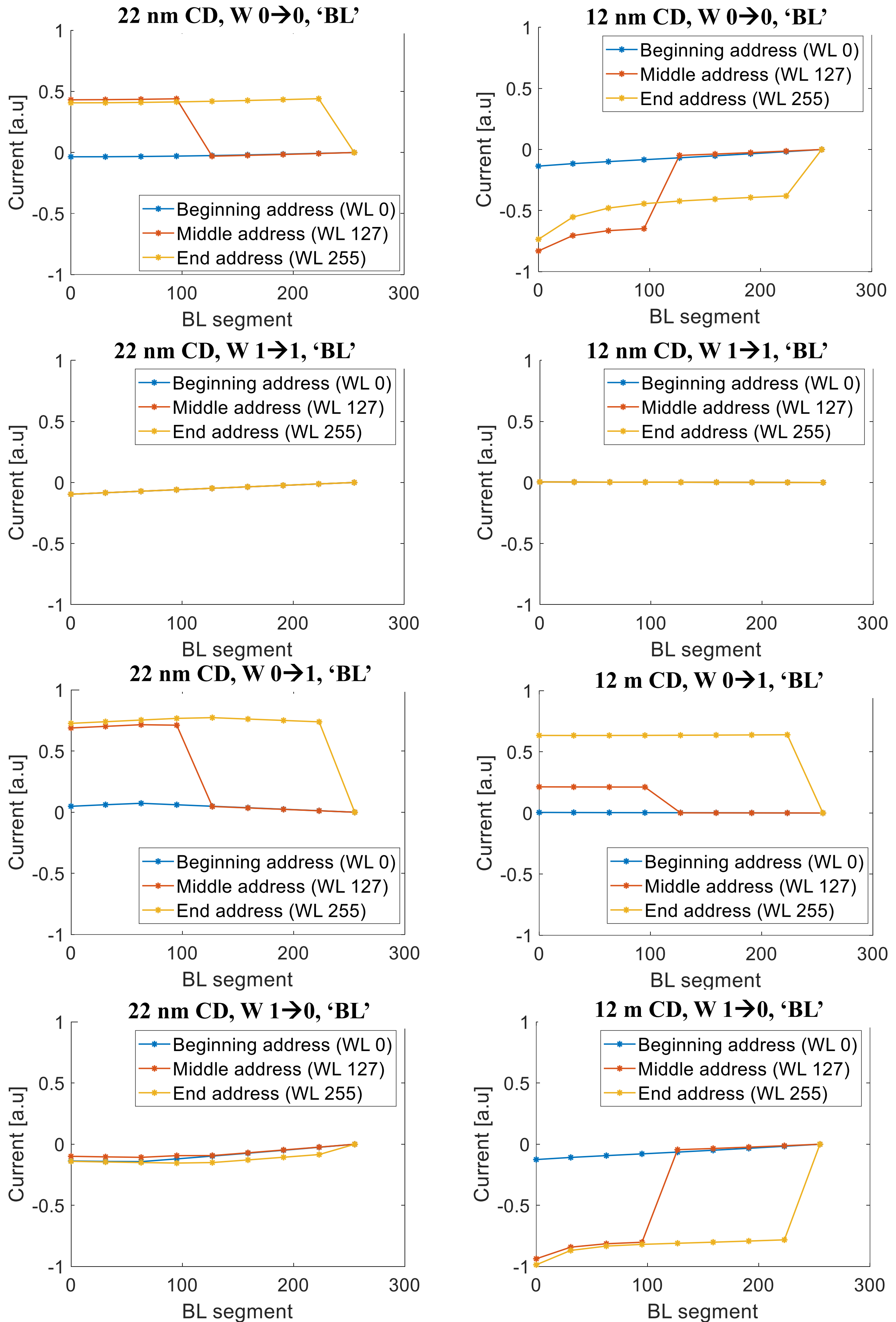


Figure 10.4.: The current versus the BL segment during the write operation in two SRAM designs based on 22 nm and 12 nm CD

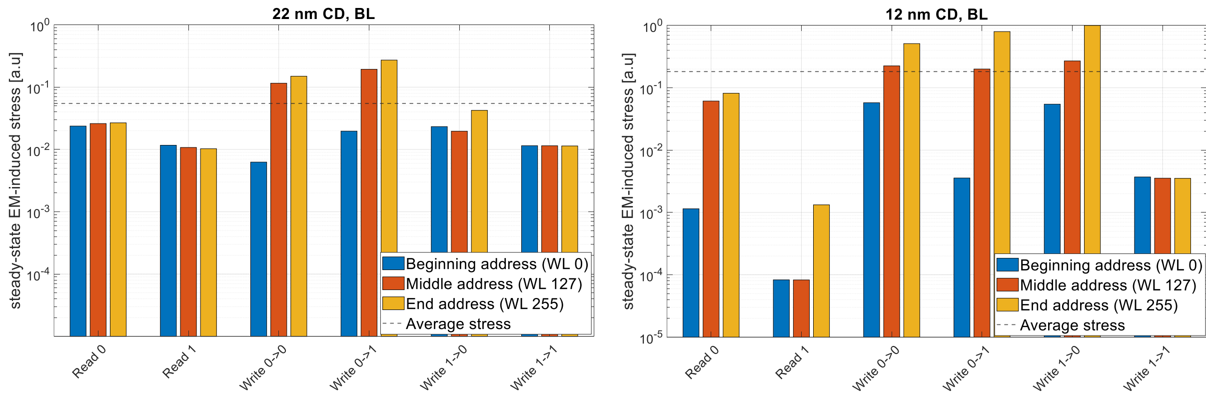


Figure 10.5.: Steady-state hydrostatic stress on the BL, based on 22 nm and 12 nm CD

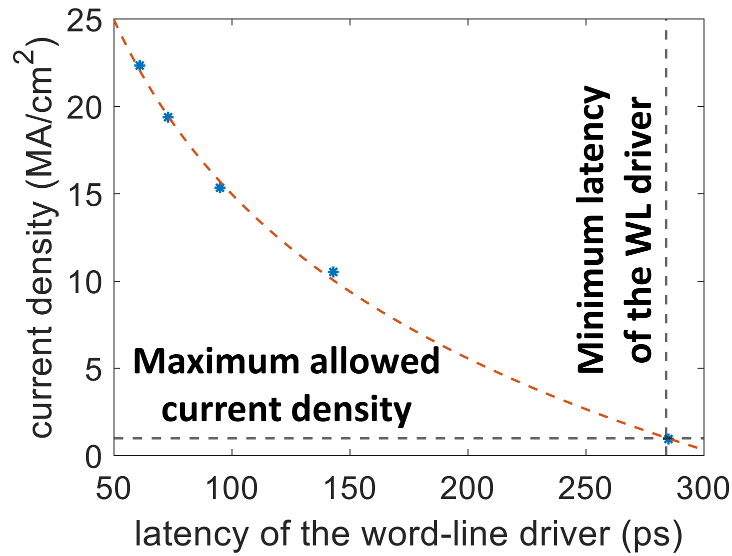


Figure 10.6.: The WL current density versus the latency of the WL driver

Table 10.3.: Summary of the comparison in SRAM based on 22 *nm* CD and 12 *nm* CD, the parameters are normalized to values based on 22 *nm* CD

Design Merit	**22 nm CD**	**12 nm CD**
SRAM cell area	1	0.214
Operating Cycle	1	0.531
Average Read Power-Delay Product (PDP)	1	0.318
Average Write PDP	1.514	0.685
EM-driven stress, BL, Read	1	1.329
EM-driven stress BL, Write	4.012	14.304
EM-driven stress WL, Read	0.187	0.433
EM-driven stress WL, Write	7.835	22.785
Coefficient of Variation (C.V) for Stress values dispersion (access dependency)	1	1.548

using the *R* and *C* model as a function of wire length and ρ_{Cu} [28]. For 22 nm CD, the wire dimensions, as well as the RC parasitic, have been extracted from NVSim [28].

10.3.1. Required periphery circuit for the SRAM sub-array

The sense amplifier circuit simulated for both the 22 nm and 12 nm CD technologies is a latch-type sense amplifier from [12], the write drivers are based on the active-high inverting tri-state buffers, and the pre-chargers are consists of a P-FET element.

10.3.2. Impact of the access dependency on the current passing through the WL and BL

As discussed in Section 10.2.2.1, the current passing through the WL and BL is access-dependent. As we will discuss, this access dependency is not similar for the WL and BL.

10.3.2.1.WL

Charging the WL is required to make the row accessible. Among the access-dependent parameters (access type, data, and address), only access type affects the current passing through the WL. Since the duration of the time that the WL is activated is not equal for the read and write operations. Figure 10.2 shows that the current versus segment for both the 22 nm and 12 nm CD technologies is *linearly* decreasing with the number of segments, i.e., by getting farther away from the WL driver.

10.3.2.2.BL

The current passing through the BL shows significant access dependency not only to the type of SRAM operation but also to the data and address. We investigate the segment dependency of the BL current for read and write operations. As discussed in Section 10.2.2.1, and also shown in Figure 10.3, the read current is typically lower. In the case of read-0, the BL current shows a sudden decrease at the activated segment which is not the case in read-1. For each SRAM operation, the BL (and BLB) are pre-charged to VDD. In the case of read-0, a current path from the BL and through the SRAM cell is forming, which results in the sudden decrease of the current at the activated segment (which is marked with the *red arrow* in Figure 10.3). Due to the non-existence of voltage drop between the BL and the SRAM cell which stored '1', the current is decreasing linearly in the case of read-1.

Figure 10.4 shows the segment-dependent current passing through the BL of the SRAM during the write operation. Similar to the read operation, in case of the existence of the voltage drop between the BL and SRAM cell, the sudden decrease of the BL current happens at the segment corresponding to the activated row. Please note that after each SRAM operation, the BL and BLB are pre-charged to VDD. Therefore, for Write $\rightarrow 0$, the BL needs to be discharged, hence the BL current for this operation is relatively higher. As already discussed in Section 10.2.1, the negative write driver has been considered for the SRAM design based on the 12 nm CD, so the BL needs to be discharged to a negative value (i.e.„ a higher voltage drop). Therefore the write current in this technology is even higher. On top of the current magnitude, the negative write driver affects the current direction. Please note the different current directions of write $0 \rightarrow 0$ in 22 nm and 12 nm CD designs which are '+' and '-', respectively. Moreover, in the 12 nm CD, transistors have lower *threshold voltage (V_{th})*, and hence, higher current. Besides the higher current, the smaller cross-sectional area of the BEoL WL and BL in the 12 nm CD results in relatively high current density (j).

10.3.3. Impact of the access dependency on the EM reliability

Figure 10.5 shows the access dependency of the EM-induced hydrostatic stress of the SRAM's BL in both 22 nm and 12 nm CD. The average EM-induced stress is significantly higher in the 12 nm CD, and typically, write operation results in higher stress. The access dependency of the EM profile is observed in both 22 nm and 12 nm CD. For instance, in both of the designs, performing the write operation on the end rows of the SRAM sub-array results in considerably higher stress on the BL, compared to performing the same operation on the beginning addresses. Although the access dependency of the EM profile exists for both 22 nm and 12 nm CD designs, the significance of its impact on the dispersion of the stress values is technology-dependent. We use the $C.V = \frac{mean\ (\mu)}{standard\ deviation\ (\sigma)}$ as the statistical metric to show the effect of the access dependency on the EM profile. The higher the C.V, the more significant the access dependency of the EM profile.

In our simulations, we focus on the individual SRAM operations, i.e., we do not consider the effect of the *access sequence* on the EM profile. After each SRAM operation, the BL and BLB are pre-charged until VDD. Such pre-charging almost de-couples the subsequent SRAM accesses. Therefore, the effect of the access sequence on the EM profile is negligible. To the best of our knowledge, such memory access-dependent EM modeling has not been performed yet in the literature, and our results (summarized in Figure 10.3, 10.4, and 10.5) incorporate these effects.

10.3.4. EM-aware DTCO analysis

To perform the EM-aware DTCO we use the methodology discussed in Section 10.2. The EM reliability for the 22 nm CD design is not crucial, since not only the current values are smaller, and BEoL cross-sectional area is larger, but also the j_{max} criteria is far more relaxed; $\sim 3.5\frac{MA}{cm^2}$ compared to $\sim 1\frac{MA}{cm^2}$ for the wire width corresponds to 12 nm CD design [69]. Our method to tune the operating cycle of the 12 nm CD SRAM is adjusting the latency of the WL driver. Figure 10.6 shows the j_{equ} of the WL with respect to the latency of the WL driver. In fact, the WL driver needs to be slowed down until the j_{max} criteria can be met. Table 10.2 shows the summary of our EM-aware DTCO analysis. In the 12 nm CD design, by increasing the delay of the WL up to $\sim 284\ ps$ (according to Figure 10.6), the current density of WL decreases to $\sim 0.884\frac{MA}{cm^2}$. In addition to ensuring the EM reliability of the WL, the same procedure for the BL needs to be repeated. However, our results (based on the parameters outlined in Table 10.1) show that by the aforementioned WL driver slow down, the current density of the BL can also meet the criteria of $j_{max} \sim 1\frac{MA}{cm^2}$, and further slowing down the BL driver (and hence, the entire 12 nm CD SRAM design) is not necessary. Besides, due to the process variation, the BEoL interconnects can be fabricated in a smaller cross-sectional area than the nominal one. Hence, in the more advanced technologies, the process variation can result in the exacerbation of the EM risk even further.

Table 10.3 shows a summary of the design merits for the SRAM designs in two earlier introduced, 22 nm and 12 nm CD. The cell area is obtained through the SRAM layout in 22 nm and 12 nm CD technologies. While the operating cycles, average read and write PDP are based on our electrical-level simulations on both the 22 nm and 12 nm CD designs. Finally, the stress values have been obtained through the steady-state stress modeling (Section 2.6.2.4). Technology scale down (from 22 nm to 12 nm CD) improves the area efficiency, latency, and energy (PDP) efficiency by 4.67x, 1.88x, and, 2.68x, respectively. However, from the EM reliability point of view, technology scale-down exacerbates the EM risk by 2.53x.

10.4. Conclusion

Due to the technology scaling, BEoL interconnect are also fabricated in tighter pitch size. The smaller cross-sectional area of the interconnect, the higher switching speed of transistors, as well as the higher

chip temperature in dense and high-performance VLSI designs, exacerbates the EM risk even for the memory WL and BL. In this work, we have performed a detailed analysis of the SRAM designs in two technologies with different wire Critical Dimension (CD). We have shown the necessity of the EM-aware DTCO for the advanced technology designs, and proposed a mechanism for it. We have shown that though scaling down the transistor size and BEoL CD is promising in terms of performance, and energy efficiency, it dramatically increases the EM risk.

11. Conclusion and Perspectives

11.1. Conclusion

In this thesis, we have investigated reliability concerns of advanced memory technologies and paradigms, with the main focus on decision failure and EM. We have shown that the true performance and energy efficiency of advanced memory technologies and paradigms are achievable only if their reliability issues can be properly addressed.

In Chapter 3, we have focused on the modeling and mitigation of the decision failure by pursuing a technology-to-circuit co-optimization approach and providing the reference adjustment as an effective means. We have continued in the line of technology-to-circuit co-optimization approach for decision failure mitigation in Chapter 4 and presented the voltage tuning as an effective mechanism. For decision failure mitigation, we have further extended our co-optimization approach to technology-to-algorithm and focused on Boolean and MAC kernels in Chapter 5 and HDC in Chapter 6.

Comprehensive investigation and modeling of the EM specifically for memory signal lines has been our main focus in Chapter 7-10. In Chapter 7 our studies have been centered on the systematic methodology for EM-aware design space exploration. In Chapter 8, we have raised the awareness to the EM concern during the MAC-based NVM-CiM. To address and mitigate the EM issue in SRAM, we have proposed a testing solution in Chapter 9, and finally an EM-aware DTCO for designing a high performance yet EM resilient SRAM module has been the main contribution to Chapter 10.

11.2. Future Perspective

The work carried out in this thesis has provided insight into cross-level co-optimization to improve system reliability without inducing vast overhead on other merits, e.g., performance and energy consumption. Besides reliability improvement, cross-level approaches can be applied to other issues as well. For instance, high resistance parasitic is a main hindrance to improving the energy efficiency of the advanced SRAM working as the cache memory. Therefore, an end-to-end System Technology Co-Optimization (STCO)-based scheme can provide a viable solution to improve the energy efficiency of the SRAM-based cache memories. The works presented in [113], [131], [132] are in this direction.

Besides, leveraging other non-volatile memory technologies such as Ferroelectric Capacitor (FeCAP) in the domain of CiM is promising. However, such emerging technologies also come along with reliability issues which need to be carefully considered. Also, in the case of long-term reliability and as a complement to EM, Time-dependent Gate Oxide Breakdown (TDDB) is another aging issue associated with the transistor elements. Investigating the workload-dependency in TDDB is another potential future work. Finally, enhancing the Electronic Design Automation (EDA) tools to be able to design using CiM primitives and also considering the reliability aspects as the design objective is another research direction that can be considered in the future.

Part III.

Appendix

Bibliography

[1] J. Black, “Electromigration—a brief survey and some recent results”, *IEEE Transactions on Electron Devices*, vol. 16, no. 4, pp. 338–347, 1969.

[2] I. A. Blech, “Electromigration in thin aluminum films on titanium nitride”, *Journal of applied physics*, vol. 47, no. 4, pp. 1203–1208, 1976.

[3] M. Korhonen, P. Bo/rgesen, K.-N. Tu, and C.-Y. Li, “Stress evolution due to electromigration in confined metal lines”, *Journal of Applied Physics*, vol. 73, no. 8, pp. 3790–3799, 1993.

[4] L. Ting, J. May, W. Hunter, and J. McPherson, “Ac electromigration characterization and modeling of multilayered interconnects”, in *31st Annual Proceedings Reliability Physics 1993*, 1993, pp. 311–316.

[5] J. Tao, J. Chen, N. Cheung, and C. Hu, “Modeling and characterization of electromigration failures under bidirectional current stress”, *IEEE Transactions on Electron Devices*, vol. 43, no. 5, pp. 800–808, 1996.

[6] E. Ogawa, K.-D. Lee, V. Blaschke, and P. Ho, “Electromigration reliability issues in dual-damascene cu interconnections”, *IEEE Transactions on Reliability*, vol. 51, no. 4, pp. 403–419, 2002.

[7] T. Sherwood and et al., “Automatically characterizing large scale program behavior”, *ACM SIGPLAN Notices*, vol. 37, no. 10, pp. 45–57, 2002.

[8] M. Durlam, P. J. Naji, A. Omair, *et al.*, “A 1-mbit mram based on 1t1mtj bit cell integrated with copper interconnects”, *IEEE Journal of Solid-State Circuits*, vol. 38, no. 5, pp. 769–773, 2003.

[9] S. S. Mukherjee, C. Weaver, J. Emer, S. K. Reinhardt, and T. Austin, “A systematic methodology to compute the architectural vulnerability factors for a high-performance microprocessor”, in *Proceedings. 36th Annual IEEE/ACM International Symposium on Microarchitecture, 2003. MICRO-36.*, IEEE, 2003, pp. 29–40.

[10] H. W. Klein, P. Hildebrant, J. Doernberg, and J. Li, *Precision analog level shifter with programmable options*, US Patent 6,717,451, Jun. 2004.

[11] J. Michelon, C. Bruynseraede, D. Tio Castro, P. Roussel, R. Hoofman, and K. Maex, “Electromigration study of sub-100nm cu-lines”, in *Advanced Metallization Conference 2004-AMC*, MRS, 2004, pp. 253–257.

[12] B. Wicht, T. Nirschl, and D. Schmitt-Landsiedel, “Yield and speed optimization of a latch-type voltage sense amplifier”, *IJSSC*, vol. 39, no. 7, pp. 1148–1158, 2004.

[13] K.-D. Lee, Y.-J. Park, and B. Hunter, “The impact of partially scaled metal barrier shunting on failure criteria for copper electromigration resistance increase in 65 nm technology”, in *2005 IEEE International Reliability Physics Symposium, 2005. Proceedings. 43rd Annual.*, 2005, pp. 31–35.

[14] N. Shibata, H. Kiya, S. Kurita, H. Okamoto, M. Tan’no, and T. Douseki, “A 0.5-v 25-mhz 1-mw 256-kb mtcmos/soi sram for solar-power-operated portable personal digital equipment-sure write operation by using step-down negatively overdriven bitline scheme”, *IEEE Journal of Solid-State Circuits*, vol. 41, no. 3, pp. 728–742, 2006.

[15] D. Tio Castro, R. Hoofman, J. Michelon, D. Gravesteijn, and C. Bruynseraede, “Void growth modeling upon electromigration stressing in narrow copper lines”, *Journal of Applied Physics*, vol. 102, no. 12, p. 123 515, 2007.

[16] J. P. Gambino, T. C. Lee, F. Chen, and T. D. Sullivan, “Reliability challenges for advanced copper interconnects: Electromigration and time-dependent dielectric breakdown (tddb)”, in *2009 16th IEEE International Symposium on the Physical and Failure Analysis of Integrated Circuits*, 2009, pp. 677–684.

[17] P. Kanerva, “Hyperdimensional Computing: An Introduction to Computing in Distributed Representation with High-Dimensional Random Vectors”, *Cognitive Computation*, vol. 1, 2009.

[18] J. Li, P. Ndai, A. Goel, S. Salahuddin, and K. Roy, “Design paradigm for robust spin-torque transfer magnetic ram (stt mram) from circuit/architecture perspective”, *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 18, no. 12, pp. 1710–1723, 2010.

[19] T. Nogami, T. Bolom, A. Simon, *et al.*, “High reliability 32 nm cu/ulk beol based on pvd cumn seed, and its extendibility”, in *2010 International Electron Devices Meeting*, IEEE, 2010, pp. 33–5.

[20] T. Böscke, J. Müller, D. Bräuhaus, U. Schröder, and U. Böttger, “Ferroelectricity in hafnium oxide: Cmos compatible ferroelectric field effect transistors”, in *IEDM*, IEEE, 2011, pp. 24–5.

[21] Y. S. Chen, H. Y. Lee, P. S. Chen, *et al.*, “Challenges and opportunities for hfox based resistive random access memory”, in *2011 International Electron Devices Meeting*, 2011, pp. 31.3.1–31.3.4.

[22] C. Christiansen, B. Li, M. Angyal, *et al.*, “Electromigration-resistance enhancement with cowp or cumn for advanced cu interconnects”, in *2011 International Reliability Physics Symposium*, IEEE, 2011, 3E–3.

[23] R. Palin, D. Ward, I. Habli, and R. Rivett, “Iso 26262 safety cases: Compliance and assurance”, 2011.

[24] Z. Zeng, P. Khalili Amiri, G. Rowlands, *et al.*, “Effect of resistance-area product on spin-transfer switching in mgo-based magnetic tunnel junction memory cells”, *Applied Physics Letters*, vol. 98, no. 7, p. 072 512, 2011.

[25] W. Zhao, T. Devolder, Y. Lakys, J.-O. Klein, C. Chappert, and P. Mazoyer, “Design considerations and strategies for high-reliable stt-mram”, *Microelectronics Reliability*, vol. 51, no. 9-11, pp. 1454–1458, 2011.

[26] X. Bi, H. Li, and X. Wang, “Stt-ram cell design considering cmos and mtj temperature dependence”, *IEEE Transactions on Magnetics*, vol. 48, no. 11, pp. 3821–3824, 2012.

[27] Y. Chen *et al.*, “Balancing set/reset pulse for> 1010 endurance”, *TED*, 2012.

[28] X. Dong, C. Xu, Y. Xie, and N. P. Jouppi, “Nvsim: A circuit-level performance, energy, and area model for emerging nonvolatile memory”, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 31, no. 7, pp. 994–1007, 2012.

[29] K. Amirkhanyan, A. Davtyan, G. Harutyunyan, *et al.*, “Application of defect injection flow for fault validation in memories”, in *East-West Design & Test Symposium (EWDTS 2013)*, IEEE, 2013, pp. 1–4.

[30] A. Fantini, L. Goux, R. Degraeve, *et al.*, “Intrinsic switching variability in hfo<inf>2</inf> rram”, in *2013 5th IEEE International Memory Workshop*, 2013, pp. 30–33.

[31] Z. Guan, M. Marek-Sadowska, and S. Nassif, “Sram bit-line electromigration mechanism and its prevention scheme”, in *International Symposium on Quality Electronic Design (ISQED)*, 2013, pp. 286–293.

[32] M.-F. Chang, J.-J. Wu, T.-F. Chien, *et al.*, “19.4 embedded 1mb reram in 28nm cmos with 0.27-to-1v read using swing-sample-and-couple sense amplifier and self-boost-write-termination scheme”, in *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, 2014, pp. 332–333.

[33] Z. Guan, M. Marek-Sadowska, and S. Nassif, “Statistical analysis of process variation induced sram electromigration degradation”, in *Fifteenth International Symposium on Quality Electronic Design*, 2014, pp. 700–707.

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

[35] X. Huang, T. Yu, V. Sukharev, and S. X.-D. Tan, “Physics-based electromigration assessment for power grid networks”, in *Proceedings of the 51st Annual Design Automation Conference*, 2014, pp. 1–6.

[36] J. Li, R. K. Montoye, M. Ishii, and L. Chang, “1 mb 0.41 μm^2 2t-2r cell nonvolatile tcam with two-bit encoding and clocked self-referenced sensing”, *IEEE Journal of Solid-State Circuits*, vol. 49, no. 4, pp. 896–907, 2014.

[37] V. Sukharev, “Beyond black’s equation: Full-chip em/sm assessment in 3d ic stack”, *Microelectronic Engineering*, vol. 120, pp. 99–105, 2014.

[38] 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.

[39] V. Seshadri, K. Hsieh, A. Boroum, *et al.*, “Fast bulk bitwise and and or in dram”, *IEEE Computer Architecture Letters*, vol. 14, no. 2, pp. 127–131, 2015.

[40] I. Ciofi, A. Contino, P. J. Roussel, *et al.*, “Impact of wire geometry on interconnect rc and circuit delay”, *IEEE Transactions on Electron Devices*, vol. 63, no. 6, pp. 2488–2496, 2016.

[41] X. Huang, V. Sukharev, T. Kim, H. Chen, and S. X.-D. Tan, “Electromigration recovery modeling and analysis under time-dependent current and temperature stressing”, in *2016 21st Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2016, pp. 244–249.

[42] M. Le Gallo, A. Athmanathan, D. Krebs, and A. Sebastian, “Evidence for thermally assisted threshold switching behavior in nanoscale phase-change memory cells”, *Journal of Applied Physics*, vol. 119, no. 2, p. 025 704, 2016.

[43] M. H. Lin and A. S. Oates, “Electromigration failure of circuit interconnects”, in *2016 IEEE International Reliability Physics Symposium (IRPS)*, 2016, 5B-2-1-5B-2–8.

[44] A. Rahimi, S. Benatti, P. Kanerva, L. Benini, and J. M. Rabaey, “Hyperdimensional biosignal processing: A case study for EMG-based hand gesture recognition”, in *IEEE ICRC*, 2016.

[45] A. Rahimi, P. Kanerva, and J. M. Rabaey, “A Robust and Energy-Efficient Classifier Using Brain-Inspired Hyperdimensional Computing”, in *IEEE ISLPED*, 2016.

[46] A. Shafiee, A. Nag, N. Muralimanohar, *et al.*, “Isaac: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars”, in *2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA)*, 2016, pp. 14–26.

[47] B. Wu, Y. Cheng, J. Yang, A. Todri-Sanial, and W. Zhao, “Temperature impact analysis and access reliability enhancement for 1t1mtj stt-ram”, *IEEE Transactions on Reliability*, vol. 65, no. 4, pp. 1755–1768, 2016.

[48] L. Zhang, Y. Cheng, W. Kang, *et al.*, “Reliability and performance evaluation for stt-mram under temperature variation”, in *2016 17th International Conference on Thermal, Mechanical and Multi-Physics Simulation and Experiments in Microelectronics and Microsystems (EuroSimE)*, 2016, pp. 1–4.

[49] A. Antonyan, S. Pyo, H. Jung, G.-H. Koh, and T. Song, “28-nm 1t-1mtj 8mb 64 i/o stt-mram with symmetric 3-section reference structure and cross-coupled sensing amplifier”, in *2017 IEEE International Symposium on Circuits and Systems (ISCAS)*, IEEE, 2017, pp. 1–4.

[50] S. Dünkel, M. Trentzsch, R. Richter, *et al.*, “A fefet based super-low-power ultra-fast embedded nvm technology for 22nm fdsoi and beyond”, in *2017 IEEE International Electron Devices Meeting (IEDM)*, IEEE, 2017, pp. 19–7.

[51] P. Huang, D. Zhu, C. Liu, *et al.*, “Rtn based oxygen vacancy probing method for ox-rram reliability characterization and its application in tail bits”, in *2017 IEEE International Electron Devices Meeting (IEDM)*, IEEE, 2017, pp. 21–4.

[52] M. Imani, A. Rahimi, D. Kong, T. Rosing, and J. M. Rabaey, “Exploring Hyperdimensional Associative Memory”, in *IEEE HPCA*, 2017.

[53] D. Kleyko, E. Osipov, A. Senior, A. I. Khan, and Y. A. Şekercioğğlu, “Holographic graph neuron: A bioinspired architecture for pattern processing”, *IEEE TNNLS*, 2017.

[54] 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), 2017*, 2017, pp. 1456–1461.

[55] A. Agrawal, A. Jaiswal, C. Lee, and K. Roy, “X-sram: Enabling in-memory boolean computations in cmos static random access memories”, *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 65, no. 12, pp. 4219–4232, 2018.

[56] 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)*, IEEE, 2018, pp. 1–4.

[57] Q. Dong, Z. Wang, J. Lim, *et al.*, “A 1-mb 28-nm 1t1mtj stt-mram with single-cap offset-cancelled sense amplifier and in situ self-write-termination”, *IEEE Journal of Solid-State Circuits*, vol. 54, no. 1, pp. 231–239, 2018.

[58] Q. Dong, Z. Wang, J. Lim, *et al.*, “A 1mb 28nm stt-mram with 2.8 ns read access time at 1.2 v vdd using single-cap offset-cancelled sense amplifier and in-situ self-write-termination”, in *2018 IEEE International Solid-State Circuits Conference-(ISSCC)*, IEEE, 2018, pp. 480–482.

[59] O. Mutlu, “Processing data where it makes sense in modern computing systems: Enabling in-memory computation”, in *Mediterranean Conference on Embedded Computing (MECO)*, 2018.

[60] O. V. Pedreira and et al., “Electromigration and thermal storage study of barrierless co vias”, in *2018 IEEE International Interconnect Technology Conference (IITC)*, IEEE, 2018, pp. 48–50.

[61] S. X.-D. Tan, H. Amrouch, T. Kim, Z. Sun, C. Cook, and J. Henkel, “Recent advances in em and bti induced reliability modeling, analysis and optimization”, *Integration*, vol. 60, pp. 132–152, 2018.

[62] S. Ben Dodo, R. Bishnoi, S. Mohanachandran 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.

[63] T. Chou, W. Tang, J. Botimer, and Z. Zhang, "Cascade: Connecting rrams to extend analog dataflow in an end-to-end in-memory processing paradigm", in *MICRO*, 2019.

[64] T. Evenblij and et al., "A comparative analysis on the impact of bank contention in stt-mram and sram based llcs", in *2019 IEEE 37th International Conference on Computer Design (ICCD)*, 2019, pp. 255–263.

[65] S. M. Nair, R. Bishnoi, M. B. Tahoori, *et al.*, "Variation-aware physics-based electromigration modeling and experimental calibration for vlsi interconnects", in *2019 IEEE International Reliability Physics Symposium (IRPS)*, 2019, pp. 1–6.

[66] K. Ni, X. Yin, A. F. Laguna, *et al.*, "Ferroelectric ternary content-addressable memory for one-shot learning", *Nature Electronics*, vol. 2, no. 11, pp. 521–529, 2019.

[67] L. Wei, J. G. Alzate, U. Arslan, *et al.*, "13.3 a 7mb stt-mram in 22ffl finfet technology with 4ns read sensing time at 0.9 v using write-verify-write scheme and offset-cancellation sensing technique", in *2019 IEEE International Solid-State Circuits Conference-(ISSCC)*, IEEE, 2019, pp. 214–216.

[68] J. Yu, H. A. Du Nguyen, M. A. Lebdeh, M. Taouil, and S. Hamdioui, "Enhanced scouting logic: A robust memristive logic design scheme", in *2019 IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH)*, IEEE, 2019, pp. 1–6.

[69] H. Zahedmanesh, O. V. Pedreira, C. Wilson, Z. Tőkei, and K. Croes, "Copper electromigration; prediction of scaling limits", in *IITC*, IEEE, 2019.

[70] W. Banerjee, I. V. Karpov, A. Agrawal, *et al.*, "Highly-stable (< 3% fluctuation) ag-based threshold switch with extreme-low off current of 0.1 pa, extreme-high selectivity of 109 and high endurance of 109 cycles", in *2020 IEEE International Electron Devices Meeting (IEDM)*, 2020, pp. 28.4.1–28.4.4.

[71] C. Bengel, A. Siemon, F. Cüppers, *et al.*, "Variability-aware modeling of filamentary oxide-based bipolar resistive switching cells using spice level compact models", *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 67, no. 12, pp. 4618–4630, 2020.

[72] R. Bishnoi, L. Wu, M. Fieback, *et al.*, "Special session – emerging memristor based memory and cim architecture: Test, repair and yield analysis", in *2020 IEEE 38th VLSI Test Symposium (VTS)*, 2020, pp. 1–10.

[73] E. M. Boujamaa, S. M. Ali, S. N. Wandji, *et al.*, "A 14.7 mb/mm 2 28nm fdsoi stt-mram with current starved read path, 52Ω/sigma offset voltage sense amplifier and fully trimmable ctat reference", in *2020 IEEE Symposium on VLSI Circuits*, IEEE, 2020, pp. 1–2.

[74] Y.-D. Chih, Y.-C. Shih, C.-F. Lee, *et al.*, "13.3 a 22nm 32mb embedded stt-mram with 10ns read speed, 1m cycle write endurance, 10 years retention at 150°c and high immunity to magnetic field interference", in *2020 IEEE International Solid-State Circuits Conference - (ISSCC)*, 2020, pp. 222–224.

[75] M. E. Fouda, S. Lee, J. Lee, G. H. Kim, F. Kurdahi, and A. M. Eltawil, "Ir-qnn framework: An ir drop-aware offline training of quantized crossbar arrays", *IEEE Access*, vol. 8, pp. 228 392–228 408, 2020.

[76] L. Ge and K. K. Parhi, "Classification Using Hyperdimensional Computing: A Review", *IEEE Circuits and Systems Magazine*, 2020, ISSN: 1558-0830.

[77] J.-M. Hung, X. Li, J. Wu, and M.-F. Chang, "Challenges and trends indeveloping nonvolatile memory-enabled computing chips for intelligent edge devices", *IEEE Transactions on Electron Devices*, vol. 67, no. 4, pp. 1444–1453, 2020.

[78] G. Karunaratne, M. Le Gallo, G. Cherubini, L. Benini, A. Rahimi, and A. Sebastian, "In-memory hyperdimensional computing", *Nature Electronics*, vol. 3, 2020.

[79] M. Le Gallo and A. Sebastian, "An overview of phase-change memory device physics", *Journal of Physics D: Applied Physics*, vol. 53, no. 21, p. 213 002, 2020.

[80] S. Lee, G. Jung, M. E. Fouda, J. Lee, A. Eltawil, and F. Kurdahi, "Learning to predict ir drop with effective training for reram-based neural network hardware", in *2020 57th ACM/IEEE Design Automation Conference (DAC)*, 2020, pp. 1–6.

[81] T. Na, S. H. Kang, and S.-O. Jung, "Stt-mram sensing: A review", *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 68, no. 1, pp. 12–18, 2020.

[82] S. M. Nair, C. Münch, and M. B. Tahoori, "Defect characterization and test generation for spintronic-based compute-in-memory", in *IEEE European Test Symposium (ETS)*, 2020.

[83] S. M. Nair, R. Bishnoi, and M. B. Tahoori, "Mitigating read failures in stt-mram", in *2020 IEEE 38th VLSI Test Symposium (VTS)*, 2020, pp. 1–6.

[84] S. M. Nair, R. Bishnoi, M. B. Tahoori, *et al.*, "Physics based modeling of bimodal electromigration failure distributions and variation analysis for vlsi interconnects", in *2020 IEEE International Reliability Physics Symposium (IRPS)*, 2020, pp. 1–5.

[85] F. N. Najm and V. Sukharev, "Electromigration simulation and design considerations for integrated circuit power grids", *Journal of Vacuum Science & Technology B, Nanotechnology and Microelectronics: Materials, Processing, Measurement, and Phenomena*, vol. 38, no. 6, p. 063 204, 2020.

[86] H. A. D. Nguyen, J. Yu, M. A. Lebdeh, M. Taouil, S. Hamdioui, and F. Catthoor, "A classification of memory-centric computing", *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, vol. 16, no. 2, pp. 1–26, 2020.

[87] S. Wiefels, C. Bengel, N. Kopperberg, K. Zhang, R. Waser, and S. Menzel, "Hrs instability in oxide-based bipolar resistive switching cells", *IEEE Transactions on Electron Devices*, vol. 67, no. 10, pp. 4208–4215, 2020.

[88] H. Zahedmanesh and K. Croes, "Modelling stress evolution and voiding in advanced copper nano-interconnects under thermal gradients", *Microelectronics Reliability*, vol. 111, p. 113 769, 2020.

[89] H. Zhou, S. Yu, Z. Sun, and S. X.-D. Tan, "Reliable power grid network design framework considering em immortalities for multi-segment wires", in *2020 25th Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2020, pp. 74–79.

[90] P. R. Genssler and H. Amrouch, "Brain-inspired computing for wafer map defect pattern classification", in *IEEE International Test Conference*, 2021.

[91] M. Giordano, K. Prabhu, K. Koul, *et al.*, "Chimera: A 0.92 tops, 2.2 tops/w edge ai accelerator with 2 mbyte on-chip foundry resistive ram for efficient training and inference", in *2021 Symposium on VLSI Circuits*, 2021, pp. 1–2.

[92] W. S. Khwa, K. Akarvardar, Y. S. Chen, *et al.*, "Mlc pcm techniques to improve nerual network inference retention time by 105x and reduce accuracy degradation by 10.8x", in *2021 Symposium on VLSI Technology*, 2021, pp. 1–2.

[93] K. Lee, D. S. Kim, J. H. Bak, *et al.*, "28nm cis-compatible embedded stt-mram for frame buffer memory", in *2021 IEEE International Electron Devices Meeting (IEDM)*, 2021, pp. 2.1.1–2.1.4.

[94] Y. Liu, M. Zhao, B. Gao, *et al.*, "Compact reliability model of analog rram for computation-in-memory device-to-system codesign and benchmark", *IEEE Transactions on Electron Devices*, vol. 68, no. 6, pp. 2686–2692, 2021.

[95] M. Mayahinia, C. Münch, and M. B. Tahoori, "Analyzing and mitigating sensing failures in spintronic-based computing in memory", in *IEEE International Test Conference (ITC)*, 2021, pp. 268–277.

[96] S. M. Nair, M. Mayahinia, M. B. Tahoori, *et al.*, "Workload-aware electromigration analysis in emerging spintronic memory arrays", *IEEE Transactions on Device and Materials Reliability*, vol. 21, no. 2, pp. 258–266, 2021.

[97] S. T. Ahmed, M. Mayahinia, M. Hefenbrock, C. Münch, and M. B. Tahoori, "Process and runtime variation robustness for spintronic-based neuromorphic fabric", in *2022 IEEE European Test Symposium (ETS)*, IEEE, 2022, pp. 1–2.

[98] M. H. Amin, M. Elbtity, and R. Zand, "Interconnect parasitics and partitioning in fully-analog in-memory computing architectures", in *2022 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2022, pp. 389–393.

[99] L. Brackmann, A. Jafari, C. Bengel, *et al.*, "A failure analysis framework of reram in-memory logic operations", in *2022 IEEE International Test Conference in Asia (ITC-Asia)*, IEEE, 2022, pp. 67–72.

[100] Y.-H. Chiang, C.-E. Ni, Y. Sung, T.-H. Hou, T.-S. Chang, and S.-J. Jou, "Hardware-robust in-rram-computing for object detection", *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 12, no. 2, pp. 547–556, 2022.

[101] W. C. Chien, L. M. Gignac, Y. C. Chou, *et al.*, "Device study on ots-pcm for persistent memory application : Ibm/macronix phase change memory joint project", in *2022 6th IEEE Electron Devices Technology & Manufacturing Conference (EDTM)*, 2022, pp. 327–329.

[102] P. R. Genssler and H. Amrouch, "Brain-inspired computing for circuit reliability characterization", *IEEE Transactions on Computers*, 2022.

[103] S. Hemaram, M. Mayahinia, and M. B. Tahoori, "Adaptive block error correction for memristive crossbars", in *2022 IEEE 28th International Symposium on On-Line Testing and Robust System Design (IOLTS)*, IEEE, 2022, pp. 1–6.

[104] A. Jafari, M. Mayahinia, S. T. Ahmed, C. Münch, and M. B. Tahoori, "Mvstt: A multi-value computation-in-memory based on spin-transfer torque memories", in *2022 25th Euromicro Conference on Digital System Design (DSD)*, 2022, pp. 332–339.

[105] S. Jung, H. Lee, S. Myung, *et al.*, "A crossbar array of magnetoresistive memory devices for in-memory computing", *Nature*, vol. 601, no. 7892, pp. 211–216, 2022.

[106] J. Krautter, M. Mayahinia, D. R. Gnad, and M. B. Tahoori, "Data leakage through self-terminated write schemes in memristive caches", in *2022 27th Asia and South Pacific Design Automation Conference (ASP-DAC)*, IEEE, 2022, pp. 666–671.

[107] H.-H. Liu, S. M. Salahuddin, D. Abdi, *et al.*, "Extended methodology to determine sram write margin in resistance-dominated technology node", *IEEE Transactions on Electron Devices*, vol. 69, no. 6, pp. 3113–3117, 2022.

[108] M. Liu, C. Zhang, S. Liu, and D. Li, "A 10-bit 2.5-gs/s two-step adc with selective time-domain quantization in 28-nm cmos", *IEEE TCAS-I: Regular Papers*, 2022.

[109] M. Mayahinia, A. Jafari, and M. B. Tahoori, "Voltage tuning for reliable computation in emerging resistive memories", in *IEEE 40th VLSI Test Symposium (VTS)*, 2022, pp. 1–7.

[110] M. Mayahinia, M. Tahoori, G. Harutyunyan, G. Tshagharyan, and K. Amirkhanyan, "An efficient test strategy for detection of electromigration impact in advanced finfet memories", in *IEEE International Test Conference (ITC)*, 2022, pp. 650–655.

[111] M. Mayahinia, M. Tahoori, M. P. Komalan, *et al.*, "Time-dependent electromigration modeling for workload-aware design-space exploration in stt-mram", *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 41, no. 12, pp. 5327–5332, 2022.

[112] M. Mayahinia, M. Tahoori, M. Perumkunnil, K. Croes, and F. Catthoor, "Analyzing the electromigration challenges of computation in resistive memories", in *2022 IEEE International Test Conference (ITC)*, IEEE, 2022, pp. 534–538.

[113] Z. Pei, M. Mayahinia, H.-H. Liu, *et al.*, "Graphene-based interconnect exploration for large sram caches for ultrascaled technology nodes", *IEEE Transactions on Electron Devices*, vol. 70, no. 1, pp. 230–238, 2022.

[114] J. Song, Y. Park, C. Lim, *et al.*, "A 9-bit 500-ms/s 2-bit/cycle sar adc with error-tolerant interpolation technique", *IEEE Journal of Solid-State Circuits*, vol. 57, no. 5, pp. 1492–1503, 2022.

[115] S. Thomann, H. L. Nguyen, P. R. Genssler, and H. Amrouch, "All-in-memory brain-inspired computing using fefet synapses", *Front. Electron. 3: 833260*, 2022.

[116] M. H. van der Veen, O. V. Pedreira, N. Jourdan, *et al.*, "Low resistance cu vias for 24nm pitch and beyond", in *IITC*, 2022, pp. 129–131.

[117] D. Wouters, L. Brackmann, A. Jafari, *et al.*, "Reliability of computing-in-memory concepts based on memristive arrays", in *2022 International Electron Devices Meeting (IEDM)*, IEEE, 2022, pp. 5–3.

[118] S. T. Ahmed, K. Danouchi, C. Münch, G. Prenat, L. Anghel, and M. B. Tahoori, "Spindrop: Dropout-based bayesian binary neural networks with spintronic implementation", *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 13, no. 1, pp. 150–164, 2023.

[119] S. T. Ahmed, M. Mayahinia, M. Hefenbrock, C. Münch, and M. B. Tahoori, "Design-time reference current generation for robust spintronic-based neuromorphic architecture", *ACM Journal on Emerging Technologies in Computing Systems*, vol. 20, no. 1, pp. 1–20, 2023.

[120] H. E. Barkam, S. Yun, H. Chen, *et al.*, "Reliable hyperdimensional reasoning on unreliable emerging technologies", in *2023 IEEE/ACM International Conference on Computer Aided Design (ICCAD)*, 2023, pp. 1–9.

[121] Y.-C. Chiu, W.-S. Khwa, C.-Y. Li, *et al.*, "A 22nm 8mb stt-mram near-memory-computing macro with 8b-precision and 46.4-160.1tops/w for edge-ai devices", in *2023 IEEE International Solid-State Circuits Conference (ISSCC)*, 2023, pp. 496–498.

[122] P. R. Genssler, H. E. Barkam, K. Pandaram, M. Imani, and H. Amrouch, "Modeling and predicting transistor aging under workload dependency using machine learning", *IEEE Transactions on Circuits and Systems I: Regular Papers (TCAS-I)*, 2023.

[123] S. Hemaram, S. T. Ahmed, M. Mayahinia, C. Münch, and M. B. Tahoori, "A low overhead checksum technique for error correction in memristive crossbar for deep learning applications", in *2023 IEEE 41st VLSI Test Symposium (VTS)*, IEEE, 2023, pp. 1–7.

[124] J. Henkel, L. Siddhu, L. Bauer, *et al.*, "Special session-non-volatile memories: Challenges and opportunities for embedded system architectures with focus on machine learning applications", in *Proceedings of the International Conference on Compilers, Architecture, and Synthesis for Embedded Systems*, 2023, pp. 11–20.

[125] W.-H. Huang, T.-H. Wen, J.-M. Hung, *et al.*, “A nonvolatile al-edge processor with 4mb slc-mlc hybrid-mode reram compute-in-memory macro and 51.4-251tops/w”, in *2023 IEEE International Solid-State Circuits Conference (ISSCC)*, 2023, pp. 15–17.

[126] S. Kengeri, “Heterogeneous integration in the ai era”, Presented as the keynote speech in 41st IEEE VLSI Test Symposium (VTS), San Diego, CA, USA, 2023.

[127] S. Kumar, S. Chatterjee, S. Thomann, Y. S. Chauhan, and H. Amrouch, “Cross-layer reliability modeling of dual-port fefet: Device-algorithm interaction”, *IEEE TCAS-I: Regular Papers*, 2023.

[128] M. Mayahinia, H.-H. Liu, S. Mishra, Z. Tokei, F. Catthoor, and M. Tahoori, “Electromigration-aware design technology co-optimization for sram in advanced technology nodes”, in *2023 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, IEEE, 2023, pp. 1–6.

[129] M. Mayahinia, M. Tahoori, G. Tshagharyan, G. Harutyunyan, and Y. Zorian, “On-chip electromigration sensor for silicon lifecycle management of nanoscale vlsi”, in *2023 IEEE European Test Symposium (ETS)*, IEEE, 2023, pp. 1–4.

[130] F. Müller, S. De, R. Olivo, *et al.*, “Multilevel operation of ferroelectric fet memory arrays considering current percolation paths impacting switching behavior”, *IEEE Electron Device Letters*, vol. 44, no. 5, pp. 757–760, 2023.

[131] Z. Pei, M. Mayahinia, H.-H. Liu, *et al.*, “Emerging interconnect exploration for sram application using nonconventional h-tree and center-pin access”, in *2023 24th International Symposium on Quality Electronic Design (ISQED)*, IEEE, 2023, pp. 1–1.

[132] Z. Pei, M. Mayahinia, H.-H. Liu, *et al.*, “Technology/memory co-design and co-optimization using e-tree interconnect”, in *Proceedings of the Great Lakes Symposium on VLSI 2023*, 2023, pp. 159–162.

[133] V. Rietz, C. Münch, M. Mayahinia, and M. Tahoori, “Timing-accurate simulation framework for nvm-based compute-in-memory architecture exploration”, *it-Information Technology*, vol. 65, no. 1-2, pp. 13–29, 2023.

[134] B. Sapui, J. Krautter, M. Mayahinia, *et al.*, “Power side-channel attacks and countermeasures on computation-in-memory architectures and technologies”, in *2023 IEEE European Test Symposium (ETS)*, IEEE, 2023, pp. 1–6.

[135] G. S. Syed, K. Brew, A. Vasilopoulos, *et al.*, “In-memory compute chips with carbon-based projected phase-change memory devices”, in *2023 International Electron Devices Meeting (IEDM)*, 2023, pp. 1–4.

[136] H. Farzaneh, J. P. C. de Lima, A. Nezhadi Khelejani, *et al.*, “Sherlock: Scheduling efficient and reliable bulk bitwise operations in nvms”, in *ACM 61st Design Automation Conference (DAC)*, 2024.

[137] P. R. Genssler, L. Alrahis, O. Sinanoglu, and H. Amrouch, “HDCircuit: Brain-inspired hyperdimensional computing for circuit recognition”, in *IEEE DATE*, 2024.

[138] P. R. Genssler, M. Mayahinia, S. Thomann, M. Tahoori, and H. Amrouch, “DropHD: Technology/algorithm co-design for reliable energy-efficient nvm-based hyperdimensional computing under voltage scaling”, in *IEEE DATE*, 2024.

[139] M. Mayahinia, H. G. Hezayyin, and M. Tahoori, “Reliability analysis and mitigation for analog computation-in-memory: From technology to application”, in *IEEE 42nd VLSI Test Symposium (VTS)*, 2024.

[140] M. Mayahinia, S. Thomann, P. Genssler, C. Münch, H. Amrouch, and M. Tahoori, “Algorithm to technology co-optimization for cim-based hyperdimensional computing”, in *2024 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2024.

[141] S. D. Spetalnick, M. Chang, S. Konno, *et al.*, “A 40-nm compute-in-memory macro with rram addressing ir drop and off-state current”, *IEEE Solid-State Circuits Letters*, vol. 7, pp. 10–13, 2024.

[142] Z. Zhang, M. Mayahinia, C. Weis, *et al.*, “Addressing the combined effect of transistor and interconnect aging in sram towards silicon lifecycle management”, in *IEEE 42nd VLSI Test Symposium (VTS)*, 2024.

List of Figures

2.1. (a) SRAM array with the required periphery circuit, (b) a 6-transistor SRAM cell. 7

2.2. NVM technologies, LRS on top and HRS on the bottom, (a) STT-MRAM, (b) ReRAM, (c) PCM, and d) FeFET 8

2.3. (a) Conventional memory architecture pressing the memory wall problem, (b) mitigation of the memory wall problem by enabling the CiM 9

2.4. a) Performing the Boolean operation using the concept of NVM-CiM on two binary vectors $\{a_1, a_2, ..., a_n\}$, $\{b_1, b_2, ..., b_n\}$, b) Realizing the threshold Boolean operation by adjusting the reference 10

2.5. Performing the MAC operation in the form of $[V_1, V_2, ..., V_n] \cdot [G_1, G_2, ..., G_n]^T$ using the concept of NVM-CiM 11

2.6. Performing the (Binary) pattern similarity measurement using the concept of NVM-CiM, using the NVM-CAM structure comparing two binary vectors $\{Q_1, Q_2, ..., Q_n\}$, $\{P_1, P_2, ..., P_n\}$ 12

2.7. Overlap regions originating from the impact of process variation resulting in an incorrect final output of NVM-CiM and its exacerbation in the case of increasing the number of operands 12

2.8. Offset issue in a) ADC and b) one-bit comparator. contributing in the increase of decision failure 13

3.1. Impact of the temperature on the resistance distributions of ‘P’ and ‘AP’ MTJ. For the setup, please refer to section 3.4 21

3.2. (a) Resistance distributions of ‘P’ and ‘AP’ MTJ and also reference resistance, box (1) and (2) show the RDF case for ‘P’ MTJ and ‘AP’ MTJ, respectively (b) equivalent RDF cases by considering $R_P - R_{Ref}$ and $R_{AP} - R_{Ref}$ distributions 22

3.3. Resistance distribution of the (Tr+R) cell model, for the setup, please refer to section 3.4 . 24

3.4. Linear approximation of the equivalent resistance of the full cell (Tr+R) vs the ‘P’ and ‘AP’ MTJs as the load at 4 different temperatures, for the setup, please refer to section 3.4 . . . 25

3.5. The conceptual design of the proposed reference resistance 26

3.6. (a) Encoding the data in the redundant cells and (b) a general structure of the reference cells for CiM operation with N number of operands and encoding the data in n_{Red} redundant cells 29

3.7. Compound reference structure (a) [8] for the standard STT-read, (b) and (c) generalizing to the 2-operand CiM for R_{Ref}^{OR} and R_{Ref}^{AND}, respectively 30

3.8. RDF probability for the proposed structure, compound, and only poly as the reference resistance. (R) model has been considered for the cell. (a) STT-read (b) CiM-2 30

3.9. RDF probability for different number of CiM operands based on (R) cell model, the reference resistance structure for STT-Read and CiM-2 are the proposed structure and for CiM-4 and CiM-8 is the compound structure 31

3.10. RDF probability with different approaches of the reference optimization and RDF estimation using (R) and (Tr+R) cell modelings 32

3.11. RDF probability for the CiM-2 operation, the cell model is considered to be (Tr+R) 32

3.12. The minimum number of the redundant (Tr+R) cells versus different numbers of the CiM operands to achieve target RDF probability of STT-read 33

4.1. a) Using the memristive array to perform CiM, b) Pre-charge SA to evaluate scouting-CiM AND (NAND) operation, <H> and <L>: HRS and LRS respectively 36
4.2. Effect of the temperature on the resistance distributions of LRS and HRS for the STT-MRAM and ReRAM bit-cells at the nominal (VDD, V_{WL}) 37
4.3. Impact of the voltage bias on the resistance distributions of LRS and HRS (at 27 $°C$ for the STT-MRAM and ReRAM bit-cells, the distributions for the nominal (**VDD,** V_{WL}) are shown by *red* 37
4.4. Impact of the voltage tuning and the temperature on the RDF probability in STT-MRAM and ReRAM technologies, STT-CiM operation: logical AND (NAND), ReRAM-CiM operation: logical OR (NOR), the measured RDF probability are shown with '*' 40
4.5. Pareto front (a) CiM-2,4,8 operations for an array of 256kB STT-CiM, (b) CiM-8 operations for an array of 256kB ReRAM-CiM, for various (VDD (V), V_{WL} (V)) combinations 40

5.1. The end-to-end flow for reliability analysis during NVM-CiM, μ and σ show the mean and standard deviation of a statistical distribution, N_{MC} shows the number of Monte Carlo simulations, $\mathbb{E}[E_{HW}^{DF}]$ and $\mathbb{E}[E_{SW}^{DF}]$ show the hardware- and software-level expected errors, respectively 46
5.2. The trade-off between energy, offset, and expected error values, for scouting Boolean NVM-CiM with two operands, crossbar size: 128×128, interconnect node: 5 nm, to achieve a comparable $\mathbb{E}[E_{HW}^{DF}]$ with an equal number of operands, the position of operands for STT-MRAM-CiM and ReRAM-CiM are closest to and farthest from the sensing circuitry, respectively 51
5.3. Heat map of $\mathbb{E}[E_{HW}^{DF}]$ under different hardware parameters during the MAC operation, X labels: S/R, STT-MRAM, ReRAM. 5/22, interconnect node. 64/128/256/512, crossbar size. Y labels: 2/4/8/12, number of operands. F/M/C, Farthest, Middle, and Closest positions . . 52
5.4. Heat map of $\mathbb{E}[E_{HW}^{DF}]$ under different hardware parameters during NVM-CiM for scouting Boolean operation, the offset of the comparator is 3%, for X and Y labels, refer to the caption of Figure 5.3 52

6.1. Golden baseline inference accuracy of language recognition 56
6.2. Overview of the proposed method. (a) Split the HVs into chunks at the algorithmic level. (b) Multi-bank AM architecture to maintain the parallel search capability. (c) The array-level realization of the CAM structure. (d) The timing diagram shows the hierarchical calculation of the similarity measure. (e). The differential structure of the resistive CAM for 1T-1NVM (PCM, STT-MRAM, and ReRAM) and for 1T structure (FeFET). (f) Inter-Bank switches for scalability of the CAM-based accelerator for HDC 56
6.3. (a-e) The expected ΔV with respect to the N_{G-bank} for different NVM technologies, (f) the distinguishability versus N_{G-bank}, (C_x, are the prototype classes, in the evaluated data set: C_1, C_4 are the most and the second most similar classes, respectively 60
6.4. Median inference accuracy loss of the tested technologies. (a) The relation of accuracy loss over block size at a hypervector (HV) dimension of 6000 bits (b) The relation of accuracy loss over HV dimension at a block size of 320 bits 61
6.5. Pareto optimal analysis of the different HV and block sizes with respect to accuracy loss and energy consumption. (a) The Pareto front of PCM at the connected dots and all non-optimal configurations in the background as a scatter plot. (b) Comparison of the different technologies 61

7.1. Change of the hydrostatic stress due to the execution of an arbitrary sequence of operations and equivalent currents through methods RCA, CAC, and CAI corresponds to this sequence, cycles without a label are idle cycles 68

7.2. The worst-case EM-induced MTTF for different workloads for ten different applications, using 3 different methods for calculating the I_{seg}: RCA, CAC and CAI 69
7.3. The impact of the BL width on the MTTF for methods RCA and CAC of calculating the I_{seg} 71

8.1. A segmented BL of a NVM crossbar consists of RC segments corresponding to each memory bit-cell 74
8.2. The proposed EM modeling based on the existing models 75
8.3. Current distribution across the BL segments for different case studies in a ReRAM-based crossbar, N_{rows} = 512, for the STT-MRAM and PCM technologies, the current trends in the three case-studies are similar, the segment 0 and 511 are the farthest and closest segments to the sensing circuitry, respectively, number of activated rows in worst-case CiM operation is 64 75
8.4. Impact of the row activation pattern on the EM-induced MTTF, N_{rows} = 256, Width = 22 nm, normalized to MTTF of the *worst-case standard (normal) memory operation* in the respective technology, *best- (worst-) case activation pattern*: activated rows are the closest (farthest) addresses to the sensing circuitry 76

9.1. Advanced Inductive Failure Analysis (AIFA) 83
9.2. The flowchart for our proposed EM test 84
9.3. The EM test generation by using the concept of D.t.O 86
9.4. The waveform as the output of the SPICE simulation on the defect injection netlist 86

10.1. The proposed EM-aware DTCO methodology for SRAM fabricated in 12 nm CD 90
10.2. The current versus the WL segment during the read and write operations in two 22 nm and 12 nm CD technologies 92
10.3. The current versus the BL segment during the read operation in two SRAM designs based on 22 nm and 12 nm CD, red arrows show the sudden decrease of the current at the activated segment during read-0 93
10.4. The current versus the BL segment during the write operation in two SRAM designs based on 22 nm and 12 nm CD 94
10.5. Steady-state hydrostatic stress on the BL, based on 22 nm and 12 nm CD 95
10.6. The WL current density versus the latency of the WL driver 95

List of Tables

3.1. TMR^{*}_{CiM-N} (normalized to TMR of single-cell STT-Read operation) for different N at $25°C$ 20
3.2. The normal distribution parameters of 'P' and 'AP' MTJs, for the setup, please refer to section 3.4 24
3.3. simulation setup tools and parameters 29
3.4. Average RDF probability and area of the different reference resistance structures for STT-Read (1 MTJ activated) and CiM-2 (2 MTJs activated) based on (R) cell modeling 31
3.5. The minimum RDF probability for different values of the RA. The TMR is 150% and the cell model is (Tr+R) 34
3.6. The minimum RDF probability for different values of the TMR. The RA is 7.5 $\Omega\mu m^2$ and the cell model is (Tr+R) 34

4.1. Simulation setup tools and parameters 41
4.2. Rate of the RDF probability degradation for ReRAM-CiM-8 operation 43
4.3. Array-level power and latency for both the STT-MRAM and ReRAM technologies in nominal and the variation-optimized voltage combination, considering the memory-read, and CiM-8 at 27 $°C$ for the array-size of 256 kB 43

5.1. Summary of the related work 45
5.2. End-to-end simulation setup and tools 49
5.3. The application-level analysis of error-masking capability and execution time for *2mm* benchmark consisting the MAC kernel, using 22 nm interconnect and crossbar size of 128×128 53
5.4. The application-level analysis of error-masking capability and execution time for *database query* consisting the Boolean kernel, using 22 nm interconnect, and crossbar size of 128×128, number of records in the database: 8192 53
5.5. Decreasing $\mathbb{E}[E^{DF}_{SW}]$ by hierarchical execution of *database query* application and its latency overhead, using STT-MRAM, 22 nm interconnect, and with the crossbar size of 128×128 . 53

6.1. Simulation setup tools and parameters 59
6.2. Summary of the results and comparison with the related work, the size of the hypervector is 10k 63

7.1. System-level parameters for *gem5* simulation 69
7.2. MTTF for different STT-MRAM arry organization 71

8.1. Simulation setup tools and parameters 76
8.2. Impact of the BL length ($N_{rows}\times$ segment's length corresponding to each cell (118.8) nm) on the EM-induced MTTF, Width = 22 nm, and (R = 11.95 Ω, C=33.26 aF), for worst-case CiM operation, the number of activated rows is $\frac{N_{rows}}{8}$ 77
8.3. Impact of the BL width on the EM-induced MTTF, $N_{rows} = 512$, for worst-case CiM operation: number of activated rows is 64, the RC parasitic values: (R = 13.86 Ω, C = 32.39 aF), (R = 11.95 Ω, C = 33.26 aF), (R = 10.26 Ω, C = 34.18 aF), (R = 9.44 Ω, C = 35.18 aF) for Width = 21..24 nm, respectively 78

9.1. Simulation setup parameters 85

9.2. Impact of the VDD of SRAM array and number of memory instructions on the early detection of the proposed EM test 87

10.1. Parameters for co-simulation of the energy and EM reliability 92
10.2. Summary of the results for performing the EM-aware DTCO on the SRAM based on 22 *nm* and 12 *nm* CD 93
10.3. Summary of the comparison in SRAM based on 22 *nm* CD and 12 *nm* CD, the parameters are normalized to values based on 22 *nm* CD 95

Acronyms

ADAS Advanced Driver Assistance Systems.

ADC Analog-to-Digital Converter.

AIFA Advanced Inductive Failure Analysis.

Al Aluminum.

AM Associative Memory.

API Application Programming Interface.

AVF Architectural Vulnerability Factor.

BEoL Back-End-of-the-Line.

BL Bit-Line.

BLB Bit-Line-bar.

C.V Coefficient of Variation.

CAM Content Addressable Memory.

CD Critical Dimension.

CiM Computation in Memory.

Cu Copper.

CVF CiM Vulnerability Factor.

DAC Digital-to-Analog Converter.

DC Direct Current.

DPPM Defective Parts Per Million.

DRAM Dynamic RAM.

DTCO Design Technology Co-Optimization.

EDA Electronic Design Automation.

EM Electromigration.

EMG Electromyography.

FeCAP Ferroelectric Capacitor.

FeFET Ferroelectric Field-Effect Transistor.

FEoL Front-End-of-the-Line.

FET Field-Effect Transistor.

FIT Failure In Time.

GAA Gate-All-Around.

GDS Graphic Design System.

HDC Hyperdimenssional Computing.

HRS High Resistive State.

HV Hypervector.

ISA Instruction Set Architecture.

KS test Kolmogorov-Smirnov test.

LRS Low Resistive State.

MAC Multiply-Accumulate.

MMU Memory Management Unit.

Mn Manganese.

MTJ Magnetic Tunnel Junction.

MTTF Mean Time To Failure.

NTC Negative Temperature Coefficient.

NVM Non-volatile Resistive Memories.

NVM-CAM NVM-based CAM.

NVM-CiM Computation in Non-volatile Resistive Memories.

PCM Phase Change Memory.

PDE Partial Derivative Equation.

PDK Process Design Kit.

PDP Power-Delay Product.

POH Power-on Hours.

PTC Positive Temperature Coefficient.

RA Resistance-Area product.

RDF Read Decision Failure.

ReRAM Redox-based RAM.

RTN Random Telegraph Noise.

SA Sense Amplifier.

SRAM Static RAM.

STCO System Technology Co-Optimization.

STT-MRAM Spin-Transfer Torque Magnetic RAM.

Ta Tantalum.

TDDB Time-dependent Gate Oxide Breakdown.

TMR Tunnel Magneto-Resistance Ratio.

WL Word-Line.